ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE
Provozně ekonomická fakulta
Katedra statistiky

Obecná metodologie – dataminingová metoda dynamické segmentace aplikovaná v Business Intelligence

Disertační práce

Ing. Tereza Bělková

Santiago de Chile
červen, 2005

SOUHRN

Obecná metodologie – dataminingová metoda dynamické segmentace aplikovaná v Business Intelligence

Jeden ze široce užívaných nástrojů v oblasti identifikace vzorů chování, Data Miningu a aplikované statistiky, tvoří algoritmy shlukování, které dovolují identifikovat shluky skryté v datech.

Shluky, získané aplikací existujících algoritmů shlukování, jsou statické, tedy přísluší fotografii hodnot atributů objektů získaných v jednom okamžiku. Avšak v případech, jako je například chování klientů v bance, se informace o objektech aktualizuje v čase (tato skutečnost má dynamický charakter). Je proto nutné dle chování klientů v průběhu času aktualizovat vzory chování shluků a tedy disponovat dynamickou metodou segmentace. Tato aktualizace by měla využít znalostí získaných z předchozí realizace shlukování, která je obecně náročná na čas a vyžaduje úzce specializovanou práci.

Cílem této práce je zkonstruovat obecnou metodologii, použitelnou pro analýzu shlukování, založenou na dynamickém chování objektů, která detektuje změny ve shlucích a aktualizuje vzory chování shluků optimálním způsobem. Její realizace bude vycházet ze struktury shluků vytvořené v předcházejících obdobích.

Pro splnění cíle je identifikován charakter dat – vlastní-li nebo nevlastní identifikátor a jsou definovány možné scénáře změny, kterými jsou: Zachování počtu shluků, Tvorba nových shluků a Zrušení shluků. Výzkum se zaměřuje na vývoj obecné metodologie pro případ dat s identifikátorem, která pojímá všechny možné scénáře změny.

Navržená obecná metodologie se skládá z pěti základních etap procesu: Etapa I: Identifikace objektů, které představují změnu, Etapa II: Rozpoznání stavu změny, Etapa III: Rozhodnutí o možnostech aktualizace shluků, Etapa IV: Zánik shluků a Etapa V: Identifikace trajektorií. Každá etapa má svoji vlastní dynamiku. Snahou je definovat každou etapu co možná nejobecnějším způsobem.

Obecná metodologie, realizovaná pomocí metody k-means, byla aplikována na data simulovaná i skutečná, která byla získána z Banky úvěru a investic (Banco de Crédito

y Inversiones se zkratkou Bci) v Santiagu, Chile. Mimo splnění očekávaných cílů bylo prohloubeno a potvrzeno poznání o jejích dalších charakteristikách, které jsou s výhodou využívány při její aplikaci v praxi. V oblasti Business Intelligence obecná metodologie umožňuje flexibilizovat a automatizovat proces dynamického dataminingového modelování maximalizujíc příležitosti obchodu, které na segmentovaném trhu dosud nebyly plně využity v důsledku dvou skutečností: neexistovala dynamická technika shlukování schopná vyjádřit reálné procesy na trhu a z toho odvozená druhá příčina, kterou je zpožďování rozhodování.

Daný problém má dlouhou trajektorii budoucího výzkumu; týká se například rozšíření řešení v jednotlivých etapách obecné metodologie identifikovaných v této práci či uplatnění rozdílných metod a kritérií v těchto etapách.

Poděkování

Obsah a formu zpracování předkládané disertační práce do značné míry ovlivnila skutečnost, že jsem měla možnost cestovat a sbírat postřehy a zkušenosti nejen životní, ale i studijní a pracovní, mimo území naší České republiky.

Poděkování, že stojím zde - na konci začátku své profesní kariéry - patří celé mé rodině:

Mamince, která mi stále připomínala, že mám ještě něco důležitého dokončit.
Dědovi, který mi svými životními zkušenostmi byl oporou po dobu celého univerzitního studia.
Babičce, která mi svými životními zkušenostmi byla oporou duševní.
Mému manželovi, Alejandrovi, o kterého se má rodina během zpracovávání této disertační práce rozrostla, patří mé poděkování největší - za jeho respektování nedostatku společně tráveného času až po starost o domácnost.

V neposlední řadě bych ráda poděkovala všem, se kterými jsem se po dobu svého univerzitního studia na ČZU v Praze setkala a kteří mi svým osobním příkladem, radami a pomocí byli tou největší motivací.
Zvláštní poděkování bych chtěla vyslovit svému školiteli - doc.Ing.Otakarovi Macháčkovi, CSc., za jeho trpělivost a poskytovanou odbornou pomoc.

Děkuji.

Ing. Tereza Bělková

.

OBSAH

Tituly tabulek

Tituly grafů

Tituly obrázků

1 ÚVOD

1.1 Oblast výzkumu

Jedna z nejčastěji řešených úloh v oblasti identifikace *vzorů (formátů) chování* [11][14], aplikované statistiky [19] a *Data Miningu* [11], spočívá v aplikaci algoritmu *shlukování (segmentace,* anglicky *clustering)* k rozdělení datového souboru do *shluků (tříd, segmentů)* a získání jejich příslušných charakteristik. Vzorem chování je obecně nazývaná jakási abstraktní reprezentace souboru dat [6].

Metody shlukování jsou ve své většině limitovány tím faktem, že analyzují objekty v jednom determinovaném okamžiku v čase. Algoritmy(*) (publikované software) po vykonaném seskupení objektů již nejsou schopny agregovat nový nebo efektivně aktualizovat již zařazený objekt a vytvořit tak novou strukturu shluků(**), která vychází ze struktury již poznané. Z právě popsaného je možné vyvodit poznání, že existující *algoritmy* segmentace mají *statický charakter*.

V této práci je věnována pozornost algoritmu, který by definoval shluky v závislosti na *dynamickém chování objektů.* Předpokládá se totiž, že kontinuálním sledováním chování objektů je možné detektovat reálné změny ve formátech chování a na tomto základě adekvátním způsobem, vlastním každému segmentu, obnovit struktury shluků. Důležitost vytvoření *dynamického algoritmu segmentace* přichází mimo jiné také s potřebou vyřešit situace vzniklé aktualizací či agregací objektů, které mohou zároveň znovu využít známou informaci o struktuře shluků v minulosti. Tento potenciál je důležitý, jelikož aplikace v Data Miningu bývají náročné na čas a specializaci modelátora a při praktickém užití vyžadují zachování kontinuity post modelačních procesů.

1.2 Získávání znalostí z báze dat a Data Mining

V současnosti, díky vývoji software a hardware, je možné skladovat informace o objektech a historii jejich chování ve velkých bázích dat. To má nutně za následek vznik

---

* V této práci: algoritmus neboli metoda.

** Vytvořit novou strukturu shluků neboli aktualizovat vzory chování znamená změnit charakteristiky shluků, definované v předešlém časovém období, na základě použité metodologie.

oddělení výzkumu a aplikací, ve kterých se rozvíjejí strategie založené na informacích vytěžených z bází dat [33]. Znalosti se z bází dat získávají procesem *KDD* (Knowledge DataBase Discovery). KDD je definován jako *nebanální proces identifikace vzorů v datech, které jsou platné, nové, potenciálně užitečné a srozumitelné* [7].

Výše popsané je definováno jako *proces* a to na ozřejmění, že nalezení znalostí často zahrnuje experimenty a jejich opakované provádění, jednání s klienty, návrhy řešení a zpracování vedoucí k uspokojení o dosaženém poznání z dat jak ze strany uživatelů tak i klientů [7]. Zde nacházejí úlohy a algoritmy Data Miningu své užití [6][7][33]. Termín Data Mining se vztahuje k aktu *dolování* vzorů chování z dat [7].

Typická reprezentace procesu KDD zahrnuje devět základních kroků [33]:

1) Pozorumění problému a ovládání aplikace software.
2) Získání dat, se kterými se bude pracovat.
3) Úprava dat.
4) Redukce a transformace dat.
5) Výběr techniky Data Miningu.
6) Výběr algoritmu Data Miningu.
7) Data Mining.
8) Analýza výsledků Data Miningu.
9) Utvrzení se v nalezeném poznání.

Ve druhém kroku je třeba vybrat soubor dostupných dat, která jsou jednotně uložena v bázi dat neboli *Data Warehouse* (sklad dat) [33]. V mnoha reálných aplikacích obsahují data v Data Warehouse různé druhy chyb, například chybějící údaje, hodnoty přesahující logický rozsah atributu nebo hodnoty, které jsou nepravděpodobné. Tento důvod vede analytika ke třetímu kroku, kdy dochází k předzpracování dat [33].

V páté fázi je vybrána specifická technika Data Miningu pro detailní analýzu. Zde je třeba rozhodnout, zda je cílem seskupovat objekty (což je úloha navrhovaná v této práci) nebo řešit další úkol v rámci modelování závislostí a nezávislostí mezi atributy a objekty. Na základě tohoto rozhodnutí mají být vybrány nejvhodnější algoritmy Data Miningu (krok 6), aby poté mohly být užity při hledání vzorů chování v datech (krok 7).

1.3 Situace ve výzkumu

Dynamický Data Mining má ve výzkumných kruzích vzrůstající atraktivitu. Uživatelé instalovaných nástrojů Data Miningu se nyní zajímají o využití technik souvisejících s jejich prací. Jakmile většina *dataminingových systémů* (systém ve smyslu modelace problému jistou technikou v rámci dataminingového nástroje) bude potřebovat aktualizaci v budoucnu[*], jejich zájem bude soustředěn právě na techniky dynamické.

Jestliže je budoucí chování systému velmi podobné tomu minulému (například výnosy plodin na jistém půdním typu nebo v jisté klimatické zóně), užití počátečního systému Data Miningu by se zdůvodňovalo. Mění-li se ovšem chování objektů v čase (například chování klientů ve finančním prostředí), stálé užívání počátečního systému by mohlo vést k nereálným výsledkům a z toho vyplývajícím neakceptovatelným rozhodnutím. Právě zde příchází na řadu oblast nového výzkumu, dynamického Data Miningu[**].

1.3.1 Současný vývoj v oblasti dynamického Data Miningu

V oblasti Data Miningu byly vyvíjeny nejrůznější techniky s cílem nalézt užitečné informace v souboru dat. Za nejdůležitější (z hlediska frekvence používání) jsou považovány *rozhodovací stromy, neuronové sítě, asociační zákony* a metody shlukování [10].

Pro každou výše zmíněnou dataminingovou techniku má aktualizace různé orientace; některé z navržených přístupů jsou zde zmíněny:

- *Rozhodovací stromy*: různorodé techniky na zvyšování stupně učení, restrukturalizace stromů [23][29][30] a identifikace konceptu drift [4].
- *Neuronové sítě*: aktualizace je často užívána ve smyslu opakovaného učení nebo zlepšení výkonnosti sítě učením se z nových příkladů předkládaných síťovému modelu [13].
- *Asociační pravidla*: rozvinutí systému pro dynamický Data Mining realizovaný technikou asociačních pravidel [25].

---

* Znalosti chování dat *vytěžené* pomocí dataminingového systému jsou založeny na analýze chování objektů v minulosti.

** Tradiční Data Mining s dynamickými prvky.

- *Segmentace*: v následujícím odstavci jsou shrnuty přístupy k dynamickému Data Miningu užívajícímu techniky segmentace, které je možné nalézt v literatuře.

Současný dynamický dataminingový systém segmentace je soustředěn na vývoj shlukovacího algoritmu ve smyslu modelovaní změn pozic objektů v segmentech pozorovaných v jistých momentech v čase [15][34] (algoritmy v Latent Gold nebo v segmentaci ve dvou fázích). Existují také výzkumy, zabývající se evolučním algoritmem na optimalizaci počtu shluků v průběhu času – dynamické dělení segmentu užitím vývojových strategií [1][5][21][24]. Byl vyvinut systém na vyjadření dynamického pohybu shluků pro objekty bez identifikace [5].

1.3.2 Logický vývoj statického algoritmu segmentace k dynamickému

Až doposud se v úlohách shlukování používají v podstatě tyto čtyři strategie:

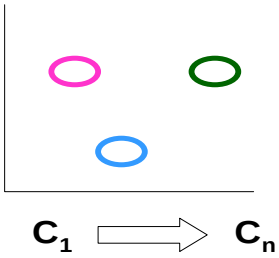
1. Uživatel nedbá na změny v prostředí a aplikuje výchozí dataminingový systém beze změn.
2. Je užíván počáteční systém a novými daty se provádí aktualizace *klasifikací* [2][9].
3. Pro každý cyklus – který závisí na konkrétní aplikaci a problému – je vyvinut nový systém užitím všech dostupných znalostí, dat a technik.
4. Aktualizace počátečního systému se provádí novými daty, kdy se zároveň mění i struktura segmentu. Ta je založena na předešlé segmentaci, ovšem bez zahrnutí dynamické složky chování objektů. Výsledky tak nevedou k optimálnímu praktickému využití znalostí získaných *aktualizací vzorů chování shluků*.

   Výše popsané strategie je možné pojmout jako logický vývoj statického algoritmu k dynamickému.

První strategie, představená obrázkem 1-1, má tu výhodu, že není náročná výpočetně, jelikož není prováděna aktualizace dataminingového systému. Mimoto nepožaduje změny v následných procesech, kterým může být projekt marketingové kampaně pro jednotlivé segmenty klientů. Její nevýhoda je ta, že neodkrývá současné tendence, nepopisuje realitu a teoreticky vzato následně nelze docházet k reálným rozhodováním. Charakteristiky segmentu ani hodnoty proměnných se v čase nemění. Není možné zařadit nový objekt.

Zde je možné při vytváření počátečního systému (s přihlédnutím k charakteru dat a praktickému využití výsledků) použít téměř jakoukoliv metodu shlukování (jednotlivé metody shlukování jsou zmíněny v *sekci 1.4*).

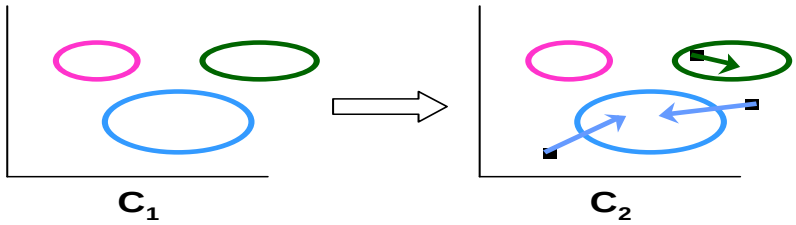
Obr 1-1 ***První strategie logického vývoje statického algoritmu k dynamickému***[*]



Druhá strategie užívá známé segmenty (vytvořené v předešlém období) a aktualizuje (pouze však klasifikuje) pozice klientů mezi statickými segmenty dle aktualizovaných hodnot proměnných (mění se hodnoty objektů, ne však struktura segmentu). Nevýhodou je, že vzory chování jsou v dynamickém prostředí s největší pravděpodobností zastaralé, neaktualizované, neodpovídají skutečnosti. Aktualizují-li se objekty, měla by se aktualizovat i struktura shluků. Není možné zařadit nové objekty. Tato strategie, znázorněná obrázkem 1-2, byla použita například k vytvoření modelu pro získaní Life Time Value bankovních klientů metodou segmentace v Latent Gold nebo jednoduše metodou *k-means*[**].

Při této strategii je nutné užít metodu, která je schopná na základě jistého kritéria (například pravděpodobnosti příslušnosti ke každému segmentu) zařazovat objekty v každém sledovaném cyklu do příslušných segmentů.

Obr 1-2 ***Druhá strategie logického vývoje statického algoritmu k dynamickému***



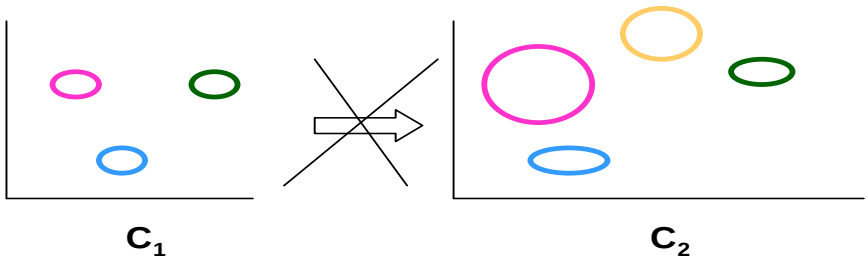
* *C* značí (v každém z prvních čtyř obrázků) časový cyklus.

** Příkazem *pouze klasifikovat*.

Ve třetí strategii vyvíjí uživatel nový systém. Nevýhodou této strategie je výpočetní a finanční náročnost a navíc neefektivita práce, jelikož analytik *ztrácí* již získané znalosti z dat. To v praxi znamená nově projektovat následné procesy jako je například marketingová kampaň, jelikož segmenty budou mít pravděpodobně značně odlišné charakteristiky od segmentů získaných v předešlém cyklu a navíc na ně nebudou navazovat. Aktualizace nemá dynamický charakter, jelikož se provádí nová segmentace, která je pouze fotografií situace v jistém časovém okamžiku (obrázek 1-3). Mění se charakteristiky shluků, je možné zařadit nové objekty.

Opět je možné použít téměř jakoukoliv metodu shlukování.

Obr 1-3 Třetí strategie logického vývoje statického algoritmu k dynamickému

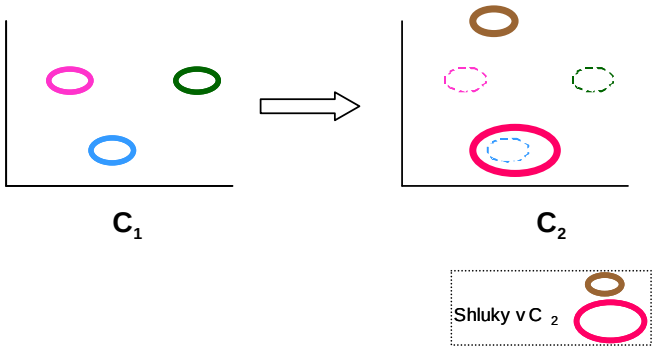


Poznámka: Přeškrtnutá šipka značí, že neexistuje návaznost modelace z jednoho období do druhého.

Strategie, uvedená jako poslední, mění v každém cyklu charakteristiky shluků na základě hodnot objektů a to opět jen v jistých momentech v čase; může nebo se nemusí zakládat na segmentaci vytvořené v předešlé etapě vývoje. Je možné zařadit nové objekty. Nevýhoda spočívá ve skutečnosti, že uživatel nového systému opět ztrácí možnost praktického využití znalostí získaných z předešlé segmentace. Po automatické aktualizaci středů shluků je totiž nutné změnit následné procesy, i když pravděpodobně tuto modifikaci není třeba provádět pro každého klienta - ne vždy je nutná aktualizace struktury segmentu, jelikož to mohou být právě jen nově zařazované objekty, které *nepřísluší* do již vytvořených shluků nebo je to jen několik objektů, které představují změnu (pohyb). Chování objektů v průběhu času zde není věnována pozornost.

Pro realizaci této strategie je používána metoda *segmentace ve dvou fázích* v SPSS. Strategie je zobrazena obrázkem 1-4.

Obr 1-4 *Čtvrtá strategie logického vývoje statického algoritmu k dynamickému*



Ani jedna z výše uvedených strategií nemá dynamický charakter, jelikož se nezakládá na dynamickém chování objektů, které je nutné v případech měnících se charakteristik prostředí zahrnout do aktualizace formátů segmentů. Tyto strategie a metody užívají dynamické elementy při aplikaci shlukové analýzy na aktualizaci statického datového souboru. Obecná metodologie Data Miningu, vyvíjena v této práci, může být aplikována na dynamický datový soubor, který je charakterizován pohyby objektů v čase.

Metodologie, navrhovaná v této práci, na základě dynamického chování objektů identifikuje, je-li nutná aktualizace charakteristik shluků a aktualizuje tyto nejvhodnějším způsobem. Aktualizace vychází z efektivního využití předchozího systému. Tento systém má několik výhod, například: je výpočetně výkonnější, finančně výhodnější, není závislý na typu aplikace, výsledky odpovídají skutečnosti a v neposlední řadě identifikace zřejmých změn v systému by mohla vést k pochopení budoucích změn v příslušném prostředí[(*)].

1.4 Shlukování objektů: koncept a příklady

V *sekci 1.1* bylo z abstraktního úhlu pohledu definováno to, čím se rozumí shlukování. Nyní je nutné tento koncept rozšířit.

Jsou stanoveny dva základní pojmy. Prvním je vlastní segmentace a druhým je klasifikace objektů do tříd [14][20][25].

---

* Další výhody jsou odkrývány v průběhu práce, především v *sekci 8* a shrnuty jsou pak v *sekci 9*.

Shluková analýza je často definována jako nalezení přirozených shluků. Konkrétně, cílem shlukování je rozdělit množinu individuí, objektů nebo pozorování (dále jen objekty) do takových shluků, kdy stupeň přirozené asociace pro každý objekt je vysoký se členy svého vlastního segmentu a nízký se členy ostatních segmentů [14]. Pravidla rozdělení množiny objektů do podmnožin jsou shrnuty následovně:

- Podobné objekty jsou seskupovány do podobných tříd; tato vlastnost se nazývá *homogenita*.
- Rozdílné objekty jsou rozdělovány do rozdílných tříd; tato vlastnost se nazývá *heterogenita*.

Klasifikace je proces zařazování objektů do třídy, kterou mají zaujímat.

Dalším z úkolů shlukování je nalézt vzory chování, které identifikují charakteristiky množiny objektů. Jedná se o sloučení objektů do *h* podmnožin, nazývaných shluky. Každá z těchto podmnožin je reprezentována jedním prototypem S_h a pro každý objekt je stanoven koeficient příslušnosti ke každému shluku [14].

Mezi metodami segmentace se nacházejí klasické statistické metody, jako je k-means, *hierarchická segmentace* [20] nebo segmentace ve dvou fázích. Metodami, vyvíjenými v oblasti umělé inteligence, jsou neuronové sítě, *automatické učení*, *metody difusní segmentace* [11] a další [11]. Vzory chování jsou determinovány v referenci na konkrétní aplikaci [14].

Aby mohly být objekty klasifikovány, musí existovat třídy a jejich identifikace [14]. V případě, kdy třídy nejsou známy *a priori*, jsou získány užitím algoritmu segmentace [20] [27]. Příkladem je segmentace klientů v bance, kdy počáteční stav tříd není znám, ale může být získán aplikací jistého algoritmu shlukování; tak se získají shluky objektů a reprezentativní vzory jejich chování. Je-li počet tříd předem determinován, jsou užívány algoritmy segmentace jako je například k-means [19]. Jestliže počet tříd není znám, je třeba jejich počet nalézt a to algoritmy jako jsou například hierarchické shlukování nebo segmentace ve dvou fázích.

Segmentace jako technika rozpoznání vzorů chování [27] je aplikována v mnoha oblastech, jako je například diagnostika a léčení v medicíně, analýzy textu, předpověď

stavu počasí, kontrola procesů, zpracování neurologických signálů, analýza fotografie pozemku, analýza přírodních zdrojů, detekce a klasifikace sonarů, rozpoznání otisků v daktyloskopii a další [14]. Její explozivní využití v současné době registruje oblast Business Intelligence; typickým příkladem jsou nejrůznější strategie segmentace klientů finančně orientovaných institucí.

1.5 Dynamické chování shluků a objektů

V této sekci je podrobně vysvětlen koncept dynamického chování, užití dynamického algoritmu a rozlišeny situace, které závisí na existenci identifikátoru.

Změny ve shlucích jsou způsobeny změnami v některých nebo ve všech proměnných studovaného objektu. Množina klientů banky periodicky informuje o změnách přijmu. Tato informace je aktualizována a proto je to jedna z dynamických charakteristik, která je v průběhu času doplňována do báze dat. Jiné charakteristiky, jako je například věk klientů, jsou aktualizovány automaticky, je-li známo jejich datum narození.

Objekty, které existují v bázi dat, se pojmenují *známé objekty*. Tyto byly v nějakém momentu podrobeny segmentaci a proto je známo, do jaké třídy patří. Objekty, které se agregují do báze dat, se nazývají *agregované objekty*. V každém časovém období jsou hodnoty proměnných obou typů objektů aktualizovány (to vyjadřuje dynamické chování objektů) a tyto mohou být pak nazývány *objekty aktualizované*.

Aby bylo možné rozlišit situaci aktualizace a agregace, je třeba vlastnit charakteristiku *registru*, to znamená adresu, která dovoluje identifikovat objekt. Tento registr je pojmenován *identifikátor*. V případě bankovních klientů je to občanské číslo nebo rodné číslo [6].

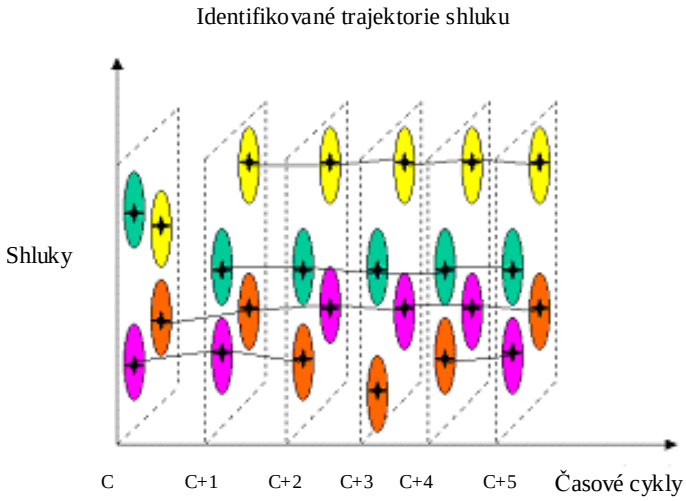
Z pohledu existence identifikátoru je třeba vzít v úvahu dvě alternativy sledování dynamického chování, které mají specifické subproblémy a vyžadují rozdílná zacházení; jsou jimi [5]:

1. Případ, kdy neexistuje identifikátor, proto neexistuje kritérium jak rozlišit mezi objekty agregovanými a známými. V obchodních domech většinou nejsou ukládána jména a identifikace klientů, kteří nakupují, a proto vždy, kdy jde klient nakupovat, je tento považován za nový objekt, který se může chovat rozdílně než v předchozí situaci

a eventuelně může být spojován s jinými shluky podle typu své denní spotřeby. Při tomto omezení nedochází k aktualizaci proměnných objektů; objekty se pouze agregují. Z konceptuálního pohledu je třeba s těmito podmínkami vytvořit interpretaci změny a docílit zahrnutí efektu změny do vzorů chování.

Analýza spojená s tímto typem problému vede ke ztrátě informace - například ve dvou rozdílných sekvenčních obdobích jsou rozpoznány jako identické třídy ty, které se mezi sebou pohybují jen málo (obrázek 1-5). U objektů není možné definovat trajektorie pohybu. Trajektorie se identifikují podle pohybu tříd.

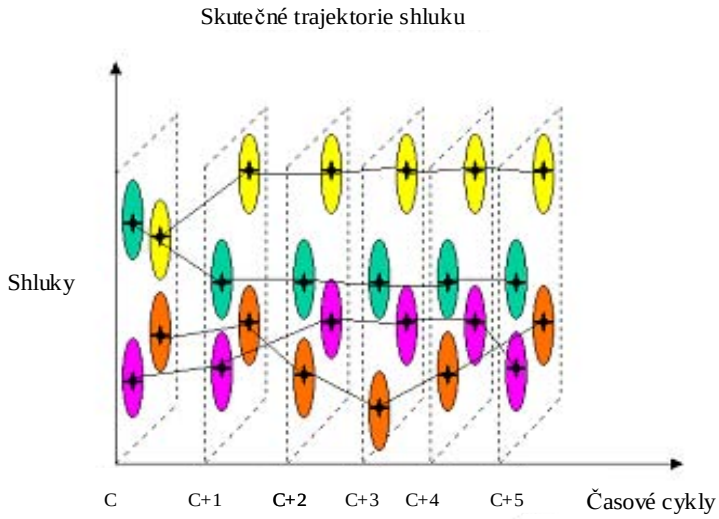
Obr 1-5 Příklad identifikovaných trajektorií vývoje struktury shluků v čase



Poznámka: C vyjadřuje časové cykly, černá spojnice vyjadřuje hypotézu o vývoji shluků v čase.

2. Příklad, kdy existuje identifikátor, tedy paměť registru pro každý objekt (obrázek 1-6). Hodnoty atributů každého objektu je možné aktualizovat a tudíž jsou sledovány skutečné trajektorie pohybu objektů vyjadřující dynamické chování. Příkladem jsou bankovní klienti, kteří mají rodné číslo a každý z nich je charakterizován proměnnými, jako je příjem nebo věk, které se aktualizují v čase. Kromě aktualizace známých objektů je možné sledovat agregované objekty a jejich vliv (změny, které produkují ve struktuře tříd) analyzovat.

Obr 1-6 Příklad skutečných trajektorií vývoje chování objektů a struktury shluků v čase



Poznámka: C vyjadřuje časové cykly, černá spojnice reprezentuje reálné trajektorie.

1.6 Definice dynamického shlukování

Tradiční metody segmentace neuvažují jako vstupní data vývoj objektů v čase [12]. Jejich algoritmy, které mají statický charakter, se aplikují na analyzované objekty v jistém determinovaném okamžiku bez zařazení dynamické formy aktualizace či agregace objektů. Naproti tomu algoritmus segmentace je dynamický, jestliže je schopen dynamickou formou zařadit aktualizace či agregace objektů vyplývající z dynamického charakteru chování objektů pro znovudefinování přirozených shluků.

Pro vytvoření dynamické segmentace je třeba stanovit podmínku, že pro každý objekt i shluk je nutné sledovat vektory proměnných v čase. Proto pod pojmem dynamické shlukování se budou rozumět takové algoritmy, které dovolí sestrojit trajektorie chování v čase[(*)] a na základě jejich analýzy aktualizovat vzory chování shluků, vycházející z aktualizace realizované v předešlém cyklu. V důsledku toho se dynamická segmentace transformuje v nový nástroj shlukování, který dovolí získat reprezentaci vzorů dynamického chování objektů či shluků; jako aplikace významně rozšiřuje kapacitu algoritmů statické segmentace a znalost o datech.

---

* V případě objektů bez identifikátoru trajektorie chování shluku v čase a v případě objektů s identifikátorem navíc trajektorie chování objektů.

Příklady aplikací, kde je důležité uvažovat trajektorie a ne jen jednotlivé body, jsou [3][12]:

- Monitorování pacientů v medicíně, speciálně v intenzivních léčeních, kde sledování jejich stavu a změn v čase je podstatné.
- Klasifikace akcií na burze. Sledování trajektorií cen a dalších charakteristik akcií směřuje k přesnějším odhadům než při uvažování pouhé aktuální ceny akcií.

2 CÍLE

2.1 Obecný cíl

Hlavním cílem je zkonstruovat obecnou metodologii, použitelnou pro analýzu shlukování, založenou na dynamickém chování objektů, která detektuje změny ve shlucích a aktualizuje vzory chování shluků optimálním způsobem. Její realizace bude vycházet ze struktury shluků vytvořené v předchozích obdobích.

2.2 Specifické cíle

- Identifikovat scénáře změny.
- Definovat etapy procesu.
- Vybrat metody a vyvinout kritéria pro detekci změn v hodnotách objektů tak jako i ve složení tříd.
- Implementovat obecnou metodologii na datech simulovaných a skutečných.
- Navrhnout a vyhodnotit způsob srovnání strategie statického shlukování s obecnou metodologií dynamického shlukování.
- Ukázat inovace, výhody a užití navrhované metodologie.

3 METODOLOGIE PRÁCE

Cílem disertační práce je zkonstruovat obecnou metodologii. Proto jako metodologie práce je považován proces vývoje etap procesu a jejich součástí, které vedou k rozhodnutí o aplikaci některého z definovaných scénářů změn.

Metodologie spočívá v návrhu algoritmu, který prostřednictvím metod a kritérií podpoří automatický sekvenční proces aktualizace struktury shluků dat v různých časových cyklech. Důležitost vývoje automatického procesu aktualizace vzorů chování spočívá ve vytvoření nástroje, který by byl schopný dle standardní metodologie analyzovat situaci a zvládnout změny prostředí.

Původní struktura shluků je získána prostřednictvím shlukování dat před první aktualizací či agregací objektů. Události prostředí předpokládají změny ve struktuře shluků, je-li chování objektů známých či agregovaných rozdílné od toho v cyklu předchozím. Relevantní úsilí obecné metodologie proto spočívá v detekci rozdílného chování a docílení adekvátní dynamické aktualizace, vycházející ze známých vzorů chování a to všechno prostřednictvím automatického procesu.

Jelikož cílem práce je vyvinout obecnou metodologii, metodologie práce bude s detaily popisována přímo při jejím vývoji v *sekci 5* a *6.*

3.1 Vymezení práce

S aplikací obecné metodologie se začíná v momentu, kdy se aktualizují objekty (po výběru optimálního počtu shluků); objekty mají identifikátor. Ještě před konstrukcí metod a kritérií užívaných v obecné metodologii je třeba *a priori* zvolit algoritmus segmentace, který je aplikován na objekty při jejich shlukování; transformuje se tak v částečné omezení problému - také volba metod a vyvíjení kritérií jsou omezeny algoritmem segmentace; tento vztah platí i naopak. Konkrétně, ke stanovení optimálního počtu shluků je vybrán algoritmus segmentace ve dvou fázích a k vytváření aktualizace formátů chování shluků je zvolena metoda segmentace k-means; k realizaci metod je využit software SPSS. Jelikož náplní práce není stanovit optimální počet tříd a povaha dat, se kterými se pracuje,

odpovídá zvolené metodě, výběr algoritmu k-means pro vývoj obecné metodologie je v této práci zdůvodněn.

Pro celou analýzu se předpokládá, že se nezařazují nové proměnné (atributy charakterizující objekt) k těm, které byly vybrány na počátku studia. Všechny užité proměnné mají spojitý charakter.

Data užitá na testování obecné metodologie pocházejí ze dvou zdrojů; jedním jsou data simulovaná a druhým jsou data získaná z Bci v Santiagu, Chile – tento zdroj bude s detaily vysvětlen v *sekci 7.2*.

Vyvíjená obecná metodologie je úvodem do dynamické segmentace - má za úkol navrhnout scénáře změn a etapy průchodem těchto scénářů, nemá však za úkol vyčerpávajícím způsobem vyřešit všechna možná řešení ani vybrat nejadekvátnější metodu či kritérium zpracování informace. Úkolem také není programovat algoritmus obecné metodologie.

Je třeba zdůraznit, že záměrem konstrukce obecné metodologie je prakticky ji využít. Proto zvolené postupy, které vedou k rozhodování o aktualizaci vzorů chování shluků (založené především na metodách a kritériích uvnitř etap) jsou konstruovány nejen s úmyslem inovovat algoritmy shlukování Data Miningu, ale především pro účely reálné aplikace obecné metodologie v praxi Business Intelligence.

4 PLÁN PRÁCE

Plán práce se skládá ze dvou částí; první, která zahrnuje permanentní revizi literatury a druhá, plán pokroku práce v čase.

4.1 Permanentní revize literatury

Permanentní revize literatury zahrnuje revize následujících časopisů:

- Analýza inteligence dat,
- IEEE Transactions on K-means,
- IEEE Transactions on Evolutionary Computation,
- IEEE Transactions on Pattern Analysis and Machine Intelligence,
- IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans,
- IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics.

Tyto časopisy představují pokroky v technickém a aplikovaném výzkumu v oblastech spojených s koncepty a algoritmy, které jsou později aplikovány v Data Miningu. Rozvíjejí postupy a metody ve statistice, informatice, umělé inteligenci a dalších odvětvích, které jsou aplikovány při rozpoznání vzorů chování.

4.2 Plán pokroku práce na disertační práci

Plán pokroku práce v čase dovoluje měřit stupeň pokroku v realizaci práce. Jako hlavní úseky zahrnuje následující etapy:

- Pravidelné schůzky s profesory na Universidad de Chile: s profesorem řízení operací a Data Miningu Jaimem Mirandou, profesorem marketingu Maximo Boschem a profesorkou statistiky Nancy Lacourly. Paralelně jsou organizovány schůzky s Luisem Aburtem (ředitelem Data Miningu v Penta/Analytics, firma zabývající se poskytováním služeb Data Miningu a Business Intelligence) a Andreou Romero (konzultorkou firmy SPSS).

- Konzultace některých etap práce s analytiky a modelátory z oddělení Inteligence obchodů, Marketingu a Korporativního Data Warehouse, Bci.
- Korespondence se školitelem doc. Ing. Otakarem Macháčkem, CSc.
- Definice pojmu změny v chování objektů a situace odvozené ze změn.
- Definice metod a kritérií, které dovolí identifikovat změny v chování skupin objektů.
- Definice obecné metodologie, jejích etap procesu a scénářů změn, které logickým sekvenčním způsobem dovolí poznat globální chování objektů a shluků a identifikovat možné změny ve strukturách shluků.
- Vývoj obecné metodologie užitím výše uvedených definic, které dovolí aktualizovat strukturu shluků na bázi dynamického chování objektů.
- Aplikace obecné metodologie na simulovaná data.
- Aplikace obecné metodologie na reálný případ.
- Pozorování chování shluků a obecných charakteristik obecné metodologie a analýza výsledku aplikace obecné metodologie včetně odhadu a sledování tendencí budoucího vývoje.
- Návrh a provedení srovnání a vyhodnocení obecné metodologie a algoritmu statického shlukování.
- Zhodnocení výhod, přínosu a užití obecné metodologie.

Každá etapa plánu práce je doplňována kontrolami referátů o částečných pokrocích.

5 DEFINICE OBECNÉ METODOLOGIE DYNAMICKÉ SEGMENTACE PRO PŘÍPAD OBJEKTŮ S IDENTIFIKÁTOREM

Změny v okolí, které se odrážejí v chování objektů, způsobují změny ve struktuře shluků definovaných prostřednictvím algoritmu segmentace. Proto úsilím při vývoji obecné metodologie je ukázat evidentní dynamické změny v objektech doprovázené změnami v charakteristikách segmentů; jsou jimi například efekty změn středů shluků, příslušnosti objektů do shluků či počet shluků. Tento problém je možné pojmout z pohledu vzorů chování – úsilí pak spočívá v přenesení změn v objektech na reprezentaci vzorů chování získanou předešlou segmentací.

Obecná metodologie, jejímž záměrem je vypořádat se se změnami v okolí a tedy přenést tyto změny do charakteristik shluků, spočívá v sérii sekvenčních etap, které dovolují čelit různým scénářům změn ve vzorech chování shluků. Tyto etapy mají být zkonstruovány z metod a kritérií, které mají efektivně a automaticky pomáhat rozhodovat.

Pro navržení a realizaci metod a kritérií je třeba v každé etapě definovat vstupy, jejichž forma je podřízena například typu dat i etapě, ve kterých budou figurovat. Uvnitř etap je třeba také identifikovat rozdílné možnosti řešení podporované metodami, které pomocí kritérií definovaných v podmínkách etap dovolí vybrat pro každou etapu nejoptimálnější řešení ze všech. Stejně tak je třeba definovat druh výstupu, který bude sloužit jako vstupní informace pro nadcházející etapy.

5.1 Scénáře změn ve vzorech chování shluků

Nezávisle na etapách procesu[(*)], metodách a kritériích řešení je třeba identifikovat scénáře změn, které mohou nastat ve vzorech chování shluků. Identifikované scénáře změn ve shlucích jsou následující (obrázek 5-1):

a) *Zachování počtu shluků*. Objekty setrvávají v některém z existujících shluků. Otázka, kterou je třeba v tomto scénáři zodpovědět je, zda-li bude realizována

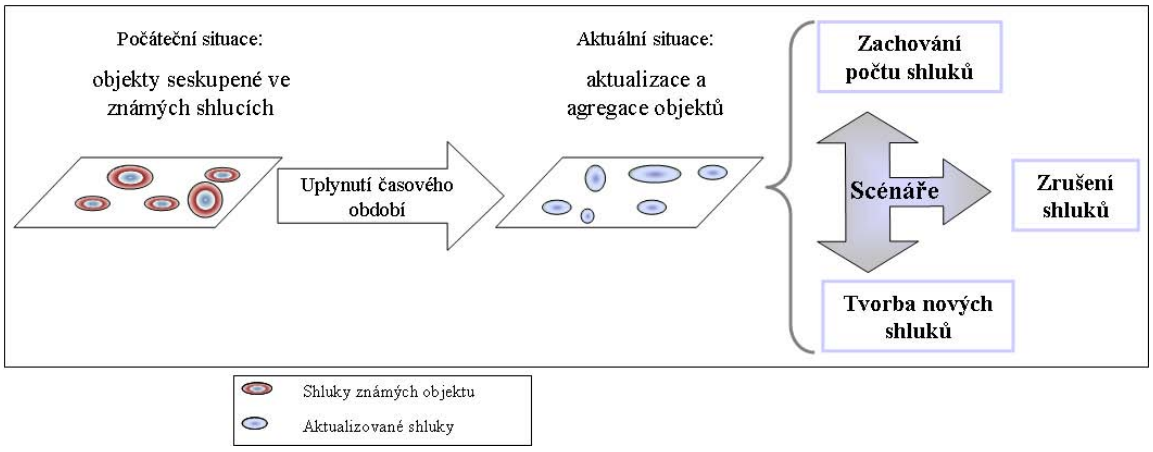
* *Etapa procesu* neboli *etapa identifikace scénářů změn* nebo jen *etapa*.

taková aktualizace shluků, která obratným způsobem využije informace aktualizovaných a/nebo agregovaných objektů; v kladném případě řešení je třeba stanovit způsob aktualizace vzorů chování.

b) *Tvorba nových shluků*. V tomto scénáři se objevují nové segmenty kromě těch již existujících. Proto agregované či aktualizované objekty se zařazují do existujících shluků či do nových shluků. Úsilím je vyřešit, v jakém případě a jak narušit strukturu tříd a jak sloučit již vytvořené vzory chování s těmi novými.

c) *Zrušení shluků*. Neexistují žádné nebo dostatečný počet objektů, které jsou zařazovány do známého shluku. Jedním z úkolů je stanovit v jakém momentu je třeba narušit strukturu shluků vyloučením těch, které počítají s nedostatečným množstvím objektů nebo nepřijímají známé či agregované objekty a co dělat s dodatečnou informací, kterou by měly pojmout eliminované shluky.

