УНИВЕРСИТЕТ "ПРОФ. Д-Р АСЕН ЗЛАТАРОВ"-БУРГАС,
БУГАС, БЪЛГАРИЯ

АВТОРЕФЕРАТ

ЗА ПРИДОБИВАНЕ НА ОБРАЗОВАТЕЛНА И НАУЧНА СТЕПЕН
"ДОКТОР"

НА **ВЕСЕЛИНА КУНЧЕВА БУРЕВА**

ТЕМА: **МОДЕЛИРАНЕ ПРОЦЕСА НА ИЗВЛИЧАНЕ НА ЗНАНИЯ ЧРЕЗ ОБОБЩЕНИ МРЕЖИ**

Област на висшето образование:	Технически науки
Професионално направление:	5.3. Комуникационна и компютърна техника

НАУЧНИ РЪКОВОДИТЕЛИ:

1. чл.кор. проф. дтн дмн Красимир Т. Атанасов
2. доц.д-р Евдокия Н. Сотирова

РЕЦЕНЗЕНТИ:

1. проф. дтн Людмил Даковски
2. проф. д-р Магдалина Тодорова

БУРГАС, 2014

Дисертационният труд е обсъден и допуснат до защита на разширено заседание на катедра “Компютърни системи и технологии”, проведено на 20.06.2014 г. в Университет “Проф. д-р Асен Златаров”-Бургас.

Дисертационният труд съдържа 116 страници, от които 49 фигури и 1 таблица. Използвани са 165 литературни източници. Резултатите са публикувани в 7 статии.

Защитата на дисертационния труд ще се състои на .....2014 г. от ..... ч. в зала ..... в Университет “Проф. д-р Асен Златаров”-Бургас на научно жури в състав:

1. доц. д-р Сотир Сотиров
2. чл.кор. проф. дтн дмн Красимир Атанасов
3. проф. дтн Людмил Даковски
4. проф. д-р Магдалина Тодорова
5. доц. д-р Олимпия Роева

Резервни членове: доц. д-р Любка Дуковска
доц. д-р Станислав Симеонов

Материалите по защитата са предоставени за заинтересуваните в кабинет 303, Органичен корпус.

Автор: Веселина Кунчева Бурева
Заглавие: Моделиране процеса на извличане на знания чрез Обобщени мрежи

Изказвам голямата си благодарност към ръководителите на дисертационния ми труд чл.кор. проф. дмн дтн Красимир Атанасов и доц. д-р Евдокия Сотирова за знанията, помощта, възможностите и контактите, които ми предоставиха.

Благодаря и на всички колеги от катедра „Компютърни системи и технологии" при Университет „Проф.д-р А.Златаров", Институт по биофизика и биомедицинско инженерство, секция Биоинформатика и математическо моделиране, БАН и
Институт по информационни и комуникационни технологии, секция Интелигентни системи, БАН
за подкрепата и съдействието.

Изследванията в дисертационния труд и публикациите в него са подпомогнати от проект по договор №ДФНИ-И-01/0006 на тема "Симулиране на поведението на горски и полски пожари".

Характеристика на дисертационния труд

Извличането на данни води своето начало от края на 1980 г. и началото 1990-те години и все още се определя и прецизира. Терминът извличане на знания от данни (*knowledge discovery in databases*) се приема за български превод на популярния термин *Data Mining,* въпреки че съществува определена разлика между тях. "*Knowledge discovery in databases*" е целият процес на избор на данни, предварителната им обработка, трансформиране на стойностити им, извличане на знания/съставяне на модел, оценяване и прилагене на знанията/модела. "*Data Mining"* е конкретната стъпка от процеса на извличане на знания, в която се откриват скрити зависимости. Извличането на знания от данни е автоматичен изследователски анализ, извличащ неочевидни, нови, полезни и разбираеми зависимости от огромни количества данни. Извличането на знания е ядрото на този процес, влючващ съответните алгоритми за изследователския анализ, конструиране на модел и извличане на предварително неизвестни зависимости. Използват се различни алгоритми в зависимост от целта на анализа. Както беше споменато, областта е в самото начало от своето развиетие и в настоящия момент тя се развива изключетелно динамично. Поради огромното разнообразие от подобластите, включени в нея, част от които са инструменти на изкуствения интелект, като машинно обучение, невронни мрежи, статистика, алгоритми, бази от данни, визуализция и други, тя достига изключетелно широк диапазон на приложение. Извличането на знания е възможно да бъде приложено при медицинска диагностика, за разкриване на измами, в биоинформатиката, търговията и на още много други места. Областта на приложение не се ограничава само до конкретна съхранена информация. Извличанието на знания е приложимо в Интернет пространството (*Web Mining* - проследяване на кликвания, създаване на профили на потребитили и др.) и при работа с текстови документи (*Text Mining, Information Retrieval*).