Obr 5-1 Aktualizace vzorů chování scénáři změn



Definovaná obecná metodologie uvažuje provedení sekvenčních etap, které dovolují realizovat globální analýzu objektů a dovést analytika k výběru optimálního scénáře změny.

5.2 Etapy identifikace scénářů změn

Dynamika každé etapy obecné metodologie (kromě té poslední) je reprezentována rozhodovacím stromem ilustrovaným v obrázku 5-2. V každé etapě je dána vstupní informace, která je tvořena daty (a jejich omezeními při jejich zpracování a užití v možných

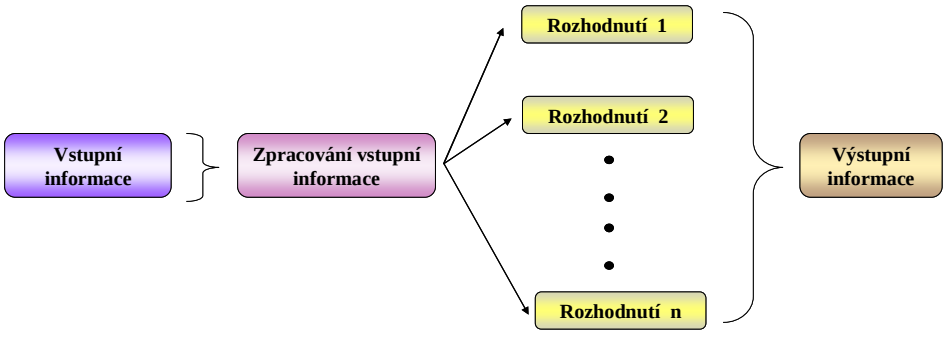
volbách rozhodování); konkrétně, vstupy tvoří charakteristiky objektů ve shlucích jako je například příslušnost a vzdálenost každého objektu k určitému shluku získaného prostřednictvím aplikace algoritmu segmentace nebo hodnoty proměnných objektů v každém sledovaném cyklu.

Po vstupu informace přijde na řadu její zpracování. Toto zpracování může být tak jednoduché jako je počítání objektů, určení pořadí informace či akce složitější jako je aplikace určité metody, která by zpracovala informaci o každém objektu. Metody doprovází podmínky řešení a jejich kritéria.

Jakmile je informace zpracována, nabízí se vějíř různých možností rozhodnutí. V každé etapě (kromě té poslední) je třeba vybrat jednu z nich. K rozhodnutí výběru nejlepší varianty může dojít buď prostým výběrem na základě logických kritérií a nebo je třeba vytvořit určitý index (který bývá uveden v definicích podmínek řešení jako kritérium vytvářené pomocí koeficientu jisté *hranice rozhodování*), který dovoluje srovnávat možnosti mezi sebou tím způsobem, že racionální rozhodovatel se rozhodne pro tu, která nabízí nejlepší či nejvhodnější způsob aktualizace. Kritériem může být například čas, který potřebuje algoritmus na zpracování informací, stabilita pohybu objektů, homogenita segmentů, kvalita segmentace a další.

Výstup z jedné etapy může tvořit vstupní informaci pro další etapu.

Obr 5-2 Schéma dynamiky etapy procesu



Aplikace sekvenčních etap dovoluje detailně poznat situaci a v posloupnosti čelit možným scénářům změn. Tento sekvenční způsob integrace poznání a rozhodování

dovoluje užít jasná kritéria, která jsou jednoduše aplikovatelná a modifikovatelná a dovolují aktualizovat současnou strukturu shluků optimálním způsobem.

Před popsáním jednotlivých etap scénářů změn je vhodné předeslat, že v této sekci je identifikována každá etapa se svým příslušným posláním jak obecně tak již se zaměřením na konkrétně aplikované metody a kritéria v *sekci 7*; definice a implementace každé etapy v algoritmu obecné metodologie je předvedena v *sekci 6 a 7*.

Sekvenční etapy procesu obecné metodologie jsou následující (obrázek 5-3):

I. Etapa: *Identifikace objektů, které představují změnu*. Tato etapa přísluší aplikaci metod a kritérií, které pro každý objekt dovolí rozhodnout, zda-li jeho chování představuje změny v relaci ke struktuře platných shluků ekvivalentně ke známým vzorům chování. Kritéria mají jako vstup hodnoty vzdálenosti objektů ke středu příslušného shluku[(*)]. Je definováno, že objekt představuje změnu, je-li jeho chování heterogenní od chování původních objektů v segmentu. Je stanovena jistá *hranice heterogenity*.

II. Etapa: *Rozpoznání stavu změny*. Vstupy jsou tvořeny počtem objektů ve shlucích, kterým byly identifikovány objekty se změnou. S touto informací je možné analyzovat objekty, které představují změnu, například metodou součtu těchto objektů. Jedna z forem stanovení kritéria na tento příklad je definovat, že nerelevantní změna nastává tehdy, kdy *počet objektů* ve stádiu změny je *nevýznamný* vzhledem k celku objektů a potom tedy relevantní změny ve známé struktuře shluků nastávají v případě, kdy počet aktualizovaných či agregovaných objektů představujících změnu je *významný*. To podněcuje k definici kritéria *hranice změny*.

III. Etapa: *Rozhodnutí o možnostech aktualizace shluků*. Po identifikaci objektů se změnou a po rozhodnutí o rozpoznání stavu změny je třeba rozhodnout o možnostech optimální reprezentace změn detekovaných v objektech vzhledem ke struktuře tříd vytvořené v období předcházejícím aktualizaci. Možnosti, které se berou v úvahu, jsou: zachovat počet tříd a pouze klasifikovat objekty bez

* Příslušným shlukem pro objekt se rozumí jemu nejbližší shluk.

speciálních modifikací současných vzorů chování, což je pojmenováno *mechanická aktualizace* – dochází k ní v případech *entropického chování objektů* ve smyslu jejich pohybu nebo/a směru; jiná alternativa v rámci zachování počtu shluků je aktualizovat využitím informace předešlé a navíc té, kterou přispívají aktualizované objekty a to prostřednictvím exkluzivní heuristiky vhodné pro každý jednotlivý případ; tato forma aktualizace je pojmenována *inteligentní aktualizace.* Inteligentní aktualizace vytváří efekt změn uvnitř shluků nazvaného *pohyb shluků*.

V případě, kdy není identifikován žádný objekt se změnou nebo není rozpoznán stav změny, je volena jedna z výše uvedených alternativ, jejíž výběr závisí na entropii chování všech objektů ve smyslu jejich pohybu a směru. Je definován *neentropický pohyb a směr* a také kritéria a koeficienty *hranice stability objektů.*

V případě, kdy je rozpoznán stav změny, je vybrána buď inteligentní aktualizace nebo druhá možnost, kterou je *tvorba nových shluků,* kdy je zvyšován počet shluků a hledána nová segmentace. Je stanoveno kritérium a koeficient *hranice stability outlierů* a *hranice tvorby nových shluků*.

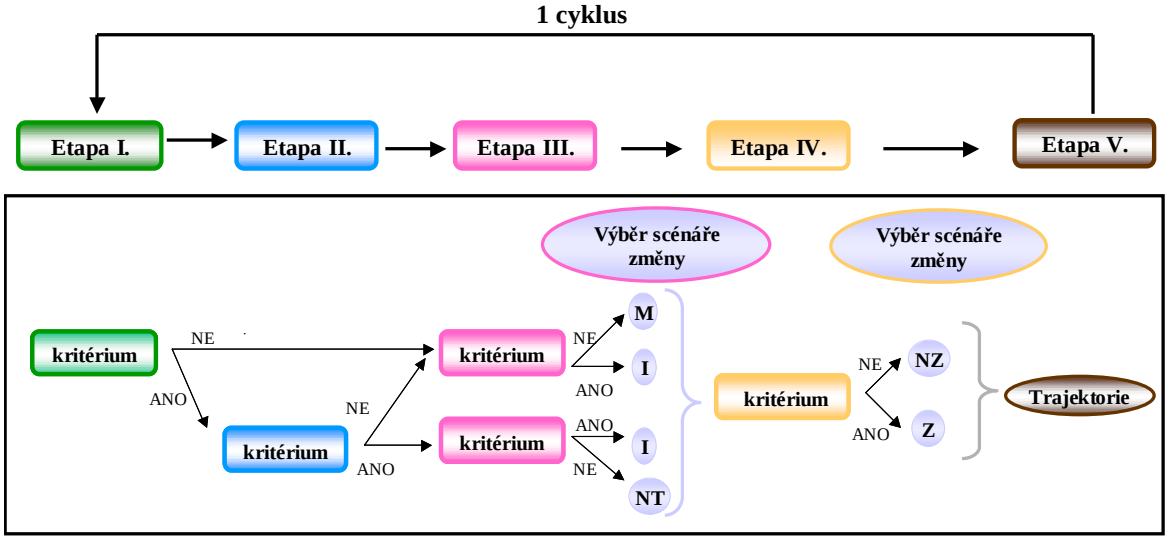
Realizace každé uvažované možnosti je omezena na metody a algoritmus shlukování, které užívá každý modelátor; například pro volbu tvorby nových shluků (jejich počtu) je možné užít algoritmus, který automaticky, na základě dat, vybere optimální uspořádání tříd a najde příslušnost objektů do nových tříd. Stejně tak v případě pohybu tříd existují metody, které dovolí docílit optimálního rozložení stávajícího počtu segmentů.

Detailní charakteristiky vstupů, metod a kritérií, které je třeba pro etapu definovat, jsou vysvětleny v *sekci 6.3.*

IV. Etapa: *Zánik shluků*. Počet objektů spolu se znalostí reálných trajektorií objektů a shluků slouží jako vstupní informace a kritérium pro zrušení tříd, které nesplňují podmínky pro *přežití* segmentu. V této práci je třída kandidátem na své vyloučení, nemá-li dostatečný počet objektů a zároveň nebyl-li jí přirazen po jistý počet cyklů aktualizace vzorů chování žádný objekt. Jsou proto definovány *hranice minimálního možného počtu objektů* a *hranice času pozorování.*

V. Etapa: *Identifikace trajektorií*. Cílem této etapy je identifikovat novou strukturu shluků s tou bezprostředně předcházející a to konstrukcí *reálných trajektorií pohybu*, které ukazují, které třídy a jakým způsobem byly aktualizovány. Vstupy jsou tvořeny předchozí a novou strukturou tříd získanou prostřednictvím algoritmu obecné metodologie. V průběhu času je možné sledovat tendenci a stabilitu pohybu shluků, což je důležitá informace pro realizaci předpovědí do budoucnosti.

Obr 5-3 Etapy procesu obecné metodologie s indikací rozhodování dle kritérií pro některý z typů aktualizace vzorů chování



Poznámka: M je mechanická aktualizace, I je inteligentní aktualizace, NT je tvorba nových tříd, N je zrušení shluků, NZ je negace zrušení shluků.

Aplikace obecné metodologie začíná v momentu, ve kterém se realizuje aktualizace či agregace objektů. Konkrétně, vytvořená obecná metodologie je aplikována na konci každého cyklu (po klasifikaci objektů) pro každý shluk zvlášť, kdy je analyzován stav a chování objektů a v rámci scénářů změn vybírány formy aktualizace formátů chování shluků.

Cyklus sledování objektů a segmentů v rámci obecné metodologie dynamické segmentace je ukončen realizací poslední, páté etapy. Tato etapa je souhrnem a výsledným stavem aplikace dynamické segmentace a tudíž všechna nutná rozhodování jsou provedena již před touto etapou (během prvních čtyř etap). V etapě III. je nalezena nejvhodnější

alternativa aktualizace shluků. Neprojde-li analytik etapou IV., nemůže být dokončen výběr scénářů změn a aktualizace vzorů chování[*]. Jak bude ukázáno v *sekci 6 a 7*, scénář Zachování počtu shluků či Tvorba nových shluků se navzájem vylučují, nevylučují však sledování a také následnou aplikaci třetího možného scénáře změny, kterým je Zrušení shluků.

[*] Termín *aktualizace shluků* se užívá pro tři alternativy aktualizace shluků definované v etapě III. v rámci dvou scénářů změn. Termín *aktualizace vzorů chování* shluků je termín obecný, který zahrnuje aplikace všech definovaných scénářů změn.

6 VÝVOJ OBECNÉ METODOLOGIE DYNAMICKÉ SEGMENTACE PRO PŘÍPAD OBJEKTŮ S IDENTIFIKÁTOREM A SHLUKOVÁNÍ REALIZOVANÉ PROSTŘEDNICTVÍM ALGORITMU K- MEANS

Tato sekce představuje vývoj obecné metodologie[*] pro případ objektů s identifikátorem aplikací algoritmu k-means.

Je připomínáno, že obecná metodologie je úvodem do dynamické segmentace - má za úkol navrhnout scénáře změn a etapy průchodem těchto scénářů, nemá však za úkol vyčerpat všechna možná řešení.

6.1 Etapa I: Identifikace objektů, které představují změnu

Změna pozice fyzického objektu je detektována v případě, je-li objekt přemístěn vzhledem k jistému setrvačnímu systému, který se pohybuje stejnou rychlostí jako systém reference sledovatele [12]. Stejným způsobem je možné vytušit, že objekty představují stádium změny, přemístí-li se vzhledem k určitému referenčnímu systému modelovaného jistou aplikací. Referenční systém je charakterizován strukturou známých tříd před inkorporací jakýchkoliv změn. Je stanoveno, že objekt je ve stavu změny vzhledem ke stanovené referenci, splní-li definovaná kritéria změny.

Původní shluky jsou tvořeny známými objekty, jelikož vytváří strukturu shluků, která je známá a známý je i koeficient jejich příslušnosti do shluků, což je jejich vzdálenost ke středům shluků. Shluky, tvořené známými objekty, jsou charakterizovány průměrnými hodnotami všech atributů objektů před momentem, kdy se produkuje první aktualizace známých objektů či agregace nových objektů. Celek známých objektů má velikost N_1 a celek agregovaných objektů má velikost N_2. Celkový počet objektů je N ($N = N_1 + N_2$).

Je definována $\mathbf{X}$ jako matice N_1 *x* M hodnot známých objektů, kde každá řádka obsahuje objekty a sloupce vyjadřují hodnoty proměnných objektu $\mathbf{X}_{i\bullet}$, jejichž příslušnost do tříd je v případě algoritmu k-means exkluzivní, tedy každý objekt je zařazen právě do

[*] Vývoj obecné metodologie je metodologií práce. Při jejím postupném sledování je vhodné konzultovat *souhrn symbolů, definic a parametrů* z *přílohy 11.2.*

jedné třídy. Potom třídy nemohou mít žádný objekt společný a všechny třídy společně musí tvořit celek objektů [12]. $\mathbf{Y}$ je matice N_2 *x* M agregovaných objektů, kde každá řádka obsahuje objekty a sloupce vyjadřují hodnoty proměnných objektu $\mathbf{Y}_{i\bullet}$; jsou to objekty agregované k celku známých objektů a tyto dva typy objektů mohou představovat změnu vzhledem k existující struktuře tříd[(*)].

Je stanoveno, že známý objekt $\mathbf{X}_{i\bullet}$ je reprezentován vektorem proměnných v M dimenzích, pak $\mathbf{X}_{i\bullet} \in \Re^M$. Tento známý objekt přispívá svými změnami k aktualizaci informace ve třídách tak jako objekt $\mathbf{Y}_{i\bullet} \in \Re^M$, který se agreguje do množiny $\mathbf{X}$. Je třeba měřit, jak odlišný je aktualizovaný známý či agregovaný objekt vzhledem k existujícím třídám a také vyřešit problém jak aktualizovat tímto novým přídělem informace, což přísluší problému aktualizace ve vzorech chování.

Shluk je označen S_h, kde $h \in \{1,.....,S\}$; jeho struktura je tvořena vektory známých objektů z matice $\mathbf{X}$, po agregaci také hodnotami objektů z matice $\mathbf{Y}$. Celek shluků je označen $CS = \{S_h / h \in \{1,...,S\}\}$, střed shluků pak s_h. Metoda k-means je založena na zařazení každého objektu do shluku, jehož střed se nachází nejblíž. Střed shluku je definován jako M rozměrný bod, který vznikl průměrováním hodnot každého objektu segmentu v každé dimenzi [22]. Je počítána vzdálenost $d(s_m, s_n)$ mezi středy tříd S_m a S_n, $\forall m,n \in \{1,...,S\}$, kde $m \neq n$. Vzdálenost mezi známým $\mathbf{X}_{i\bullet}$ či agregovaným objektem $\mathbf{Y}_{i\bullet}$ a středem shluku je definována jako $d(\mathbf{X}_{i\bullet}, s_h)$ nebo $d(\mathbf{Y}_{i\bullet}, s_h)$. Vzdálenosti jsou v této práci vypočteny *Eukleidovou vzdáleností* (v *příloze 11.1* jsou uvedeny ostatní metody výpočtu vzdáleností a případy jejich použití), druhou odmocninou ze sumy čtverců rozdílu mezi hodnotami proměnných objektů a středu shluku (či středů shluků navzájem). V dvourozměrném prostoru je to přepona pomyslného trojúhelníku, jehož odvěsny se získají jako rozdíly hodnot proměnných z kontrastovaných objektů [12]. Obecná forma zápisu vzdálenosti mezi objektem $\mathbf{X}_{i\bullet}$ a středem shluku s_h ve formě Eukleidovy

---

* Objekty $\mathbf{X}_{i\bullet}$ a $\mathbf{Y}_{i\bullet}$ jsou ve skutečnosti vektory objektů *i* přes všechny proměnné (zde pojmenovány pouze objekty).

vzdálenosti je následující:

$$d(\mathbf{X}_{i\bullet}, s_h) = \sqrt[2]{\sum_{j=1}^{M} (\mathbf{X}_{ij} - \mathbf{X}_{hj})^2} \quad .$$

Příslušnost objektu $\mathbf{X}_{i\bullet}$ k příslušnému shluku S_{h^*} je označena a definována následovně: $\mathbf{X}_{i\bullet} \in S_{h^*}$, jestliže $d(\mathbf{X}_{i\bullet}, s_{h^*}) < d(\mathbf{X}_{i\bullet}, s_h)$, kde s_{h^*} je střed konkrétního segmentu a s_h je střed jakéhokoliv segmentu (platí i pro $\mathbf{Y}_{i\bullet}$).

6.1.1 Definice(*) identifikace objektů představujících změnu

Definice 1: Identifikace objektů, které představují změnu.

Jako objekty, které představují změnu, jsou identifikovány objekty známé i agregované, které splňují podmínku 1.

Podmínka 1: ***definice kritéria hranice heterogenity***

Je dán shluk S_h, cyklus C_c a období t_l. Je definováno $d_{\max_{s_h}^{t_{poc}^c}} = \max_i d\left\{\mathbf{X}_{i\bullet}^{t_{poc}^c}, s_h^{t_{poc}^c}\right\}$.

Potom $\forall S_h, \forall \mathbf{X}_{i\bullet} \in S_h$ jsou objekty představující změnu identifikovány podmínkou 1:

$$\text{Podmínka} \quad 1: \quad d(\mathbf{X}_{i\bullet}^{t_{konc}^c}, s_h^{t_{poc}^c}) > d_{\max_{s_h}^{t_{poc}^c}} \quad .$$

Vzdálenost $d_{max_{s_h}^{t_{poc}^c}}$ reprezentuje hranici heterogenity. Je to vzdálenost nejvzdálenějšího známého objektu $\mathbf{X}_{i\bullet}$ ke středu s_h svého příslušného shluku S_h v období před modifikací t_{poc}. $d(\mathbf{X}_{i\bullet}^{t_{konc}^c}, s_h^{t_{poc}^c})$ je vzdálenost aktualizovaného $\mathbf{X}_{i\bullet}$ (nebo agregovaného $\mathbf{Y}_{i\bullet}$) objektu v období t_{konc} od středu s_h příslušného shluku S_h vytvořeného v t_{poc} (pojmy cyklus a období jsou vysvětleny v *sekci 6.3*). Podmínka 1, vztahující se k první definici,

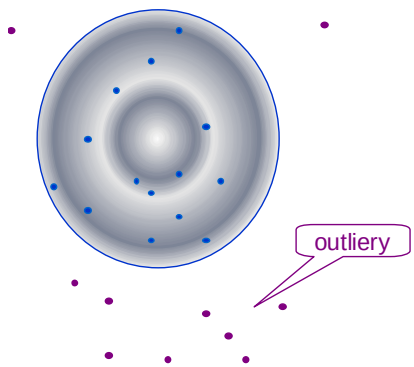
---

* Všechny definice se vztahují k poslednímu období každého cyklu. Jelikož záměrem je nepřehltit text a tak nerozptýlit uživatele této disertační zatížením symbolů všemi indexy, které by tyto měly v přísném režimu vlastnit, nejsou všechny koeficienty do textu zařazeny s výjimkou případů, kdy analitik považuje přesný zápis za nezbytně nutný. Toto ustanovení se vztahuje na celou disertační práci.

vyplývá z faktu, že aktualizovaný nebo agregovaný objekt přestane příslušet do shluku a tak představuje změnu, překročí-li v hodnoceném období jeho vzdálenost od středu shluku hranici nejvzdálenějšího známého objektu od středu téže třídy sledované na počátku cyklu.

Aktualizované a agregované objekty se v posledním období aktuálního cyklu klasifikují do nejbližších existujících shluků. Až teprve po zařazení objektů se provádí aplikace kritérií stanovených podmínkou 1: jsou počítány vzdálenosti mezi objekty a středy příslušných shluků a porovnávány s $d_{\max}^{t_{poc}^{c}}{}_{s_h}$. Objekty, jejichž vzdálenost ke středu shluku přesahuje hranici heterogenity, jsou objekty představující změnu a pojmenovány *outliery* (zobrazeny v obrázku 6-1). Outliery jsou zapisovány $\mathbf{X}_{i\bullet}^{o\ (*)}$.

Obr 6-1 Objekty shluku



Outliery jsou objekty, které nemají stejné chování jako zbytek pozorování; jsou to objekty s heterogenním chováním. Nejsou typickými objekty pro populaci. Analytik může outliery eliminovat nebo je může zařadit do analýzy, jelikož mohou obsahovat reprezentativní interpretaci jistého dominantního segmentu. V této etapě bude každý outlier - determinovaný podmínkou z definice 1 - zařazen do analýzy a vyhodnocen v dalších etapách. Existence outlieru je v posloupnosti obecné metodologie první známkou možné změny ve struktuře shluku.

---

* Ta samá podmínka platí i pro objekty $\mathbf{Y}_{i\bullet}$ a jejich outliery, zapisované $\mathbf{Y}_{i\bullet}^{o}$.

Vstupy pro tuto etapu jsou tvořeny hodnotami vzdáleností všech objektů od středů svých příslušných shluků před i po modifikaci příslušnosti objektů ke shlukům. Metodou pro vyhodnocení kritéria na identifikaci objektů se změnou je výpočet a porovnání Eukleidových vzdáleností.

V této etapě existují dvě možnosti rozhodnutí. V případě, že nejsou identifikovány žádné objekty, které představují změnu, dochází k mechanické či inteligentní aktualizaci shluků (přechází se do *sekce 6.3).* V případě identifikace objektů se změnou se postupuje do *sekce 6.2*.

6.2 Etapa II: Rozpoznání stavu změny

Dalším úkolem ve vývoji obecné metodologie je návrh metodiky, která na základě analýzy heterogenních objektů rozpozná stádium změny ve struktuře existujících tříd.

Ne automaticky je informace o objektech, které jsou identifikovány jako objekty ve stavu změny, zařazena do vzorů chování. Jelikož úsilím je udržet co možná nejstabilnější marketingovou kampaň (jako jakýkoliv proces post segmentace, který je definován na základě charakteristik shluků), záměrem je provádět pouze nevyhnutelné změny ve vzorech chování; zároveň je však nutné respektovat reálné a podstatné změny v okolí a přenášet je do charakteristik shluků.

Stádium změny ve struktuře shluků je rozpoznáno v tom případě, je-li *počet outlierů* vzhledem k celkovému počtu objektů *významný*. Počet outlierů v segmentu S_h bude označen $e^o_{S_h}$. V tomto případě budou upřednostňovány záměry marketingu (před statistickými kritérii analýzy) a bude uvažována změna marketingové kampaně a tedy i změna v charakteristikách shluků v případě, budou-li náklady na změnu vykompenzovány výnosností kampaně[*].

Jedna z forem, jak vyjádřit kritéria rozhodování této etapy je pomocí podmínky 2.

[*] Předpokládá se, že výnosnost kampaně bude v podstatné míře záviset na objemu reakcí.

6.2.1 Definice rozpoznání stavu změny

Definice 2: Rozpoznání stavu změny.

Je rozpoznán stav změny, jestliže počet objektů představujících změnu vzhledem k celkovému počtu objektů ve shluku je větší než jistá hranice změny.

Podmínka 2: ***definice kritéria hranice změny***

$\forall S_h, \forall \mathbf{X}^o_{i\bullet} \in S_h$ platí:

$$\text{Podmínka } 2: \quad \frac{e^o_{S_h}}{N_{S_h}} > \alpha, \quad \text{kde} \quad 0 < \alpha < 1.$$

Koeficient α představuje hranici změny.

Definice je vytvořena na základě faktu, že ke skutečné změně nedochází, neexistuje-li dostatečná příčina změny vzorů chování, tedy dostatečný počet objektů se změnou.

Pro podmínku je třeba definovat koeficient α, kterým může být například 20%: je-li v segmentu počet outlierů vyšší než 20% všech analyzovaných objektů, tj. objektů známých či agregovaných včetně outlierů, je splněna podmínka 2 a rozpoznán stav změny.

K podmínce je nutné dodat, že ne ve všech proměnných, charakterizujících objekt, musí docházet ke změně. Jsou-li objekty charakterizovány mnoha proměnnými, jsou vybrány jen některé z nich, které mají určující význam ve sledování objektů. Ve většině úloh je segmentace tvořena nízkým počtem atributů a tak nepřináší chaos do segmentace. Existují případy, kdy shluk je charakterizován mnoha atributy, z nichž většina je považována za doprovodné - při sledování pohybu shluků jsou zanedbané, jelikož při tvorbě následných procesů je třeba se soustředit na jednoznačná kritéria. Bude záviset na marketerovi, analytikovi a aplikaci, jak bude v takovém případě rozhodnuto (toto tvrzení se týká i rozhodování v *sekci 6.3*).

Vstupy pro tuto etapu jsou tvořeny počtem všech objektů a zvlášť outlierů. Jsou analyzovány pouze shluky vlastnící objekty ve stavu změny. Metodou na stanovení podmínky 2 je prostý součet objektů se změnou a celkového počtu objektů pro každý segment.

Při nesplnění podmínky 2 přichází v úvahu inteligentní nebo mechanická aktualizace shluků. Je-li splněna podmínka 2, je potvrzen stav změny, tedy změny ve formátech chování. To však automaticky neznamená, že je nutné zvýšit počet tříd, jelikož ten aktuální může dobře reprezentovat novou informaci. Na vyřešení navržených problémů formy aktualizace segmentů je nutné postoupit do následující, třetí etapy.

6.3 Etapa III: Rozhodnutí o možnostech aktualizace shluků

V této etapě je vybrán jeden ze dvou scénářů změn: Zachování počtu shluků nebo Tvorba nových shluků. Rozhodnutí spočívá ve výběru optimální možnosti aktualizace shluků založené na daných kritériích přiřazených ke každé možnosti.

Celá myšlenka definovat různé formy aktualizace formátů chování vyplývá ze zahrnutí konceptu dynamiky chování objektů ve shlucích. Vzory chování definovaných shluků nejsou porušeny, není-li na základě jistých kritérií prokázána tendence v pohybu a směru *významného počtu objektů*. V případech, kdy je detektován *významný počet objektů*, které vykazují neentropický pohyb a směr, mění se vzory chování segmentu. Navíc, je-li chování *významného počtu objektů* shluku heterogenní bez prokázání evidentní tendence v chování povšechného[(*)] počtu objektů stabilních v pohybu, je třeba tyto objekty oddělit a vyjádřit tak rozdílnost jejich chování porušením aktuálních tříd a vytvořením tříd nových.

Výše popsané v podstatě zahrnuje analyzovat objekty ve smyslu heterogenity a entropie - pohybu a směru - jejich chování.

6.3.1 Definice aktualizace klasifikací do tříd

V rámci možností zachování počtu shluků je možné zvolit mechanickou aktualizaci, která spočívá v klasifikaci agregovaných a aktualizovaných objektů do příslušných shluků, přičemž nedochází ke změně středů shluků. Takto je rozhodnuto pro shluky, kterým nejsou identifikovány objekty se změnou a zároveň povšechné chování objektů v těchto shlucích je *chaotické*. Druhý případ, kdy se analytik přiklání k mechanické aktualizaci, je

* Termín *povšechný* zahrnuje do analýzy všechny objekty shluku, tj. outliery i objekty uvnitř shluku.

v případech, kdy jsou identifikovány objekty se změnou, ovšem jejich *počet* vzhledem k celku objektů není *významný* (tzn. není rozpoznán stav změny) a zároveň povšechné chování objektů v tomto shluku je entropické.

Podmínkou pro volbu mechanické aktualizace je konstatování o chování objektů: je entropické (chaotické). Zde přichází na řadu dynamický pohled na segmentaci. Jelikož podmínky prostředí se mění a chování objektů (klientů na finančních trzích) je převážně dynamické, je třeba provádět jejich analýzu v průběhu času. Fotografie stavu objektů v jednom okamžiku přináší do informací o chování objektů náhodnou složku. Ať slouží jako názorný příklad transakce klientů v bance: jistý klient provádí každý měsíc v průměru 20 bankovních transakcí. Přijde měsíc, ve kterém cestuje do zahraničí a v zemi své rezidence nemá aktivní chování; nevykazuje žádnou transakci. V tomto případě by bylo nesprávné vzít tento nulový údaj jako údaj jedné proměnné pro aktualizaci segmentace, jelikož chování klienta v tom okamžiku bylo výjimečné a nepodobá se jeho typickému chování.

Pro definování chování objektů, stability jejich pohybu a směru, je třeba provádět sledování po trajektoriích po jistá období, která tvoří cykly. Jeden cyklus vyjadřuje čas, po který je zkoumána dynamika objektů. Na jeho konci je provedena příslušná aktualizace vzorů chování. Matematický zápis pro cyklus je C_c, kde $c \in \{1,...,C\}$; pro období t_l cyklu C_c je to t_l^c. Cyklus je složen z období: $C_c = \{t_k^c / t_1^c \le k \le t_{konc}^c\}$. Celek období je zapisován $CT = \{t_l^c / l \in \{1,...,konc\}\}$ a celek cyklů $CC = \{C_c / c \in \{1,.....,C\}\}$. Dynamickou segmentaci je možné provozovat pro jakýkoliv počet cyklů, který je dán konkrétním případem a podmínkami zpracování. Stejně tak délka období je dána odvětvím, které je analyzováno a dispozicí a typem dat, které definují jednotlivé proměnné. Přijatelné období z hlediska pořízení dat a možných změn ve většině finančně zaměřených firem je jeden měsíc. V *sekci 7*, aplikace na data simulovaná a skutečná, je sledován pohyb objektů měsíčně a trimestrálně vyhodnocován. První cyklus C_1 je tvořen obdobím t_{poc}^1, což je stav před jakoukoliv modifikací a je ukončen obdobím t_3^1, což je poslední období prvního cyklu(*).

---

* V prvním cyklu je $t_{poc}^1 = t_1^1$. V dalších cyklech je $t_{poc}^{c+1} = t_{konc}^c$. t_{konc}^c je období t_3^c po aktualizaci shluků.

Je nutné definovat neentropický pohyb.

Definice neentropického pohybu(*)

$\forall S_h, \forall \mathbf{X}_{i\bullet} \in S_h, \forall j,$ kde $CT = \left\{ t_l^c / l \in \{1,...,konc\} \right\}$ a $\forall t > \left(\frac{t_{konc}^c}{2} \right)^*$ je definován pohyb(**):

a) $\mathbf{X}_{ij}^{t_{l+1}^c} - \mathbf{X}_{ij}^{t_l^c} > 0$ jako *stabilní pozitivní,*

b) $\mathbf{X}_{ij}^{t_{l+1}^c} - \mathbf{X}_{ij}^{t_l^c} < 0$ jako *stabilní negativní,*

c) $\mathbf{X}_{ij}^{t_{l+1}^c} - \mathbf{X}_{ij}^{t_l^c} = 0$ jako *stabilní konstantní.*

Poznámka: $\left(\frac{t_{konc}^c}{2} \right)^*$ *je počet uvažovaných období k definici neentropického chování; je definován následovně:*

- $\left\lceil \frac{t_{konc}^c}{2} \right\rceil$, *jestliže* t_{konc}^c *je liché číslo,*
- $\frac{t_{konc}^c}{2} + 1$, *jestliže* t_{konc}^c *je sudé číslo .*

V praktickém užití je možné konstantní stav zařadit buď ke stabilnímu pozitivnímu nebo stabilnímu negativnímu pohybu. Posuzovány jsou objekty známé i agregované, $\mathbf{X}_{i\bullet}$ i $\mathbf{Y}_{i\bullet}$. Definice je platná i pro outliery. Je-li objekt stabilní v pohybu, bude označen $\mathbf{X}_{i\bullet}^{s}$. Objekt nestabilni v pohybu $\mathbf{X}_{i\bullet}^{ns}$ je objekt s *entropickým* neboli *chaotickým pohybem*.

Je také vytvořena definice pro trajektorie pohybu skupin objektů, jelikož stabilní objekty mohou vykazovat - tak jak je vyjádřeno šipkami v obrázku 6-2 III. (dvourozměrný prostor) - stabilní a) pozitivní, stabilní b) negativní, stabilní konstantní nebo kombinovaný c) a d) pozitivní a negativní pohyb v celku proměnných a pohybovat se tak sice stabilně, ovšem všemi směry.

---

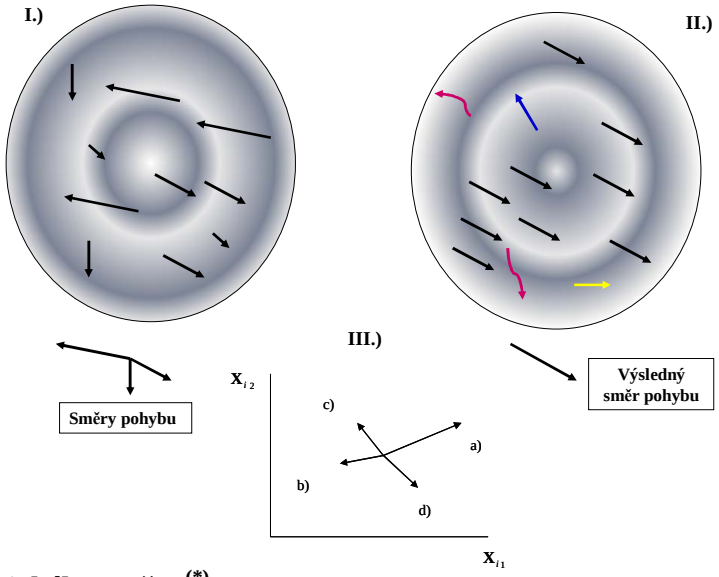
* Neentropický, zde i stabilní nebo tendenční.

** V případě konstanty je definován stav.

Jsou-li všechny objekty, zakreslené v obrázku 6-2 I., stabilní v pohybu, pak pouhým pohledem na směry pohybu objektů je možné konstatovat, že objekty mají entropický směr, jelikož se pohybují všemi směry, aniž by existovala skupinka objektů vykazující převažující směr pohybu.

V obrázku 6-2 II. jsou pozorovatelné dva objekty nestabilní v pohybu (červené šipky). Dále je sledována skupinka objektů s převažujícím směrem pohybu (černé šipky); existuje jeden stabilní objekt (modrá šipka), který se pohybuje v jiném směru, než je trajektorie identifikované většinové skupinky objektů. Je pozorovatelný i jeden konstantní objekt v jedné proměnné (žlutá šipka).

Obr 6-2 Entropie chování objektů



Definice neentropického směru(*)

$\forall S_h$, $\forall \mathbf{X}^s_{i\bullet} \in S_h$, $\forall j$, kde $CT = \{t^c_l / l \in \{1,...,konc\}\}$ a $\forall t > \left(\frac{t^c_{konc}}{2}\right)^*$ je definována matice trajektorií $\mathbf{U}^c$, jejíž prvky jsou vyjádřeny následovně:

$$\mathbf{U}^c_{ij} = \begin{cases} 1 & , \quad \textit{je-li pohyb stabilní pozitivní,} \\ 0 & , \quad \textit{je-li pohyb stabilní konstantní,} \\ -1 & , \quad \textit{je-li pohyb stabilní negativní.} \end{cases}$$

* Neentropický, zde i stabilní nebo tendenční.

Pro $\forall \mathbf{X}_{i\bullet}^{s}$ je definován vektor $\mathbf{U}_{k\bullet}^{c}$ jako trajektorie typu *k*. Pod trajektorií se rozumí celek hodnot $\mathbf{U}_{ij}^{c}$ v jednotlivých obdobích (trajektorie představují řádky v matici $\mathbf{U}^{c}$). Dodatečně je definován celek trajektorií jako $K = \{k / k \ \textit{je trajektorie}\}$. Počet $\mathbf{X}_{i\bullet}^{s}$ shluku S_h, které mají trajektorii typu *k*, je vyjádřen jako $e_{S_h}^{s,k}$; navíc $E_{S_h}^{s} = \max_{k \in K}\{e_{S_h}^{s,k}\}$ reprezentuje největší počet $\mathbf{X}_{i\bullet}^{s}$ shluku S_h se stejnou trajektorií; jejich trajektorie (a také trajektorie shluku S_h) je označena $\mathbf{U}_{S_h}^{c}$. Skupinky objektů mající stabilní pohyb a pohybující se po stejné trajektorii mají neentropický směr.

Definice 3: Aktualizace klasifikací do tříd.

Mechanická aktualizace shluků je aplikována v případech, kdy analýza chování objektů ve shlucích prokáže povšechnou entropii pohybu a/nebo směru objektů a zároveň buď nejsou identifikovány objekty se změnou nebo není rozpoznán stav změny.

Je vytvořena následující podmínka:
Podmínka 3: ***definice kritéria hranice nestability objektů***
$\forall S_h, \forall \mathbf{X}_{i\bullet}^{s} \in S_h$ platí:

$$\text{Podmínka } 3: \ \frac{E_{S_h}^{s}}{N_{S_h}} \le \beta, \quad \text{kde} \quad 0 < \beta < 1.$$

Koeficient β přísluší hranici stability objektů.

Aby byla splněna definice stabilního chování, objekty shluku se musí pohybovat stabilně, ne chaoticky. Nezávisí na velikosti pohybu, tedy vzdálenosti od stavu v předešlém období, ale na tendenci pohybu. Vykazuje-li objekt konstatní stav nebo pohyb mající buď stálou pozitivní a nebo stálou negativní tendenci a to ve více než v například 50% stanoveného počtu posledních navazujících obdobích cyklu, pak je splněna první část definice a zároveň je definován stabilní, neentropický pohyb objektů. Například, jsou-li objekty sledovány v průběhu jednoho cyklu, který je tvořen třemi obdobími, proměnnými

transakce a saldo účtu, pak v případě, že klient po všechna období zvyšuje (snižuje) své transakce a snižuje (zvyšuje) své saldo nebo vykazuje konstantní stav, je splněna definice stabilního pohybu. Pohyb může být jak stejno tak protisměrný. Definicí stabilního směru jsou identifikovány skupinky objektů stabilních v pohybu pohybujících se stejným směrem(*).

Neexistuje-li žádná skupinka, která by vlastnila objekty stabilní v pohybu pohybující se v jednotlivých proměnných (pohyb může být jak pozitivní, negativní tak kombinovaný) v $\left(\frac{t^c_{konc}}{2}\right)^*$ stejným směrem, jejíž velikost by byla $> \beta * N_{S_h}$, pak je splněna podmínka rozhodující o mechanické aktualizaci. V *sekci 7.1* bude mít koeficient β hodnotu 50%.