Настоящият дисертационен труд е посветен на изследването на техниките за извличане на знания и моделирането им чрез апарата на Обобщените мрежи. Понятието „обобщена мрежа“ възниква преди 30 години като математически обект и като средство за моделиране на реални процеси. Чрез моделиране на реално протичащите процеси по извличане на знания е възможно тяхното по-добро възприемане и

оптимизация. Направен е обзор на техниките за извличане на знания. Задълбочено са изследвани методите за извличане на асоциативни правила (*frequent pattern mining and association rule mining*), извличане на последователни зависимости (*sequence pattern mining*) и процеса на клъстеризация (*clustering*). Конструирани са 8 обобщеномрежови модела. Проучени са наличните софтуерни продукти и езици за обработка на данни с цел извличане на знания. Избрана е софтуерната среда *RapidMiner* във версията си под свободен лиценз и статистическия език *R* за извършване на реални тествания върху данни за откриване на закономерности между тях или за конструиране на модел.

Апробация на резултатите

Апробацията на резултатите е осъществена в рамките на представяния на доклади на няколко международни конференции, в статии в научни списания и тематични сборници. Изследванията се явяват резултати и по един научен проект, в чиито колектив авторът е член:

- Проект с Фонд „Научни изследвания“ №ДФНИ-И-01/0006 на тема “Симулиране на поведението на горски и полски пожари”.

Съдържание на дисертационния труд

Дисертационният труд е в обем от 116 страници и се състои от увод, три глави, заключение, декларация за оригиналност на резултатите, списък на публикациите по дисертационния труд, библиография. Дисертационният труд включва 49 фигури и 1 таблица, а библиографията към него – 165 заглавия.

Глава 1. Въведение в обобщените мрежи и извличане на знания от данни

В първа глава са дадени обзор на областта на извличане на знания от данни и основните дефиниции на обобщените мрежи.

1.1 Извличане на знания от данни

Тук са представени основни понятия и техники, свързани с извличането на знания от данни.

1.2. Обобщени мрежи

Тук са дадени основни дефиниции, които са необходими за изложението по-нататък и кратки исторически бележки за теорията на обобщените мрежи.

Цел и задачи на дисертационния труд

Цел на дисертационния труд е да се изследват различни процеси от теорията на извличането на знания от данни (*Data mining*) чрез моделирането им с помощта на обобщени мрежи и програмната им реализация.

За да се постигне тази цел са поставени следните **задачи**:

1. Да се анализират методите за извличане на закономерности от данни;
2. Да се анализират алгоритмите за откриване/извличане на асоциативни правила;
3. Разработване на обобщеномрежов модел на процеса на създаване на асоциативни правила чрез *Apriori* алгоритъм;
4. Разработване на обобщеномрежов модел на процеса на създаване на асоциативни правила чрез *Eclat* алгоритъм;
5. Разработване на обобщеномрежов модел на процеса на създаване на асоциативни правила чрез *FP-Growth* алгоритъм;
6. Разработване на йерархичен обобщеномрежов модел на процеса на създаване на асоциативни правила;

7. Разработване на обобщеномрежов модел на процеса на извличане на последователни зависимости чрез *GSP* алгоритъм;
8. Разработване на обобщеномрежов модел на конструиране на дърво на решението.
9. Разработване на йерархичен обобщеномрежов модел на отделните стъпки на процеса на клъстеризация и йерархичен подмодел на избора на клъстеризиращ метод.
10. Програмна реализация и тестване на основни алгоритми.

Глава 2. Приложения на обобщените мрежи в областта на извличане на знания от данни

2.1. Обобщеномрежови модели в областта на извличане на знания от данни чрез съставяне на асоциативни правила

В обобщеномрежовите модели, описани в тази глава основно място заемат техниките за извличане на асоциативни правила. Те откриват често повтарящи се закономерности, които могат да послужат за предсказване на бъдещи такива. В зависимост от типа на данните и вида на асоциативните правила могат да бъдат използвани различни алгоритми за откриване на асоциативни правила.