Objekty, které byly agregovány v období t_{l+1}, kdy $t_{l+1} > t_l$, jsou sledovány pouze po dobu zbývajících období cyklu. Aktualně agregované objekty, které není možné sledovat ve více než v 50% obdobích cyklu, jsou definovány z defektu jako nestabilní, jelikož není jak prokázat stabilitu jejich pohybu v posledních $\left(\frac{t^c_{konc}}{2}\right)^*$ obdobích.

Dynamickou formou je v posledním období aktuálního cyklu postupováno od posouzení definice stabilního pohybu a směru až k vyjádření podmínky 3 této etapy. Realizace této volby spočívá v klasifikaci každého aktualizovaného či agregovaného objektu do shluku, pro nějž obdrží nejnižší koeficient příslušnosti (vzdálenost), bez zařazení informace spojené se změnami do formátů chování současných shluků:

$$s_{h^*} = \underset{S_h}{\operatorname{argmin}}\{d(\mathbf{X}_{i\bullet}, s_h)\}.$$

Důvodem pro toto rozhodnutí je konstatování, že chaotické chování ve smyslu pohybu a směru není příčinou změny formátů chování a z toho vyplývajících následných procesů, jako je například marketingová kampaň.

---

* V jedné nebo ve více proměnných sledovaných zvlášť nebo zároveň; definice závisí na analytikovi a konkrétním případu.

Je-li splněna podmínka 3, je aplikována mechanická aktualizace. V jiném případě přichází v úvahu další možnost aktualizace shluků, která již předpokládá změnu formátů chování současných shluků. Předpoklad, že aktualizované objekty mají být inteligentně aktualizovány ve známé struktuře tříd bez změny jejich počtu, je možné realizovat prostřednictvím metody jako je například ta, která je popsána v následující *sekci 6.3.2.*

Vstupy pro *sekce 6.3.1* a *6.3.2* jsou tvořeny příslušností objektů ke shlukům včetně hodnot jejich proměnných v každém období cyklu pro všechny shluky, jimž nebyly identifikovány objekty se změnou nebo kterým nebyl rozpoznán stav změny. Metodou pro realizaci definice neentropického pohybu a směru je prostá registrace stability po trajektoriích pro každý objekt a jeho proměnné v každém determinovaném období.

Klasifikací objektů v rámci aplikace mechanické aktualizace byl zvolen i jeden ze tří scénářů změn, které mohou nastat ve vzorech chování. To však nebrání v rozhodnutí o relizaci scénáře změny Zrušení shluků, jelikož tento scénář se nevylučuje s rozhodnutím o aplikaci některého ze zbývajících dvou scénářů změny. Proto shluky, pro které bylo rozhodnuto o aktualizaci formou mechanické aktualizace, postupují sekvenčně do etapy IV., Zrušení shluků.

6.3.2 Definice aktualizace pohybem tříd

Je analyzován problém aktualizace shluků, které buď nevlastní outliery nebo u kterých nebyl rozpoznán stav změny a zároveň obě uvedené situace se charakterizují podmínkou povšechného stabilního chování objektů ve shlucích z hlediska pohybu i směru. V tomto případě jsou objekty ve stavu možnosti vnést změny do struktury tříd. Tyto změny mohou být s inteligencí zařazeny do vzorů chování tím způsobem, že zároveň využijí známou strukturu tříd v kombinaci s informací, kterou vlastní a přinášejí aktualizované a agregované objekty. Popsaný způsob změny vzorů chování je nazýván inteligentní aktualizace.

Dle následující definice jsou aktualizovány ty objekty, které splnily opačnou podmínku ze *sekce 6.3.1*.

Definice 4: Aktualizace pohybem tříd.

Inteligentní aktualizace je aplikována na případy, kdy analýza chování objektů shluku neprokáže povšechnou entropii ve smyslu směru a pohybu objektů a zároveň buď nejsou identifikovány objekty se změnou nebo není rozpoznán stav změny.

Podmínka 4: ***definice kritéria hranice stability objektů***

$\forall S_h, \forall \mathbf{X}_{i\bullet}^{s} \in S_h$ platí:

$$\text{Podmínka} \quad 4: \quad \frac{E_{S_h}^{s}}{N_{S_h}} > \beta, \quad \text{kde} \quad 0 < \beta < 1.$$

V *sekci 7.1* je volena hranice 50 % pro koeficient hranice stability objektů z toho důvodu, že 50% stabilních objektů je považováno za kritérium dostatečně *silné*, aby mohlo byl mluveno o pohybu shluků a na druhou stranu dostatečně *volné*, aby mohlo být s objekty v segmentech zacházeno jako s prostředkem metody předpovědi chování shluků do budoucnosti; není třeba čekat na zvýšení hranice stability, nýbrž je třeba včas organizovat a realizovat konkurenční procesy na podpoření či postavení se tendenčnímu chování objektů. Volba koeficientu β však obecně závisí na konkrétně řešeném příkladu.

Předpoklad shlukovacího algoritmu je, že středy tříd reprezentují jejich homogenitu [22] a proto se v popsané situaci nabízí s agregovanými a aktualizovanými objekty realizovat aktualizaci středů shluků. Inteligentní aktualizace středů tříd pro případ k-means je realizována následujícím způsobem: nejprve se všechny objekty klasifikují do příslušných shluků. Poté jsou přepočítány středy shluků - jako nejvhodnější aktualizace, spočívající v pohybu tříd, je ta, která má nejmenší sumu čtverců chyb vzdáleností mezi objekty a středy shluků. Výpočet souřadnic středu shuku je prováděn *metodou nejmenších čtverců* (*MNČ*). Formální zápis MNČ pro segmentaci operující na základě Eukleidovy vzdálenosti je následující:

$$\min\left\{\sum_{i=1}^{N}\left(d(\mathbf{X}_{i\bullet}, s_{h^*})\right)\right\}.$$

Tento výpočet je proveden pro každý shluk a stanoven pro celek shluků, pro které je řešením pohyb tříd[*].

Stejně tak jako po realizaci mechanické aktualizace i zde je nutné následně projít etapou IV a analyzovat možný zánik shluků.

V této chvíli zbývá analyzovat případ, kdy dochází k rozpoznání stavu změny (je prokázána existence dostatečného počtu heterogenních objektů - *sekce 6.2.1* a *6.2.2*). Potom zbývá rozhodnout, zda shluky po rozpoznání stavu změny se budou pohybovat nebo je třeba některé shluky rozdělit a vytvořit tak třídy nové.

Při nerozpoznání stavu změny a splnění podmínky z definice 4 se automaticky aktualizuje inteligentní formou. V úvodu *sekce 6.3* bylo uvedeno, že je-li chování *významného počtu objektů* heterogenní bez prokázání evidentní tendence v chování povšechného počtu stabilních objektů ve shlucích s outliery, je třeba tyto objekty oddělit a vyjádřit tak rozdílnost jejich chování a to porušením aktuálních tříd a vytvořením nových. Rozpoznání stavu změny automaticky neznamená, že by bylo nutné zvýšit počet tříd, jelikož ten aktuální počet může dobře reprezentovat novou informaci. V závislosti na rozhodnutí uživatele, analytika a aplikaci může být volba tvorby nových shluků dle definice kritérií nahrazena jinou možností, která by předpokládala, že jednu z možností aktualizace tříd v případě rozpoznání stavu změny tvoří aktualizace pohybem tříd, která byla popsána v *sekci 6.3.2***.** Je proto třeba postoupit do *sekce 6.3.3* na definice aktualizací a kritérií rozhodování.

6.3.3 Definice aktualizace tvorbou nových shluků

V případech, kdy nebyl rozpoznán stav změny, byly outliery vyhodnocovány společně se zbývajícími objekty shluku. Analýza této etapy se bude chováním outlierů zabývat podrobněji.

---

[*] V tomto případě pořadí objektů v souboru dat může ovlivnit výsledné hodnoty středů shluků nerozhodne-li se pro aktualizaci až po dokončení zařazení všech objektů.

Jelikož se předpokládá, že obecná metodologie dynamické segmentace by měla na základě chování objektů předvídat aktualizace a stav segmentů do budoucnosti, je výhodné pracovat s outliery za jistých podmínek ne jako s objekty heterogenními (v důsledku jejich vzdálenosti ke středu shluku), nýbrž jako s objekty, které svým chováním vyjadřují budoucí chování segmentu (a objektů v něm).

V této etapě bude rozhodováno, zda se shluky s *významným počtem* heterogenních *objektů* budou dělit mezi nové třídy (pro zajištění homogenity chování objektů uvnitř shluků) nebo bude aplikována možnost inteligentní aktualizace. V případě tvorby nových tříd je aktualizace realizována vytvářením nových tříd a přirazováním příslušnosti každého objektu ke svému shluku. Je třeba vyřešit problém optimálního počtu tříd a zařazení objektů do nich. Optimalní počet tříd je až do současnosti tématem výzkumu. Aplikace, které jsou na tento problém připraveny, vlastní kritéria na výběr optimálního počtu tříd (metoda k-means tento algoritmus nevlastní). Kromě zvolené techniky na stanovení optimálního počtu shluků se na tomto rozhodnutí významnou měrou podílí analytik a uživatel. Je třeba také stanovit maximální možný počet shluků, což opět závisí na konkrétním marketingovém případu. Je třeba zdůraznit, že úkolem této práce není stanovit technicky optimální počet shluků. Počet shluků je sledován spíše z praktického hlediska, jelikož uživatel není ochoten ztratit potenciální, nově se objevující celky objektů. Bude-li v *sekci 7* rozhodnuto pro scénář Tvorby nových tříd, shluk bude rozdělen do počtu tříd pouze o jeden stupeň vyšší a zároveň bude stanoven maximální počet shluků: 10. Třída bude vytvořena v případě, bude-li vlastnit přinejmenším 3% známých objektů sledovaných na počátku aktuálního cyklu.

Situace, ve kterých byl rozpoznán stav změny a které jsou v tomto duchu analyzovány, mají jako vstupy shluky s rozpoznaným stavem změny, konkrétně všechny informace, které pomohou vyhodnotit stanovené podmínky a na tomto základě rozhodnout, dojde-li k tvorbě nových tříd nebo bude zvolena aktualizace inteligentní. Metodou a kritérii jsou opět sledování pohybu objektů ve smyslu stability po trajektoriích, nyní však vyhodnoceny způsobem příslušným této etapě, který je následující: shluk může vlastnit outliery pohybující se různým směrem. Jestliže shluk vlastní několik skupinek stabilně se

pohybujících outlierů po různých trajektoriích a alespoň jedna ze skupinek outlierů charakterizující se svou trajektorií stabilně se pohybujících objektů reprezentuje více než například 80% všech outlierů shluku, pak je aplikováno podobné kritérium jako v *sekci 6.3.2* a sledováno, zda více než 50% všech objektů (včetně outlierů) se pohybuje v tom samém směru jako největší skupinka outlierů (mající více než 80% všech outlierů). V případě splnění popsaných podmínek je rozhodnuto ve prospěch pohybu segmentu. Hranice 80% je volena vysoká z toho důvodu, že pohyb segmentu s *významným počtem* heterogenních *objektů* bez vytvoření nových tříd může nastat jen v případě evidentního stabilního pohybu většiny objektů (jak outlierů tak objektů uvnitř třídy) po trajektorii.

V případě, kdy outliery netvoří početnou skupinku více než 80% stabilních a stejnosměrných outlierů shluku nebo tyto nejsou podpořeny stabilním stejnosměrným (mající stejnou trajektorii jako největší skupinka stabilních stejnosměrných outlierů) pohybem více než 50% objektů segmentu, je rozhodnuto pro vytvoření nových tříd.

Objekty, které se pohybují v jiném směru než je pohyb shluků nebo jsou příliš vzdálené nově vytvořeným třídám jsou po definici a realizaci aktualizace všech shluků vyhodnocovány jako případné skutečné(*) outliery. Aby segment nebyl příliš rozšířen, jsou tyto klasifikovány do segmentu, k jehož středu mají po provedení všech aktualizací nejblíž a to aniž by ovlivnily formáty chování těchto segmentů.

(*)*Poznámka: Jedním z problémů metod, založených na minimalizaci rozptylu (k-means), je efekt, který mají atypické objekty na výpočet sumy čtverců chyb. V sekci 6.3 je možné pracovat s outliery v jejich pravém slova smyslu, tedy sledovat, které objekty definované ze sekce 6.1 jako outliery jsou skutečnými outliery (objekty s extrémními hodnotami vzdáleností) a pro výpočet MNČ je možné je z analýzy vyloučit. Je proto nutné zrevidovat data, analyzovat každý outlier a rozhodnout o jeho osudu.*

V dvourozměrném prostoru je možné outliery detektovat například z ***grafu disperse****. Další možností je vytvořit* ***graf boxplot****, je-li úkolem identifikovat vzdálená pozorování pro jisté prioritní proměnné. Pomocí* ***histogramu*** *je možné považovat 5% nejnižších a nejvyšších hodnot jako outliery. V mnohorozměrném prostoru je možné užít* ***Mahalanobisovu vzdálenost****, která je mírou vzdálenosti každého objektu k průměru všech pozorování. Vysoká hodnota tohoto ukazatele identifikuje objekt, který má extrémní hodnoty v jedné nebo ve více proměnných.*

Definice trajektorií outlierů a objektů (stabilních v pohybu) ve shlucích s rozpoznaným stavem změny:

Ať $\mathbf{U}^{o}_{S_h}$ je trajektorie $\mathbf{X}^{o,s}_{i\bullet}$ shluku S_h a $E^{s,o}_{S_h}$ je největší skupina $\mathbf{X}^{o,s}_{i\bullet}$ (z celkového počtu $e^{o}_{S_h}$ outlierů shluku S_h) s trajektorií $\mathbf{U}^{o}_{S_h}$. Je definováno, že $e^{s}_{S_h}$ je počet $\mathbf{X}^{s}_{i\bullet}$ shluku S_h se stejnou trajektorií jako je trajektorie $E^{o,s}_{S_h}$, to znamená $\mathbf{U}^{e^{s}_{S_h}}_{S_h} = \mathbf{U}^{o}_{S_h}$.

Definice 5: Aktualizace tvorbou nových tříd.

Shluky ve stavu změny jsou rozděleny mezi větší počet tříd v případě, kdy pohyb většiny objektů v segmentu je entropický a/nebo není směrově identický pohybu outlierů, které jsou tvořeny skupinkou významného počtu stabilních outlierů majících identickou trajektorii.

Podmínka 5a: ***definice kritéria hranice stability outlierů***

$\forall S_h, \forall \mathbf{X}^{o,s}_{i\bullet} \in S_h$ platí:

$$\text{Podmínka} \quad 5\text{a}: \quad \frac{E^{o,s}_{S_h}}{e^{o}_{S_h}} \le \gamma, \quad \text{kde} \quad 0 < \gamma < 1.$$

Koeficient γ přísluší hranici stability outlierů.

nebo

Podmínka 5b: ***definice kritéria hranice tvorby nového shluku***

$\forall S_h, \forall \mathbf{X}^{s}_{i\bullet}, \forall \mathbf{X}^{o,s}_{i\bullet} \in S_h$ platí:

$$\text{Podmínka} \quad 5\text{b}: \quad \frac{E^{s,o}_{S_h} + e^{s}_{S_h}}{N_{S_h}} \le \delta, \quad \text{kde} \quad 0 < \delta < 1.$$

Koeficient δ přísluší hranici tvorby nových shluků.

Je-li splněna podmínka 5a nebo 5b definice 5 - eventuelně mohou být splněny obě podmínky současně - je zvýšen počet tříd. Metodou tvorby nových středů tříd, tedy nových vzorů chování, je opět MNČ, která byla popsána v *sekci 6.3.2*; v aplikaci k-means je definován determinovaný počet tříd a metodou minimální sumy čtverců chyb je stanoveno optimální rozložení středů shluků.

V jiném případě, při splnění obou opačných podmínek z této etapy, je uchován počet tříd a metodou MNČ dochází k pohybu středů shluků a k rozptýlení objektů do shluku, k jehož středu mají nejmenší vzdálenost. Důvodem pro úvahu možnosti pohybu segmentu v případě existence heterogenních objektů je následující hypotéza: jestliže se pohybuje většina objektů stabilně a určitým směrem a některé *(významný počet)* z těchto objektů jsou heterogenní těm původním, pak je velmi pravděpodobné, že v příští etapě se budou další objekty posunovat tímto směrem. Úmyslem je nejen mapovat současnou situaci, ale také předvídat budoucí chování a toto chování podpořit či naopak mu zabránit a využít tak konkurenční výhody obchodu. Tak se vytvářejí specifické kampaně, definované na základě charakteristik objektů reprezentovaných středem shluku.

Stanovení nižšího počtu tříd - ve smyslu sloučení tříd - než je ten současný, již není obsahem této práce (nebyl definován takový scénář změny). Dle logiky problému by se jednalo o zkoumání možného sloučení homogenních objektů. Ke sloučení tříd může dojít, ovšem tato situace se projeví jako klasifikace objektů do již existujících tříd a zánik té třídy, která nebude splňovat definice hranice minimálního možného počtu objektů a hranice času pozorování, definované v etapě IV. Sloučení tříd a například také možnost překročení stanoveného počtu shluků je tématem budoucího výzkumu.

Po průchodu prvními třemi etapami obecné metodologie a po aktualizaci vzorů chování užitím jedné ze dvou možností navržených scénářů změn jsou znovu posouzeny všechny objekty ve smyslu jejich vzdálenosti ke všem nově vytvořeným formátům chování. Může totiž nastat situace, že některé objekty budou patřit do některého ze zbývajících segmentů z důvodu menší vzdálenosti k jeho středu, jelikož došlo ke změně vzorů chování. Je proto znovu počítána vzdálenost každého objektu ke středům nově aktualizovaných shluků; k přesunu objektů dochází na základě porovnání vzdálenosti ke každému ze středů shluků.

Po stanovení jedné ze dvou forem aktualizace shluku je třeba, na kompletaci možností scénářů změn, automaticky přejít do etapy IV.

6.4 Etapa IV: Zánik shluků

Vždy existuje informace, která po uplynutí jistého časového období již není relevantní; další formou, jak pohlížet na tento jev, je připustit možnou změnu ve vzorech chování, jelikož přestávají existovat ty současné. To přísluší v dynamické shlukové analýze scénáři Zrušení tříd, které nemají dostatečný počet objektů na to, aby byly považovány za podklad pro následné kampaně (jsou to nově vzniklé třídy s malým počtem objektů nebo třídy, u nichž byl redukován počet objektů; může nastat i situace, že objekty jednoduše v příštím cyklu přestanou existovat, například při úmrtí nebo změně bydliště klientů v bance) a navíc po několik sekvenčních cyklů (období) jim nebyl agregován takový počet objektů, aby jejich počet spolu se známými objekty byl *dostačující*.

V této etapě obecné metodologie je třeba vyřešit několik problémů najednou; jsou to otázky: kdy zrušit třídu - s čímž souvisí definice pojmu hranice času pozorování a hranice minimálního možného počtu objektů - a jak zacházet, v případě že existuje, se zbytkovou informací.

Je třeba definovat hranici minimálního možného počtu objektů, pod kterou již počet známých a agregovaných objektů není dostatečný na *přežití* segmentu ve smyslu jeho praktického využití. Hranice času pozorování spočívá v počtu období, kdy není pozorována agregace objektů známých ani nových ve třídách s nedostatečným počtem objektů. Období a počet objektů jsou spojeny s každým individuálním problémem a aplikací. Toto období je definováno jako C_{max}, což je maximální počet cyklů (období), po které třída s nedostatečným počtem objektů udržuje paměť, zatímco jí není přiřazen objekt. Minimální možný počet objektů je jistá hranice, kterou je možné charakterizovat určitým procentem z celkového počtu objektů; je definována koeficientem ω pojmenovaným hranice minimálního možného počtu objektů.

Na základě popsaného je třeba vytvořit a udržovat jednoduchou paměť. Tato paměť má zachovat informaci po určený počet cyklů (období) a navíc registrovat pro třídu s menším než minimálním možným počtem objektů počet známých (aktualizovaných) a agregovaných objektů během každého určeného cyklu (období). Na zformalizování právě popsaného bude tato jednoduchá paměť označena *paměť registru* a je definována formou

matice $\mathbf{P}^t$ (*S x t*), která registruje existenci objektů $\{\mathbf{X}_{i\bullet}\}_i$ třídy S_h v každém období cyklu. Je zapsána hodnota *1*, jestliže byly agregovány objekty do třídy S_h s nedostatečným počtem objektů; v opačném případě je zaznamenána hodnota *0*.

Shrnutí definic pro třídu, která bude sledována ve smyslu její možného zániku, je následující:

Definice 6: Zánik shluků

Ke zrušení shluku dochází v případě, kdy shluk neobsahuje dostatečné množství objektů a po jistá časová období (cykly) nepřijímá žádný objekt, tak jako je stanoveno podmínkou 6:

Podmínka 6a: ***definice hranice minimálního možného počtu objektů***

$\forall S_h$, $\forall \mathbf{X}_{i\bullet} \in S_h$ platí:

$$\text{Podmínka 6a}: \quad \frac{e_{S_h}}{N} < \omega, \quad \text{kde} \quad 0 < \omega < 1.$$

Koeficient ω je hranice minimálního možného počtu objektů.

a

Podmínka 6b: ***definice hranice času pozorování***

$\mathbf{X}_{i\bullet} \in S_{h^*}$ jestliže je splněno: $d(\mathbf{X}_{i\bullet}, s_{h^*}) < d(\mathbf{X}_{i\bullet}, s_h)$, $\forall S_h$, $\forall \mathbf{X}_{i\bullet} \in S_h$.

Dodatečně se požaduje, že je-li $\forall \mathbf{X}_{i'\bullet} \notin S_{h^*}$ v t^c_{konc}, pak:

$$\text{Podmínka 6b}: \quad \mathbf{X}_{i'\bullet} \notin S_{h^*} \text{ v } C_{c+1} \text{ až } C_{\max}.$$

Hranici času pozorování a minimální počet objektů stanoví analytik s uživatelem. Měla by být shodná s determinací počtu cyklů, po které je sledován dynamický pohyb objektů před rozhodnutím o jejich aktualizaci.

V případě, že počet objektů není dostatečný a nebyly agregovány nové ani známé objekty po určená časová období, shluk je zrušen a objekty mohou být vyloučeny nebo sloučeny s již existujícími shluky. V této práci se analytik přiklání k jejich klasifikaci do

nejbližšího segmentu a v případě, tvoří-li skutečné outliery, nebudou ovlivňovat vzory chování.

Vstupy do této etapy jsou tvořeny počty objektů v každé třídě. Metodou je vytvoření paměti registru, která sleduje počet objektů známých i agregovaných po jistá časová období pro určené segmenty (s nedostatečným počtem objektů).

Zvolenou aktualizací shluků jsou eliminovány outliery, kterými byly před aktualizací označeny objekty vzdálené shlukům. Objekty, které by - po průchodu čtyřmi etapami obecné metodologie a po posouzení aplikace všech determinovaných scénářů změn - tvořily skutečné outliery, je možné z analýzy buď eliminovat nebo je zařadit do nejbližšího segmentu, aniž by ovlivnily výpočet nových formátů chování (mohlo by totiž dojít k umělému rozšíření shluků).

Poté je opakovaně provedena kontrola, zda-li všechny objekty přísluší do shluků, od jehož středu mají nejmenší vzdálenost. V případě, že s ohledem na vzdálenost objektů ke středům shluků již nedochází k dalšímu přesunu objektů mezi shluky, nevytváří se reálné outliery a neobjeví se shluk, který by měl nedostatečný počet objektů, je možné tento stav považovat za ukončení cyklu. Tím je také ukončena aktualizace vzorů chování realizovaná výběrem navržených scénářů změny.

6.5 Etapa V: Identifikace trajektorií

K implementaci obecné metodologie je přidán problém identifikace struktury tříd před a po jejich aktualizaci.

Vždy, kdy se provádí aktualizace, je získána příslušnost každého objektu k jednomu z existujících shluků v následujícím cyklu. Spojením současných shluků, definovaných svými formáty chování (ať již aktualizovanými nebo zachovanými z minulého cyklu), s těmi, které existovaly v předcházejícím období, je vytvořena spojnice mezi středy shluků pojmenovaná *trajektorie vzorů chování* mezi cykly; je to spojnice vektoru středů shluků. Matematicky zapsáno, $\mathbf{T}^c$ je matice *S x C*, kde každá řádka reprezentuje souřadnice středu

každého shluku v čase, to znamená v posledním období každého cyklu. $\mathbf{T}^{c}_{S_h}$. přísluší vektoru, který obsahuje trajektorie shluku S_h v čase, tedy souřadnice svého středu v čase.

Je připomenuto ustanovení, že segment může být identický tomu z předešlého období; další možností je pohyb středů shluků, vytvoření nových shluků nebo zánik některého ze shluků.

Uživatel by měl mít jasný záměr strategie ve smyslu aktualizace formátů chování shluků reprezentovaných svými vzory chování. Jinými slovy - existují segmenty, které je třeba podporovat a doplňovat dalšími objekty a na druhou stranu segmenty, které obsahují objekty, které jsou obecně řečeno nerentabilní. Existují marketingové kampaně, které mají za úkol pobízet přesun objektů ve smyslu požadovaného směru pohybu. Vývojové tendence, které jsou představovány právě popsanými trajektoriemi, konstruují podkladové údaje, na kterých se zakládají marketingové strategie.

Tento způsob analýzy sledování objektů po trajektoriích v průběhu časových cyklů je významný také z toho důvodu, že dovoluje pochopit budoucí změny a tendence v příslušném prostředí.

7 APLIKACE OBECNÉ METODOLOGIE PRO PŘÍPAD S IDENTIFIKÁTOREM ALGORITMEM K- MEANS

7.1 Aplikace na data simulovaná

Data použitá v následující aplikaci jsou simulována a zpracována v souladu s následující dynamikou: výchozí stav v prvním období t_1 prvního cyklu C_1 je tvořen čtyřmi třídami, kde se aktualizují a agregují objekty po dva navazující cykly (každý cyklus je složen ze tří období: první t_1^c, meziobdobí t_2^c a třetí t_3^c) do maximálně šesti shluků způsobem, že v prvním cyklu jsou dva shluky rozděleny mezi nové třídy, jeden je aktualizován pohybem a do jednoho jsou pouze klasifikovány objekty. Na konci druhého cyklu existuje třída, která neobsahuje žádný objekt a třída, která má jen jeden nový objekt; žádnému ze shluků nejsou obnoveny vzory chování.

Vektory proměnných (dále jen proměnné), které definují shluky, jsou dva: $\mathbf{X}_{\bullet 1}^{t_1^c}$ a $\mathbf{X}_{\bullet 2}^{t_1^c}$. Analýza objektů ve smyslu jejich dynamiky je sledována mezi t_1^c a t_3^c v obou dvou proměnných zároveň.

Vytvořená obecná metodologie je aplikována pro každý shluk zvlášť na konci každého cyklu, kdy je analyzován stav a chování objektů a v rámci scénářů změn vybírány formy aktualizace formátů chování shluků.

Stálá kritéria, aplikovaná během dvou analyzovaných cyklů, jsou následující[(*)]:

- Výchozí stav tříd: $h \in \{1,2,3,4\}$.
- V etapě I, Identifikace objektů, které představují změnu,
  jsou tyto v segmentu S_h identifikovány výpočtem a porovnáním Eukleidovy vzdálenosti následovně:

$$\mathbf{X}_{i\bullet} \Rightarrow \mathbf{X}_{i\bullet}^{o}, \text{ jestliže } d(\mathbf{X}_{i\bullet}^{t_{konc}^c}, s_h^{t_{poc}^c}) > d_{\max s_h}^{t_{poc}^c} \text{ (platí i pro } \mathbf{Y}_{i\bullet}).$$

---

* Přehled symbolů, definic a parametrů a také souhrn etap s podmínkami a jejich kritérii je uveden v *přílohách 11.2 a 11.3.*

- V etapě II, Rozpoznání stavu změny,

  je tento v segmentu S_h rozpoznán, jestliže $e^o_{S_h} > 0.2 \ * N_{S_h}$.

- V etapě III., Rozhodnutí o možnostech aktualizace shluků,

  je provedena aktualizace klasifikací do shluků S_h, jestliže $E^s_{S_h} \leq 0.5 \ *(N_1)_{S_h}$.

  V případě opačném, kdy $E^s_{S_h} > 0.5 \ *(N_1)_{S_h}$, dochází k aktualizaci pohybem shluků. Kritéria jsou stanovena zároveň pro obě dvě proměnné – tzn., že obě proměnné se podílejí na konstrukci trajektorie pohybu ($E^s_{S_h}$ je kombinací stabilních objektů majících identický směr pohybu v obou proměnných). Tyto aktualizace jsou provedeny pro situace, kdy není rozpoznán stav změny.

  Při rozpoznání stavu změny, kdy $E^{o,s}_{S_h} \leq 0.8 * e^o_{S_h}$ nebo $E^{o,s}_{S_h} + e^s_{S_h} \leq 0.5 * N_{S_h}$, dochází k vytvoření nových tříd. V případě splnění obou dvou opačných kritérií je aplikována aktualizace pohybem shluků.

  Aktualizace středů tříd při změně formátů chování se provádí MNČ.

- V etapě IV, Zrušení shluků,

  dochází k eliminaci shluku, jestliže $e_{S_h} < 0.03 \ * N_1$ a zároveň pro tento shluk a $\forall \mathbf{X}_{i'\bullet} \notin S_{h^*}$ v t^c_{konc} platí:

  $$\mathbf{X}_{i'\bullet} \notin S_{h^*} \text{ v } C_{c+1} \text{ až } C_{c+2} \text{ (platí pro každé období cyklu).}$$

- V etapě V, Identifikace trajektorií,

  jsou tyto vyjádřeny spojnicí $\mathbf{T}^c_{S_h\bullet}$, která přísluší vektoru obsahujícímu trajektorie shluku S_h v čase, tedy souřadnice svého středu mezi cykly.

Po průchodu prvními třemi etapami je pro každý shluk vybrána příslušná možnost aktualizace shluků definovaná v etapě III., ovšem až po čtvrté etapě je ukončen cyklus identifikace možných scénářů změn. Poslední, pátá etapa, shrnuje konečné charakteristiky shluků při ukončení každého časového cyklu a promítá stav vývoje dynamické segmentace, což umožňuje sledovat konkrétní stavy shluků a jejich variace po reálných trajektoriích svých středů v čase.

V následující části je předveden postup aplikace obecné metodologie dynamické segmentace. Grafy a tabulky reprezentují mezivýsledky a stavy rozhodování v jednotlivých

etapách. Na konci každého cyklu je prezentován graf vývoje dynamické segmentace a shrnuty charakteristiky jednotlivých shluků.

Poznámka: V *sekci 7 a 8* v některých tabulkách, ale i v některých obrázcích a grafech, nebylo možné nastavit volbu jazyka a ortografii češtiny ani používat symboliku definovanou v *sekci 6* (tento problém se týká outputů instalace SPSS).

7.1.1 Cyklus první

0) Počáteční počet tříd v t_1^1: $S = |4|$

Technikou segmentace ve dvou fázích je vybrán jako optimální výchozí stav shluků počet 4. Proměnné jsou v prvním období t_1^1 pojmenovány $\mathbf{X}_{\bullet 1}^{t_1^1}$ a $\mathbf{X}_{\bullet 2}^{t_1^1}$. Počet původních objektů je 100 (soubor všech objektů s jejich identifikátorem, který slouží také jako pořadové číslo v souboru dat, s příslušností ke svému shluku v t_1^1 a se vzdáleností ke shluku je uveden v *příloze 11.4,* sloupce 1, 2 a 3). V prvním cyklu je $t_{poc}^1 = t_1^1$. V dalších cyklech je $t_{poc}^{c+1} = t_{konc}^c$. t_{konc}^c je období t_3^c po aktualizaci vzorů chování.

V následujících třech tabulkách 7-1, 7-2 a 7-3 jsou shrnuty některé charakteristiky původních tříd: jejich středy, počet objektů spadajících do každého shluku a vzdálenosti středů shluků od sebe navzájem. Celá situace je definována pro stav t_1^1 (první období prvního cyklu), tedy stav před jakoukoliv aktualizací či agregací objektů. Graf 7-1 znázorňuje všechny objekty období t_1^1 a barevně vyjadřuje přislušnost objektů ke svým shlukům, které jsou reprezentovány středy.

Tab 7-1 ***Středy shluků v t_1^1***

	Středy shluků			
Proměnné	s1	s2	s3	s4
X.1	10	30	31	5
X.2	30	55	7	11

Tab 7-2 ***Počet objektů ve shlucích v t_1^1***

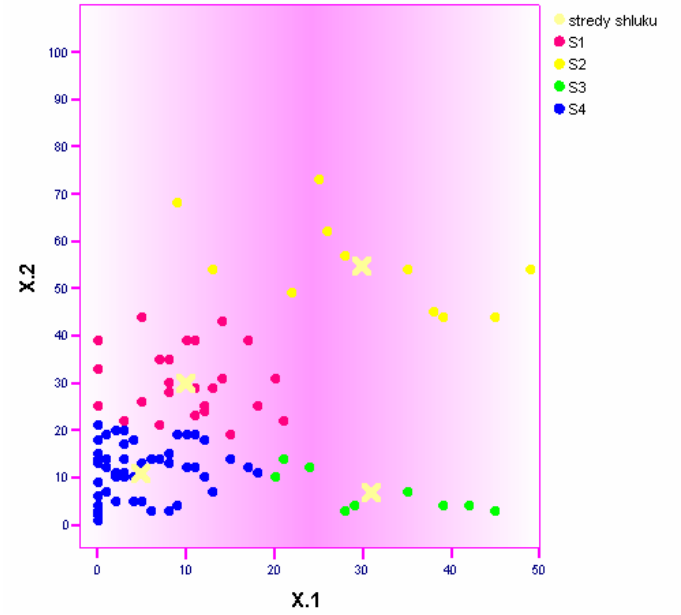
Shluky	N1
S1	25
S2	11
S3	9
S4	55
Celkem	100

Tab 7-3 ***Vzdálenosti středů shluků navzájem v*** t_1^1

Vzdálenosti mezi středy shluků	S1	S2	S3	S4
S1	0			
S2	32,0	0		
S3	31,2	48,1	0	
S4	19,6	50,6	26,4	0

Nejvzdálenější jsou si shluky S_2 a S_4 a nejblíže k sobě mají ty nejpočetnější segmenty, kterými jsou S_1 a S_4.

Graf 7-1 ***Distribuce objektů v*** t_1^1 ***mezi shluky v*** t_1^1 (*)



Poznámka: Bodem je označen objekt, křížem střed shluku.
Ve všech grafech sekce 7.1 je platná následující symbolika: bodem je označen objekt, křížem střed shluku, čísla vyjadřují identifikátor a také pořadí objektů v archívu dat (pro unifikaci objektů a lepší orientaci), čtverečky vyjadřují outliery, čtverečky s tečkou skutečné outliery, objekty vyjádřené fialovym písmem jsou agregované objekty.

Pro vyjádření distribuce objektů uvnitř každého shluku z pohledu jejich vzdálenosti ke středu shluku slouží tabulka 7-4 a graf 7-2.

---

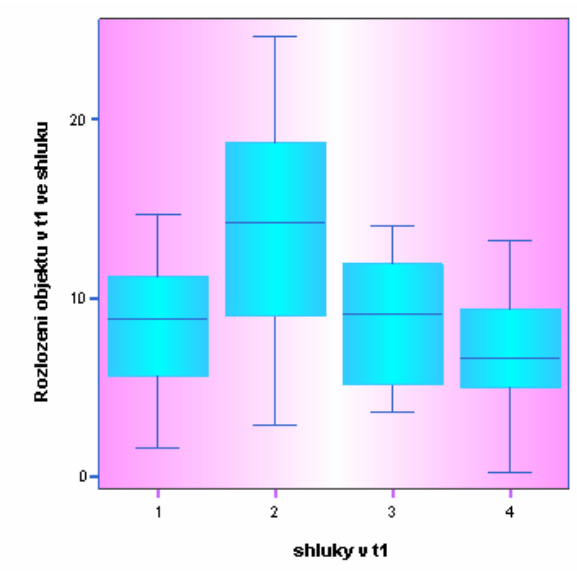
* *Objekty v* t_1^1 znamená objekty *sledované* v prvním období prvního cyklu. *Shluky v* t_1^1 znamená shluky *vytvořené* v prvním období prvního cyklu. Všechny názvy grafů a tabulek v *sekci 7* sledují právě popsanou logiku.

Tab 7-4 ***Vzdálenost objektů v*** t_1^1 ***ke středu příslušného shluku v*** t_1^1

Shluky	Průměr	Min	**Max**	N1
S1	8,3	1,6	14,6	25
S2	13,7	2,8	24,7	11
S3	8,8	3,6	14,1	9
S4	6,8	,2	13,2	55

Dle charakteristik shluků, pozorovaných v tabulce 7-4, je možné vyslovit jisté hypotézy. Shluky s největší průměrnou (a maximální) vzdáleností objektů od středu shluku, které mají navíc relativně malý počet objektů, jsou *shluky široké (rozptýlené)*, *nehomogenní* v objektech a tudíž *nestabilní*. Hypotézou o jejich chování v budoucích obdobích je jejich rozdělení mezi více tříd; v analyzovaném případě toto platí pro shluk 2, dále pak shluk 3. Naopak, shluky s malou průměrnou (a maximální) vzdáleností a s velkým počtem objektů jsou *shluky úzké, homogenní a stabilní.* Hypotézou o jejich budoucím chování je pohyb nebo klasifikace objektů. Těmito charakteristikami disponují především shluk 4, dále pak i shluk 1. O zrušení shluků zatím není možné vyslovit žádnou hypotézu, jelikož obecná metodologie postupuje sekvenčně přes etapu III., kde jsou vybrány možnosti aktualizace a teprve poté je postoupeno do etapy IV., která na základě aktualizovaného stavu shluků přechází k hodnocení možnosti zrušení shluků.

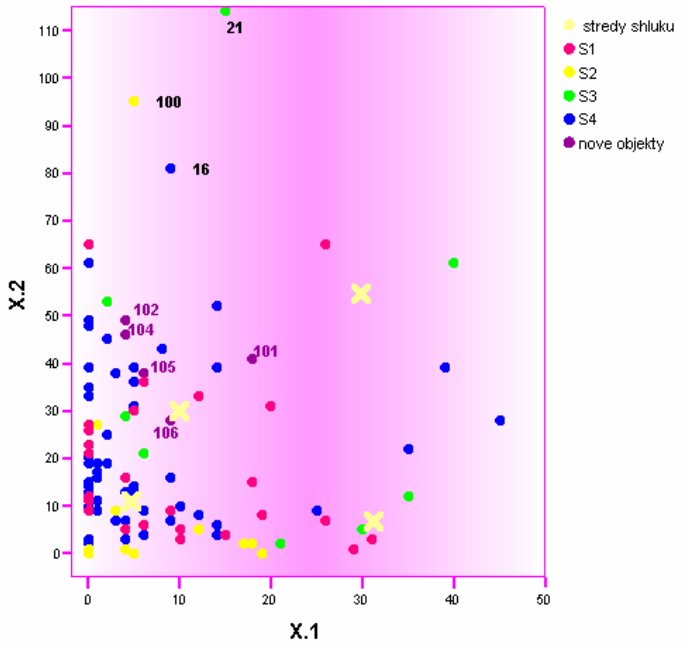
Graf 7-2 ***Rozložení objektů v*** t_1^1 ***uvnitř shluků v*** t_1^1



Pro sledování reálných outlierů je vybrána technika, kterou reprezentuje graf typu boxplot, ve kterém není sledován žádný skutečný outlier definovaných shluků.

Stav objektů v období t_3^1 je možné pozorovat v grafu 7-3. Jsou přidány nové objekty; počet objektů je nyní 105. Všechny objekty si s sebou přinášejí svou barvu z grafu v období t_1^1 z důvodu lepší orientace při sledování jejich přemístění a příslušnosti ke shluku. Červené jsou objekty, které patřily v období t_1^1 do shluku 1, žluté do shluku 2, zelené do shluku 3 a modré do shluku 4. Vzory chování zůstávají zatím beze změny, shodné jsou tudíž i středy shluku, opět vyjádřené křížem.

Graf 7-3 Distribuce objektů v t_3^1 mezi shluky v t_1^1



Již pouhým pohledem na graf je možné dedukovat, že došlo k výraznému pohybu objektů mezi shluky. Jsou vidět i evidentní outliery, jako je například objekt číslo 21, 100 nebo 16. Zatím však ještě není známo, které objekty budou patřit do jakého segmentu a tudíž kterého shluku budou zmíněné objekty outliery. Proměnné mají v tomto období, které je posledním obdobím prvního cyklu, hodnoty $\mathbf{X}_{\bullet 1}^{t_3^1}$ a $\mathbf{X}_{\bullet 2}^{t_3^1}$.

1) Etapa I.: Identifikace objektů, které představují změnu

Dynamika etapy:

Aby byly objekty zařazeny do svých příslušných tříd, jsou Eukleidovou větou počítány vzdálenosti každého objektu od středu každého shluku a objekty jsou zařazeny do toho shluku, ke kterému mají nejmenší vzdálenost (hodnoty proměnných $\mathbf{X}_{\bullet 1}^{t_l^1}$ a $\mathbf{X}_{\bullet 2}^{t_l^1}$ v t_1^1, t_2^1 a t_3^1, Eukleidovy vzdálenosti objektů t_3^1 k příslušnému shluku v t_1^1 a shluk, do kterého je klasifikován každý objekt v t_3^1, jsou v *příloze 11.4*, sloupce 4 – 11).