2.1.1. Обобщеномрежов модел за анализ на процеса на конструиране на дърво на решението

Обобщеномрежовият модел на процеса на съставяне на дърво на решението е съставен, използвайки алгоритъм "отгоре-надолу". Алгоритъмът за дървото на решението е широко използван в практиката за класифициране на целите и в някои приложения, за да подкрепи регресията. Класификационната техника работи с обучаващ алгоритъм, за да определи най-добрия модел.

Обобщеномрежовият модел, показан на фиг. 1, моделира процеса на съставяне на дърво на решението, използвайки алгоритъма на Hunt (*Hunt's algorithm*), който е основата на алгоритмите *ID*3, *C*4.5 и *CART*. В началото всички обучаващи записи са в корена. Всеки запис съдържа множество от атрибути и един от атрибутите е от класа.

Следващата стъпка е рекурсивно разделяне на записите чрез избиране на разделящ атрибут всеки път.

Обобщеномрежовият модел от фиг. 1 съдържа 9 прехода и 28 позиции. Преходите представят:

Z_1 – предварителна обработка и почистване на данните;

Z_2 – разделяне на данните в обучаващо, валидиращо и тестово множества;

Z_3 – избиране на метод (критерий) за разделяне на обучаващото множество;

Z_4 – определяне на "най-добрия атрибут" за разделяне на обучаващото множество;

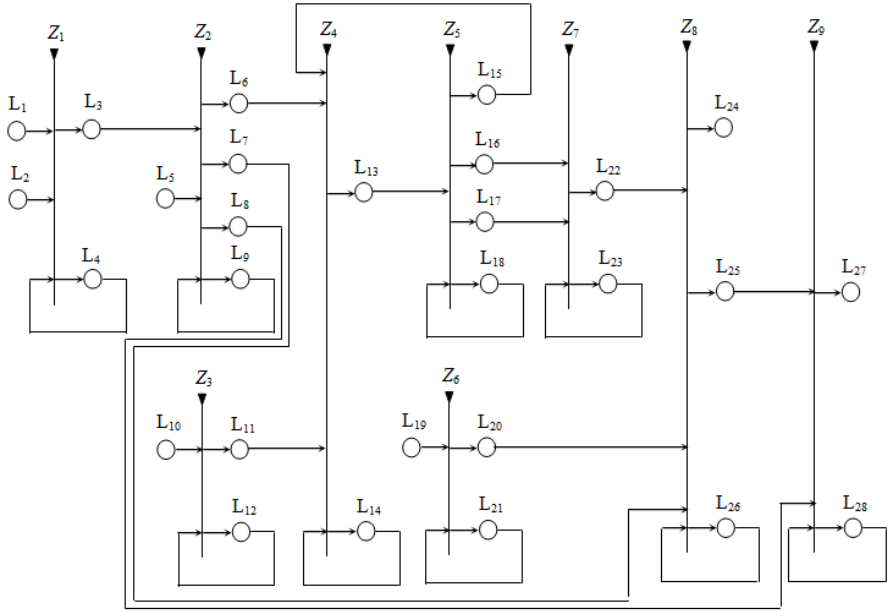
Z_5 – разделяне на данните;

Z_6 – дейности с методи за окастряне;

Z_7 – съставяне на дървото на решението;

Z_8 – валидиране на модела;

Z_9 – тестване на модела.



Фиг. 1: ОМ -модел на процеса на създаване на дърво на решението

Дадено е подробно описание на модела.

2.1.2. Обобщеномрежов модел на процеса на създаване на асоциативни правила чрез алгоритъм Apriori