Poté jsou pro každý shluk sledovány outliery: vzdálenost $d(\mathbf{X}_{i\bullet}^{t_3^1}, s_h^{t_1^1})$ každého objektu v t_3^1 (platí také pro $\mathbf{Y}_{i\bullet}^{t_3^1}$) od svého příslušného shluku v t_1^1 je srovnávána se vzdáleností $d_{\max_{s_h}^{t_1^1}}$. Objekty, které přesahují tuto vzdálenost, jsou označeny jako outliery (*příloha 11.4*, sloupec 12, outlier označen číslem ***1***).

Následuje tabulkovou a grafickou formou souhrn celkového stavu pro všechny shluky v období t_1^1 a t_3^1: počet objektů v každé třídě a jejich přesun mezi třídami (tabulka 7-5), průměrné, největší a nejmenší vzdálenosti objektů v t_3^1 od středů shluků definovaných v t_1^1 (tabulka 7-6). Situaci doplňuje graf 7-4, který zobrazuje také reálné outliery. Na závěr je ukázán stav rozhodování v této etapě (tabulka 7-7).

Tab 7-5 Počet a přemístění objektů mezi shluky mezi t_1^1 a t_3^1

	Shluky v t1					
Shluky v t3	S1	S2	S3	S4	S0	N
S1	7	1	3	14	5	30
S2	2	1	2	4	0	9
S3	4	3	3	3	0	13
S4	12	6	1	34	0	53
N	25	11	9	55	5	**105**

Poznámka: S_0 je třída pro agregované objekty vložené do analýzy v obdobích po t_1^1.

Analýzou grafu 7-3 bylo sledováno *významné přemístění objektů* mezi shluky. V tabulce 7-5 je kromě prostého stavu objektů v t_1^1 a t_3^1 možné tento pohyb detailně sledovat. Celkový počet objektů v každém jednotlivém shluku se příliš nezměnil: nejpočetnější je S_4, nejmenší S_2 a S_3. Z S_1 bylo více než 2/3 objektů přesunuto do jiných

tříd, především do nejbližšího S_4. Nejvíce objektů získal S_1 přesunem také právě z tohoto shluku. Z pouhé výměny objektů mezi S_1 a ostatními shluky a bez detailní analýzy dynamiky objektů v každém období zatím není možné sledovat tendenci objektů a tudíž ani celého S_1. Tento shluk vlastní všechny nové, agregované objekty, které by mohly být objekty heterogenními, jelikož byly klasifikovány do S_1 z důvodu nejmenší vzdálenosti od jeho středu – ta by ovšem mohla být *významná*. S_2 a S_3 také registrují nepravidelné přemístění objektů mezi ostatními shluky. Tyto mají navíc malý počet objektů, z čehož je možné dedukovat budoucí tvorbu nových tříd. V S_4 setrvala většina známých objektů. K podstatné výměně mezi tímto největším shlukem došlo s S_1 a vzhledem k relativnímu počtu objektů také s S_2 (směr pohybu objektů z S_2 do S_4 je v obou proměnných negativní).

Tab 7-6 ***Vzdálenost objektů v t_3^1 ke středu příslušného shluku v t_1^1***

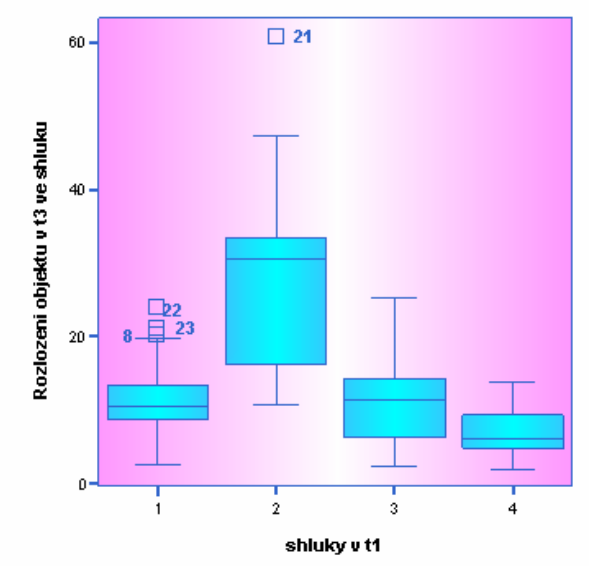
Shluky	Průměr	Min	**Max**	N
S1	11,3	2,5	24,1	30
S2	29,0	10,8	60,9	9
S3	10,7	2,3	25,2	13
S4	6,8	2,0	13,8	53

Hypotézy o typu aktualizace shluků, získané v průběhu analýzy z tabulek 7-4 a 7-5, podporuje tabulka 7-6: pro S_1 je to pohyb objektů nebo jejich klasifikace, pro S_2 a S_3 vytvoření nových tříd a pro S_4 nejpravděpodobněji klasifikace objektů.

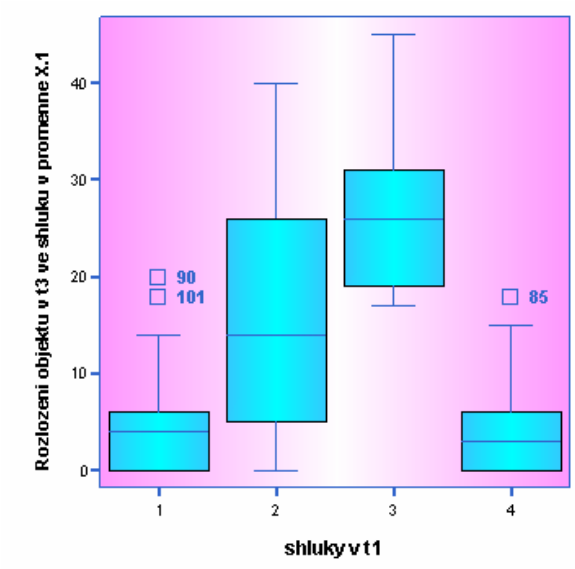
Prostým zařazením objektů do shluku se zvyšuje průměrná vzdálenost mezi objekty a středy shluků; segment se rozšiřuje. Rozšíření shluku po prosté klasifikaci objektů bez aktualizace vzorů chování potvrzuje graf 7-4. V grafu 7-4a) je možné sledovat reálné outliery shluku 1 (objekty 8, 22 a 23) a 2 (objekt 21), posuzované dle vzdálenosti objektů ke středu shluku, vyhodnocené pro obě dvě proměnné. Je-li hodnocena proměnná $\mathbf{X}_{\bullet 1}^{t_3^1}$ zvlášť (7-4b), jsou nalezeny reálné outliery pro segment 1 a 4. Při sledování proměnné $\mathbf{X}_{\bullet 2}^{t_3^1}$ (7-4c) jsou identifikovány reálné outliery v segmentu 2 a 3.

Graf 7-4 Rozložení objektů v t_3^1 a znázornění reálných outlierů uvnitř shluků v t_1^1

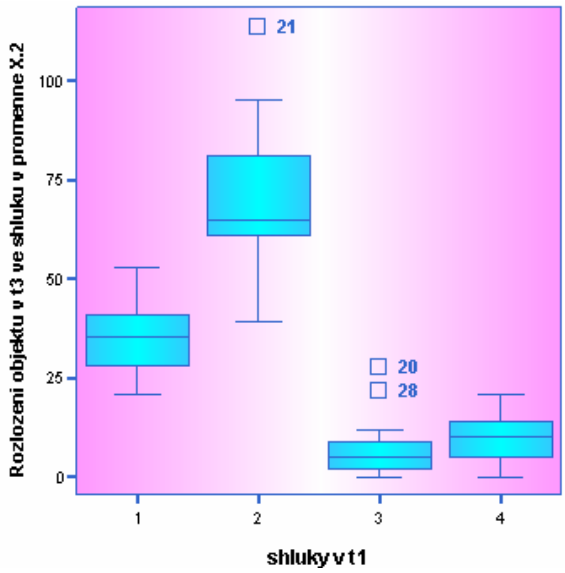
a) pro obě proměnné



b) pro $\mathbf{X}_{\bullet 1}^{t_3^1}$



c) *pro* $\mathbf{X}_{\bullet 2}^{t_3^1}$



Každý shluk bude mít outlier, jelikož všechny maximální vzdálenosti pro všechny shluky jsou v tomto období vyšší než v období t_1^1 (tabulka 7-6). Aktualizace charakteristik shluků poslouží k úpravě shluků tak, aby existoval nízký počet tříd a zároveň žádný objekt nebyl příliš vzdálený typickému objektu shluku, který je jeho středem.

Všechny shluky, jelikož vlastní objekty představující změnu, by mohly být kandidáty na vytvoření nové třídy (tabulka 7-7).

Tab 7-7 ***Stav rozhodování v první etapě*** C_1

Shluky	N	Existence outlieru	Stav rozhodování
S1	30	**ANO**	M,I,NT
S2	9	**ANO**	M,I,NT
S3	13	**ANO**	M,I,NT
S4	53	**ANO**	M,I,NT
Celkem	105		

Poznámka: V celé sekci 7.1 je označena symbolem M mechanická aktualizace, symbolem I inteligentní aktualizace a symbolem NT nové třídy.

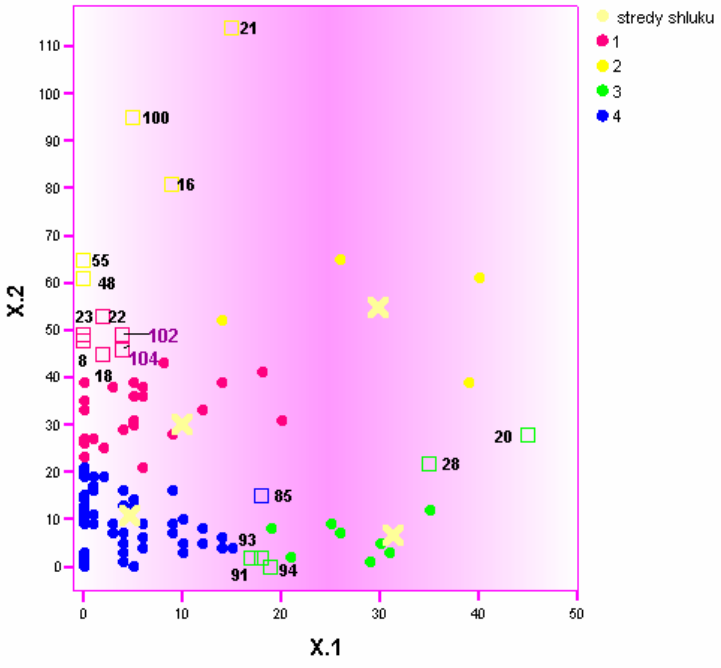
2) Etapa II.: Rozpoznání stavu změny

Dynamika etapy:

Jsou sledovány všechny shluky, kterým byly v předchozí etapě identifikovány objekty ve stavu změny. Vstupními daty jsou počet známých i agregovaných objektů a zvlášť outlierů s příslušností ke svému segmentu. Aby bylo rozhodnuto ve prospěch stavu změny, musí být počet outlierů v příslušném shluku v t_3^1 větší než 0,2 * N_{S_h} .

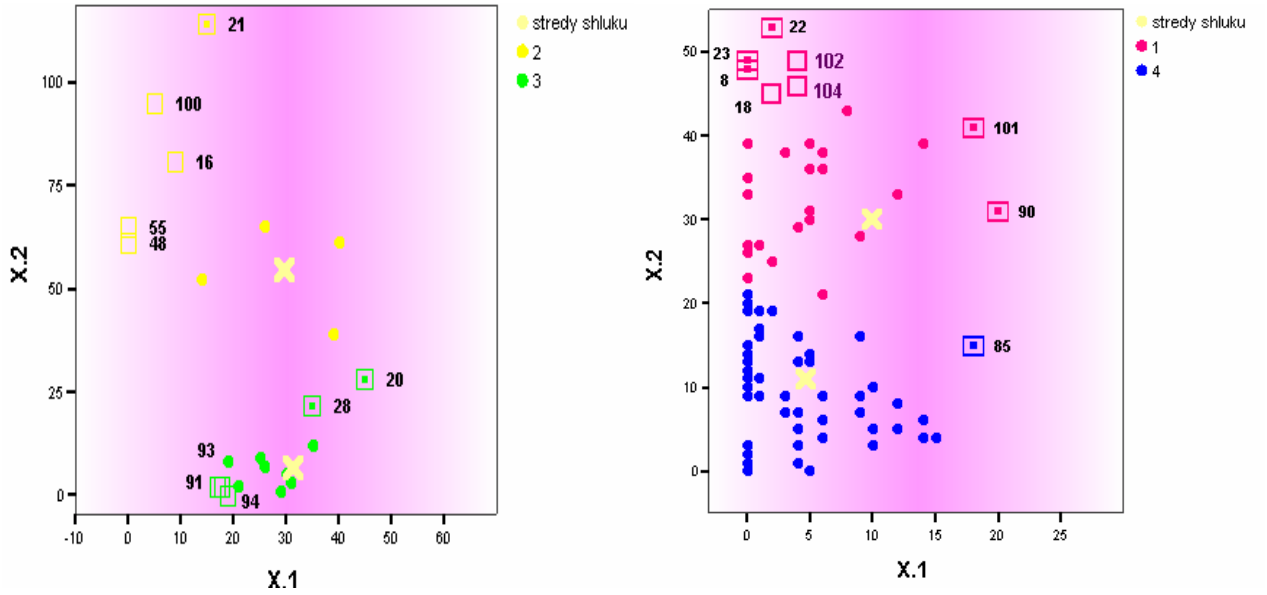
Grafy 7-5 a 7-6 zobrazují klasifikované známé a agregované objekty v t_3^1 do svých shluků v t_1^1; zvlášť jsou znázorněny outliery identifikované obecnou metodologií .

Graf 7-5 ***Objekty v*** t_3^1 ***klasifikované do shluků v*** t_1^1



Graf 7-6 Distribuce outlierů mezi zbývajícími objekty v t_3^1

a) ve shlucích 2 a 3 *b) ve shlucích 1 a 4*



Poznámka: v grafech jsou znázorněny outliery definované dle obecné metodologie i outliery skutečné dle analýzy grafu boplot v obou proměnných zároveň i v proměnných analyzovaných zvlášť. Je-li objekt identifikován jako reálný outlier (může být zároveň i outlier dle obecné metodologie), je označen čtverečkem s tečkou.

Po porovnávání vzdálenosti všech objektů v t_3^1 ke středu svých příslušných shluků v t_1^1 s největšími původními vzdálenostmi objektů od středů shluků v t_1^1 bylo zjištěno, že existuje celkem 17 outlierů v celkovém počtu N objektů v t_3^1 (tabulka 7-8).

Situace rozhodování v průběhu obecné metodologie, konkrétně na konci druhé etapy, je následující (tabulka 7-8):

Tab 7-8 Stav rozhodování ve druhé etapě C_1

Shluky	Objekty uvnitř shluků		**Outliery**		Celkem	**Stav rozhodování**
	N	% N	N	% N	N	
S1	24	80,0%	6	20,0%	30	M, I
S2	4	44,4%	5	55,6%	9	NT
S3	8	61,5%	5	38,5%	13	NT
S4	52	98,1%	1	1,9%	53	M, I
Celkem	88	88,3%	17	16,2%	105	

Rozhodování o vytvoření nové třídy se zúžilo na druhý a třetí shluk. Pro první a čtvrtý připadají v úvahu (z důvodu nevýznamného počtu outlierů) mechanická nebo inteligentní aktualizace vzorů chování.

3) Etapa III.: Rozhodnutí o možnostech aktualizace shluků

Dynamika etapy:

Ve třetí etapě je rozhodnuto ve prospěch mechanické aktualizace klasifikací do tříd, jestliže analýza chování objektů ve shluku prokáže (alespoň pro jednu proměnnou) povšechnou entropii směru a/nebo pohybu, tzn. počet všech objektů shluku (včetně outlierů) se stabilním pohybem a směrem pro obě proměnné současně je menší nebo roven koeficientu hranice stability objektů násobeným počtem známých objektů; koeficient hranice stability objektů je v tomto případě 0,5. Inteligentní aktualizace je aplikována na případy, kdy analýza povšechného chování objektů shluku pro obě dvě proměnné zároveň neprokáže entropii, tedy počet objektů se stabilním pohybem a stejnou trajektorií je větší než 50% známých objektů a zároveň buď nejsou identifikovány objekty se změnou nebo není rozpoznán stav změny, což platí i pro rozhodnutí ve prospěch možnosti mechanické aktualizace. Aktualizace pohybem tříd se provádí v souladu s MNČ.

Pohyb objektů je stabilní, i když je protisměrný v obou proměnných. Stabilní konstantu je možné dle předchozího ustanovení přiřadit k objektům buď s pozitivním nebo s negativním stabilním pohybem. Agregované objekty, tvořící outliery, budou v těchto podmínkách (třech obdobích tvořících cyklus) vždy nestabilní, jelikož neexistuje dostatečný počet období na potvrzení stability pohybu.

Kritériem na stanovení tvorby nových tříd je směr a stabilita pohybu objektů a outlierů. Shluky ve stavu změny jsou rozděleny mezi větší počet tříd v případě, kdy neexistuje skupinka stabilních outlierů větší než 80% (což je hranice stability outlierů) všech outlierů s identickou trajektorií nebo v případě, kdy trajektorie většiny objektů (včetně outlierů) se stabilním pohybem (koeficient hranice tvorby nových shluků je 0,5) není identická trajektorii outlierů, které mají stabilní pohyb a tvoří skupinku > 80% všech outlierů toho samého shluku. Nastává tvorba nových tříd, kdy se počet shluků zvyšuje o jeden stupeň, v analyzovaném případě do maximálně deseti tříd.

V případě splnění obou dvou opačných kritérií jsou shluky ve stavu změny aktualizovány inteligentní aktualizací. Aktualizace vzorů chování se provádí taktéž MNČ.

Pro třídu 1 a 4 dle stavu rozhodování z předchozí etapy případá v úvahu klasifikace či pohyb tříd. Je tedy vyhodnocována stabilita objektů těchto shluků a to prostým sledováním a součtem objektů, které prokazují stabilitu jak pohybu, tak i směru. Kritéria pohybu a směru jsou posuzována pro obě dvě proměnné zároveň.

Pro shluky 2 a 3 je třeba navíc sledovat počet a chování outlierů a srovnávat ho s povšechným chováním shluku.

Počet objektů stabilních v pohybu je celkem 38, z toho třída jedna jich vlastní 17, třída dva 4, třída tři 7 a třída čtyři 10 (tabulka 7-9). Proměnné, které zaznamenávají stabilitu, jsou uvedeny v *příloze 11.4*, sloupce 13 – 16; stabilita je označena číslem *-1, 0* nebo *1*.

Tab 7-9 ***Stabilita pohybu objektů ve shlucích v*** C_1

	Stabilita pohybu objektů		
Shluky	nestabilní	**stabilní**	N
S1	13	17	30
S2	5	4	9
S3	6	7	13
S4	43	10	53
Celkem	67	**38**	105

Povšechnou stabilitu objektů je možné sledovat v tabulce 7-10, která zpřehledňuje všechny shluky, ovšem slouží především pro ty, kterým nebyl rozpoznán stav změny či nebyly identifikovány objekty se změnou – těmi jsou třídy 1 a 4.

Tab 7-10 ***Stabilita pohybu a směru objektů v proměnných*** $\mathbf{X}_{\bullet 1}$ ***a*** $\mathbf{X}_{\bullet 2}$ ***v*** C_1

Stabilita objektů		**X.2 + Y.2**				**X.2 + Y.2**						
		S1	S2	S3	S4	S1		S2	S3		S4	
		99	99	99	99	1	-1	1	1	-1	1	-1
X.1 + Y.1	99	13	5	6	43	0	0	0	0	0	0	0
	1	0	0	0	0	2	0	2	2	2	0	0
	-1	0	0	0	0	11	1	2	0	3	0	8
	0	0	0	0	0	2	1	0	0	0	1	1

Poznámka: 99 značí nestabilní pohyb, 1 značí pohyb stabilní pozitivní, -1 stabilní negativní a 0 stabilní konstantu. Stejná symbolika platí pro celou sekci 7.1.

Například, shluk 1 má třicet objektů, z toho 17 stabilních. 2 objekty mají pozitivní tendenci u obou proměnných, 11 objektů pozitivní u proměnné $\mathbf{X}_{\bullet 2}$ *a negativní u proměnné* $\mathbf{X}_{\bullet 1}$*, 2 objekty stabilní konstantu* $\mathbf{X}_{\bullet 1}$ *a pozitivní* $\mathbf{X}_{\bullet 2}$ *atd. Celkově 15 objektů v proměnné* $\mathbf{X}_{\bullet 2}$ *roste, klesají dva objekty. U proměnné* $\mathbf{X}_{\bullet 1}$ *12 objektů vykazuje pokles, 2 růst a 3 konstantní stabilní stav.*

Detailní pohled na jednotlivé shluky

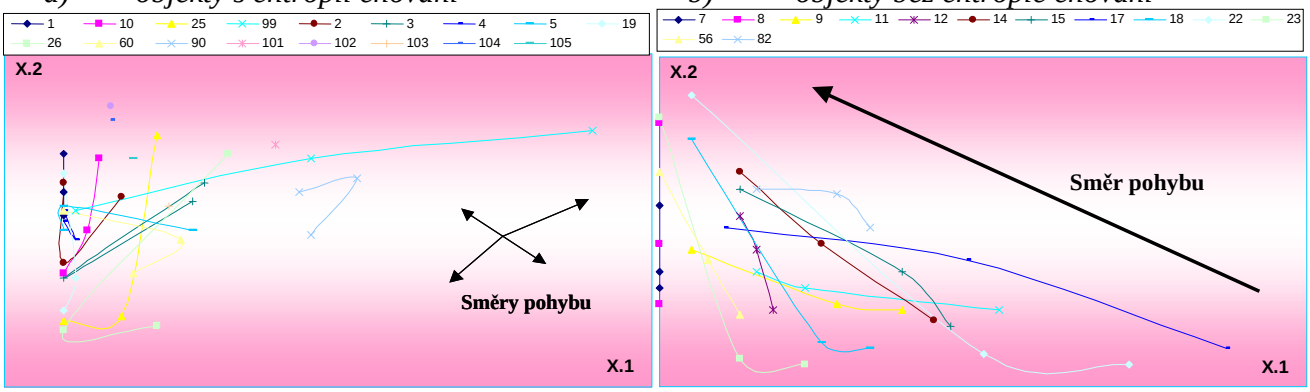
A) Shluk 1

Při sledování stability pohybu a směru všech objektů segmentu ve všech třech obdobích je z tabulek 7-9 a 7-10 vidět, že shluk 1 má 13 nestabilních a 17 stabilních objektů v pohybu. Z těchto 17 objektů je pohyb 11 objektů u proměnné $\mathbf{X}_{\bullet 2}$ pozitivní a zároveň u proměnné $\mathbf{X}_{\bullet 1}$ negativní. Konečný směr pohybu objektů a tedy jako trajektorie shluku bude definován ten směr, kterým se pohybuje většina stabilních objektů. 11 objektů + 2 konstanty ze 30 objektů ve shluku 1 tvoří 43% z celkového počtu objektů tohoto shluku. Jelikož však nové objekty nemohou vykazovat stabilitu, nejsou proto zahrnuty do vyčíslení kritéria podmínky této etapy. 11 + 2 objekty se stabilním směrným pohybem tvoří 52% objektů shluku 1, kde počet objektů bez těch nových je 25. Při tomto způsobu vyčíslení kritéria na rozhodnutí o klasifikaci objektů či pohybu segmentu je počet stabilních objektů s jistou trajektorií větší než 50% všech známých objektů a proto je rozhodnuto pro pohyb tohoto shluku, realizovaného pomocí MNČ.

Pro přehled stability pohybů a trajektorií objektů shluku 1 jsou tyto znázorněny v grafu 7-7.

Graf 7-7 Trajektorie shluku 1 v průběhu C_1

a) objekty s entropií chování *b) objekty bez entropie chování*



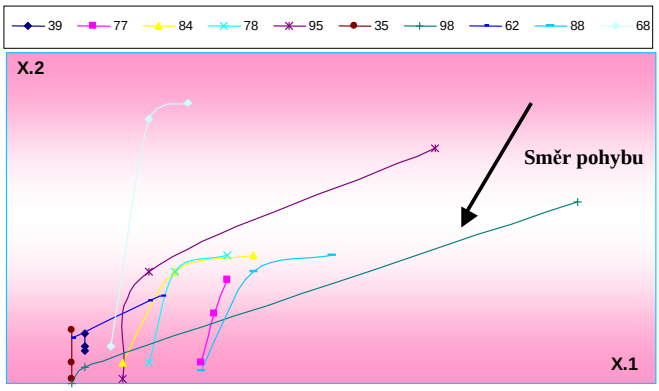
Poznámka: Šipky vyznačují povšechné směry pohybu. V legendě jsou objekty označeny číslem identifikátoru. Ta samá symbolika platí i pro ostatní grafy tohoto typu v sekci 7.1.

Graf 7-7a) vyjadřuje entropii chování objektů (v pohybu nebo směru). Naproti tomu graf 7-7b) zobrazuje objekty, které mají nejenom stabilní pohyb, ale i stejnou trajektorii pohybu. 11 objektů je stabilních v pohybu i směru a 2 objekty jsou stabilní konstanty (konstanty v jedné proměnné; ve druhé mají tendenci pohybu 11ti stabilních objektů).

B) Shluk 4

Shluk 4 obsahuje 10 stabilních objektů v pohybu z celkových 53 objektů, což již bez další analýzy trajektorie směru objektů značí, že dojde k prosté klasifikaci objektů do tříd. Pro ilustraci je předveden pohyb a směr stabilních objektů tohoto segmentu v grafu 7-8.

Graf 7-8 ***Pohyb objektů stabilních v pohybu a jejich trajektorie ve shluku 4 v průběhu*** C_1



Jak již bylo napsáno v úvodu této etapy, pro shluky 2 a 3 je třeba provést hlubší analýzu, která se týká srovnávání chování stabilních outlierů s povšechným chováním objektů těchto shluků (tabulka 7-11).

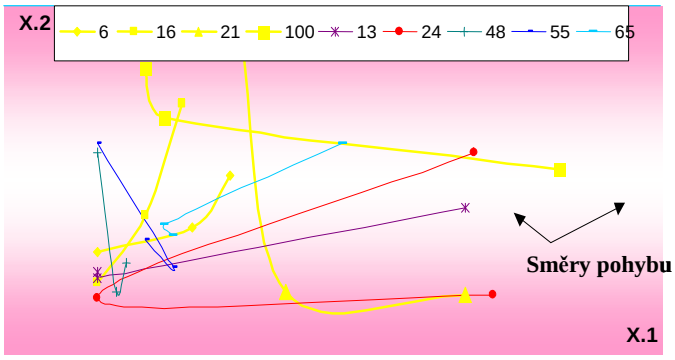
Tab 7-11 ***Stabilita pohybu a směru všech objektů a zvlášť' outlierů v proměnných*** $\mathbf{X}_{\bullet 1}$ ***a*** $\mathbf{X}_{\bullet 2}$ ***v*** C_1 ***ve shlucích 2 a 3***

Shluky		X.1	X.2		
			99	1	-1
2	objekty uvnitř shluku	99	3	0	0
		1	0	1	0
	outliery	99	2	0	0
		1	0	1	0
		-1	0	2	0
3	objekty uvnitř shluku	99	5	0	0
		1	0	0	2
		-1	0	0	1
	outliery	99	1	0	0
		1	0	2	0
		-1	0	0	2

C) Shluk 2

Druhý shluk obsahuje 4 stabilní objekty, z nichž 3 jsou outliery. Celkový počet objektů v segmentu je 9 (graf 7-9). Neexistuje skupinka stabilních směrných outlierů větší než 80% všech outlierů toho shluku, jelikož dva z těchto outlierů (67%) se pohybují jiným směrem než třetí outlier. Již z této skutečnosti je možné vyvodit rozhodnutí o aktualizaci shluku, jímž bude rozdělení shluku mezi dvě nové třídy. Navíc, uvnitř shluku existuje jen jeden stabilní objekt, tzn. že ani v případě stejnosměrného pohybu všech stabilních objektů shluku by nebylo docíleno hranice 50% stabilních stejnosměrných objektů a tudíž by nebylo rozhodnuto ve prospěch druhé alternativy rozhodování, kterou je pohyb shluku.

Graf 7-9 Pohyb a trajektorie objektů shluku 2 v průběhu C_1

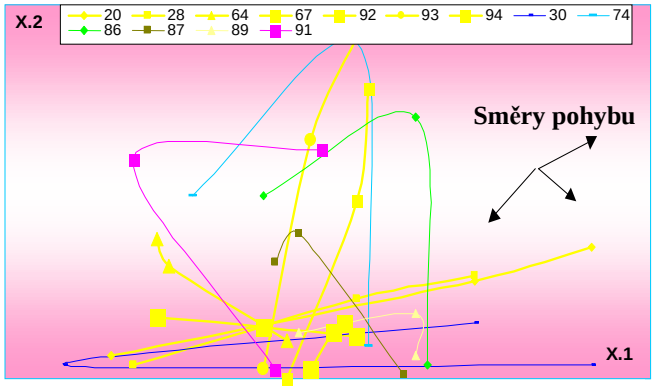


Poznámka: žlutou barvou jsou znázorněny objekty stabilní v pohybu - šipky ukazují jejich trajektorie.

D) Shluk 3

Shluk číslo 3 obsahuje celkem 13 objektů, z nich 4 jsou outliery stabilní v pohybu: 2 a 2 se pohybují rozdílným směrem (pozitivní, negativní). Není třeba provádět hlubší analýzu objektů uvnitř segmentu, jelikož neexistuje většinová skupinka outlierů segmentu mající stabilní směrný pohyb. I kdyby analytik pokračoval v analýze, zjistil by, že ani uvnitř segmentu neexistuje významný počet objektů pohybujících se po stejné trajektorii stabilním pohybem (graf 7-10). Je přijata volba tvorby nových tříd metodou nejmenších čtverců.

Graf 7-10 Pohyb a trajektorie objektů shluku 3 v průběhu C_1



Poznámka: žlutou barvou jsou znázorněny objekty stabilní v pohybu - šipky ukazuji jejich směr.

Stav rozhodování na konci této etapy je zpřehledněn v tabulce 7-12.

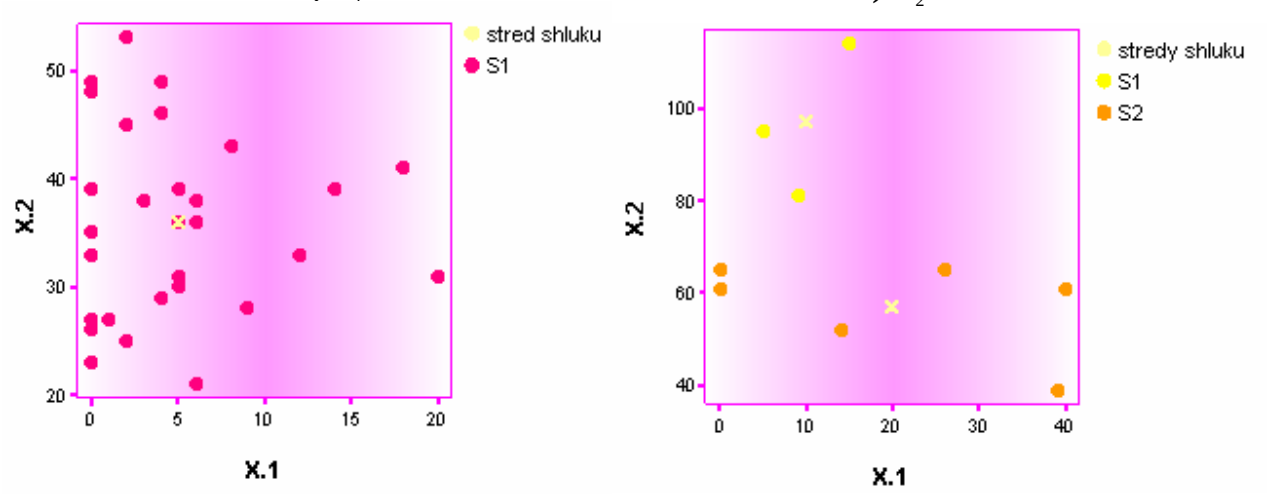
Tab 7-12 Stav rozhodování ve třetí etapě C_1

	počet známých objektů	počet všech objektů	všechny objekty shluku			outliery			všechny objekty shluku		
				Největší skupina stabilních objektů s trajektorií k:	(%)		největší skupinka stabilních outlierů s trajektorií k:	(%)	největší skupinka stabilních objektů s trajektorií k:	(%)	Stav rozhodování
				$E^{\,s}_{S_h}$	$\frac{E^{s}_{S_h}}{(N_1)_{S_h}}$		$E^{\,o,s}_{S_h}$	$\frac{E^{o,s}_{S_h}}{e^{o}_{S_h}}$	$E^{\,s,o}_{S_h} + e^{s}_{S_h}$	$\frac{E^{s,o}_{S_h} + e^{s}_{S_h}}{N_{S_h}}$	
Shluky	N1	N	stabilní			stabilní					
S1	25	30	17	13	**52%**				0	0	**I**
S2	9	9				3	2	**67%**	2	22%	**NT**
S3	13	13				4	2	**50%**	2	15%	**NT**
S4	53	53	10	9	**17%**				0	0	**M**
N	100	105									

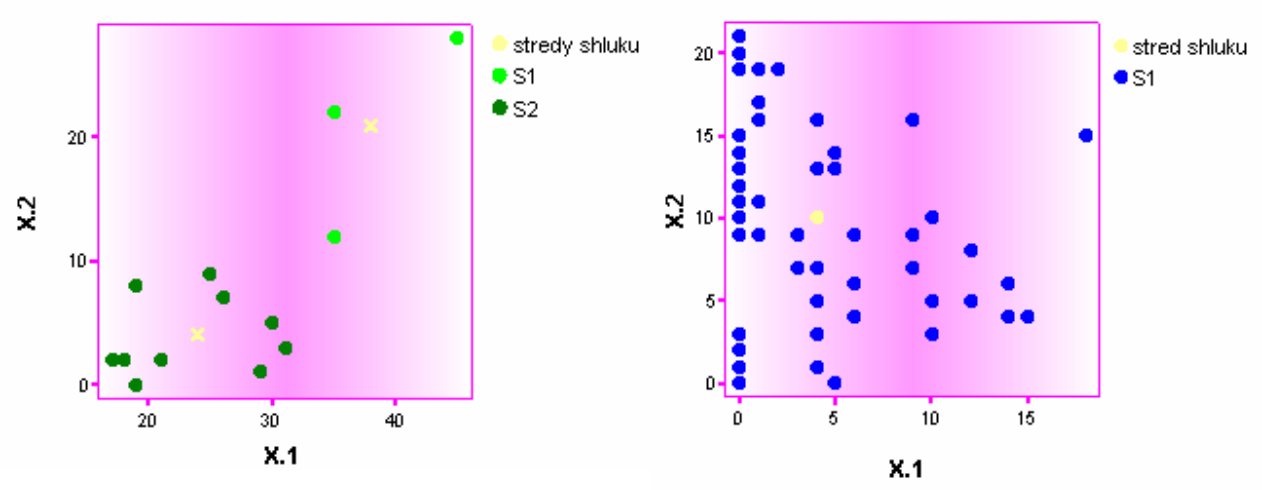
Nyní je realizována aktualizace vzorů chování shluků dle tabulky 7-12 a to ze čtyř současných do šesti navržených metodou nejmenších čtverců (graf 7-11).

Graf 7-11 Aktualizace vzorů chování v C_1

a) S_1 ***b)*** S_2



c) S_3 ***d)*** S_4



Shluk 1 je inteligentně aktualizován. Shluk 2 je rozdělen mezi dva shluky, které budou označeny S_2 a S_3. Shluk 3 je rozdělen mezi shluky S_4 a S_5 a shluk S_4, aktualizovaný klasifikací objektů, je přejmenován na S_6.

Nyní jsou znovu posouzeny všechny objekty ve smyslu jejich vzdálenosti ke všem nově vytvořeným formátům chování. Takto je 7 objektů přesunuto z jedné třídy do jiné, k jejímuž středu mají blíž (tabulka 7-14, graf 7-12) Poté jsou opět MNČ stanoveny středy shluků. Opakovaným výpočtem vzdálenosti každého objektu ke každému ze shluků je potvrzeno jejich zařazení do příslušného shluku, tak jako charakteristiky vzorů chování

shluků na konci třetí etapy (*příloha 11.4*: příslušnost objektů do aktualizovaných shluků, vzdálenost objektů k aktualizovaným shlukům, přesunované objekty - ve sloupci je uvedn shluk přesunu, oprava příslušnosti objektů k aktualizovaným shlukům a opravená vzdálenost ke středům shluků jsou uvedeny ve sloupcích 17 - 21).

Následující tři tabulky (7-13, 7-14, 7-15) shrnují stavy před a po aktualizaci, přesuny objektů až k vyjádření konečného stavu po všech opravách příslušností objektů ke shlukům.

Tab 7-13 Počet objektů ve shlucích před a po aktualizaci v t_3^1 ***a*** t_{konc}^1

		objekty v t3 klasifikované do shluků vytvořených v t0				
		S1	S2	S3	S4	N
shluky po aktualizaci v t_konc	S1	30	0	0	0	30
	S2	0	3	0	0	3
	S3	0	6	0	0	6
	S4	0	0	3	0	3
	S5	0	0	10	0	10
	S6	0	0	0	53	53
N		30	9	13	53	105

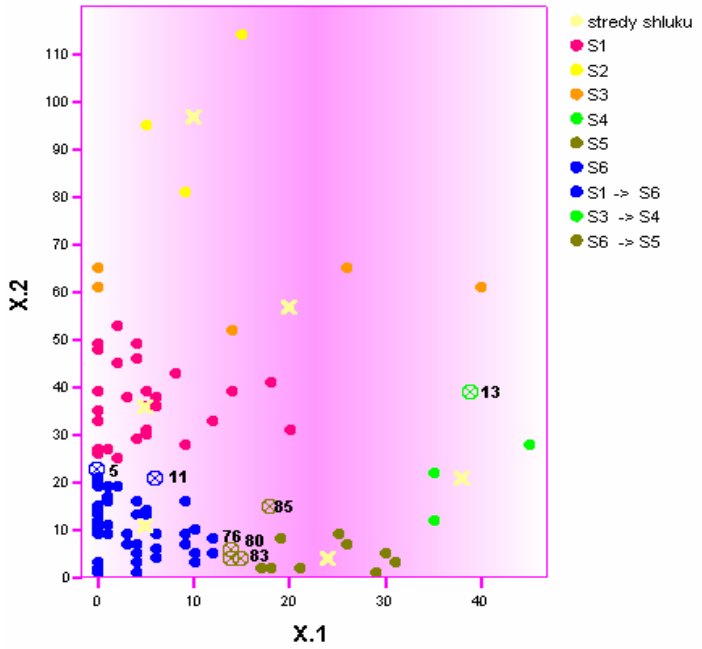
Tab 7-14 Počet a umístění opravených objektů ve shlucích po jejich aktualizaci v t_{konc}^1

		přesun objektů do shluku			
		S4	S5	S6	N
přesun objektů ze shluku	S1	0	0	2	2
	S3	1	0	0	1
	S6	0	4	0	4
N		1	4	2	**7**

Tab 7-15 Konečný počet objektů ve shlucích po opravách jejich zařazení v t_{konc}^1

		objekty v t3 klasifikované do shluků vytvořených v t0				
		S1	S2	S3	S4	N
shluky po aktualizaci a po přesunu opravených objektů	S1	28	0	0	0	**28**
	S2	0	3	0	0	**3**
	S3	0	5	0	0	**5**
	S4	0	1	3	0	**4**
	S5	0	0	10	4	**14**
	S6	2	0	0	49	**51**
N		30	9	13	53	**105**

Graf 7-12 Oprava po aktualizaci a finální zařazení objektů do aktualizovaných shluků v t^1_{konc}

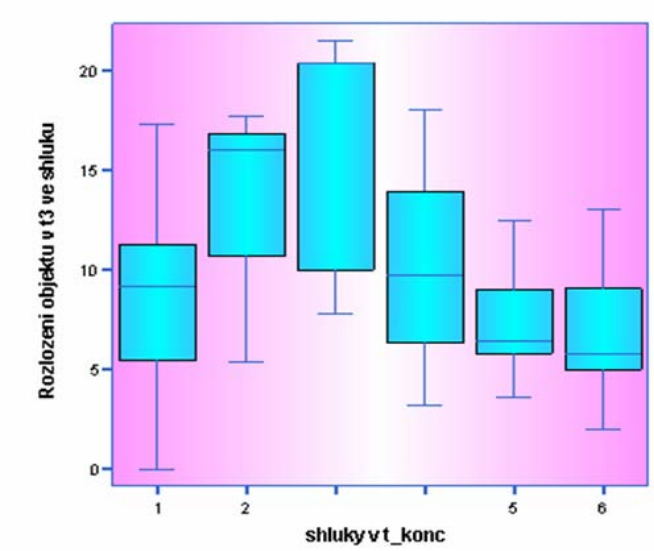


Poznámka: Objekty označené kolečky s křížky jsou objekty, které se přesunují mezi shluky (oprava po aktualizaci). Barva kolečka vyjadřuje novou třídu příslušnosti.

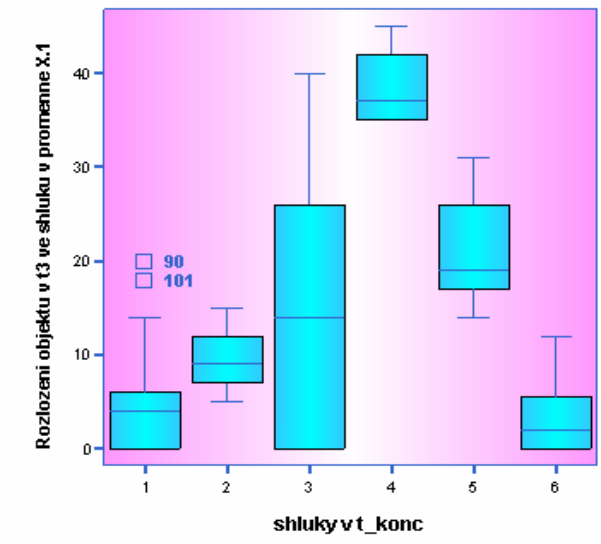
Zvolenými aktualizacemi jsou sníženy vzdálenosti objektů ke středům příslušných shluků (srovnat tabulky 7-6 a 7-19) a jsou eliminovány outliery, kterými byly před aktualizací označeny objekty vzdálené shlukům. Objekty, které by nyní tvořily skutečné outliery (graf 7-13), je možné z analýzy buď eliminovat nebo je zařadit do nejbližšího segmentu, aniž by ovlivnily výpočet nových formátů chování (vytvářely by nereálné středy shluků).

Graf 7-13 Rozložení objektů v t_3^1 a znázornění reálných outlierů uvnitř shluků v t_{konc}^1

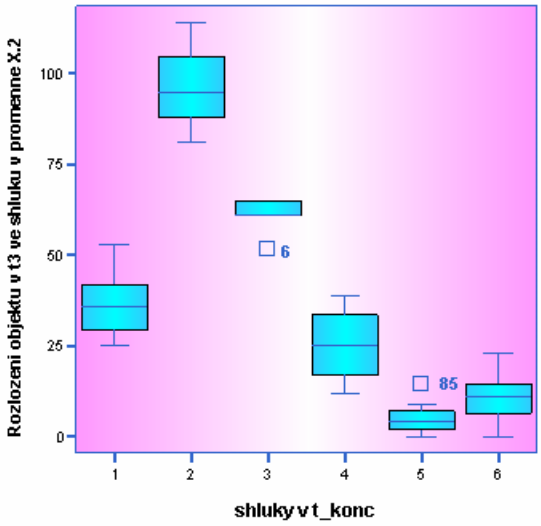
a) pro obě proměnné



b) pro $\mathbf{X}_{\bullet 1}^{t_{konc}^1}$



c) pro $\mathbf{X}_{\bullet 2}^{t_{konc}^1}$



Jak ukazují grafy boxplot, jsou-li sledovány obě proměnné najednou, nejsou identifikovány reálné outliery (na rozdíl v období t_3^1 před aktualizací vzorů chování). Proto žádný objekt nebude eliminován a všechny budou zařazeny do tvorby nových vzorů chování.

4) Etapa IV.: Zánik shluků

K zániku shluku dochází v případě, vlastní-li tento méně objektů, než je definovaná hranice minimálního možného počtu objektů, zde 3% ze známých objektů všech shluků; navíc, během dvou následujících navazujících cyklů (a všech jeho období) nepřijímá žádný agregovaný ani žádný ze známých objektů.

Jelikož je procházeno prvním cyklem, nejsou zatím sledovány žádné shluky, které by měly nedostatečný počet objektů. Ani v této fázi není takový shluk detektován, jelikož minimální počet objektů ve shluku jsou 3, což tvoří 3% ze známých objektů a tento počet nesplňuje podmínku definující zánik shluků.

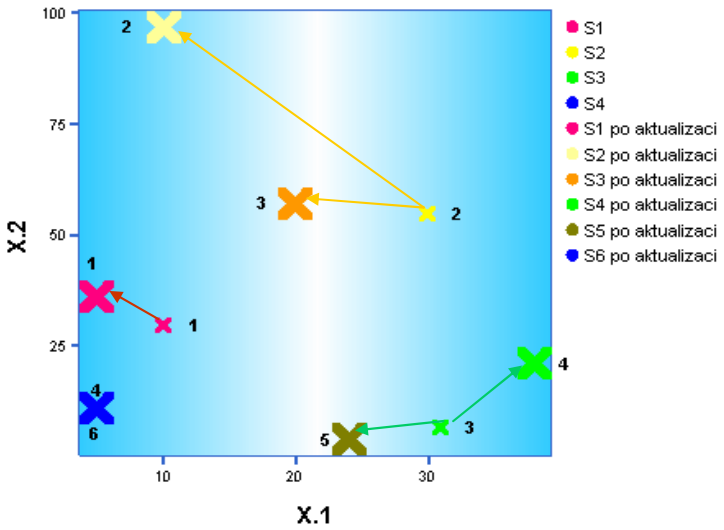
V případě, že s ohledem na vzdálenost objektů ke středům shluků již nedochází k dalšímu přesunu objektů mezi shluky, je tento stav považován za ukončení prvního cyklu aktualizace vzorů chování.

Touto etapou je také ukončen cyklus výběru možných scénářů změn.

5) Etapa V.: Identifikace trajektorií

Pro identifikaci trajektorií, tedy směru pohybu segmentů, je třeba spojit jejich středy na počátku a na konci realizovaného cyklu (stav na konci jednoho cyklu je i počátečním stavem následujícího cyklu). Z grafu 7-14 je vidět, jak jsou shluky aktualizovány.

Graf 7-14 Reálné trajektorie pohybu shluků mezi t^1_{poc} ***a*** t^1_{konc}



Poznámka: Malým křížkem jsou označeny původní shluky z t^1_{poc}. *Velkými kříži jsou označeny shluky* t^1_{konc} *(neboli z* t^2_{poc}*). Střed shluku 4 v* t^1_{poc} *se překrývá se středem shluku 6 v* t^1_{konc}.

V tabulkách 7-16, 7-17, 7-18 a 7-19 jsou shrnuty charakteristiky aktualizovaných tříd na konci prvního cyklu: jejich středy, počet objektů spadajících do každého shluku, vzdálenosti středů shluků od sebe navzájem a vzdálenosti objektů od středu svých shluků. Tento stav je výchozím stavem pro druhý cyklus (t^2_{poc}).

Tab 7-16 Středy shluků v t^1_{konc} (t^2_{poc})

	Středy shluků					
Proměnné	s1	s2	s3	s4	s5	s6
X.1	5	10	20	38	24	5
X.2	36	97	57	21	4	11

Tab 7-17 Počet objektů ve shlucích v t^1_{konc} (t^2_{poc})

Shluky	**N**
S1	28
S2	3
S3	5
S4	4
S5	14
S6	51
Celkem	105

Tab 7-18 Vzdálenosti středů shluků navzájem v t^1_{konc} (t^2_{poc})

Vzdálenosti mezi středy shluků	S1	S2	S3	S4	S5	S6
S1	,0					
S2	61,2	,0				
S3	25,8	41,2	,0			
S4	36,2	81,0	40,2	,0		
S5	37,2	94,0	53,2	22,0	,0	
S6	25,0	86,1	48,4	34,5	20,2	,0

Tab 7-19 Vzdálenost objektů v t^1_3 ke středu příslušného shluku v t^1_{konc} (t^2_{poc})

Shluky	Průměr	Min	**Max**	N
S1	8,5	,0	17,3	28
S2	13,0	5,4	17,7	3
S3	16,0	7,8	21,5	5
S4	10,2	3,2	18,0	4
S5	7,1	3,6	12,5	14
S6	6,6	2,0	13,0	51

7.1.2 Cyklus druhý

0) Počáteční počet tříd v t^2_{poc}: $S = |6|$

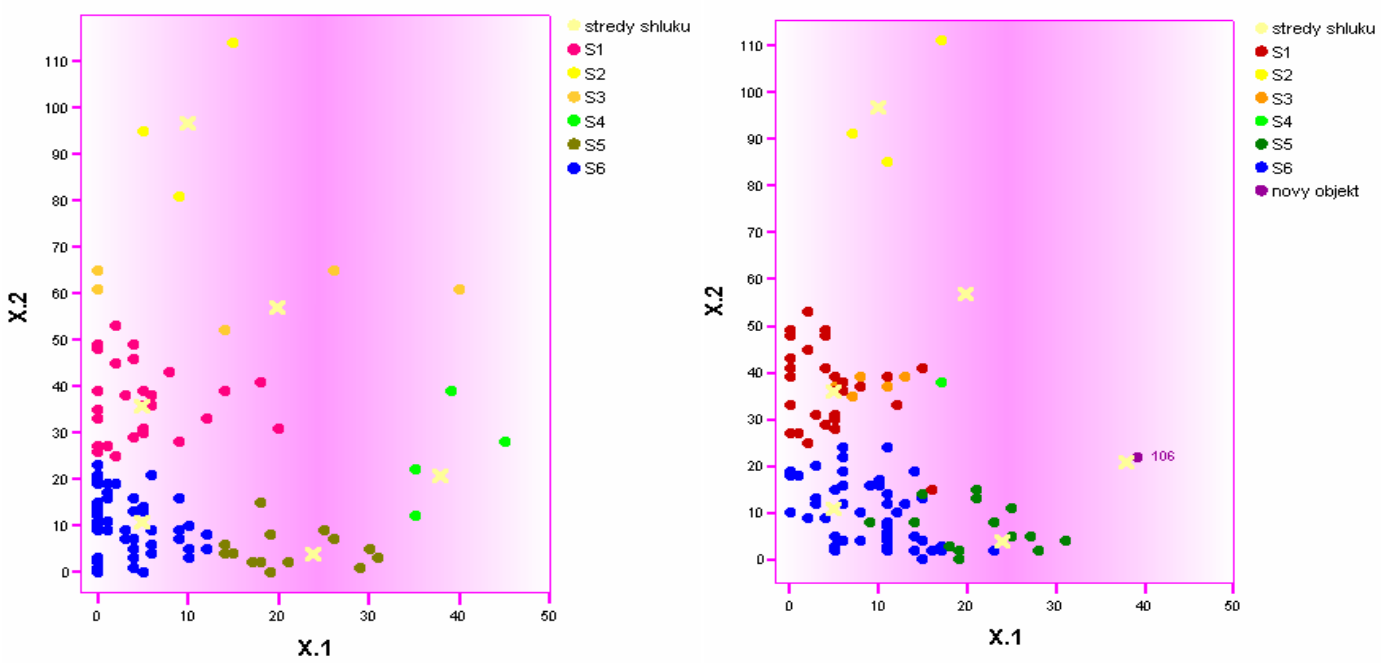
Ve druhém cyklu se vychází ze segmentace vytvořené v cyklu prvním; souhrn charakteristik původních tříd v tomto období je identický těm, které jsou uvedeny v tabulkách v koncovém období prvního cyklu. Počet shluků je zvýšen o dva. Shluky, které nebyly rozděleny mezi nové třídy (nyní 1 a 6) mají téměř stejný počet objektů jako na počátku prvního cyklu; došlo k přemístění některých objektů. Nejvzdálenější shluk všem ostatním je shluk 2. Nejpočetnější shluk 6 je shlukem nejvyrovnanějším, má nejmenší průměrnou vzdálenost objektů ke svému středu. Naopak, segmentem s největším rozptylem objektů je shluk 3. Malými shluky co do počtu objektů jsou shluky 2, 3 a 4, které by mohly být adepty na zrušení shluků. Původní rozložení objektů ve shlucích na počátku t^2_{poc} je zobrazeno v grafu 7-15 a).

Proměnné v prvním období t^2_1 jsou pojmenovány $\mathbf{X}^{t^2_1}_{\bullet 1}$ a $\mathbf{X}^{t^2_1}_{\bullet 2}$. Počet původních objektů je 105.

Stav objektů ve třetím období C_2 je ukázán v grafu 7-15b). Je možné předpokládat, že třída 3 bude zrušena (oranžové objekty) a objekty klasifikovány do shluku 1. Stejně tak v případě třídy 4 – v grafu je možné pozorovat pouze jeden - světle zelený - objekt, což znamená, že zbylé objekty třídy 4 přestaly existovat. Třída 1 udržuje relativně stejné rozložení objektů, stejně tak třídy 5 a 6. Objekt, označený vínovou barvou, je novým objektem přidaným v tomto cyklu do analýzy. Patřil by zřejmě do bývalé třídy 4; protože je však v této třídě objektem osamoceným, mohl by být buď eliminován nebo klasifikován do nejbližší třídy. Obecně je možné říci, že objekty nejpočetnějších shluků zůstaly téměř beze změny příslušnosti tříd.

Graf 7-15 Distribuce mezi shluky v t^2_{poc} objektů

a) v t^2_{poc} ***b) v t^2_3***



1) Etapa I.: Identifikace objektů, které představují změnu

Dynamika etapy:

Objekty jsou nejprve klasifikovány do svých příslušných tříd, tak jak je vidět v tabulce 7-20, grafu 7-17 a v *příloze 11.4* (hodnoty proměnných $\mathbf{X}^{t_l^2}_{\bullet 1}$ a $\mathbf{X}^{t_l^2}_{\bullet 2}$ v t_1^2, t_2^2 a t_3^2 druhého cyklu, vypočtená Eukleidova vzdálenost a příslušnost ke svému shluku v t_1^2 pro objekty v t_3^2 jsou ve sloupcích 22 - 29).

Tab 7-20 Počet a přemístění objektů mezi shluky mezi t^2_{poc} a t^2_3

	Shluky v t_poc							
Shluky v t3	S0	S1	S2	S3	S4	S5	S6	N
S1	0	27	0	5	1	0	2	35
S2	0	0	3	0	0	0	0	3
S4	1	0	0	0	0	0	0	1
S5	0	0	0	0	0	11	8	19
S6	0	1	0	0	0	3	41	45
N	1	28	3	5	1	14	51	103

Třída 3 neobsahuje žádný objekt, třída 4 obsahuje jeden a to agregovaný objekt. Počet objektů na počátku cyklu je 105. V průběhu cyklu je do shluku 4 zařazen jeden nový objekt (v t_2^2), ovšem již v t_1^2 přestávají existovat tři objekty toho samého shluku. Ze třídy 3 jsou všechny objekty přesunuty do třídy 1 a tak v t_3^2 nevlastní žádný objekt. Počet objektů v t_3^2 je 103.

Dle kritéria hranice heterogenity jsou sledovány outliery; porovnáním maximálních vzdáleností z tabulek 7-21 a 7-19 není prokázána existence outlierů s výjimkou shluku 1. Navíc, většina průměrných vzdáleností je ve srovnání s koncem minulého cyklu redukována. To znamená, že objekty a tím i shluky se v tomto cyklu zkonsolidovaly – snižuje se počet shluků a objekty jsou blíže typickým objektům shluků.

Tab 7-21 ***Vzdálenost objektů v t_3^2 ke středu příslušného shluku v t_{poc}^2***

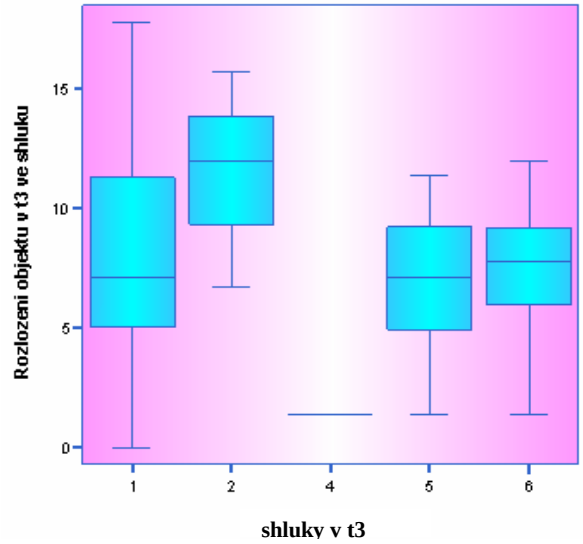
Shluky	Průměr	Min	**Max**	N
S1	7,7	,0	17,8	35
S2	11,5	6,7	15,7	3
S4	1,4	1,4	1,4	1
S5	6,8	1,4	11,4	19
S6	7,2	1,4	12,0	45

V grafu 7-16 je možné sledovat reálné outliery.

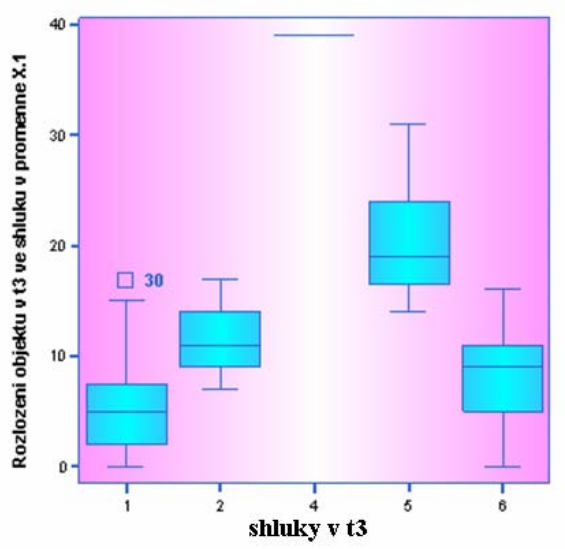
Jediný shluk, který by mohl být kandidátem na vytvoření nových tříd, je shluk 1. Tabulka 7-22 zobrazuje výsledný stav rozhodovaní pro první etapu.

Graf 7-16 ***Rozložení objektů v t_3^2 a znázornění reálných outlierů uvnitř shluků v t_{konc}^2***

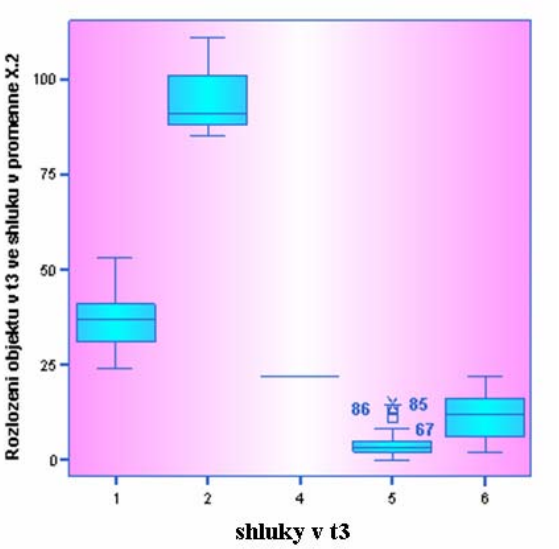
a) *pro obě proměnné*



b) *pro* $\mathbf{X}_{\bullet 1}^{t_{konc}^1}$



c) *pro* $\mathbf{X}_{\bullet 2}^{t_{konc}^1}$



Tab 7-22 ***Stav rozhodování v první etapě C_2***

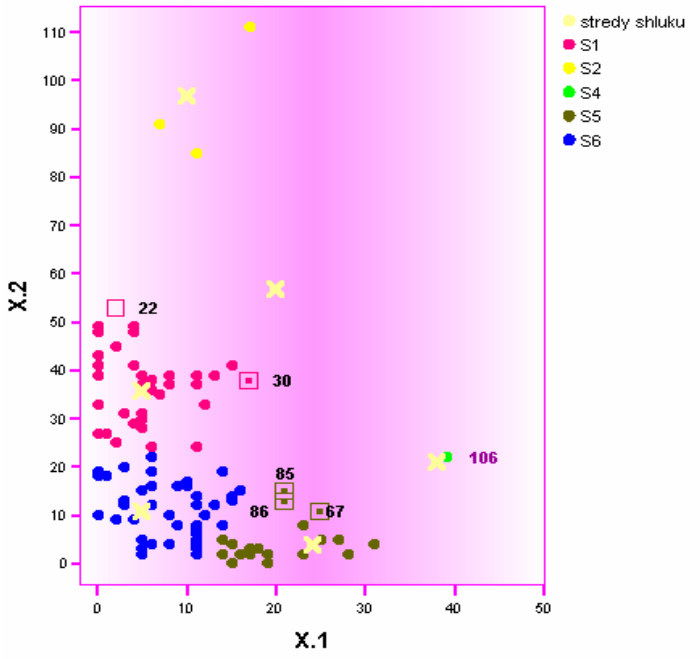
Shluky	N	Existence outlieru	Stav rozhodování
S1	35	**ANO**	M,I,NT
S2	3	**NE**	M,I
S4	1	**NE**	M,I
S5	19	**NE**	M,I
S6	45	**NE**	M,I
Celkem	103		

2) Etapa II.: Rozpoznání stavu změny

Dynamika etapy:

Jelikož je nalezen pouze jeden outlier a to v prvním shluku (*příloha 11.4*, sloupec 30), (graf 7-17), který má 35 objektů, není rozpoznán stav změny.

Graf 7-17 Outliery v t_3^2 mezi klasifikovanými objekty do shluků v t_{poc}^2



To znamená, že všechny shluky budou aktualizovány pouze mechanickou nebo inteligentní aktualizací, jak ukazuje tabulka 7-23.

Tab 7-23 Stav rozhodování ve druhé etapě C_2

Shluky	Objekty uvnitr shluku		Outliery		Celkem	Stav rozhodovani
	N	% N	N	% N	N	
S1	34	97,1%	1	2,9%	35	M, I
S2	3	100,0%	0	0.0%	3	
S4	1	100,0%	0	0.0%	1	M, I
S5	19	100,0%	0	0.0%	19	M, I
S6	45	100,0%	0	0.0%	45	M, I
Celkem	102	99,0%	1	1,0%	103	

3) Etapa III.: Rozhodnutí o možnostech aktualizace shluku

Dynamika etapy:

V tomto cyklu je možné pozorovat pouze 7 objektů stabilních v pohybu. Ve shluku jedna existuje jeden takový objekt tak jako ve shluku pět. Shluk šest jich má 5, zbývající shluky žádné (tabulky 7-24 a 7-25).

Tab 7-24 Stabilita pohybu objektů ve shlucích v C_2

	Stabilita pohybu objektů		
Shluky	nestabilní	**stabilní**	N
S1	34	1	35
S2	3	0	3
S4	1	0	1
S5	18	1	19
S6	40	5	45
Celkem	96	7	103

Tab 7-25 Stabilita pohybu a směru objektů v proměnných $\mathbf{X}_{\bullet_1}$ ***a*** $\mathbf{X}_{\bullet_2}$ ***v*** C_2

		X.1 + Y.1					X.1 + Y.1			
		S1	S2	S4	S5	S6	S1	S5	S6	
Stabilita objektů		99	99	99	99	99	0	0	1	0
X.2 + Y.2	99	34	3	1	18	40	0	0	0	0
	1	0	0	0	0	0	0	0	0	0
	-1	0	0	0	0	0	0	1	1	4
	0	0	0	0	0	0	1	0	0	0

V tomto cyklu žádný ze segmentů nemá více než 50% objektů stabilních v pohybu, proto by ani nebylo nutné sledovat jejich směr na vyslovení závěru, že všechny třídy budou aktualizovány mechanickou aktualizací (tabulka 7-26).

Tab 7-26 Stav rozhodování ve třetí etapě C_2

	počet známých objektů	počet všech objektů	všechny objekty shluku			
				největší skupinka všech stabilních objektů v pohybu s trajektorií k: $E^{s}_{S_h}$	(%) $\frac{E^{s}_{S_h}}{(N_1)_{S_h}}$	**Stav rozhodování**
Shluky	N1	N	stabilní			
S1	35	35	1	1	**2,9%**	**M**
S2	3	3				**M**
S4	0	1				**M**
S5	19	19	1	1	**5,3%**	**M**
S6	45	45	5	5	**11,1%**	**M**

Jelikož došlo pouze k mechanické aktualizaci a žádný segment není zrušen, stav na konci C_2 je opět 6 tříd se stejnými vzory chování jako na konci C_1. Změnila se pouze distribuce některých objektů mezi shluky (graf 7-15 b, tabulka 7-27 a *příloha 11.4*, sloupce 29 a 30).

Tab 7-27 Počet objektů ve shlucích před a po aktualizaci v t_3^2 ***a*** t_{konc}^2

		objekty v t3 klasifikované do shluků vytvořených v t0					
		S1	S2	S4	S5	S6	N
shluky po aktualizaci v t_konc	S1	35	0	0	0	0	35
	S2	0	3	0	0	0	3
	S5	0	0	1	19	0	**20**
	S6	0	0	0	0	45	45
N		35	3	1	**19**	45	103

Poznámka: V tabulce je již vyjádřen přesun osamoceného objektu shluku 4 do nejbližšího shluku - shluku 5, ve kterém figuruje tento objekt jako outlier a není započítáván do charakteristik shluku 5.

4) Etapa IV.: Zánik shluků

Tato analýza se týká třídy 3, která nevlastní žádný objekt a třídy 4 s jedním objektem.

Objekt ze segmentu 4 bude pro ostatní shluky nejspíš outlierem, proto bude buď eliminován nebo zařazen do nejbližšího shluku, ovšem nebude mít vliv na změnu vzorů chování a tím ani na realizaci následných procesů post aktualizace vzorů chování shluku.

Pro ujasnění termínu paměť registru a jejího operování je v tabulce 7-28 ukázáno, jak je do této paměti zapisována informace o třídách, které nevlastní dostatečný počet objektů.

Třída 3 neměla ve druhém cyklu žádný objekt (všechny její objekty byly klasifikovány do jiného shluku). Je třeba sledovat, zda po dva následující cykly tento shluk nepřijme žádný nový ani známý objekt – v tomto případě by došlo na konci čtvrtého cyklu k jeho zániku. Ovšem ve skutečnosti ve třetím cyklu (který v *sekci 7.1* není analyzován) je do shluku 3 znovu zařazen jeden jeho bývalý objekt (objekt číslo 48). To znamená, že nyní sice vlastní menší počet objektů, než je definovaná hranice minimálního možného počtu objektů, ovšem cyklus sledování paměti registru je přerušen zařazením objektu 48 a je proto nutné opakovat ten samý proces a registrovat stejnou informaci po další dva cykly.

Ve shluku 4 přestávají tři objekty existovat, jeden objekt se přesunuje do jiné třídy a jeden objekt je agregován. Tento objekt je v paměti registru po dva následující cykly registrován jako jediný; bude proto na konci čtvrtého cyklu rozhodnuto o zrušení čtvrtého shluku.

Tab 7-28 Paměť registru mezi cykly C_1 až C_4 pro shluky 3 a 4

			Existence objektů mezi cykly			
Shluk	**Objekty**	**Identifikátor**	**C1 t_konc**	**C2 t_konc**	**C3 t_konc**	**C4 t_konc**
3	$\{X_{i\bullet}\}_i$	24	1	0	0	0
		6	1	0	0	1
		48	1	0	1	0
		55	1	0	0	0
		65	1	0	0	0
	$\{Y_{i\bullet}\}_i$		0	0	0	0
4	$\{X_{i\bullet}\}_i$	30	1	0	0	0
		13	1	0	0	0
		20	1	0	0	0
		28	1	0	1	1
	$\{Y_{i\bullet}\}_i$	106	0	1	1	1

Poznámka: Paměť registruje objekty mezi cykly po jednotlivých obdobích. V tabulce 7-28 jsou však pro zjednodušení uvedeny informace po cyklech. ***1*** *znamená existence objektů,* ***0*** *znamená negace existence.*

Neprovádí se opravy zařazení objektů, jelikož nejsou změněny vzory chování.

Tento stav je považován za ukončení druhého cyklu aktualizace vzorů chování.

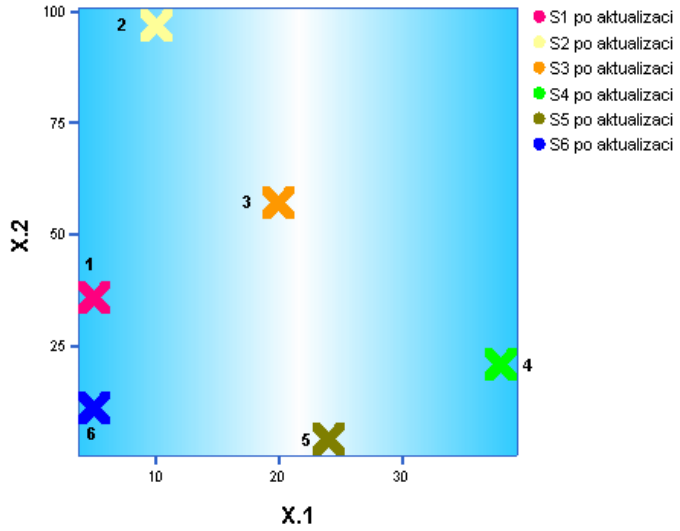
Touto etapou je také ukončen cyklus výběru možných scénářů změn.

6) Etapa V.: Identifikace trajektorií

Objekt s identifikátorem 106, který by byl ve shluku 4 jediným objektem, je klasifikován do shluku 5, aby nedošlo k jeho eliminaci. Neovlivňuje ovšem strukturu shluku 5, proto průměrná i maximální vzdálenost shluku 5 pro další cyklus bude platná ta, která byla získána v období t_3^2. Shluk 4 jako takový je však stále v platnosti, v tomto cyklu se neruší. To samé platí pro shluk 3.

Reálné trajektorie aktualizovaných shluků jsou zobrazeny v grafu 7-18.

Graf 7-18 Reálné trajektorie pohybu shluků mezi t_{poc}^2 ***a*** t_{konc}^2



Poznámka: Středy shluků v t_{konc}^2 *jsou identické středům shluků z* t_{konc}^1.

Jelikož nejsou aktualizovány vzory chování mezi prvním a druhým cyklem, středy shluků zůstávají identické těm na počátku C_2.

Charakteristiky všech shluků na konci druhého cyklu jsou shrnuty v tabulkách 7-29, 7-30, 7-31 a 7-32 (středy shluků, počet objektů spadajících do každého shluku, vzdálenosti středů shluků od sebe navzájem a vzdálenosti objektů od středu svých shluků). Tento stav je výchozím stavem pro třetí cyklus (t_{poc}^3).

Tab 7-29 Středy shluků v t^2_{konc} (t^3_{poc})

	Středy shluků					
Proměnné	s1	s2	s3	s4	s5	s6
X.1	5	10	20	38	24	5
X.2	36	97	57	21	4	11

Tab 7-30 Počet objektů ve shlucích v t^2_{konc} (t^3_{poc})

Shluky	**N**
S1	35
S2	3
S3	0
S4	0
S5	20
S6	45
Celkem	103

Tab 7-31 Vzdálenosti středů shluků navzájem v t^2_{konc} (t^3_{poc})

Vzdálenosti mezi středy shluků	S1	S2	S3	S4	S5	S6
S1	0					
S2	61,2	0				
S3	25,8	41,2	0			
S4	36,2	81,0	40,2	0		
S5	37,2	94,0	53,2	22,0	0	
S6	25,0	86,1	48,4	34,5	20,2	0

Průměrná vzdálenost objektů od středů shluků v t^2_{konc} se oproti t^2_3 nemění.

Tab 7-32 Vzdálenosti objektů v t^2_3 ke středu příslušného shluku v t^2_{konc} (t^3_{poc})

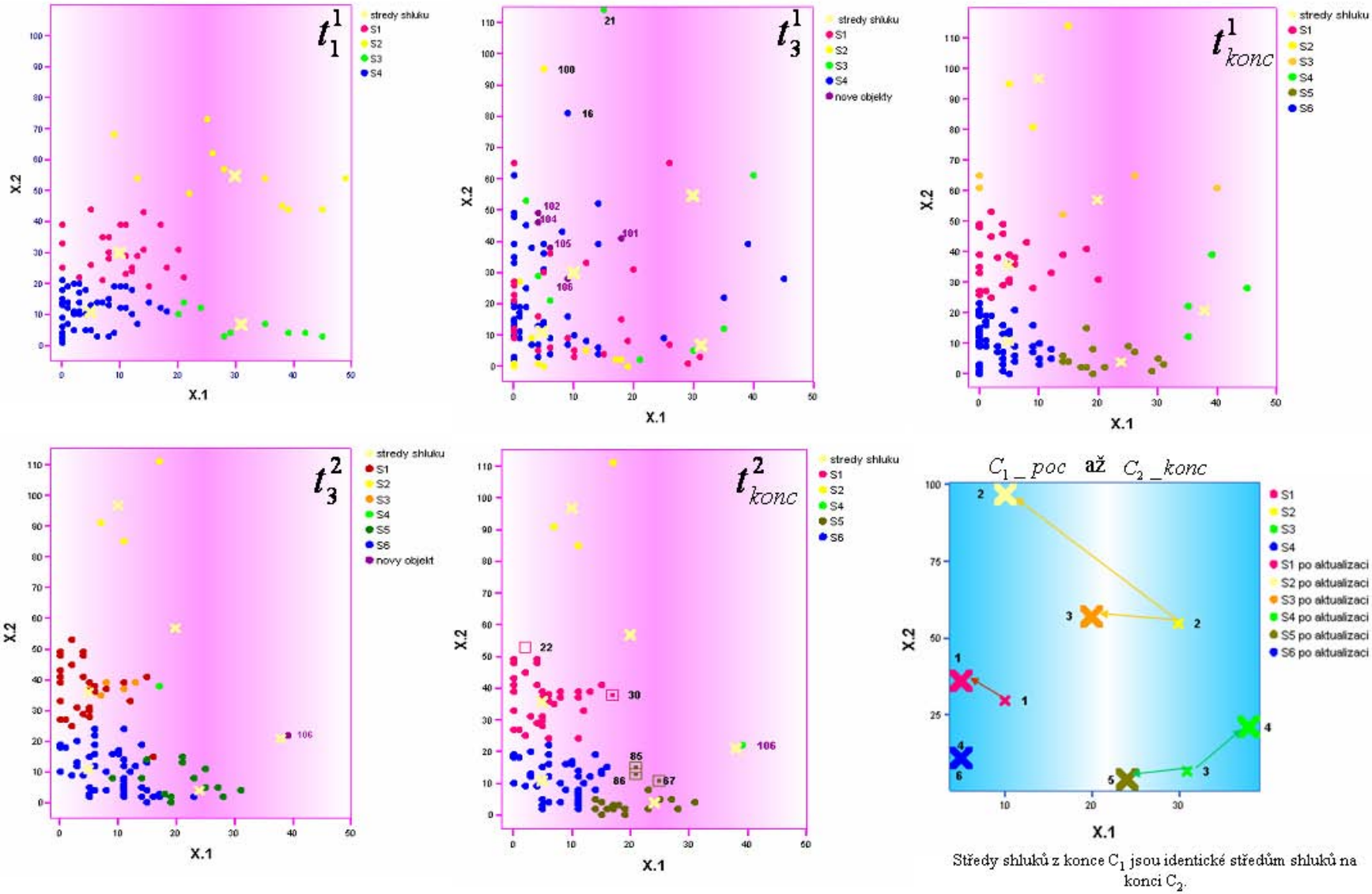
Shluky	Průměr	Min	**Max**	N1
S1	7,7	,0	17,8	35
S2	11,5	6,7	15,7	3
S3	-	-	-	0
S4	1,4	1,4	1,4	0
S5	6,8	1,4	11,4	20
S6	7,2	1,4	12,0	45

Na uzavření sekce aplikace obecné metodologie pro případ s identifikátorem algoritmem k-means na datech simulovaných je v tabulce 7-33 shrnut stav nejdůležitějších charakteristik prvních dvou etap sledovaných shluků dvou analyzovaných cyklů a v grafu 7-19 je sledován pohyb objektů, vývoj středů shluků mezi obdobími C_1 a C_2 a zobrazeny reálné trajektorie pohybu shluků.

Tab 7-33 Charakteristiky shluků v průběhu aplikace obecné metodologie v C_1 a C_2

			N		souřadnice S_h		$d(X_{i\bullet}, S_h)$		$d(s_m, s_n)$						Typ aktualizace
C_c	t_l	S_h	N_1	N_2	$X_{\bullet 1}$	$X_{\bullet 2}$	průměr	max	S_1	S_2	S_3	S_4	S_5	S_6	
C_1	t_{poc}	S1	25		10	30	8,3	14,6							
		S2	11		30	55	13,7	24,7	32,0						
		S3	9		31	7	8,8	14,1	31,2	48,1					
		S4	55		5	11	6,8	13,2	19,6	50,6	26,4				
N			100												
C_1	t_3	S1	25	5			11,3	24,1							inteligentní
		S2	9	0			29,0	60,9							nové třídy
		S3	13	0			10,7	25,2							nové třídy
		S4	53	0			6,8	13,8							mechanická
N			100	5											
C_1	t_{konc}	S1	28		5	36	8,5	17,3							
		S2	3		10	97	13,0	17,7	61,2						
		S3	5		20	57	16,0	21,5	25,8	41,2					
		S4	4		38	21	10,2	18,0	36,2	81,0	40,2				
		S5	14		24	4	7,1	12,5	37,2	94,0	53,2	22,0			
		S6	51		5	11	6,6	13,0	25,0	86,1	48,4	34,5	20,2		
N			105												
C_2	t_3	S1	35	0			7,7	17,8							mechanická
		S2	3	0			11,5	15,7							mechanická
		S3	0	0											
		S4	0	1			1,4	1,4							mechanická
		S5	19	0			6,8	11,4							mechanická
		S6	45	0			7,7	24,0							mechanická
N			102	1											
C_2	t_{konc}	S1	35		5	36	7,7	17,8							
		S2	3		10	97	11,5	15,7	61,2						
		S3	0		20	57			25,8	41,2					
		S4	0		38	21	1,4	1,4	36,2	81,0	40,2				
		S5	20		24	4	6,8	11,4	37,2	94,0	53,2	22,0			
		S6	45		5	11	7,2	12,0	25,0	86,1	48,4	34,5	20,2		
N			103												

Graf 7-19 Přehled reálných trajektorií pohybu shluků mezi t^1_{poc} a t^2_{konc}



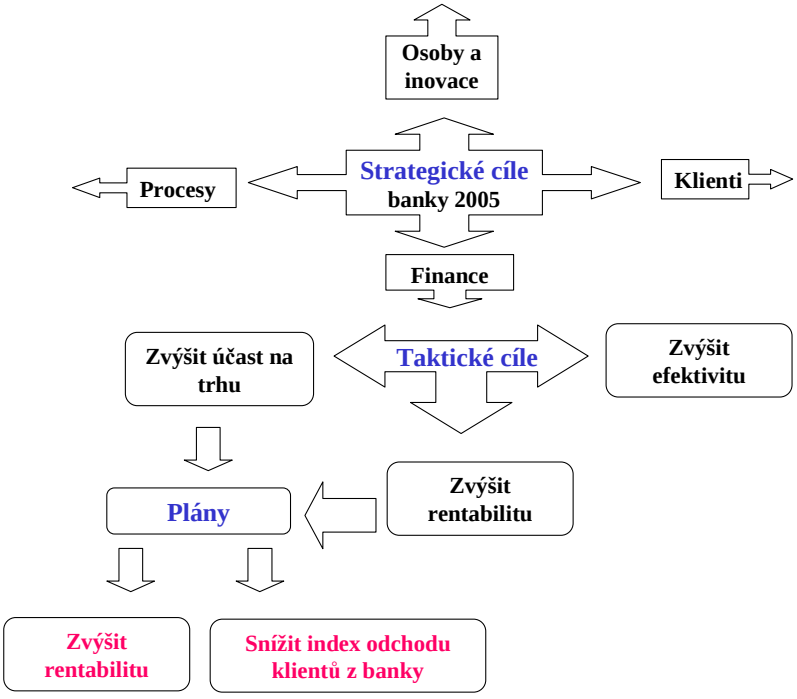
Středy shluků z konce C_1 jsou identické středům shluků na konci C_2.

7.2 Aplikace na data skutečná(*)

7.2.1 Analyzovaný problém

Bankovní oddělení Řízení korporativní báze dat má za úkol v rámci BIO'S (Business Improvement Opportunities) zrentabilizovat nový projekt Data Warehouse, kterým je právě probíhající migrace dat do nově zakoupeného a implementovaného logického modelu dat. S úmyslem vybrat takový projekt rentability, který by korespondoval s ročními cíly korporace jsou nejprve zrevidovány strategické cíle banky na rok 2005. Cíle se týkají několika okruhů, kterými jsou: osoby a inovace, procesy, klienti a finance. Do finančního okruhu každoročně patří úkoly zvýšení zisku, zvýšení účasti na trhu a zvýšení efektivity. Plány, kterými jmenované bankovní oddělení může v rámci finančních záměrů přispět ke splnění některých cílů roku 2005 a zároveň se přímo podílet na rentabilitě Data Warehouse využíváním a analýzou jeho informací, jsou zobrazeny v obrázku 7-1.