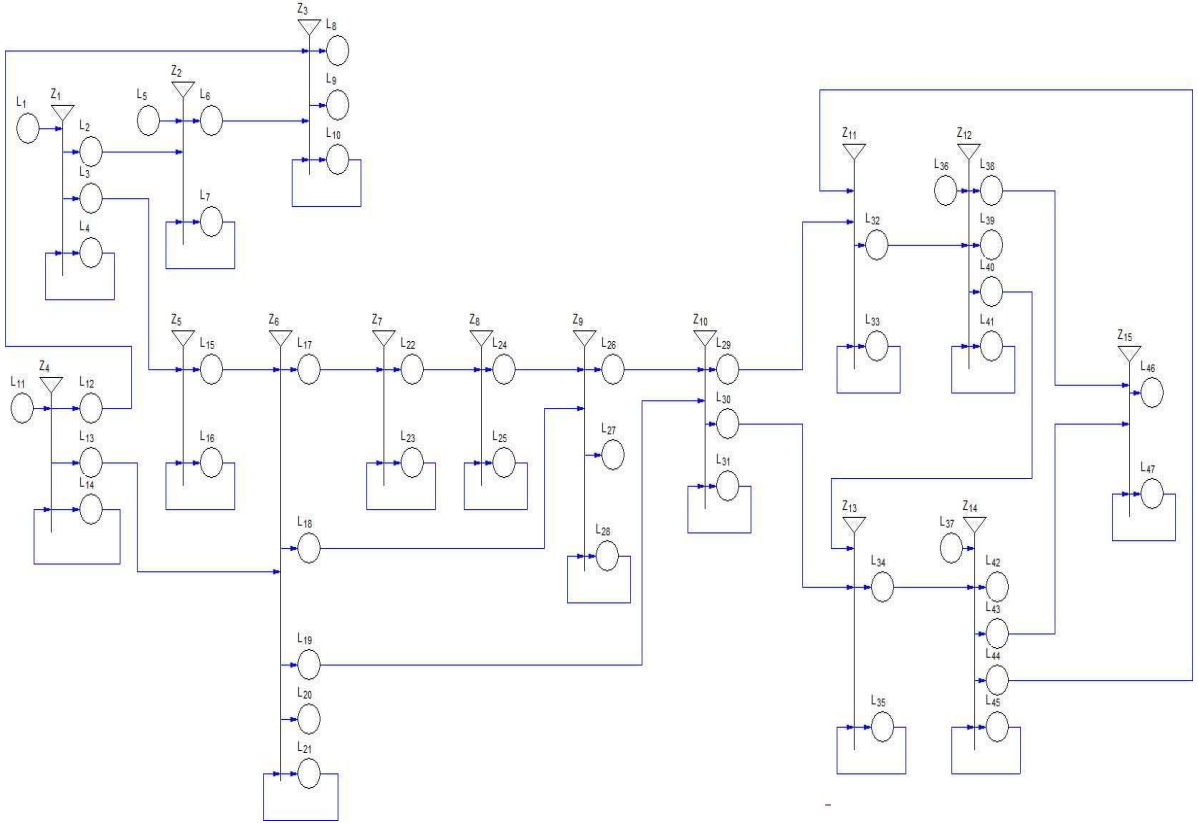
Асоциативният анализ представлява техника за извличане на знания от данни чрез обучение без наблюдение (без учител). Входната информация е записана във вид на транзакции и най-често е в релационна база данни или хранилище. Част от алгоритмите, използвани за асоциативен анализ са: *Apriori, sampling* (вземане на проби), *partitioning* (разделяне), както и алгоритми за паралелна и разпределена обработка.

Ще наричаме едно множество *често множество* или накратко, *често*, ако се появява в данните над предварително определен праг.

Обобщеномрежовият модел, описващ процеса по създаване на асоциативни правила чрез *Apriori* е показан на фиг. 2. Алгоритъмът се основава на свойството, че ако едно множество е често, то и неговите съставни ще бъдат чести.

Обобщената мрежа има 15 прехода и 47 позиции (фиг. 2). Преходите описват следните процеси:

- Z_1 – постъпването на транзакции от база данни или хранилище
- Z_2 – преобразуване на постъпилите транзакции в подходящ формат за намиране на честите единични елементи (в табличен формат)
- Z_3 – генериране на чести единични елементи
- Z_4 – задаване на честота за множествата
- Z_5 – създаване на всички възможни кандидати от двуелементни комбинации от всички елементи
- Z_6 – намиране на честите двуелементни множества
- Z_7 – създаване на кандидат - триелементни множества
- Z_8 – разделяне на всяко триелементно множество на три двуелементни
- Z_9 – проверяване на честотата на всички три двуелементни множества от едно триелементно (окастряне)
- Z_{10} – генериране на асоциативни правила



фиг. 2: Обобщена мрежа, описваща процеса по създаване на асоциативни правила чрез Apriori алгоритъм

- Z_{11} – изчисляване на подкрепата за асоциативните правила
- Z_{12} – задаване на минимален праг на подкрепа за асоциативните правила
- Z_{13} – изчисляване на доверието на асоциативните правила
- Z_{14} – задаване на минимален праг на доверие за асоциативните правила
- Z_{15}– записване на крайни (силни) асоциативни правила, удовлетворяващи минималните критерии за подкрепа и доверие.