Obr 7-1 Cíle a některé plány Bci na rok 2005

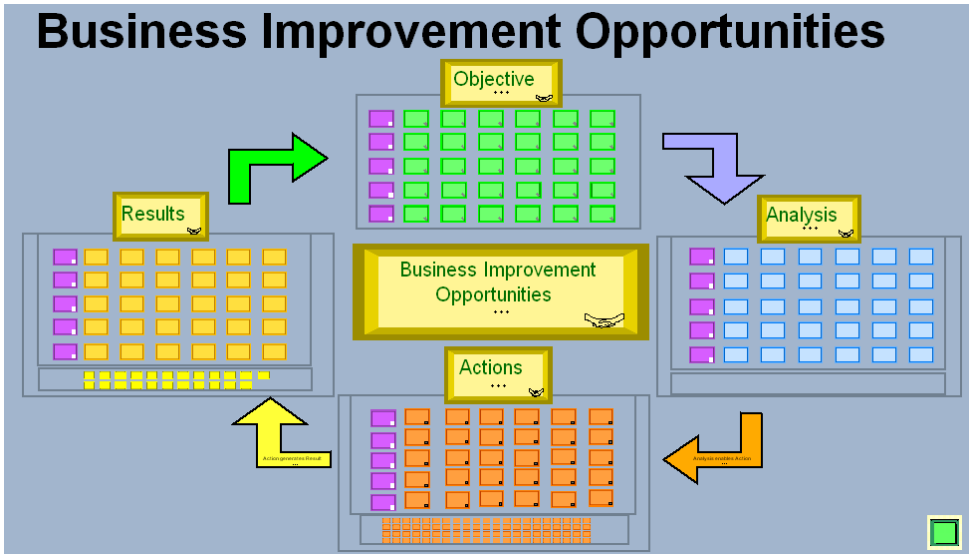


* *Sekce 7.2* je ve srovnání se *sekcí 7.1*, kde byla s detaily vysvětlena aplikace obecné metodologie, méně podrobná. Cílem sekce *Aplikace na data skutečná* již není vysvětlit aplikaci a porozumět jí, nýbrž sledovat její funkčnost na datech skutečných a ukázat tak její praktické použití.

Konkrétním úkolem, který je k řešení vybrán, je identifikovat potenciálně rentabilní klienty s vyšší tendencí zavřít svůj běžný účet (což je vyjádřeno *indexem odchodu*) a posléze v kooperaci s oddělením marketingu zabránit těmto klientům v odchodu z banky. Klienti, na které se analýza zaměří, budou mikropodnikatelé zemědělského sektoru (vinice, rybolov a mořské plody, ovocnářství, export zemědělských produktů a produkce mléka).

K řešení problému je vybrána metodologie práce BIO´S (obr. 7-2, patentovaná firmou NCR), která bude řídit proces plnění cíle. Spočívá v sekvenčním průchodu čtyř etap: stanovení cíle, analýza problému (definice *otázek obchodu* a klíčových indikátorů, extrakce dat a aplikace dataminingových a statistických technik na vyvinutí deskriptivního nebo prediktivního modelu), komerční akce a zhodnocení výsledků sledováním metrických indexů. Metodologie, vyvinutá v této práci, spadá do etapy analýzy problému, konkrétně modelace problému deskriptivními a prediktivními technikami. Jsou definovány tři základní otázky obchodu: Kdo je klientem s potenciální rentabilitou? Kteří klienti mají tendenci v nejbližším období odejít z banky? Jaké je chování skupin klientů, segmentovaných dle hodnot potenciální rentability a indexu odchodu?

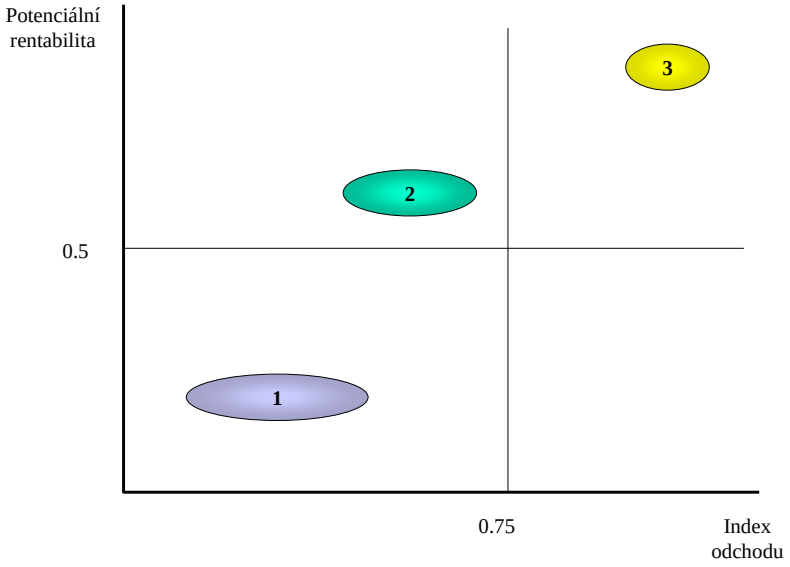
Obr 7-2 Schéma metodologie práce BIO‘S



Zdroj: NCR

Po vymodelování úlohy a zhodnocení chování klientů přijdou na řadu obchodní akce, jako například: pro klienty, kteří mají vysokou (vyšší) tendenci odejít z banky a vysokou potenciální rentabilitu (obrázek 7-3, shluk 3) bude vytvořen plán jejich zadržení a to nabídkou produktu, který jim chybí, zlepšením kvality služeb či zrušením poplatků za jisté služby. Klientům se středně vysokou potenciální rentabilitou a střední tendencí odchodu (obrázek 7-3, shluk 2) bude poslán dárek či zatelefonováno s blahopřáním k narozeninám. Pro klienty s nízkým koeficientem odchodu a nízkou rentabilitou (obrázek 7-3, shluk 1) bude nabídnut jistý kreditní produkt s úmyslem je zrentabilizovat. Pro klienty z posledního kvadrantu obr. 7-3 není v tomto momentě připraven žádný program.

Obr 7-3 ***Sloučení klientů do tříd dle hodnot potenciální rentability a indexu odchodu***



K modelaci navrhovaného problému je vybrána technika segmentace, konkrétně algoritmus obecné metodologie, která je díky svým charakteristikám považována za nejvhodnější ke splnění stanoveného úkolu. Záměrem je totiž provádět konstantní pozorování pohybu klientů a dle aktuálního stavu a chování klientů v minulosti předpovědět jejich budoucí vývoj. Užitím obecné metodologie dynamické segmentace tak bude možné vyvíjet post segmentační procesy *manipulace* s vybranými klienty dle jejich tendenčního chování v čase, tzn. realizovat komerční kampaně *šité jim na míru* v okamžiku, kdy je to nejvhodnější se záměrem zabránit odchodu klientů s potenciálem. Využitím dynamické segmentace a její schopnosti předpovědět budoucí tendence není nutné čekat, až

se klient přesune z jednoho segmentu do jiného a teprve potom vyvíjet příslušné aktivity; s výhodou před konkurencí je možné využít příležitosti obchodu a to regulací zákazníkova pohybu - pobízením přesunu do segmentu rentabilního s nízkým indexem odchodu a naopak zabránit pohybu do rizikového a nerentabilního segmentu. O každém klientovi je k dispozici pohled na jeho chování v průběhu času a tak i na aplikované logicky navazující komerční kampaně, jelikož analýza v každém období je postavena na segmentaci vytvořené v období předchozím. Užití výsledků aplikace je široké, proto kromě hlavního cíle je možné například i rentabilizovat některé skupiny klientů. Popsaná strategie zvýhodní banku ve flexibilnějším, rychlejším, ekonomicky a konkurenčně výhodnějším rozhodování.

7.2.2 Data

Z pohledu modelátora je v první řadě třeba analyzovat vybraný problém z perspektivy disponibilních dat. Sledováním hodnot proměnných jsou také definovány koeficienty kritérií etap procesu obecné metodologie.

Data přísluší informaci 1.056 mikropodnikatelů zemědělského sektoru Bci se svým identifikátorem a *rut*[(*)]. Tato skupina je pořízena výběrem z celku klientů Bci *Banka Mikropodnikatelé*. Počáteční počet tříd je volen technikou segmentace ve dvou fázích. Je pracováno pouze se dvěma proměnnými[(**)]. Proměnná potenciální rentabilita je však kombinací čtyř atributů: segment (vysoký, střední, nízký), index *mateřské banky* (dluh uvnitř banky/dluh ve finančním systému), idex *křížení produktů* a existence služeb *platby závazků za dodavateli*. Vytvořením proměnné potenciální rentabilita je zodpovězena první definovaná otázka obchodu. Techniky použité na vytvoření popsané proměnné jsou analýza hlavních komponent, scoring a logit, který na základě údajů z minulosti předpoví pro jisté budoucí období, bude-li klient v tomto období rentabilní či nikoliv. Tato kvantitativní spojitá proměnná bude převedena do intervalu hodnot od 0 do 1 (tabulka 7-34).

Vysvětlením druhé otázky obchodu je proměnná tendence odchodu klientů z banky a tak jako první proměnná, potenciální rentabilita, není pevným atributem přímo získaným z báze dat; je třeba jej předpovědět. K jeho předpovědi bude užita technika neuronové sítě

---

* *Rut* je identifikační číslo každého občana v Chile srovnatelné s OP Čechů.

** Bylo ověřeno, že pro více atributů operuje obecná metodologie stejným způsobem.

kombinující proměnné demografického charakteru a bankovního chování (proměnné zařazené do modelu, architektůra sítě, metoda učení, stop parametry a stav *trainingu* jsou uvedeny v *příloze 11*.5). Proměnná bude kvantitativní spojitá a bude dosahovat hodnot od 0 do 1. Obě proměnné jsou formovány z měsíčních informací, získaných v jednom determinovaném momentu v čase počínajíc dubnem 2004. Jsou vybrána data historická se záměrem provést vyhodnocení analýzy (jak tendence chování, tak efektivity předpovědi odchodu klientů z banky) na skutečných datech[(*)].

Tab 7-34 Převod skutečných hodnot potenciální rentability do intervalu (0,1)

Decily	Potenciální rentabilita_ duben 2004	Potenciální rentabilita_duben 2004 v rozmezí (0,1)
1	-32.466	,0503
2	1.965	,1502
3	5.434	,2501
4	9.418	,3499
5	14.317	,4498
6	20.912	,5502
7	27.225	,6501
8	38.155	,7499
9	60.017	,8503
10	257.437	,9502

Z tabulky 7-34 je možné srovnávat hodnoty z intervalu (0,1) a jim ekvivalentní skutečné hodnoty potenciální rentability.

7.2.3 Aplikace obecné metodologie

Data použitá v následující aplikaci jsou zpracována v souladu s následující dynamikou: výchozí stav v počátečním období t_{poc} prvního cyklu C_1 je tvořen pěti třídami, kde se aktualizují objekty po čtyři navazující cykly (každý cyklus je složen ze tří období: první t_1^c, meziobdobí t_2^c a třetí t_3^c) do maximálně šesti shluků způsobem, že v prvním cyklu je jeden shluk rozdělen mezi dvě nové třídy, jeden je aktualizován pohybem a do zbývajících jsou pouze klasifikovány objekty. V dalších cyklech jsou shluky aktualizovány mechanickou a především inteligentní aktualizací (pohyb v obou nebo jen v jedné proměnné). Neexistuje

* Tak se také provádí hodnocení výkonnosti prediktivních modelů v praxi.

třída, která by obsahovala nedostatečný počet objektů a proto žádný ze shluků nebude zrušen ani nebude kandidátem na svou eliminaci.

Vektory proměnných, které definují shluky, jsou dva: : $\mathbf{X}_{\bullet 1}^{t_l^c}$ a $\mathbf{X}_{\bullet 2}^{t_l^c}$. Analýza objektů ve smyslu jejich dynamiky je sledována mezi t_1^c a t_3^c v obou proměnných odděleně.

Vytvořená obecná metodologie je aplikována na každý shluk zvlášť na konci každého cyklu, kdy je analyzován stav a chování objektů a v rámci scénářů změn vybírány formy aktualizace formátů chování shluků.

V *příloze 11.6.* jsou uvedeni sledovaní klienti Bci, mikropodnikatelé zemědělského sektoru, se svým identifikátorem a rut, zařazením do původních shluků a s hodnotami původních proměnných po všechna sledovaná období čtyř cyklů.

Stálá kritéria aplikovaná během čtyř analyzovaných cyklů jsou následující:

- Výchozí stav tříd: $h \in \{1,2,3,4,5\}$.
- V etapě I, Identifikace objektů, které představují změnu,

  jsou tyto v segmentu S_h identifikovány výpočtem a porovnáním Eukleidovy vzdálenosti následovně:

  $$\mathbf{X}_{i\bullet} \Rightarrow \mathbf{X}_{i\bullet}^{o}, \text{ jestliže } d(\mathbf{X}_{i\bullet}^{t_{konc}^c}, s_h^{t_{poc}^c}) > d_{\max s_h^{t_{poc}^c}} \text{ (platí i pro } \mathbf{Y}_{i\bullet}).$$

- V etapě II, Rozpoznání stavu změny,

  je tento v segmentu S_h rozpoznán, jestliže $e_{S_h}^{o} > 0.2 \ * N_{S_h}$.

- V etapě III., Rozhodnutí o možnostech aktualizace shluků,

  je provedena aktualizace klasifikací do shluku S_h, jestliže pro obě proměnné sledované zvlášť platí $E_{S_h}^{s} \le 0.25 \ * (N_1)_{S_h}$ nebo alespoň pro jednu proměnnou platí $(E_{S_h}^{s}) > 0.25 \ * (N_1)_{S_h}$ a zároveň pro ni platí $0{,}5 * (E_{S_h}^{s,k}) \le (e_{S_h}^{s,(-k)})$.

  V případě opačném, kdy alespoň jedna proměnná splňuje $E_{S_h}^{s} > 0.25 \ * (N_1)_{S_h}$ a zároveň splňuje také $0{,}5 * (E_{S_h}^{s,k}) > (e_{S_h}^{s,(-k)})$, dochází k aktualizaci pohybem shluků (v proměnné, která splňuje uvedená kritéria). Tyto aktualizace jsou provedeny pro situace, kdy není rozpoznán stav změny.

Při rozpoznání stavu změny, kdy alespoň pro jednu proměnnou platí $E^{o,s}_{S_h} \leq 0.8 * e^{o}_{S_h}$ nebo $E^{o,s}_{S_h} + e^{s}_{S_h} \leq 0.25 * N_{S_h}$, dochází k vytvoření nových tříd. V případě splnění obou dvou opačných kritérií je aplikována aktualizace pohybem shluků.

Aktualizace středů tříd při změně formátů chování alespoň v jedné proměnné se provádí MNČ.

- V etapě IV, Zrušení shluků,
  dochází k eliminaci shluku, jestliže $e_{S_h} < 0.03 * N_1$ a zároveň pro tento shluk a $\forall \mathbf{X}_{i\bullet} \notin S_{h^*}$ v t^{c}_{konc} platí:

  $$\mathbf{X}_{i\bullet} \notin S_{h^*} \text{ v } C_{c+1} \text{ až } C_{c+2} \text{ (platí pro každé období cyklu).}$$

- V etapě V, Identifikace trajektorií,
  jsou tyto vyjádřeny spojnicí $\mathbf{T}^{c}_{S_h\bullet}$, která přísluší vektoru, který obsahuje trajektorie shluku S_h v čase, tedy souřadnice svého středu v čase.

Koeficient α rovný 20% je poměrně vysoce zvolená hranice změny, pro řešený případ však vhodná. Existuje totiž dostatečný počet tříd a úmyslem není tento počet zvyšovat.

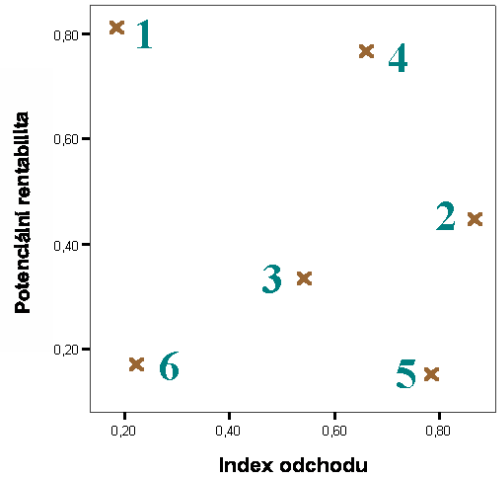
Co se týče chování objektů: i když jsou nalezeny objekty stabilní v pohybu v jedné nebo druhé proměnné, stabilních objektů v pohybu, majících stejnou trajektorii v obou proměnných zároveň, je málo. Je proto volena strategie inteligentní aktualizace v případě, je-li ve shluku počet objektů se stabilním chováním alespoň v jedné proměnné $> 0.25 * (N_1)_{S_h}$ a zároveň počet objektů shluku stabilních v pohybu s jistou trajektorií v této proměnné je větší více než o 50% než je počet objektů pohybujících se ve stejné proměnné stabilním protisměrným pohybem. To znamená, že trajektorie jsou konstruovány na základě alespoň jedné proměnné. Hranice 25% pro koeficient stability objektů je volena z toho důvodu, že 25% stabilních objektů je považováno za kritérium dostatečně *silné*, aby mohlo být mluveno o pohybu shluku a na druhou stranu dostatečně *volné*, aby mohlo být s objekty v segmentech zacházeno jako s prostředkem metody předpovědi chování objektů do budoucnosti; pro realizaci post segmentačních aktivit není třeba čekat na splnění kritérií s vyšší hranicí stability, nýbrž je třeba předem včas organizovat a realizovat následné konkurenční procesy na podpoření či postavení se tendenčnímu chování objektů.

V řešeném příkladě nedochází k agregaci nových klientů; jejich celkový počet mezi cykly se ovšem mění. Důvodem je chybějící údaj některé ze základních proměnných v některém z cyklů (období).

V průběhu analýzy není třeba měnit na počátku definovaná kritéria, jelikož vyhovují[(*)] jak při sledování hetoregenních objektů, tak tendence v pohybu a směru. V opačném případě by byla upravována kritéria etap procesu (procenta koeficientů).

Počáteční shluky před jakoukoliv aktualizací vzorů chování shluků zobrazuje graf 7-20.

Graf 7-20 Shluky v t^1_{poc}



V souhrnné tabulce 7-35 aplikace obecné metodologie jsou ukázány hlavní charakteristiky, které v každé etapě (kromě třetí etapy, jelikož její vyjádření je komplexní) determinují rozhodování o výsledné aktualizaci. Aktualizované středy shluků jsou zobrazovány pro každý cyklus v posledních dvou sloupcích tabulky.

* Nedochází k chaosu a je sledovatelná tendence v pohybu.

Tab 7-35 Charakteristiky shluků v průběhu aplikace obecné metodologie v C_1 až C_4

		Středy t_poc									Středy t_konc		
Cykly	Shluky	IO	PO	N_1	Outliery	Rozhodování I. etapa	% Outlierů	Rozhodování II. etapa	Rozhodování III. etapa	Aktualizovaná proměnná	IO	PO	
1 cyklus	1	0,163	0,826	253	15	I,M,N	6%	I,M	M		0,163	0,826	
	2	0,859	0,449	281	24	I,M,N	9%	I,M	M		0,859	0,449	
	3	0,256	0,110	170	15	I,M,N	9%	I,M	I	IO	0,238	0,110	
	4	0,530	0,747	182	25	I,M,N	14%	I,M	M		0,530	0,747	
	5	0,621	0,224	170	42	I,M,N	25%	I,N	N	z 5 => 5 a 6	0,732	0,137	
	Celkem			1056	121		11%				0,544	0,278	
2 cyklus	1	0,163	0,826	228	2	I,M,N	1%	I,M	I	IO	0,169	0,826	
	2	0,859	0,449	192	1	I,M,N	1%	I,M	I	IO	0,869	0,449	
	3	0,238	0,110	126	1	I,M,N	1%	I,M	M		0,238	0,110	
	4	0,530	0,747	178	5	I,M,N	3%	I,M	I	IO	0,566	0,747	
	5	0,732	0,137	112	8	I,M,N	7%	I,M	I	IO	0,764	0,137	
	6	0,544	0,278	136	17	I,M,N	13%	I,M	M		0,544	0,278	
	Celkem			972	34		3%						
3 cyklus	1	0,169	0,826	210	1	I,M,N	0%	I,M	I	PR	0,169	0,812	
	2	0,869	0,449	178	1	I,M,N	1%	I,M	M		0,869	0,449	
	3	0,238	0,110	133	0	I,M	0%	I,M	I	IO,PR	0,225	0,182	
	4	0,566	0,747	192	2	I,M,N	1%	I,M	I	PR	0,566	0,766	
	5	0,764	0,137	99	0	I,M	0%	I,M	I	IO,PR	0,786	0,149	
	6	0,544	0,278	143	0	I,M	0%	I,M	I	PR	0,544	0,335	
	Celkem			955	4		2%						N
4 cyklus	1	0,169	0,812	243	0	I,M	0%	I,M	I	IO	0,186	0,812	253
	2	0,869	0,449	159	2	I,M,N	1%	I,M	M		0,869	0,449	148
	3	0,225	0,182	152	0	I,M	0%	I,M	I	PR	0,225	0,172	152
	4	0,566	0,766	180	1	I,M,N	1%	I,M	I	IO	0,662	0,766	174
	5	0,786	0,149	127	1	I,M,N	1%	I,M	I	PR	0,786	0,152	125
	6	0,544	0,335	148	1	I,M,N	1%	I,M	M		0,544	0,335	156
	Celkem			1009	5		3%						1 008

Poznámka: IO vyjadřuje index odchodu, PO potenciální rentabilitu.

Poslední, pátá etapa procesu, shrnuje konečné charakteristiky shluků při ukončení každého časového cyklu a promítá stav vývoje dynamické segmentace. To umožňuje sledovat konkrétní stavy shluků a jejich variace po reálných trajektoriích svých středů v čase, což přísluší zodpovězení třetí otázky obchodu.

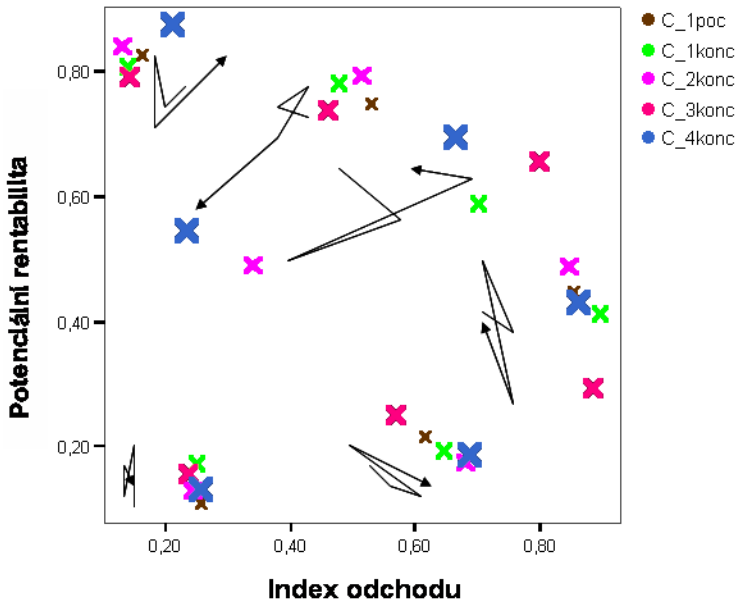
I když k řešení případu je jako optimální vybrána obecná metodologie, ke srovnání a budoucí analýze (provedené v *sekci 8*) jsou shluky paralelně aktualizovány dalšími dvěma metodami (strategiemi) segmentace, kterými jsou klasická metoda k-means a segmentace ve dvou fázích. V *příloze 11.7* je však možné sledovat pět variant výsledků. Jelikož počet vzniklých přírodních tříd se u jednotlivých technik liší, pro efekty jejich srovnávání je třeba v každém cyklu vyrovnat počet tříd (za základ je volen počet tříd, který vytvořila obecná metodologie). Tak vzniká *k-means vyrovnaná* a *segmentace ve dvou fázích vyrovnaná*.

Následně jsou tři výše popsané *vyrovnané* techniky segmentace prezentovány ve vývojových grafech 7-21 a) až c), které dovolují sledovat trajektorie shluků a výkyvy (chaotické chování) ve shlukování. Pro vyhodnocení kvality předpovědi vývojových tendencí stavů shluků (ve smyslu tendence zavřít běžný účet a tendence hodnot potenciální

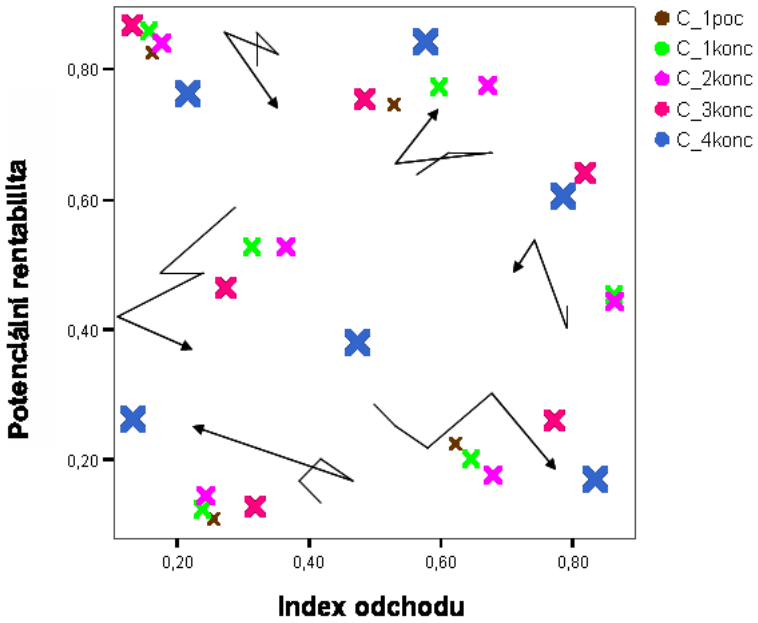
rentability) a jejich srovnání se skutečně realizovanými středy shluků v budoucnu jsou užity dvě období, a to květen a červenec 2005 (cyklus 4 končil březnem 2005). V grafu 7-21 c) je realizováno popsané pozorování obecnou metodologií.

Graf 7-21 Trajektorie tendence chování shluků v C_1 až C_4

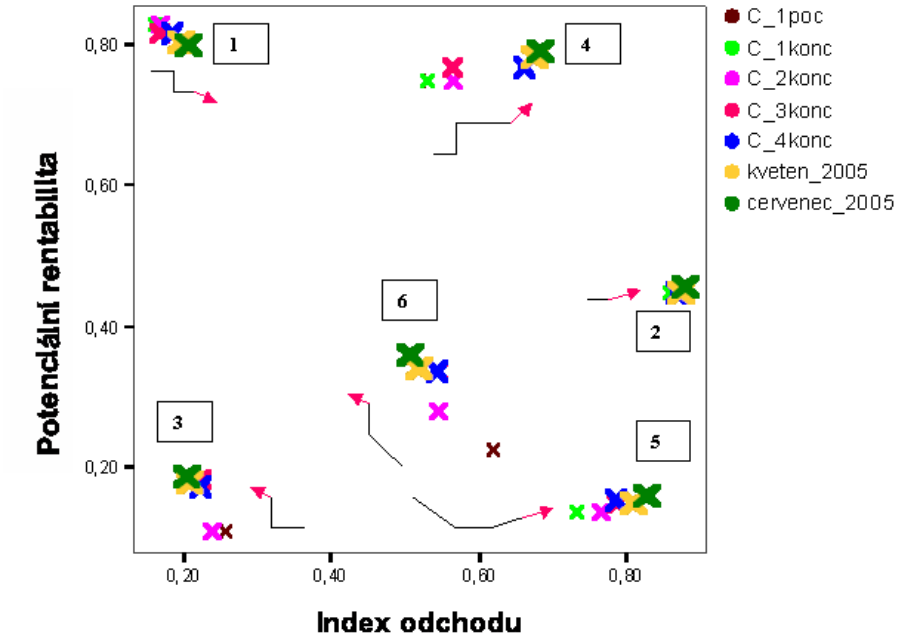
a) metodou segmentace ve dvou fázích vyrovnaná



b) metodou k-means vyrovnaná



c) obecnou metodologií se znázorněním skutečných shluků v budoucím období



Poznámka: Křížky označují středy shluků, které jsou očíslovány. Posloupnost cyklů je uvedena barevně v legendě - například: C_1poc vyjadřuje t^1_{poc} *pro každý ze šesti sledovaných shluků.*
Černé křivky vyznačují reálné (v případě obecné metodologie) nebo identifikované (v případě zbylých dvou strategií) trajektorie aktualizace vzorů chování shluků a naznačují tendenci chování. Červené šipky vyjadřují skutečné stavy středů shluků v předpovězeném období.
Tendence je trajektoriemi a šipkami pouze naznačena, tak jako i velikost pohybu.

Po aktualizaci charakteristik shluků na konci posledního sledovaného cyklu jsou pro každý shluk navrhnuty marketingové strategie komerčních kampaní, které jsou představeny v tabulce 7-36.

Tab 7-36 Marketingové strategie v t^{4}_{konc}

		Proměnné	
Shluk	**Analýza**	**Index odchodu**	**Potenciální rentabilita**
1	stav	nízká	extrémně vysoká
	tendence	roste	klesá
	strategie	**Udržovat**	
2	stav	extrémně vysoká	průměrná
	tendence	roste	konstantní průměrná
	strategie	**Bez akce**	
3	stav	velmi nízká	nízká
	tendence	klesá	fluktuuje
	strategie	**Bez akce**	
4	stav	**vyšší**	**vysoká**
	tendence	**roste**	**roste**
	strategie	**Maximální pozornost**	
5	stav	vysoká	velmi nízká
	tendence	roste	roste
	strategie	**Sledovat**	
6	stav	průměrná	podprůměrná
	tendence	klesá	roste
	strategie	**Udržovat**	

Analýza a vyhodnocení užitých technik (především z grafu 7-21) tak jako návrh pilotního projektu komerčních kampaní, zakládajících se na výsledných aktualizacích, jsou provedeny v *sekcích 8.1 a 8.2***.**

8 VÝSLEDKY

V *sekci 7* ukázal analytik na datech simulovaných a skutečných aplikaci a funčnost obecné metodologie - všech navržených etap a scénářů změn. Automatickým sekvenčním procesem procházel prostřednictvím metod a kritérií etapami procesu, které dovolily poznat charakteristiky a chování objektů a shluků. Výsledkem aplikace obecné metodologie jsou optimálně aktualizované vzory chování. Proces vycházel ze struktury segmentů vytvořené před první aktualizací a poté v každém cyklu z aktualizované struktury shluků vytvořené v předchozím období.

Až doposud neexistovala metoda segmentace založená na dynamickém chování objektů. Teoreticky vzato by proto měla obecná metodologie vyjadřovat chování a stavy v reálném (dynamickém) prostředí lépe než algoritmy statické právě díky zahrnutí konceptu dynamiky do shlukování. Na potvrzení této hypotézy je třeba zavést konkrétní kritérium a za jeho použití vyhodnotit kvalitu aktualizace, zdůraznit rozdíly ve výsledcích aplikace obecné metodologie a segmentace využívající existujících algoritmů shlukování a analýzou výsledků vyzdvihnout přednosti použití obecné metodologie v analýze shlukování v praxi.

8.1 Konstrukce kritérií na srovnávání algoritmů

Rozdíl mezi statickými a dynamickými algoritmy segmetnace je sledován ze dvou perspektiv. První je *technický přístup*, který je zajímavý a důležitý především pro samotné modelátory, vybírající pro svou práci nejvhodnější techniky a jejich metodologie. Vhodnost techniky posuzují pro každý konkrétní případ s úmyslem vybrat tu, kterou by došli k co nejkvalitnějším výsledkům, které by kopírovaly skutečnost, byly by replikovatelné a použitelné v praxi. Použitelnost vytvořeného dataminingového modelu závisí také na jeho snadné a rychlé periodické proveditelnosti.

Nejenom pomocí technických atributů je možné ukázat rozdíly a kvalitu segmentací prováděných různými algoritmy. Přínos dataminingových modelací je v praxi měřen především jejich přímým finančním přínosem.

Na simulovaná data bude aplikováno technické kritérium srovnání rozdílných algoritmů segmentace a na skutečných datech bude předveden *ekonomický přístup*

vyhodnocení obecné metodologie z pohledu zisku vyplývajícího z předvídání budoucího stavu a chování shluků.

Srovnávanými algoritmy jsou obecná metodologie, reprezentující dynamický algoritmus shlukování, a dvě strategie se statickým algoritmem, zastoupené technikami k-means a segmentace ve dvou fázích, představující logický vývoj statického algoritmu k dynamickému[*].

8.1.1 Technické kritérium

V průběhu práce je zdůrazňována důležitost kontinuity vytvářených segmentací a proto potřeba v každém novém cyklu vycházet z aktualizace vytvořené v předchozím cyklu. Kromě zajištění návaznosti procesů post segmentace, jako jsou marketingové kampaně, faktem, že není vyvíjen úplně nový systém segmentace, se šetří čas a práce specialisty a tím i finance.

Požadavkem na výslednou segmentaci je také její kvalita ve smyslu přesnosti (nebo naopak chyby) v konstruované segmentaci. Tento požadavek je možné vyjádřit průměrnou a maximální vzdáleností objektů ke středu svého shluku.

Jak již bylo předesláno, díky analýze dynamického chování objektů je možné poznat a sledovat významné rysy v pohybu a směru objektů a tak definovat tendenci v chování shluků. To dovoluje předpovídat budoucí vývoj a stav jak objektů, tak i shluků a konstruovat tak nejoptimálněji definované shluky, které dovolí v relaci s post shlukovými aktivitami ovládat pohyb objektů: podporovat požadovaný směr či brzdit nežádoucí chování. Tak je možné být připravený (na rozdíl od konkurence) na budoucí trend a mít náskok ve vyvozování včasných a reálných rozhodování.

V návaznosti na právě popsané požadavky charakteristik algoritmu je možné zkonstruovat tři atributy pro srovnávání obecné metodologie (OM) s každou ze strategií (S): prvním je kvalita (chyba) shlukování; druhým atributem je čas potřebný na vykonání shlukování (implikuje také finanční náročnost) a posledním atributem je dynamika, která vypovídá o schopnosti strategie vyjadřovat skutečnost, použitelnost v praxi a kapacitu předpovídat budoucí vývoj shluků.

---

[*] Strategie I., III. a IV. ze *sekce 1.3.2*.

Technickým kritériem srovnávání algoritmů je následující formule:

$$v\ (OM,\ S) = \sum_{j=1}^{M} \left[\left(OM_j - S_j \right) * P_j \right],\ \text{kde}\quad 0 < P_j < 1.$$

Symbol v značí vzdálenost mezi hodnotami atributů OM a jednotlivých strategií, oceňovanými zvolenými vahami P_j; symbol j značí atribut (zde $j \in \{1,2,3,4\}$).

Kvalita shlukování je vyjádřena průměrnou a maximální Eukleidovou vzdáleností objektů ke středu shluku, čas v týdnech a dynamika binominální proměnnou s kvalitativními hodnotami *ano* a *ne*[(*)]. Z důvodu unifikace jednotek atributů pro účely srovnávání jsou tyto vystupňovány v rozsahu (–1, 0). Čím menší je chyba v kvalitě shlukování, menší čas potřebný na vývoj aktualizovaných vzorů chování a čím *dynamičtější* je strategie, tím zápornější je hodnota atributů.

Je-li výsledkem kritéria v srovnávaných algoritmů v celkovém porovnání všech zvolených atributů číslo blízké -1, pak vítězí OM nad ostatními strategiemi; naopak, číslo blízké 0 vyrovnává rozdíly mezi OM a S a kladné číslo upřednostňuje některou z hodnocených strategií. K posouzení citlivosti výsledku na jednotlivé atributy jsou zavedeny váhy.

Volba konkrétního technického kritéria je přikládána následujícím zdůvodněním: kritérium umožňuje studovat atributy kvalitativní i kvantitativní a to navíc v různých měrných jednotkách, jelikož jejich hodnoty je možné vystupňovat mezi určitý interval a pracovat tak s hodnotami z tohoto intervalu. Jsou uvažována různá hlediska pohledu na jeden problém. Uvedené technické kritérium zahrnuje také subjektivní přístup, který je vykompenzován přítomnými vahami, které navíc dovolují měřit citlivost výsledku na míře zastoupení jednotlivých atributů. Nezanedbatelnou výhodou je způsob interpretace výsledku – v procentech.

---

* Hodnoty některých atributů jsou přisouzeny na základě subjektivního vyhodnocení.

8.1.2 Ekonomické kritérium

Před rozhodnutím o implementaci metodologie je nutné ověřit (na skutečných datech z minulosti) funkčnost obecné metodologie z pohledu identifikace skutečných a předpovězených stavů a pohybů klientů. Následně je třeba analyzovat její finanční přínos a proto realizovat pilotní projekt využívající výsledky aplikace. Za tímto účelem je definován plán aktivit na zadržení klientů v bance a strategie hodnocení plánu. Cílem je soustředit zdroje proaktivního zadržení klientů a obsloužit s prioritou tu skupinu klientů, která je charakterizována vyšší tendencí odchodu z banky s tendencí růstu a zároveň vysokými hodnotami aktuální a potenciální rentability. Následuje kvantitativní odhad výnosu projektu.

K hodnocení je vybrán shluk číslo čtyři, jelikož tento shluk obsahuje klienty s vysokou potenciální rentabilitou a vyšší rostoucí mírou odchodu; bude mít prioritní pozornost v plánu zadržení klientů v bance, kterou jsou kontakty klientů s nabídkami produktů z rodiny úvěrů; pro klienty s extra vysokou potenciální rentabilitou a vysokým indexem úniku existuje nabídka zrušení poplatků na rok 2006 a pro všechny klienty shluku čtyři je v plánu vyšetřit jejich dojem a spokojenost se službami celé korporace. Post segmentačními procesy je tak možné manipulovat tendenční chování klientů shluku (podpořit rentabilitu či/i zabránit odchodu) a tak konkurenčně využít příležitosti obchodu.

Je-li posouzen současný stav distribuce klientů ve shluku ve smylu kvality (průměrné a maximální vzdálenosti objektů ke středu shluku) a spolehlivě předpovězen budoucí pohyb, je možné aplikovat plán činnosti na zadržení klientů a počítat s jeho finančním přínosem, který je ohodnocen dle následující strategie:

1. Jsou uvažovány dva scénáře hodnocení:
    a. Pesimistický scénář, ve kterém klienti udržují svou aktuální rentabilitu (nezvyšuje se index cross-sellingu ani up-sellingu).
    b. Očekávaný scénář, ve kterém jsou zadržení klienti rentabilizováni na úroveň průměrné hodnoty klienta segmentu, k jejíž hodnotě směřují s rostoucí tendencí.
2. Cílem plánu je snížit index uzávěrky běžného účtu o 50%.
3. Předpoklady pro hodnocení:

a. Současná hodnota klienta, který setrvává v průměru 4 roky v bance, pro časový horizont 3 let bankovní sazbou 15% ročně je $1.067.740 ($40.465 měsíčně; pesimista) nebo $1.279.204 ($48.479 měsíčně; očekáváno).
b. Kontaktabilita klientů telemarketingem a jinými kanály remoto je 60% (proaktivní outbound).

8.2 Vyhodnocení kritérií srovnávání algoritmů

8.2.1 Vyhodnocení technického kritéria

Tabulka 8-1 představuje atributy v jejich původních jednotkách a také v hodnotách vystupňovaných do intervalu od -1 do 0.

Tab 8-1 Původní a vystupňované hodnoty atributů strategií

Atributy/Strategie	Hodnoty strategií			OM	Vystupňované hodnoty strategií			
	I.	III.	IV.	V.	I.	III.	IV.	V.
kvalita_průměrná vzdálenost	30,5	7,1	10,5	11,4	0	-1	- 0,85	- 0,82
kvalita_maximální vzdálenost	66,5	15,8	23,1	21,4	0	-1	- 0,86	- 0,89
čas	0	8	2	4	-1	0	- 0,50	- 0,25
dynamika	NE	NE	NE/ANO	ANO	0	0	- 0,25	-1

Poznámka: Strategie I. je k-means statická, strategie III. je k-means nově utvářená v každém požadovaném cyklu, strategie IV. je segmentace ve dvou fázích, strategie V. (OM) je obecná metodologie. Atribut dynamika je ohodnocen na bázi výsledků ze sekce 7.

Tabulka 8-2 předkládá výsledky srovnávání technického kritéria pro různé hodnoty vah atributů.

Tab 8-2 Výsledky srovnání strategií technickým kritériem pro rozdílné váhy

Atributy	Váhy_1	Výsledky vah_1			Váhy_2	Výsledky vah_2			Váhy_3	Výsledky vah_3		
		V. x I.	V. X III.	V. x VI.		V. x I.	V. X III.	V. x VI.		V. x I.	V. X III.	V. x VI.
j_1	0,15	-0,12	**0,03**	**0,01**	0,20	-0,16	**0,04**	**0,01**	0,30	-0,24	**0,06**	**0,01**
j_2	0,15	-0,13	**0,02**	-0,01	0,20	-0,18	**0,02**	-0,01	0,30	-0,27	**0,03**	-0,01
j_3	0,30	**0,23**	-0,08	**0,08**	0,20	**0,15**	-0,05	**0,05**	0,10	**0,08**	-0,03	**0,03**
j_4	0,40	-0,40	-0,40	-0,30	0,40	-0,40	-0,40	-0,30	0,30	-0,30	-0,30	-0,23
	v	**-0,43**	**-0,43**	**-0,22**	***v***	**-0,59**	**-0,39**	**-0,25**	***v***	**-0,74**	**-0,24**	**-0,20**

Atributy	Váhy_4	Výsledky vah_4			Váhy_5	Výsledky vah_5			Váhy_6	Výsledky vah_6		
		V. x I.	V. X III.	V. x VI.		V. x I.	V. X III.	V. x VI.		V. x I.	V. X III.	V. x VI.
j_1	0,00	0,00	0,00	0,00	0,25	-0,20	**0,05**	**0,01**	0,33	-0,27	**0,06**	**0,01**
j_2	0,00	0,00	0,00	0,00	0,25	-0,22	**0,03**	-0,01	0,33	-0,30	**0,04**	-0,01
j_3	1,00	**0,75**	-0,25	**0,25**	0,25	**0,19**	-0,06	**0,06**	0,33	**0,25**	-0,08	**0,08**
j_4	0,00	0,00	0,00	0,00	0,25	-0,25	-0,25	-0,19	0,00	0,00	0,00	0,00
	v	**0,75**	**-0,25**	**0,25**	***v***	**-0,49**	**-0,24**	**-0,12**	***v***	**-0,32**	**0,01**	**0,08**

Na základě výsledků *v* z tabulky 8-2 je možné shrnout, že pro většinu kombinací strategií a vah atributů[(*)] je hodnota formule záporná, což znamená vítězství OM nad ostatními strategiemi. Nejoptimálnější výsledky podává OM díky zahrnutí a vyhodnocení ukazatele dynamiky. V případech nízké hodnoty váhy atributu dynamiky je výsledek OM velmi podobný výsledku ostatních strategií, především strategii segmentace ve dvou fázích (kladné hodnoty strategií v dolní části tabulky 8-2). Vysoké kladné hodnoty pro některé strategie podává kritérium v případě vah_4, kde je uvažován pouze atribut čas. Je-li dána všem atributům stejná síla (váhy_5), je možné konstatovat, že OM je v souhrnu působení čtyř determinovaných atributů o 12% nad S IV., o 24% nad S III. a o 49% nad S I.

Technické kritérium je indikátorem kvality technických parametrů obecné metodologie. Důraz je však v praxi přikládán především ekonomickému kritériu, které uvádí obecnou metodologii do praxe a hodnotí konkrétní finanční přínos.

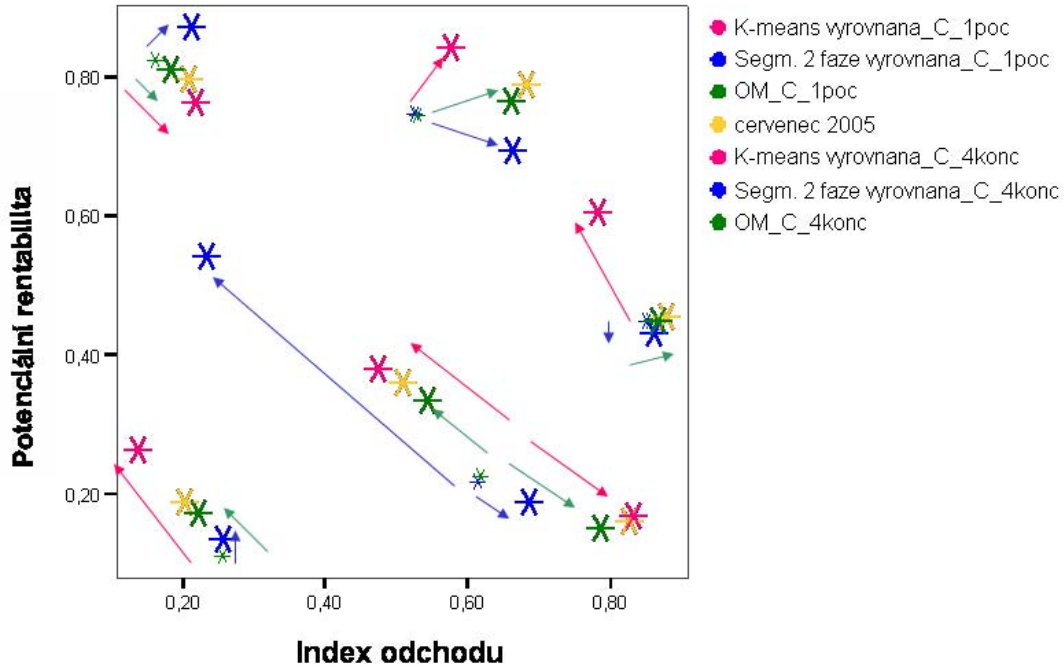
8.2.2 Vyhodnocení ekonomického kritéria

Aby komerční aktivity byly efektivní, je třeba aplikovat v *pravém momentu* tu *správnou kampaň* pro *vybranou skupinu* klientů. Při práci s daty historickými je aplikací obecné metodologie identifikována evidentní růstová tendence obou proměnných shluku 4 a tato je ověřena opakovanou kontrolou ve dvou následných obdobích: střed shluku se nacházel v předpovězeném směru tendenčního pohybu. Při sledování charakteristik obecné metodologie, dynamiky a schopnosti předpovědět budoucí vývoj na dalších dvou analyzovaných technikách segmentace, z grafů 7-21 a 8-1, je zřejmý chaotický pohyb středů každého ze shluků v průběhu cyklů; navíc není možné přesně zobrazit reálné trajektorie pohybu – pouze trajektorie identifikované. Shluky jsou tak pouhou fotografií hodnot proměnných objektů v jednom momentu, což nevyjadřuje typické chování objektů; není vyjádřena tendence v pohybu a tudíž není možné předpovědět chování objektů (shluků) do budoucnosti. Proto navazující post segmentační aktivity uplatňované na shluky aktualizované metodami k-means či segmentace ve dvou fázích budou často kontraproduktivní a tak neefektivní.

---

* V dataminingové terminologii nazýváno *Trade Off*.

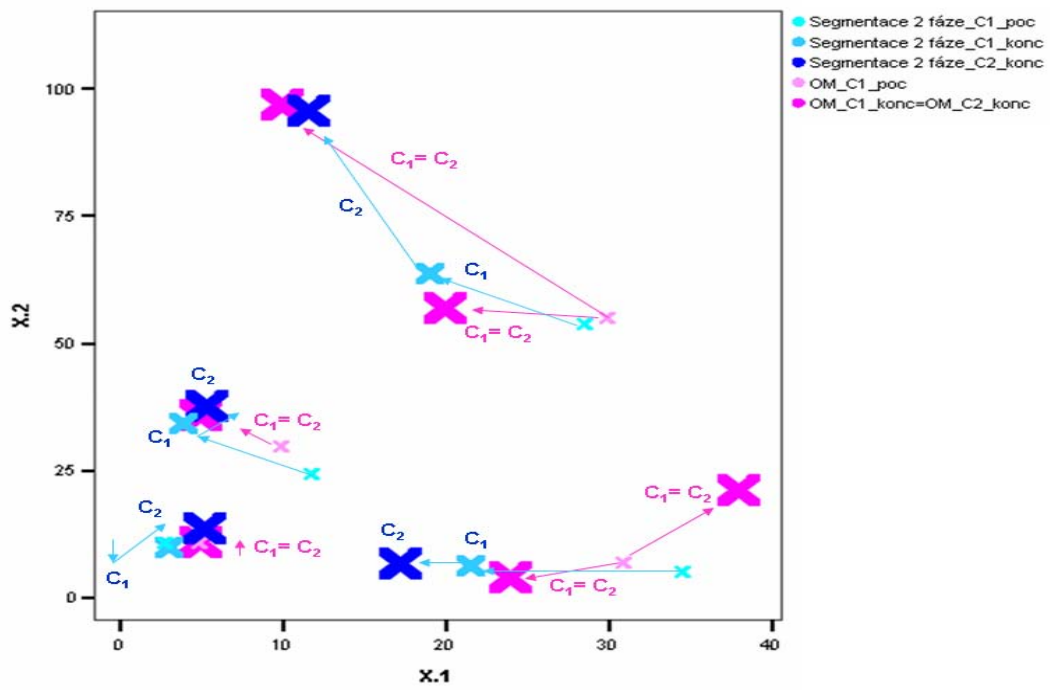
Graf 8-1 ***Tendence chování shluků vyjádřené obecnou metodologií, k-means vyrovnanou a segmentace ve dvou fázích vyrovnanou a skutečné středy shluků v budoucích obdobích***



Poznámka: zobrazené cykly jsou redukovány na t^1_{poc}, t^4_{konc} *a červenec 2005 (z důvodu přehlednosti). Zde například: C_1poc vyjadřuje* t^1_{poc}. *Všechny uvedené techniky mají počátek téměř ve stejném bodě, proto se barvy malých křížků, vyjadřujících první cyklus, překrývají. Velké křížky vyjadřují čtvrtý cyklus. Budoucím obdobím, kam by měly tendenčně směřovat strategie, je čevenec 2005 - zobrazeno žlutým křížkem.*

Na podpoření závěrů analýz je přiložen graf 8-2 vývoje středů shluků vyjádřených obecnou metodologií a segmentací ve dvou fázích aplikovaných na simulovaná data. Z grafu 8-2 je možné shrnout, že obecná metodologie vytváří v C_1 shluky přibližně na stejném místě, kde je vytváří segmentace ve dvou fázích, ovšem o jeden cyklus déle, tedy v C_2. Důvodem je etropický pohyb shluků vytvářených segmentací ve dvou fázích, který nesleduje dynamiku chování objektů. Proto v některých případech dochází k protisměrným pohybům shluků mezi cykly, v jiných případech se shluk realizuje se zpožděním za shluky vytvořenými obecnou metodologií.

Graf 8-2 Tendence chování vyjádřené obecnou metodologií a segmentací ve dvou fázích na simulovaných datech



Poznámka: velikost hvězdiček se zvyšuje se zvyšujícím se počtem cyklů. Zde například: C_1_poc vyjadřuje t^1_{poc} *. OM vyjadřuje obecnou metodologii.*

Pro kontrolu kvality distribuce klientů ve shluku je využit jeden z atributů technického kritéria a to průměrná a maximální vzdálenost objektů ke středům shluků. Hodnoty sledované v tabulce 8-3 potvrzují kvalitu sloučení.

Tab 8-3 Kvalita shlukování v C_4 ***vyjádřená obecnou metodologií, k-means vyrovnanou a segmentací ve dvou fázích vyrovnanou***

Shluky	Segmentace ve dvou fázích vyrovnaná			K-means vyrovnaná			Obecná metodologie		
	Vzdálenost		N	Vzdálenost		N	Vzdálenost		N
	Průměr	Maximum		Průměr	Maximum		Průměr	Maximum	
1	0,164	0,322	196	0,165	0,344	174	0,165	0,304	253
2	0,168	0,293	137	0,158	0,424	168	0,131	0,282	148
3	0,207	0,359	137	0,160	0,317	202	0,178	0,330	152
4	0,151	0,367	186	0,152	0,372	122	0,178	0,396	174
5	0,160	0,423	209	0,157	0,327	166	0,138	0,314	125
6	0,170	0,282	158	0,143	0,322	195	0,150	0,306	156
Průměr	**0,170**	**0,341**		**0,156**	**0,351**		**0,157**	**0,322**	

Je tak možné ocenit, že obecná metodologie je schopná předpovědět tendenci chování shluků do budoucnosti a produkovat kvalitní výsledné aktualizace, na kterých je možné

postavit stabilní, logicky navazující komerční kampaně (v tomto případě se záměrem zabránit odchodu z banky potenciálně rentabilním klientům).

V tabulce 8-4 je možné ocenit, že zadržení plánovaného počtu klientů přinese dle očekávaného scénáře roční užitek 30 miliónů chilských pesos[(*)]. Jelikož shluk čtyři by měl ve skutečnosti při hodnocení celé báze 4.524 mikropodnikatelů, majících běžný účet v Bci, druhý sloupeček v tabulce 8-4 zpřehledňuje výnos zadržení plánovaného počtu klientů, který je v případě očekávaného scénáře $875 miliónů. Při neexistenci plánu zadržení klientů by minimálně tato částka byla odečtena ze zisku korporace.

Tab 8-4 Zhodnocení plánu post segmentačních aktivit obecné metodologie jako ekonomické kritérium

		Shluk 4	Rozšířený shluk 4	Namátkou	
Kontaktovaní klienti		174	4 524	174	4 524
Index kontaktability		60%	60%	60%	60%
Klienti skutečně kontaktovaní		104	2 714	104	2 714
Index úniku v příštím měsíci		4	114	0,42	11
Index zábrany úniku klientů		50%	50%	50%	50%
Pesimistický scénář	Současná hodnota klienta zadrženého na tři roky	1 067 740	1 067 740	**Jeden jediný scénář**	
	Hodnota zadržených úniků v následujícím měsíci	2 135 480	60 861 183		
	Hodnota ročních zadržení ($)	25 625 761	730 334 193		
Očekávaný scénář	Současná hodnota klienta zadrženého na tři roky	1 279 204	1 279 204	633 282	633 282
	Hodnota zadržených úniků v následujícím měsíci	2 558 407	72 914 600	132 989	6 966 102
	Hodnota ročních zadržení ($)	30 700 884	874 975 196	1 595 871	83 593 224

Poznámka: Rozšířený shluk čtyři obsahuje počet všech klientů z Banky Mikropodnikatelů, ne jen vybranou skupinu klientů zemědělského sektoru.

Ve shluku 4 je index úniku 4,2% (dle dat historických). Ve shluku vybraného namátkou z celé populace klientů o stejném počtu klientů jako rozšířený shluk 4, je index odchodu klientů 0,4%. Toto ocenění vyjadřuje přínos, který je docílen vytvořením a užíváním prediktivního modelu. Porovnává efektivitu aktivit pro případ práce s prediktivní obecnou metodologií a práce bez modelace případu dataminingovou technikou.

V případě neexistence techniky, modelující tendence chování skupin klientů dle indexu odchodu a potenciální rentability, efektivita plánu zadržení klientů (se stejnými podmínkami jako bylo popsáno) by byla pro počet 4.524 klientů $83 miliónů (oproti $875 miliónům získaným technikou obecné metodologie).

* 100 chilských pesos je přibližně 4,2 Kč; jsou označovány $.

9 ZÁVĚR

Na závěr práce lze konstatovat, že bylo dosaženo zkonstruování standardní, obecné metodologie pro analýzu shlukování dovolující optimálním způsobem, založeným na **dynamickém chování objektů**, aktualizovat objekty spolu se vzory chování shluků. Každý následný časový cyklus navazuje na **aktualizaci provedenou v předešlém cyklu. Obecná metodologie vytváří, flexibilizuje a automatizuje procesy dataminingového modelování maximalizujíc příležitosti obchodů, které se ztrácely jednak v důsledku neexistence dynamické techniky shlukování, schopné vyjádřit reálné procesy na trhu a také zpožďováním rozhodování.** Vytvořená obecná metodologie je přínosem především pro odvětví dynamického a prediktivního **Data Miningu a Business Intelligence.**

9.1 Pohled na obecné charakteristiky obecné metodologie

Zlepšení a inovace, kterých obecná metodologie v odvětví segmentace docílila, jsou znatelné při její aplikaci a využity v praxi. Jsou jimi charakteristiky popsané v následujících odstavcích.

Jednou z výhod je **automatický sekvenční proces** průchodu etapami a výběru scénářů změny. Tato charakteristika dovoluje užívat obecnou metodologii a provádět dynamickou segmentaci i osobám, které nejsou experty v modelaci situací technikami Data Miningu. Tento fakt spolu s tím, že aktualizace **navazuje** v každém cyklu **na aktualizaci provedenou v předešlém cyklu, šetří** nejen **čas**, ale **i finance** instituci, která ji bude periodicky používat.

Obecná metodologie je vytvořena na základě obecných charakteristik. Je vhodná pro jakýkoliv případ shlukování, popisující reálné změny na trhu (kde existuje dynamika), jelikož je prakticky **šitá na míru,** tedy na každý jednotlivý případ. Příčina právě řečeného je následující: obecná metodologie **není omezena** aplikací, metodami, parametry ani kritérii rozhodování, jelikož tyto jsou snadno modifikovatelné. Je možné ji provádět jakoukoliv technikou segmentace za použití libovolného dataminingového software. Stejně tak není omezena počtem proměnných vložených do analýzy.

Je **nadstavbou** pro případy, kdy **objekty**, které přicházejí do báze dat, **nevlastní žádnou identifikaci**. Její využití je především pro objekty **identifikované,** jejichž chování

je možné sledovat v průběhu času; nedochází tak ke ztrátě informace. Navíc je možné zařazovat v průběhu analýzy i objekty **nové**, tak jako je možné další objekty v případě potřeby **eliminovat**.

Ne automaticky je informace o objektech, které jsou identifikovány jako objekty ve stavu změny, zařazena do formátů chování. Jelikož úsilím je udržet co možná nejstabilnější marketingovou kampaň (jako jakýkoliv proces, ke kterému dochází v průběhu života shluku), účelem je provádět pouze nevyhnutelné změny ve formátech chování; zároveň je však nutné respektovat reálné a podstatné změny v okolí a přenášet je do charakteristik shluku[(*)]. Je nezbytné v každé rozdílné situaci být flexibilní a schopný rychle a přesně rozhodovat ve smyslu jisté nabídky určité skupině klientů. Proto výhoda, uvedená jako následující, by měla být považována za jednu z nejpodstatnějších. Je jí schopnost obecné metodologie **předpovídat chování objektů do budoucnosti**. Existují statistické a dataminingové techniky a metody, pomocí kterých je možné předvídat chování objektů do budoucnosti, což je také jejich exkluzívním účelem. Segmentace je řazena mezi analýzy deskriptivní, což znamená, že nevlastní charakteristiku předpovědi do budoucnosti. Přesto však obecná metodologie dokáže využít techniky segmentace k právě popsanému účelu, což je pro všechny instituce ve stále se zostřujícím konkurenčním prostředí životně důležité. Obecná metodologie zprostředkuje detailní poznání chování objektů a identifikuje evidentní změny v systému pro výběr nejvhodnější formy aktualizace. Tím je schopná předpovědět tendenci chování shluků do budoucnosti - produkuje kvalitní výsledné aktualizace, na kterých je možné postavit **stabilní, logicky navazující komerční kampaně**. Je založena na segmentaci z předchozího cyklu, proto nedochází ke ztrátě informace; je možné sledovat kontinuální vývoj procesu post segmentace. V aktualizaci vzorů chování shluků je postupováno tendenčně, bez entropie, což **snižuje počet** nejrůznějších nadbytečných **post segmentačních procesů, které bývají chaotické, drahé a v průběhu času nemají reálný efekt.**

Touto výhodou předvídání budoucího stavu objektů a konstrukcí nejvhodnější aktualizace vzorů chování je v časové posloupnosti vyjadřování stavu a pohybu shluků

---

* Marketingové kampaně jsou vázány na mnoho faktorů, kterými je období nabídky, prostředky a mediální kanály kontaktu klientů, nabídka dalších produktů, konkurenční prostředí a mnoho dalších.

obecná metodologie o cyklus nebo několik cyklů napřed oproti až doposud nejsofistikovanějšímu algoritmu segmentace ve dvou fázích. Nezanedbatelnou výhodou obecné metodologie je také její **kvalita (přesnost)** z pohledu povšechné vzdálenosti objektů ke středu svého shluku a proto post shlukovací procesy budou vhodně reprezentovat všechny objekty shluku, pro které jsou vytvořeny.

9.2 Technický a ekonomický pohled na obecnou metodologii

Z výsledků realizace aplikace algoritmů na simulovaných a skutečných datech je možné shrnout **technické a ekonomické výhody** obecné metodologie srovnávané se statickými algoritmy segmentace.

Technické kritérium je indikátorem kvality technických parametrů obecné metodologie. Analýza *Trade Off* poskytuje kvalitní výsledky ve smyslu kvality shlukování a dynamiky. Důraz je však přikládán ekonomickému kritériu, které uvádí obecnou metodologii do praxe a hodnotí konkrétní finanční přínos.

Při práci s daty z minulosti byly aplikací obecné metodologie identifikovány a ověřeny evidentní tendence v chování objektů. Post segmentačními procesy je tak možné manipulovat předpokládané tendenční chování (podpořit rentabilitu či/i zabránit odchodu) a tak rychle a konkurenčně využít příležitosti obchodu. Při sledování charakteristik obecné metodologie - dynamiky a schopnosti předpovědět budoucí vývoj - na dalších dvou analyzovaných metodách statické segmentace, je zřejmý chaotický pohyb středů každého ze shluků v průběhu cyklů; shluky nevyjadřují tendenci v pohybu a tudíž není možné předpovědět chování objektů (shluků) do budoucnosti. Proto navazující post segmentační aktivity, uplatňované na shluky aktualizované technikou k-means či segmentace ve dvou fázích, budou často kontraproduktivní a tak neefektivní.

Jelikož byla ověřena kvalita stavu distribuce objektů ve shluku (průměrné a maximální vzdálenosti objektů ke středu shluku) a spolehlivě předpovězen budoucí pohyb, je možné aplikovat plán činnosti na zadržení klientů a počítat s vyčísleným finančním přínosem, který je $875 miliónů ročně (pro 4.524 mikropodnikatelů zemědělského sektoru Banka Mikropodnikatelé, což je shluk s vysokou potenciální rentabilitou a vyšším indexem odchodu).

9.3 Marketingový pohled na obecnou metodologii

Jelikož vytváření obecné metodologie dynamické segmentace kladlo důraz na její praktický význam a užití, z **marketingového hlediska** je významný poznatek o tvorbě a počtu nových tříd. Bylo předesláno, že stanovení počtu shluků je v oblasti shlukové analýzy až dodnes významným tématem výzkumu. V této práci byl dán v podstatě volný průběh vytváření nových tříd – bylo řečeno, že v případě volby scénáře tvorby nových tříd bude jedna třída rozdělena mezi dvě nové a maximální počet tříd bude 10. Navíc, nový shluk byl vytvořen na základě *náročných* kritérií (koeficient hranice stability outlierů γ a koeficient hranice tvorby nových shluků δ). Po zhodnocení názorů produkt a segment manažerů na toto téma je možné konstatovat, že **uživatelé nejsou ochotni ztratit dynamicky potenciálně se vytvářející třídy**, které mohou být zajímavé a které trh jako celek neregistruje; tím může vzniknout konkurenční výhoda. Je proto ustanoveno pravidlo stanovení počtu nových tříd na základě kritérií podporovaných obecnou metodologií a marketingem (ne statisticky).

Další komentář marketingu se týká shluků, které mají menší než minimální možný počet objektů (koeficient ω). Příčinou takového stavu může být například skutečnost, že **shlukům nebyla věnována dostatečná pozornost** a proto tyto případně významné shluky **zanikají.** Z toho vyplývá poznání, že je třeba stanovit rozdílnou hodnotu pro každý shluk a takzvaně *získat* významné segmenty, ať již jsou jimi nové shluky či shluky s kolísavým (nízkým) počtem objektů. V návaznosti na to je nutné periodicky sledovat reakci shluků na komerční nabídky a případně vytvořit podskupiny v samotných shlucích na oddělení různých forem efektů odpovědí na nabídku.

Je třeba podotknout, že je výhodnější provádět post segmentační aktivity pro **segmenty**, které jsou sledovány v průběhu času, než pro **osamocené objekty** a to z následujících důvodů: je monitorováno chování skupin, ne osamocených objektů a je tak možné ujednotit akce procesu post segmentace a tak snížit náklady; vytvořením skupin jsou detektovány potenciálně zajímavé nově se vytvářející shluky a také shluky zanikající. Detekcí tendenčního chování skupin objektů je možné provádět aktivity, které zabrání nepožadovanému vývoji chování shluků. Vytváření a sledování dynamických shluků má mnohonásobné užití. Navíc, má stejnou efektivitu jako stávající prediktivní model Bci

uzavření běžného účtu pro jednotlivé klienty (efektivita, sledovaná z pohledu tendence odchodu klientů a poté jejich rozdělení do skupin).

9.4 Hodnocení výkonnosti obecné metodologie jako metody dynamické segmentace Data Miningu

Při implementaci obecné metodologie do praxe se doporučuje sledovat její výkonnost. Kromě hodnocení technických kritérií je třeba mít na mysli, že efektivitu obecné metodologie není možné srovnávat s žádným historickým studiem, jelikož neexistuje. Segmenty nemají žádný koeficient na srovnávání vhodnosti modelu[(*)]. Kvalita shlukování se posuzuje dle praktického využití a funkčnosti modelu v celku následných procesů. Jedná se o prokázání, že objekty zařazené do shluků, aktualizovaných obecnou metodologií, reagují na nabídku (snižuje se index odchodu, zvyšuje se rentabilita); efektivitu výsledku reakcí na nabídku je možné srovnávat s efektivitou kampaní realizovaných v minulosti.

Při realizaci aplikace obecné metodologie jsou vyslovovány jisté hypotézy o budoucím chování shluků (na základě počtu objektů, vzdáleností objektů ke středům shluků a přemíťování objektů). Tyto hypotézy mohou být porovnávány (potvrzovány) s výsledkem řešení, získaného z aplikace kritérií a parametrů obecné metodologie, který je promítán v průběhu cyklů v tabulkách stavu rozhodování na konci každé etapy. Tento způsob srovnávání je možné považovat jednak za jeden způsob kontroly funkčnosti obecné metodologie (druhým způsobem jsou technická kritéria na srovnávání statického a dynamického algoritmu) a také za formu předvídání (prognóz) aktualizací formátů segmentů mezi obdobími, které jsou vyslovovány dále také na základě sledování reálných trajektorií pohybu shluků mezi cykly.

9.5 Budoucí výzkum, alternativní návrhy řešení

Obecná metodologie je úvodem do dynamické segmentace - má za úkol zkonstruovat scénáře a etapy průchodem těchto scénářů, nemá však za úkol vyčerpávajícím způsobem

---

[*] Srovnávání vhodnosti modelu neprobíhá jako u regrese nebo neuronové sítě (například RMS MAX Training y Test Errors) nebo u fuzzy C-means (RMS MAX Training y Test Errors, Partition coefficient).

poskytnout všechna možná řešení. Proto daný problém má dlouhou trajektorii budoucího výzkumu, která se týká například rozšíření řešení v jednotlivých etapách obecné metodologie identifikovaných v této práci či uplatnění rozdílných metod a kritérií v těchto etapách procesu; některé návrhy jsou popsány v následujících odstavcích.

Například v etapě I., Identifikace objektů představujících změnu, v kritériu definice 1 (*sekce 6.1.1*) je možné místo užívání hranice heterogenity, vyjádřené Eukleidovou vzdáleností, realizovat difusní celky (pomocí Fuzzy C-Means) a tak charakterizovat příslušnost objektů ke shlukům. Další forma, jak vyřešit tento problém, je zapracovat neuronální síť, která by klasifikovala do dvou stavů: *se změnou* a *beze změny*. Vzorem pro objekty *beze změny* by byly takové objekty, které by po jisté časové období nezměnily příslušnost a vzdálenost ke středu shluku a vzorem pro objekty *se změnou* by byly objekty vzdálené sloučeným celkům. Jednodušší alternativou na zlepšení by mohlo být užívání vzdálenosti Mahalanobis.

V etapě III., Rozhodnutí o možnostech aktualizace shluků, je možné pokračovat ve výzkumu například v případě, že by měl být překročen maximálně stanovený počet shluků. V tomto případě by došlo ke sloučení shluků. Pravidlem na sloučení by mohla být kombinace například následujících kritérií, jejichž analýza v práci nebyla prohloubena: dojde ke sloučení dvou shluků, popřípadě většího počtu segmentů, jejichž středy jsou si nejblíž; zároveň bude brán v úvahu směr pohybu a průměrná vzdálenost objektů ke středu svého shluku. Problému počtu shluků je možné předejít výběrem optimálního počtu shluků pomocí koeficientu BIC[*] a tímto způsobem vybrat ten počet tříd, který by měl nejvyšší hodnotu uvedeného indexu.

Existují i negativní názory na skutečnost, že obecnou metodologii je možné modifikovat na míru každému jednotlivému případu. To totiž vyžaduje úpravu parametrů kritérií etap procesu dle konkrétních podmínek. Tato charakteristika pak nemusí dovolit užívat obecnou metodologii a provádět dynamickou segmentaci osobám, které nejsou experty ve zmíněné modelaci nebo nemají možnost spolupracovat s oddělením marketingu. Marketingový analytik by se měl ovšem vždy zúčastnit definic a přípravy realizace obecné metodologie a tak rozhodovat například o tom, jak dlouhý cyklus je třeba vyvinout ke

[*] Bayesian Information Criterium.

sledování chování, aby bylo možné emitovat kvalitní a spolehlivé předpovědi. V případě transcendentní události v modelované skutečnosti je nutné vyvíjet a analyzovat segmenty po několik cyklů před vyslovením závěru o tendenci chování shluků[*].

9.6 Doporučení

V případě, že obecná metodologie nepodává požadované výsledky ve smyslu identifikace tendence pohybu nebo nesplňuje kritérium kvality shlukování ve smyslu přesnosti, alternativy zlepšení funkčnosti obecné metodologie pro praktického uživatele mohou být následující:

- Přehodnotit hodnoty používaných parametrů alfa až omega.
- Zvýšit počet sledovaných období na prodloužení cyklu.
- Přehodnotit užití navrhovaných proměnných.
- Zdokonalit akce post segmentace, které vedou objekty požadovaným směrem.
- Kombinovat obecnou metodologii či metody v jednotlivých etapách s dalšími inteligentními dataminingovými technikami - například těmi, které slouží vyhradně pro předpověď.

* Tak obecně operují prediktivní modely.

10 LITERATURA

[1] Antonsson E.K., Lee C.Y.: Dynamic Partition Clustering Using Evolution Strategy. Third Asia Pacific Conference on Simulated Evolution and Learning. Nagoya, Japan říjen 2000.

[2] Barni M., Cappellini V., Mecocci A.: Comments on a Possibilistic Approach to Clustering. IEEE Transactions on Fuzzy Systems, srpen 1996, svazek 4, číslo 3, str. 393–396.

[3] Bautist R., Sanclemente H. Harrison: Descubrimiento de Conocimiento en Bases de Datos Médicas. XXIV Conferencia Latinoamericana de Informática, Quito – Ecuador, 19. až 24. října 1998, svazek 2, str. 1037 - 1048.

[4] Black M.M., Hickey R.J.: Maintaining the Performance of a Learned Classifier under Concept Drift. Intelligent Data Analysis, prosinec 1999, svazek 3, číslo 6, str. 453-474.

[5] Crespo F.: Agrupamiento dinámico con lógica difusa. Memória para optar al Título de Ingeniero Civil Industrial. Universidad de Chile, Chile 2001.

[6] Dingsøyr T., Lidal M.E.: An Evaluation of Data Mining Methods and Tools. http://www.idi.ntnu.no/~dingsoyr/proyect/report.html.

[7] Fayyad U.M.: Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, Intelligent Systems & Their Applications, říjen 1996, svazek 11, číslo 5, str. 20 -25.

[8] Frédéric D., Le Barzic J.F.: Analyse des données évolutives, méthodes et applications. Edition Technip, Paříž 1996.

[9] Geman S., Geman D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of image. IEEE transactions on Pattern Analysis and Machine Intelligence, 1984, svazek 6, str. 721-741.

[10] Han J., Kamber M.: Data Mining: Concepts and Techniques. Morgan Kaufmann publishers, 2001.

[11] Jain A.K., Duin R., Jianchao Mao R.: Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, leden 2000, svazek 22, číslo 1, str. 4 – 37.

[12] Joentgen A., Mikenina L., Weber R., Zimmermann H.J.: Dynamic Data Analysis: Problem Description and Solution Approaches. W. Brauer (ed.), Fuzzy – Neuro Systems ’98 – Computational Intelligence. Infix, Sankt Augustin, str. 282 - 289.

[13] Joentgen A., Mikenina L., Weber R., Zimmermann H.J.: Dynamic Data Analysis: Similarity Between Trajectories, W. Brauer (ed.), Fuzzy – Neuro Systems ’98 – Computational Intelligence. Infix, Sankt Augustin, str. 98 - 105.

[14] Kandel A.: Fuzzy Techniques in Pattern Recognition. John Wiley & Sons, New York 1982.

[15] Karypis G., Han E.H., Kumar V.: Chamaleon: Hierarchical Clustering Using Dynamic Modeling. Computer, Computer Society IEEE, srpen 1999, svazek 32, číslo 8, str. 68 – 75.

[16] King Sun Fu: Digital Pattern Recognition, Second Corrected and Updated Edition. Communication and Cybernetics, svazek 10, Berlín 1980.

[17] Krishnapuram R., Keller J.M.: The Possibilistic C-Means Algorithm: Insights and Recommendations. IEEE Transactions on Fuzzy Systems, srpen 1996, svazek 4, číslo 3, str. 385 – 393.

[18] Krishnapuram R., Kim J.: A Note on the Gustafson-Kessel and Adaptive Fuzzy Clustering Algorithms. IEEE Transactions on Fuzzy Systems, srpen 1999, svazek 7, číslo 4, str. 453 – 461.

[19] Krzanowski W.J., Marriott F.H.C.: Multivariate Analysis. Part 2: Classification, covariance structures and repeated measurements. Londres: Arnold 1995.

[20] Lebart L., Morineau A., Piron M.: Statistique Exploratoire Multidimensionnelle. Paříž: Dunod 1995.

[21] Li R., Mukaidono M.: A Maximun-Entropy Approach to Fuzzy Clustering. International Join Conference of Second International Fuzzy Engineering Symposium and the Fourthe IEEE Intternational Conference Fuzzy Systtems and (FUZZ/IEEE-IFES), Japan, Yokohama březen 1995, str. 2227 – 2232.

[22] Mangin L., Varela Mallou J.P.: Análisis multivariable para las ciencias sociales., Madrid: Prentice Hall, España 2003, str. 417 – 449.

[23] Medel F.R.: Creación y prueba de una metodología para árboles de decisión en Data Mining dinámico. Tesis para optar al grado de magíster en gestión de operaciones. Universidad de Chile, Chile 2002.

[24] Pal K.R., Bezdek J.C.: On Cluster Validity for the Fuzzy C-Means Model, IEEE Transactions on Fuzzy Systems, srpen 1995, svazek 3, číslo 3, str. 370 – 379.

[25] Raghavan V., Hafez A.: Dynamic Data Mining. 13th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, USA 2000.

[26] Saporta G.: Probabilités, Analyse des Donées et Statistique. Éditions Technip, Paříž 1990.

[27] Tou J.T., González R.C.: Pattern Recognition Principles. Addison-Wesley Publishing Company, Massachusetts 1974.

[28] Utgoff P.E.: ID5: an incremental ID3. Proceedings of de Fifth Internacional Conference on Machine Learning, San Mateo, CA: Morgan Kaufmann 1988, str. 107-120.

[29] Utgoff P.E.: Incremental Induction of decision trees. Machine Learning, 4, 1989, str. 161- 186.

[30] Utgoff P.E., Berkman N. C., Clouse J.A.: Decision tree induction based on efficient tree restructuring. Machine Learning, 29, 1997, str. 5-44.

[31] Wang L.X.: A Course in Fuzzy Systems and Control. Prentice Hall, New Jersey 1997.

[32] Wasserman P.D.: Advanced Methods in Neural Computing. Van Nostrand Reinhold, New York 1993.

[33] Weber, R.: Data Mining en la Empresa y en las Finanzas Utilizando Tecnologías Inteligentes. Revista Ingeniería de Sistemas, Departamento de Ingeniería Industrial, Universidad de Chile, červen 2000, svazek XIV, číslo 1, str. 61 - 78.

[34] Weber R.: Fuzzy ID3: A class methods for automatic knowledge acquisition. Proceedings of the 2nd International Conference on Fuzzy Logic and Neural Networks, Iizuka, Japan 1992, str. 265-268.

11 PŘÍLOHY

11.1 METODY VÝPOČTU VZDÁLENOSTI (PODOBNOSTI) A PŘÍPADY JEJICH POUŽITÍ (str. 35)

typ proměnné	**binární**
metody na výpočet podobnosti	jednoduchá souvztažnost
	koeficient Sorenson
	koeficient Jaccard
	koeficient Sokal y Sneath
typ proměnné	**spojité**
metody na výpočet podobnosti	Eukleidova vzdálenost
typ proměnné	**nominální**
metody na výpočet podobnosti	transformovat na binární proměnné
	jednoduchá souvztažnost
typ proměnné	**ordinální**
metody na výpočet podobnosti	pracovat s nimi jako by byly spojité
typ proměnné	**směs**
metody na výpočet podobnosti	ignorovat různorodost
	realizovat různorodé analýzy
	transformovat kvantitativní proměnné na kvalitativní
	koeficient Gower

11.2 SOUHRN SYMBOLŮ, DEFINIC A PARAMETRŮ (str. 57)

Poznámka: symboly, definice a parametry jsou uvedeny dle jejich posloupnosti v textu.

N_1	počet známých objektů.
N_2	počet agregovaných objektů.
N	celkový počet objektů; $N = N_1 + N_2$.
M	počet proměnných.
$\mathbf{X}$	matice N_1 x M s hodnotami známých objektů, kde každý sloupec obsahuje hodnoty proměnných objektu $\mathbf{X}_{i\bullet}$.
$\mathbf{Y}$	matice N_2 x M s hodnotami agregovaných objektů, kde každý sloupec obsahuje hodnoty proměnných objektu $\mathbf{Y}_{i\bullet}$; (objekty $\mathbf{X}_{i\bullet}$ a $\mathbf{Y}_{i\bullet}$ jsou ve skutečnosti vektory objektu *i* přes všechny proměnné - zde pojmenovány pouze objekty).
S_h	shluk S_h, kde $h \in \{1,.....,S\}$.
s_h	střed shluku S_h.
s_{h^*}	střed konkrétního shluku S_{h^*}. Objekt $\mathbf{X}_{i'\bullet}$ neptatří do S_{h^*} : $\mathbf{X}_{i'\bullet} \notin S_{h^*}$.
$d(s_m, s_n)$	vzdálenost mezi středy shluků S_m y S_n; $\forall m,n \in \{1,...,S\}$, kde $m \neq n$.
$d(\mathbf{X}_{i\bullet}, s_h)$	vzdálenost mezi známým objektem $\mathbf{X}_{i\bullet}$ a středem shluku s_h.
$d_{\max_{s_h}^{t_{poc}^c}}$	hranice heterogenity; vzdálenost nejvzdálenějšího známého objektu $\mathbf{X}_{i\bullet}$ ke středu s_h svého příslušného shluku S_h v období před modifikací t_{poc}.
$\mathbf{X}_{i\bullet}^o$	outlier známého objektu $\mathbf{X}_{i\bullet}$.
$e_{S_h}^o$	počet outlierů shluku S_h.
C_c	cyklus C, kde $c \in \{1,...,C\}$.
t_l^c	období t_l cyklu c. c je složen z časových období: $C_c = \{t_k^c / t_1^c \leq k \leq t_{konc}^c\}$.
$\mathbf{X}_{i\bullet}^s$	objekt $\mathbf{X}_{i\bullet}$ stabilní, neetropický v pohybu.
$\mathbf{X}_{i\bullet}^{ns}$	objekt $\mathbf{X}_{i\bullet}$ nestabilní, entropický (chaotický) v pohybu.
$\left(\frac{t_{konc}^c}{2}\right)^*$	počet uvažovaných období k definici neentropického chování.

Poznámka: $\left(\frac{t^c_{konc}}{2}\right)^*$ *je definováno následovně:*

- $\left\lceil \frac{t^c_{konc}}{2} \right\rceil$, *jestliže* t^c_{konc} *je liché číslo,*
- $\frac{t^c_{konc}}{2}+1$, *jestliže* t^c_{konc} *je sudé číslo .*

$\mathbf{U}^c$ matice *N x M* trajektorií objektu $\mathbf{X}^s_{i\bullet}$ v cyklu c .

$\mathbf{U}^c_{ij}$ prvek matice $\mathbf{U}^c$.

$\mathbf{U}^c_{k\bullet}$ trajektorie typu *k* objektu $\mathbf{X}^s_{i\bullet}$ v cyklu c .

$e^{s,k}_{S_h}$ počet $\mathbf{X}^s_{i\bullet}$ shluku S_h , které mají trajektorii typu *k.*

$E^s_{S_h}$ největší počet $\mathbf{X}^s_{i\bullet}$ shluku S_h se stejnou trajektorií (stejným směrem); Matematicky, $E^s_{S_h} = \max_{k\in K}\left\{e^{s,k}_{S_h}\right\}$.

$\mathbf{U}^c_{S_h}$ trajektorie, po které se pohybuje $E^s_{S_h}$ shluku S_h a také trajektorie shluku S_h v cyklu c .

MNČ metoda nejmenších čtverců. Výpočítává souřadnice středů shuků. Matematicky v této práci MNČ ➔ $\min\left\{\sum_{i=1}^{N}\left(d(\mathbf{X}_{i\bullet}, s_{h^*})\right)\right\}$.

$E^{o,s}_{S_h}$ největší skupina $\mathbf{X}^{o,s}_{i\bullet}$ shluku S_h s trajektorií $\mathbf{U}^o_{S_h}$.

$\mathbf{U}^o_{S_h}$ trajektorie outlierů shluku S_h , kterou se pohybuje $E^{o,s}_{S_h}$ shluku S_h .

$e^s_{S_h}$ počet $\mathbf{X}^s_{i\bullet}$ shluku S_h se stejnou trajektorií jako je trajektorie $E^{o,s}_{S_h}$.

$\mathbf{U}^{e^s_{S_h}}_{S_h}$ trajektorie $e^s_{S_h}$; $\mathbf{U}^{e^s_{S_h}}_{S_h} = \mathbf{U}^o_{S_h}$.

$\mathbf{P}^t$ paměť registru; matice *S x t* s objekty $\mathbf{P}^t_{S_h\bullet}$, která registruje existenci objektů $\{X_{i\bullet}\}_i$ třídy S_h v každém časovém období t_l .

$\mathbf{P}^t_{S_h\bullet}$ vektor shluku S_h v časových obdobích t_l (obsahuje hodnoty 0 a 1).

$\mathbf{T}^c$ trajektorie vzorů chování; matice *S x C*, jejíž objekty tvoří souřadnice středů shluků v časových cyklech *c*.

$\mathbf{T}^c_{S_h\bullet}$ vektor trajektorií; obsahuje souřadnice středu S_h v časových cyklech.

Celky

$CS = \{S_h / h \in \{1,...,S\}\}$ celek shluků.

$CT = \{t_l^c / l \in \{1,...,konc\}\}$ celek časových období.

$CC = \{C_c / c \in \{1,.....,C\}\}$ celek cyklů.

Poznámka: V prvním cyklu je $t_{poc}^1 = t_1^1$. *V dalších cyklech je* $t_{poc}^{c+1} = t_{konc}^c$. t_{konc}^c *je období* t_3^c *po aktualizaci shluků.*

$K = \{k / k \ \ je\ trajektorie\}$ celek hodnot $\mathbf{U}_{ij}^c$ v jednotlivých časových odobích.

Parametry

α	hranice změny.
β	hranice stability objektů.
γ	hranice stability outlierů.
σ	hranice tvorby nových shluků.
ω	hranice minimálního možného počtu objektů.
C_{max}	maximální počet cyklů (období), po které třída s nedostatečným počtem objektů udržuje paměť, zatímco jí není přiřazen objekt.

Poznámka: Symboly,definice a parametry se vztahují také na agregované objekty $\mathbf{Y}_{i\bullet}$.

11.3 ETAPY SE SVÝMI PODMÍNKAMI, KRITÉRII A PARAMETRY (str. 57)

Etapa I. Identifikace objektů, které představují změnu.

Podmínka 1: ***definice kritéria hranice heterogenity***

Je dán shluk S_h, cyklus C_c a období t_l. Je definováno $d_{\max_{s_h}^{t^c_{poc}}} = \max_i d\left\{\mathbf{X}_{i\bullet}^{t^c_{poc}}, s_h^{t^c_{poc}}\right\}$.

Potom $\forall S_h, \forall \mathbf{X}_{i\bullet} \in S_h$ jsou objekty představující změnu identifikovány podmínkou 1:

$$\text{Podmínka } 1: \; d(\mathbf{X}_{i\bullet}^{t^c_{konc}}, s_h^{t^c_{poc}}) > d_{\max_{s_h}^{t^c_{poc}}} .$$

Vzdálenost $d_{max_{s_h}^{t^c_{poc}}}$ reprezentuje hranici heterogenity.

Etapa II. Rozpoznání stavu změny.

Podmínka 2: ***definice kritéria hranice změny***

$\forall S_h, \forall \mathbf{X}^o_{i\bullet} \in S_h$ platí:

$$\text{Podmínka } 2: \; \frac{e^o_{S_h}}{N_{S_h}} > \alpha, \quad \text{kde} \quad 0 < \alpha < 1.$$

Koeficient α představuje hranici změny.

Definice neentropického pohybu

$\forall S_h, \forall \mathbf{X}_{i\bullet} \in S_h, \forall j$, kde $CT = \left\{t_l^c / l \in \{1,...,konc\}\right\}$ a $\forall t > \left(\frac{t^c_{konc}}{2}\right)^*$ je definován pohyb:

a) $\mathbf{X}_{ij}^{t^c_{l+1}} - \mathbf{X}_{ij}^{t^c_l} > 0$ jako *stabilní pozitivní,*

b) $\mathbf{X}_{ij}^{t^c_{l+1}} - \mathbf{X}_{ij}^{t^c_l} < 0$ jako *stabilní negativní,*

c) $\mathbf{X}_{ij}^{t^c_{l+1}} - \mathbf{X}_{ij}^{t^c_l} = 0$ jako *stabilní konstantní.*

Definice neentropického směru

$\forall S_h$, $\forall \mathbf{X}^s_{i\bullet} \in S_h$, $\forall j$, kde $CT = \left\{t_l^c / l \in \{1,...,konc\}\right\}$ a $\forall t > \left(\frac{t^c_{konc}}{2}\right)^*$ je definována matice trajektorií $\mathbf{U}^c$, jejíž elementy jsou vyjádřeny následovně:

$$\mathbf{U}_{ij}^c = \begin{cases} 1, & \textit{je-li pohyb stabilní pozitivní,} \\ 0, & \textit{je-li pohyb stabilní konstantní,} \\ -1, & \textit{je-li pohyb stabilni negativní.} \end{cases}$$

Pro $\forall \mathbf{X}_{i\bullet}^{s}$ je definován vektor $\mathbf{U}_{k\bullet}^{c}$ jako trajektorie typu *k*. Pod trajektorií se rozumí celek hodnot $\mathbf{U}_{ij}^{c}$ v jednotlivých obdobích (trajektorie představují řádky v matici $\mathbf{U}^{c}$). Dodatečně je definován celek trajektorií jako $K = \{k / k \ \ je\ trajektorie\}$. Počet $\mathbf{X}_{i\bullet}^{s}$ shluku S_h, které mají trajektorii typu *k*, je vyjádřen jako $e_{S_h}^{s,k}$; navíc $E_{S_h}^{s} = \underset{k \in K}{max}\{e_{S_h}^{s,k}\}$ reprezentuje největší počet $\mathbf{X}_{i\bullet}^{s}$ shluku S_h se stejnou trajektorií; jejich trajektorie (a také trajektorie shluku S_h) je označena $\mathbf{U}_{S_h}^{c}$. Skupinky objektů mající stabilní pohyb a pohybující se po stejné trajektorii mají neentropický směr.

Etapa III. Rozhodnutí o možnostech aktualizace shluků.

Aktualizace klasifikací do tříd.

Podmínka 3: ***definice kritéria hranice nestability objektů***

$\forall S_h, \forall \mathbf{X}_{i\bullet}^{s} \in S_h$ platí:

$$\text{Podmínka } 3: \ \frac{E_{S_h}^{s}}{N_{S_h}} \leq \beta, \quad \text{kde} \quad 0 < \beta < 1.$$

Koeficient β přísluší hranici stability objektů.

Aktualizace pohybem tříd.

Podminka 4: ***definice kritéria hranice stability objektů*** *(dodatek; opak k podmínce 3)*

$\forall S_h, \forall \mathbf{X}_{i\bullet}^{s} \in S_h$ platí:

$$\text{Podmínka } 4: \ \frac{E_{S_h}^{s}}{N_{S_h}} > \beta, \quad \text{kde} \quad 0 < \beta < 1.$$

Definice trajektorií outlierů a objektů (stabilních v pohybu) ve shlucích s rozpoznaným stavem změny:

Ať $\mathbf{U}_{S_h}^{o}$ je trajektorie $\mathbf{X}_{i\bullet}^{o,s}$ shluku S_h a $E_{S_h}^{s,o}$ je největší skupina $\mathbf{X}_{i\bullet}^{o,s}$ (z celkového počtu $e_{S_h}^{o}$ outlierů shluku S_h) s trajektorií $\mathbf{U}_{S_h}^{o}$. Je definováno, že $e_{S_h}^{s}$ je počet $\mathbf{X}_{i\bullet}^{s}$ shluku S_h se stejnou trajektorií jako je trajektorie $E_{S_h}^{o,s}$, to znamená $\mathbf{U}_{S_h}^{e_{S_h}^{s}} = \mathbf{U}_{S_h}^{o}$.

Aktualizace tvorbou nových tříd.

Podmínka 5a: ***definice kritéria hranice stability outlierů***

$\forall S_h, \forall \mathbf{X}_{i\bullet}^{o,s} \in S_h$ platí:

$$\text{Podmínka} \quad 5\text{a}: \quad \frac{E^{o,s}_{S_h}}{e^{o}_{S_h}} \leq \gamma, \quad \text{kde} \quad 0 < \gamma < 1.$$

Koeficient γ přísluší hranici stability outlierů.

nebo

Podmínka 5b: ***definice kritéria hranice tvorby nového shluku***

$\forall S_h, \forall \mathbf{X}^{s}_{i\bullet}, \forall \mathbf{X}^{o,s}_{i\bullet} \in S_h$ platí:

$$\text{Podmínka} \quad 5\text{b}: \quad \frac{E^{s,o}_{S_h} + e^{s}_{S_h}}{N_{S_h}} \leq \delta, \quad \text{kde} \quad 0 < \delta < 1.$$

Koeficient δ přísluší hranici tvorby nových shluků.

Poznámka: Eventuelně se mohou splnit zároveň obě podmínky 5a y 5b. Při splnění obou opačných podmínek je realizován pohyb tříd.

Etapa IV. Zánik shluků.

Podmínka 6a: ***definice hranice minimálního možného počtu objektů***

$\forall S_h, \forall \mathbf{X}_{i\bullet} \in S_h$ platí:

$$\text{Podmínka} \quad 6\text{a}: \quad \frac{e_{S_h}}{N} < \omega, \quad \text{kde} \quad 0 < \omega < 1.$$

Koeficient ω je hranice minimálního možného počtu objektů.

a

Podmínka 6b: ***definice hranice času pozorování***

$\mathbf{X}_{i\bullet} \in S_{h^*}$ jestliže je splněno: $d(\mathbf{X}_{i\bullet}, s_{h^*}) < d(\mathbf{X}_{i\bullet}, s_h)$, $\forall S_h$, $\forall \mathbf{X}_{i\bullet} \in S_h$.

Dodatečně se požaduje, že je-li $\forall \mathbf{X}_{i'\bullet} \notin S_{h^*}$ v t^{c}_{konc}, pak:

$$\text{Podmínka} \quad 6\text{b}: \quad \mathbf{X}_{i'\bullet} \notin S_{h^*} \text{ v } C_{c+1} \text{ až } C_{\max}.$$

Etapa V. Identifikace trajektorií.

$\mathbf{T}^{c}$ je matice *S x C*, kde každá řádka reprezentuje souřadnice středu každého shluku v čase, to znamená v posledním období každého cyklu. $\mathbf{T}^{c}_{S_h\bullet}$ přísluší vektoru, který obsahuje trajektorie shluku S_h v čase, tedy souřadnice svého středu v čase.

Poznámka: Všechny uvedené vztahy se se vztahují také na agregované objekty $\mathbf{Y}_{i\bullet}$. *Podmínky se aplikují v* t^{c}_{konc} *každého cyklu.*

11.4 DATA SIMULOVANÉHO PŘÍPADU (str. 59 a dále)

Poznámka: Existuje 32 sloupců a 106 objektů. Nejprve je uvedeno prvních 16 sloupců a 70 objektů. Na další straně pokračují ty samé sloupce se zbývajícími objekty. Následují sloupce 17 až 30 s rozdělením objektů na prních 30 a zbývajících 31 až 106 objektů.
Některé sloupce vlastní chybějící údaje – je možné je nahradit hodnotou 0.
Index u shluků vyjadřuje časové období a cyklus pro objekty.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Identifikátor	$(S_h)_{t^1_{proc}}$	$d(\mathbf{X}^{t^1_1}_{i\bullet}, s^{t^1_1}_h)$	$\mathbf{X}^{t^1_1}_{i1}$	$\mathbf{X}^{t^1_2}_{i1}$	$\mathbf{X}^{t^1_3}_{i1}$	$\mathbf{X}^{t^1_1}_{i2}$	$\mathbf{X}^{t^1_2}_{i2}$	$\mathbf{X}^{t^1_3}_{i2}$	$(S_h)_{t^1_3}$	$d(\mathbf{X}^{t^1_3}_{i\bullet}, s^{t^1_1}_h)$	$\mathbf{X}^{o(t^1_3)}_{i\bullet}$	$\mathbf{X}^{o,s(t^1_3)}_{i\bullet}$	$\mathbf{X}^{s(t^1_3)}_{i1}$	$\mathbf{X}^{s(t^1_3)}_{i2}$	$\mathbf{X}^{s(t^1_3)}_{i\bullet}$
1	1	13,29	0	0	0	39	31	26	1	10,90	0	0	0	-1	1
2	1	10,37	0	0	5	33	16	30	1	5,00	0	0	99	99	0
3	1	1,59	11	0	12	29	13	33	1	3,40	0	0	99	99	0
4	1	11,29	0	1	0	25	21	27	1	10,50	0	0	99	99	0
5	1	7,31	11	0	0	23	28	23	1	12,30	0	0	99	99	0
6	4	11,08	0	10	14	21	31	52	2	16,20	0	0	1	1	1
7	4	8,47	0	0	0	18	21	33	1	10,40	0	0	0	1	1
8	4	6,22	0	0	0	15	26	48	1	20,40	1	1	0	1	1
9	4	10,67	15	11	2	14	15	25	1	9,60	0	0	-1	1	1
10	4	5,63	0	2	3	14	23	38	1	10,50	0	0	1	1	1
11	3	12,70	21	9	6	14	18	21	1	10,10	0	0	-1	1	1
12	4	3,74	7	6	5	14	25	31	1	5,10	0	0	-1	1	1
13	4	5,17	0	0	39	13	11	39	2	18,30	0	0	99	99	0
14	4	12,28	17	10	5	12	26	39	1	10,10	0	0	-1	1	1
15	4	13,24	18	15	5	11	21	36	1	7,60	0	0	-1	1	1
16	4	5,17	0	5	9	9	36	81	2	33,40	1	1	1	1	1
17	3	3,56	35	19	4	7	23	29	1	6,10	0	0	-1	1	1
18	4	9,16	13	10	2	7	8	45	1	16,80	1	1	-1	1	1
19	4	6,91	0	1	0	6	13	35	1	11,10	0	0	99	99	0
20	4	6,05	4	35	45	5	21	28	3	25,20	1	1	1	1	1
21	3	8,05	39	20	15	4	5	114	2	60,90	1	1	-1	1	1
22	3	3,70	29	20	2	4	6	53	1	24,10	1	1	-1	1	1
23	4	8,18	9	5	0	4	5	49	1	21,30	1	1	-1	1	1
24	3	10,91	42	0	40	4	3	61	2	11,80	0	0	99	99	0
25	4	8,47	0	5	8	4	5	43	1	12,90	0	0	1	1	1
26	4	8,63	8	0	14	3	2	39	1	9,60	0	0	99	99	0
27	3	5,11	28	0	0	3	4	11	4	4,80	0	0	99	99	0
28	4	8,09	6	25	35	3	17	22	3	15,60	1	1	1	1	1
29	4	9,31	0	0	10	3	2	10	4	5,30	0	0	99	99	0
30	3	14,07	45	0	35	3	3	12	3	6,30	0	0	99	99	0
31	4	10,18	0	0	5	2	6	14	4	3,00	0	0	99	99	0
32	4	10,18	0	0	2	2	2	19	4	8,50	0	0	99	99	0
33	4	10,18	0	0	1	2	4	11	4	3,80	0	0	99	99	0
34	4	10,18	0	0	1	2	2	19	4	8,80	0	0	99	99	0
35	4	11,08	0	0	0	1	5	13	4	5,20	0	0	0	1	1
36	4	11,08	0	0	0	1	1	15	4	6,20	0	0	99	99	0
37	4	8,84	1	2	5	19	40	13	4	2,00	0	0	99	99	0
38	4	4,81	1	0	1	14	12	17	4	7,10	0	0	99	99	0
39	4	3,89	1	1	1	12	8	9	4	4,30	0	0	0	-1	1
40	4	5,49	1	0	0	7	0	10	4	4,90	0	0	99	99	0
41	4	9,41	2	1	1	20	20	16	4	6,30	0	0	99	99	0
42	4	2,76	2	4	3	11	10	7	4	4,40	0	0	99	99	0
43	4	2,94	2	4	0	10	12	19	4	9,30	0	0	99	99	0
44	4	6,61	2	3	4	5	16	3	4	8,00	0	0	99	99	0
45	1	10,81	3	4	4	22	20	16	4	5,10	0	0	99	99	0
46	4	9,17	3	3	0	20	19	14	4	5,60	0	0	99	99	0
47	4	6,25	3	1	5	17	17	14	4	3,00	0	0	99	99	0
48	4	6,25	3	2	0	17	5	61	2	30,50	1	0	99	99	0
49	4	3,48	3	4	0	14	10	12	4	4,90	0	0	99	99	0
50	4	1,76	3	0	4	11	5	13	4	2,10	0	0	99	99	0
51	4	2,03	3	6	0	10	25	15	4	6,20	0	0	99	99	0
52	4	7,04	4	2	0	18	10	13	4	5,20	0	0	99	99	0
53	4	1,26	4	5	0	10	13	13	4	5,20	0	0	99	99	0
54	1	14,64	5	8	0	44	92	12	4	4,90	0	0	99	99	0
55	1	6,56	5	8	0	26	15	65	2	31,60	1	0	99	99	0
56	4	2,01	5	3	0	13	23	39	1	13,30	0	0	-1	1	1
57	4	1,03	5	5	0	12	15	3	4	9,30	0	0	99	99	0
58	4	0,24	5	19	0	11	51	2	4	10,20	0	0	99	99	0
59	4	6,00	5	11	0	5	11	20	4	10,20	0	0	99	99	0
60	4	3,24	6	10	0	14	21	27	1	10,50	0	0	99	99	0
61	1	5,63	7	10	0	35	54	9	4	5,20	0	0	99	99	0
62	1	9,71	7	6	0	21	20	11	4	4,80	0	0	-1	-1	1
63	1	5,16	8	7	0	35	40	21	4	11,10	0	0	99	99	0
64	1	2,01	8	9	19	30	24	8	3	12,50	0	0	1	-1	1
65	1	3,00	8	7	26	28	32	65	2	10,80	0	0	99	99	0
66	4	5,15	8	14	12	15	23	8	4	7,80	0	0	99	99	0
67	4	3,80	8	17	25	13	11	9	3	6,80	0	0	1	-1	1
68	2	24,67	9	6	3	68	64	9	4	2,70	0	0	-1	-1	1
69	4	9,05	9	9	6	19	17	9	4	2,40	0	0	99	99	0
70	1	8,76	10	13	9	39	47	9	4	4,70	0	0	99	99	0

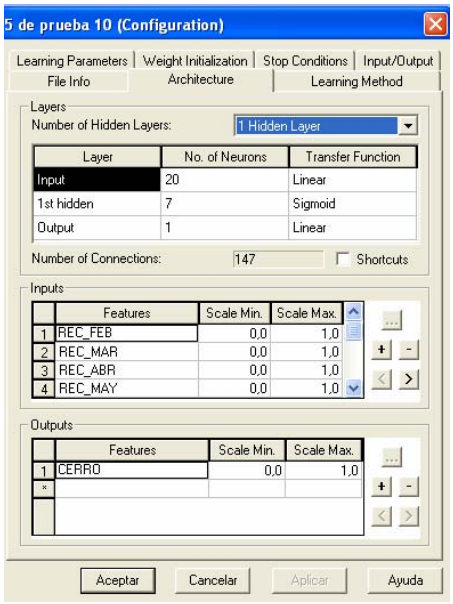
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Identifikátor	$(S_h)_{t^1_{poc}}$	$d(\mathbf{X}^{t^1_1}_{i\bullet}, s^{t^1_1}_h)$	$\mathbf{X}^{t^1_1}_{i1}$	$\mathbf{X}^{t^1_2}_{i1}$	$\mathbf{X}^{t^1_3}_{i1}$	$\mathbf{X}^{t^1_1}_{i2}$	$\mathbf{X}^{t^1_2}_{i2}$	$\mathbf{X}^{t^1_3}_{i2}$	$(S_h)_{t^1_3}$	$d(\mathbf{X}^{t^1_3}_{i\bullet}, s^{t^1_1}_h)$	$\mathbf{X}^{o(t^1_3)}_{i\bullet}$	$\mathbf{X}^{o,s(t^1_3)}_{i\bullet}$	$\mathbf{X}^{s(t^1_1)}_{i1}$	$\mathbf{X}^{s(t^1_3)}_{i2}$	$\mathbf{X}^{s(t^1_3)}_{i\bullet}$
71	4	9,56	10	14	9	19	24	9	4	4,70	0	0	99	99	0
72	4	5,33	10	6	4	12	17	7	4	4,10	0	0	99	99	0
73	4	5,33	10	10	9	12	28	7	4	5,80	0	0	99	99	0
74	1	8,82	11	25	26	39	72	7	3	5,40	0	0	99	99	0
75	4	10,14	11	12	6	19	26	4	4	7,10	0	0	99	99	0
76	4	6,32	11	4	14	12	5	6	4	10,50	0	0	99	99	0
77	1	5,61	12	11	10	25	17	5	4	8,00	0	0	-1	-1	1
78	1	6,55	12	8	6	24	20	6	4	5,20	0	0	-1	-1	1
79	4	10,07	12	11	9	18	16	16	4	6,60	0	0	99	99	0
80	4	7,31	12	20	14	10	14	4	4	11,60	0	0	99	99	0
81	2	16,93	13	16	12	54	33	5	4	9,40	0	0	99	99	0
82	1	3,25	13	11	6	29	35	36	1	7,00	0	0	-1	1	1
83	1	13,37	14	18	15	43	76	4	4	12,40	0	0	99	99	0
84	1	4,07	14	8	4	31	27	5	4	6,00	0	0	-1	-1	1
85	1	12,30	15	17	18	19	20	15	4	13,80	1	0	99	99	0
86	1	11,21	17	30	31	39	56	3	3	3,80	0	0	99	99	0
87	1	9,56	18	20	29	25	31	1	3	6,30	0	0	99	99	0
88	1	10,03	20	14	10	31	27	3	4	9,60	0	0	-1	-1	1
89	3	11,89	20	30	30	10	14	5	3	2,30	0	0	99	99	0
90	1	13,74	21	25	20	22	34	31	1	10,00	0	0	99	99	0
91	2	9,87	22	6	18	49	47	2	3	14,30	1	0	99	99	0
92	3	9,09	24	23	21	12	10	2	3	11,50	0	0	-1	-1	1
93	2	18,75	25	21	17	73	51	2	3	15,20	1	1	-1	-1	1
94	2	8,10	26	25	19	62	38	0	3	14,20	1	1	-1	-1	1
95	2	2,83	28	6	4	57	27	1	4	10,00	0	0	-1	-1	1
96	2	5,17	35	0	0	54	57	1	4	11,10	0	0	99	99	0
97	2	12,79	38	6	5	45	53	0	4	11,00	0	0	99	99	0
98	2	14,20	39	1	0	44	4	0	4	12,00	0	0	-1	-1	1
99	2	18,62	45	21	1	44	38	27	1	9,60	0	0	-1	-1	1
100	2	19,11	49	7	5	54	75	95	2	47,20	1	1	-1	1	1
101	.	.	.	.	18	.	.	41	1	13,40	0	0	99	99	0
102	.	.	.	.	4	.	.	49	1	19,70	1	0	99	99	0
103	.	.	.	.	9	.	.	28	1	2,50	0	0	99	99	0
104	.	.	.	.	4	.	.	46	1	16,90	1	0	99	99	0
105	.	.	.	.	6	.	.	38	1	8,70	0	0	99	99	0
106	.	.													

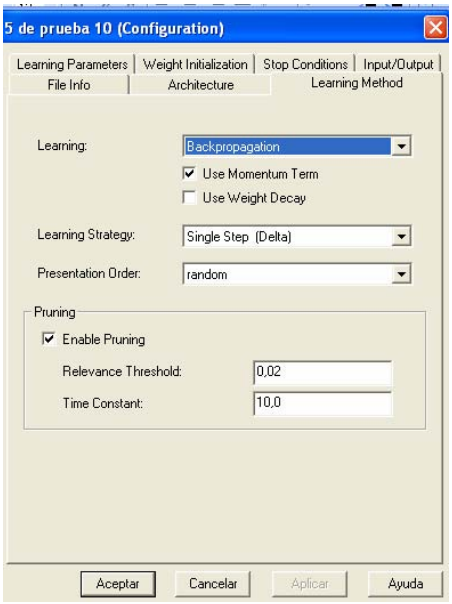
1	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Identifikátor	$(S_h)_{t^1_{konc}}$	$d(\mathbf{X}^{t^1_{konc}}_{i\bullet}, s^{t^1_3}_h)$	přesuny	$(S_h)_{t^1_{konc}}$	$d(\mathbf{X}^{t^1_{konc}}_{i\bullet}, s^{t^1_{konc}}_h)$	$\mathbf{X}^{t^2_1}_{i1}$	$\mathbf{X}^{t^2_2}_{i1}$	$\mathbf{X}^{t^2_3}_{i1}$	$\mathbf{X}^{t^2_1}_{i2}$	$\mathbf{X}^{t^2_2}_{i2}$	$\mathbf{X}^{t^2_3}_{i2}$	$(S_h)_{t^2_3}$	$d(\mathbf{X}^{t^2_3}_{i\bullet}, s^{t^2_1}_h)$	$\mathbf{X}^{o(t^2_3)}_{i\bullet}$
1	1	10,60	.	1	11,20	4	0	0	29	29	43	1	8,60	.
2	1	5,51	.	1	6,00	5	0	5	21	21	30	1	6,00	.
3	1	7,72	.	1	7,60	6	0	12	24	24	33	1	7,60	.
4	1	9,71	.	1	10,30	0	2	0	27	27	27	1	10,30	.
5	1	13,35	6	6	13,00	0	0	0	20	20	18	6	8,60	.
6	3	7,79	.	3	7,80	13	1	13	41	41	39	1	8,50	.
7	1	5,32	.	1	5,80	0	0	0	33	33	33	1	5,80	.
8	1	13,35	.	1	13,00	0	1	0	48	48	48	1	13,00	.
9	1	10,84	.	1	11,40	2	4	2	25	25	25	1	11,40	.
10	1	3,02	.	1	2,80	0	4	4	28	28	41	1	5,10	.
11	1	14,56	6	6	10,00	8	3	11	23	23	24	1	13,40	.
12	1	4,51	.	1	5,00	4	4	5	30	30	31	1	5,00	.
13	3	26,41	4	4	18,00	.			.	.	.			.
14	1	3,51	.	1	3,00	5	1	5	12	39	39	1	3,00	.
15	1	0,58	.	1	-	5	2	5	11	36	36	1	-	.
16	2	15,68	.	2	16,00	10	4	11	9	83	85	2	12,00	.
17	1	6,54	.	1	7,10	0	0	4	7	32	29	1	7,10	.
18	1	9,88	.	1	9,50	8	6	2	7	53	45	1	9,50	.
19	1	4,73	.	1	5,10	6	2	0	6	41	41	1	7,10	.
20	4	9,91	.	4	9,90	5			5	.	.			.
21	2	18,14	.	2	17,70	16	8	17	4	113	111	2	15,70	.
22	1	17,71	.	1	17,30	2	8	2	4	53	53	1	17,80	1
23	1	14,29	.	1	13,90	2	3	0	4	49	49	1	13,90	.
24	3	20,53	.	3	20,40	15	5	5	4	24	37	1	1,00	.
25	1	8,19	.	1	7,60	4	19	8	4	41	37	1	3,20	.
26	1	9,94	.	1	9,50	13	11	11	3	25	39	1	6,70	.
27	6	21,50	.	6	5,00	7	10	11	3	6	4	6	9,20	.
28	4	3,59	.	4	3,20	10	.	.	3	.	.	.	.	.
29	6	20,00	.	6	5,10	4	6	15	3	15	13	6	10,20	.
30	4	9,29	.	4	9,50	25	7	17	3	28	38	1	12,20	.

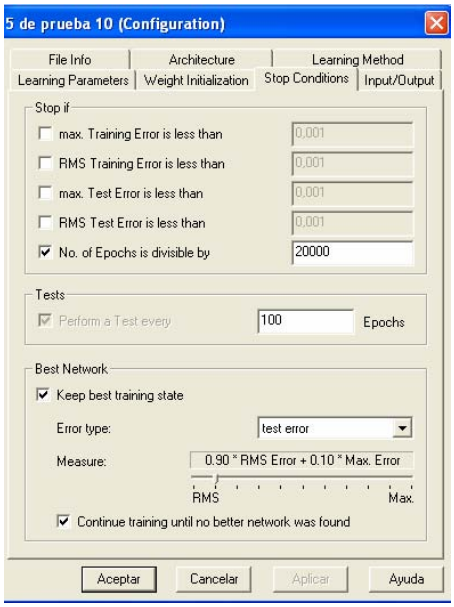
31	6	16,78	.	6	3,00	7	9	1	2	20	18	6	8,10	.
32	6	13,64	.	6	8,50	11	7	13	2	14	12	6	8,10	.
33	6	21,05	.	6	4,00	4	14	5	2	17	15	6	4,00	.
34	6	14,26	.	6	8,90	5	17	6	2	24	22	6	11,00	.
35	6	19,76	.	6	5,40	6	6	11	1	8	6	6	7,80	.
36	6	18,07	.	6	6,40	5	9	6	1	21	19	6	8,10	.
37	6	17,74	.	6	2,00	9	13	10	19	18	16	6	7,10	.
38	6	15,85	.	6	7,20	7	14	12	14	12	10	6	7,10	.
39	6	22,87	.	6	4,50	2	6	3	12	14	12	6	2,20	.
40	6	22,39	.	6	5,10	7	10	11	7	5	3	6	10,00	.
41	6	16,68	.	6	6,40	4	25	6	20	21	19	6	8,10	.
42	6	24,06	.	6	4,50	5	12	8	11	12	10	6	3,20	.
43	6	14,91	.	6	9,40	12	4	11	10	14	12	6	6,10	.
44	6	27,67	.	6	8,10	7	11	15	5	5	4	5	9,00	.
45	6	15,26	.	6	5,10	1	8	3	22	18	20	6	9,20	.
46	6	18,90	.	6	5,80	11	11	11	20	9	7	6	7,20	.
47	6	16,78	.	6	3,00	8	20	10	17	19	17	6	7,80	.
48	3	20,20	.	3	20,40	9	16	11	17	43	37	1	6,10	.
49	6	20,62	.	6	5,10	8	11	11	14	7	5	6	8,50	.
50	6	18,05	.	6	2,20	5	18	9	11	18	16	6	6,40	.
51	6	18,07	.	6	6,40	5	8	11	10	10	8	6	6,70	.
52	6	19,76	.	6	5,40	10	17	11	18	8	6	6	7,80	.
53	6	19,76	.	6	5,40	0	30	0	10	21	19	6	9,40	.
54	6	20,62	.	6	5,10	6	20	6	44	18	16	6	5,10	.
55	3	21,32	.	3	21,50	5	14	7	26	51	35	1	2,20	.
56	1	5,86	.	1	5,80	8	30	0	13	17	39	1	5,80	.
57	6	28,82	.	6	9,40	6	25	6	12	6	4	6	7,10	.
58	6	29,76	.	6	10,30	5	6	5	11	7	5	6	6,00	.
59	6	14,19	.	6	10,30	5	23	6	5	21	24	1	12,00	.
60	1	9,71	.	1	10,30	0	21	3	14	30	31	1	5,40	.
61	6	23,29	.	6	5,40	11	25	11	35	4	2	6	10,80	.
62	6	21,50	.	6	5,00	11	6	11	21	6	4	6	9,20	.
63	6	13,50	.	6	11,20	11	0	11	35	16	14	6	6,70	.
64	5	6,09	.	5	6,40	10	6	9	30	8	8	6	5,00	.
65	3	9,97	.	3	10,00	17	1	8	28	57	39	1	4,20	.
66	6	22,08	.	6	7,60	11	21	6	15	14	12	6	1,40	.
67	5	5,32	.	5	5,10	25	7	25	13	10	11	5	7,10	.
68	6	22,16	.	6	2,80	11	14	14	68	3	2	5	10,20	.
69	6	21,39	.	6	2,20	13	17	17	19	7	2	5	7,30	.
70	6	21,03	.	6	4,50	5	3	3	39	15	13	6	2,80	.
71	6	21,03	.	6	4,50	6	3	3	19	15	13	6	2,80	.
72	6	23,79	.	6	4,10	5	15	15	12	6	0	5	9,80	.
73	6	23,02	.	6	5,70	7	5	5	12	4	2	6	9,00	.
74	5	3,98	.	5	3,60	25	23	23	39	5	8	5	4,10	.
75	6	26,32	.	6	7,10	15	17	17	19	6	3	5	7,10	.
76	6	24,32	5	5	10,20	14	14	14	12	10	8	6	9,50	.
77	6	25,00	.	6	7,80	7	4	4	25	11	9	6	2,20	.
78	6	24,34	.	6	5,10	7	0	0	24	12	10	6	5,10	.
79	6	14,04	.	6	6,40	15	14	14	18	21	19	6	12,00	.
80	6	26,30	5	5	10,00	19	31	31	10	5	4	5	7,00	.
81	6	25,07	.	6	9,20	17	23	23	54	3	2	5	2,20	.
82	1	1,39	.	1	1,00	5	6	6	29	41	36	1	1,00	.
83	6	26,46	5	5	9,00	15	15	15	43	16	14	6	10,40	.
84	6	25,73	.	6	6,10	3	2	2	31	11	9	6	3,60	.
85	6	16,97	5	5	12,50	19	21	21	19	15	15	5	11,40	.
86	5	7,55	.	5	7,10	25	21	21	39	5	13	5	9,50	.
87	5	6,22	.	5	5,80	13	25	25	25	6	5	5	1,40	.
88	6	27,00	.	6	9,40	11	14	14	31	9	5	5	10,00	.
89	5	6,59	.	5	6,10	29	27	27	10	5	5	5	3,20	.
90	1	15,95	.	1	15,80	19	16	16	22	26	15	6	11,70	.
91	5	5,82	.	5	6,30	18	18	18	49	7	3	5	6,10	.
92	5	3,14	.	5	3,60	12	28	28	12	4	2	5	4,50	.
93	5	6,77	.	5	7,30	11	19	19	73	1	2	5	5,40	.
94	5	5,95	.	5	6,40	19	19	19	62	0	0	5	6,40	.
95	6	29,63	.	6	10,00	5	8	8	57	6	4	6	7,60	.
96	6	30,70	.	6	11,20	6	11	11	54	8	6	6	7,80	.
97	6	30,43	.	6	11,00	15	16	16	45	1	2	5	8,20	.
98	6	31,64	.	6	12,10	5	5	5	44	5	3	6	8,00	.
99	1	9,27	.	1	9,80	6	1	1	44	23	27	1	9,80	.
100	2	18,14	.	2	5,40	6	7	7	54	93	91	2	6,70	.
101	1	14,39	.	1	13,90	17	15	15	.	31	41	1	11,20	.
102	1	13,52	.	1	13,00	4	4	4	.	49	49	1	13,00	.
103	1	8,65	.	1	8,90	8	5	5	.	28	28	1	8,00	.
104	1	10,52	.	1	10,00	4	4	4	.	47	48	1	12,00	.
105	1	2,82	.	1	2,20	7	6	6	.	25	38	1	2,20	.
106						21	39	39	.	19	22	4	1,40	.

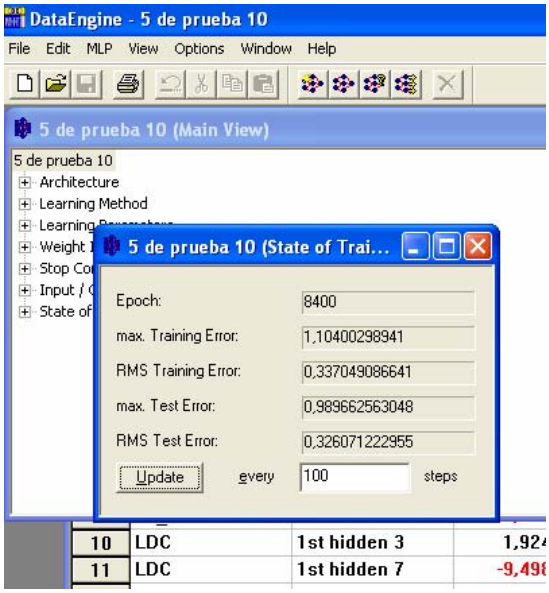
11.5 DEFINICE NEURONOVÉ SÍTĚ PRO PŘEDPOVĚĎ INDEXU ODCHODU (str.96)

Proměnné použité v neuronové síti
disponibilita úvěrového limitu
přesah úvěrového limitu
počet transakcí šeky měsíc_1
počet transakcí šeky měsíc_2
počet transakcí šeky měsíc_3
počet vkladů na běžný účet měsíc_1
počet vkladů na běžný účet měsíc_2
počet vkladů na běžný účet měsíc_3
počet automatických plateb z běžného účtu měsíc_1
počet automatických plateb z běžného účtu měsíc_2
počet automatických plateb z běžného účtu měsíc_3
reklamace měsíc_1
reklamace měsíc_2
reklamace měsíc_3
reklamace měsíc_4
reklamace měsíc_5
podíl salda měsíce 1 a měsíce 2
podíl salda měsíce 2 a měsíce 3
podíl salda měsíce 3 a měsíce 4
podíl salda měsíce 4 a měsíce 5









11.6 DATA REÁLNÉHO PŘÍPADU (str. 97)

Sledovaní klienti Bci, mikropodnikatelé zemědělského sektoru, se svým identifikátorem a rut, zařazením do původních shluků, Eukleidovou vzdáleností každého objektu k původnímu příslušnému shluku a s hodnotami původních proměnných po všechna sledovaná období čtyř cyklů jsou uvedeny na CD přiloženému k disertační práci (je umístěné na deskách na konci práce).

11.7 STŘEDY SHLUKŮ V C_1 AŽ C_4 VYTVOŘENÉ TŘEMI TECHNIKAMI SEGMENTACE (str. 100)

Shluky v	IO_K-means_ vyrovnaná	PO_K-means_ vyrovnaná	IO_K-means	PO_K-means	IO_S. 2 fáze_vyrovnaná	PO_S. 2 fáze_vyrovnaná	IO_S. 2 fáze	PO_S. 2 fáze	IO_OM	PO_OM
C^1_{poc}	0,530	0,750	0,530	0,750	0,260	0,110	0,260	0,110	0,530	0,750
	0,160	0,830	0,160	0,830	0,620	0,220	0,620	0,220	0,160	0,830
	0,860	0,450	0,860	0,450	0,850	0,450	0,850	0,450	0,860	0,450
	0,620	0,220	0,620	0,220	0,160	0,830	0,160	0,830	0,620	0,220
	0,260	0,110	0,260	0,110	0,530	0,750	0,530	0,750	0,260	0,110
C^1_{konc}	0,870	0,380	0,870	0,380	0,860	0,490	0,860	0,490	0,530	0,750
	0,200	0,800	0,200	0,800	0,250	0,170	0,250	0,170	0,260	0,110
	0,560	0,240	0,560	0,240	0,530	0,730	0,530	0,730	0,860	0,450
	0,220	0,150	0,220	0,150	0,160	0,810	0,160	0,810	0,160	0,830
	0,650	0,720	0,650	0,720	0,670	0,220	0,670	0,220	0,620	0,220
C^2_{konc}	0,860	0,440	0,800	0,440	0,520	0,790	0,820	0,470	0,870	0,450
	0,240	0,150	0,260	0,770	0,680	0,180	0,450	0,170	0,760	0,140
	0,370	0,530	0,380	0,190	0,250	0,130	0,290	0,760	0,540	0,280
	0,680	0,170			0,340	0,490			0,240	0,110
	0,670	0,770	0,260	0,790	0,850	0,490	0,300	0,170	0,570	0,750
	0,180	0,840	0,800	0,480	0,130	0,840	0,260	0,780	0,170	0,830
			0,380	0,190			0,760	0,430		
C^3_{konc}	0,130	0,870			0,890	0,290			0,540	0,340
	0,320	0,130	0,380	0,200	0,460	0,740	0,240	0,760	0,870	0,450
	0,770	0,260	0,800	0,480	0,140	0,790	0,460	0,170	0,170	0,810
	0,480	0,760	0,260	0,780	0,570	0,250	0,780	0,530	0,790	0,150
	0,270	0,460			0,800	0,660			0,230	0,180
	0,820	0,640			0,240	0,160			0,570	0,770
C^4_{konc}	0,470	0,380			0,670	0,690			0,790	0,150
	0,580	0,840			0,240	0,540			0,870	0,450
	0,140	0,260			0,210	0,870			0,190	0,810
	0,790	0,610			0,690	0,190			0,660	0,770
	0,220	0,760			0,260	0,130			0,230	0,170
	0,840	0,170			0,860	0,430			0,540	0,340

Poznámka: *IO vyjadřuje index odchodu, PO potenciální rentabilitu.*
S. 2 fáze = Segmentace ve dvou fázích
OM = obecná metodologie