  Дадено е подробно описание на модела.

2.1.3. Обобщеномрежов модел на процеса на извличане на асоциативни правила чрез frequent pattern-growth алгоритъм

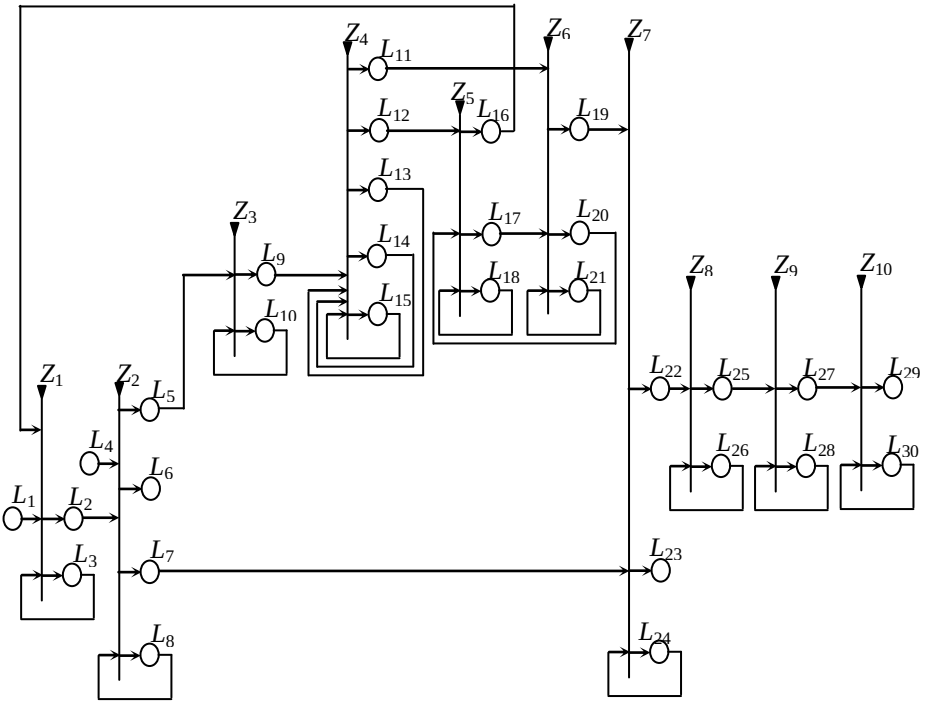
Алгоритъмът, конструиращ *Frequent pattern tree* не се нуждае от създаването на всички възможни кандидат-множества на елементите, за да извлече асоциативните правила. Той представя входните данни в компресиран вид. Конструира се чрез сканиране на транзакциите една по една и записването им като пътища в дървото, като са необходими само две преминавания през данните. Алгоритъмът обработва дървото чрез рекурсия. При нанасянето на транзакциите по дървото, ако са налични едни и същи елементи в транзакциите, то техните пътища се припокриват. По-този начин се получава по-кратко представяне на входната информация.

Обобщеномрежовият модел, описващ процеса на генериране на асоциативни правила чрез *Frequent Pattern-Growth* е показан на фиг. 3.

Обобщената мрежа има 10 прехода и 30 позиции (фиг. 3). Преходите описват следните процеси:

- Z_1 – постъпване на транзакции в хранилището
- Z_2 – задаване на поддръжка (честота)
- Z_3 – сортиране на елементите в транзакциите по низходящ ред
- Z_4 – конструиране на *Fp-Tree*
- Z_5 – обхождане на *Fp-Tree* "отдолу - нагоре"
- Z_6 – съставяне на дърво с възможните пътища до елемент

- Z_7 – записване на списък с пътищата до елемент (*F-List*)
- Z_8 – съставяне на условно дърво (*Fp-Growth*)
- Z_9 – рекурсивно извличане на честите подмножества от условното дърво (*Fp-Growth*)
- Z_{10} – запис на честите елементи и техните подмножества.



Фиг. 3 Обобщеномрежов модел, описващ генерирането на асоциативни правила чрез алгоритъм Fp-Tree

Дадено е подробно описание на модела.

2.1.4. Обобщеномрежов модел на процеса на извличане на асоциативни правила чрез алгоритъм Eclat използвайки метеорологични бази от данни

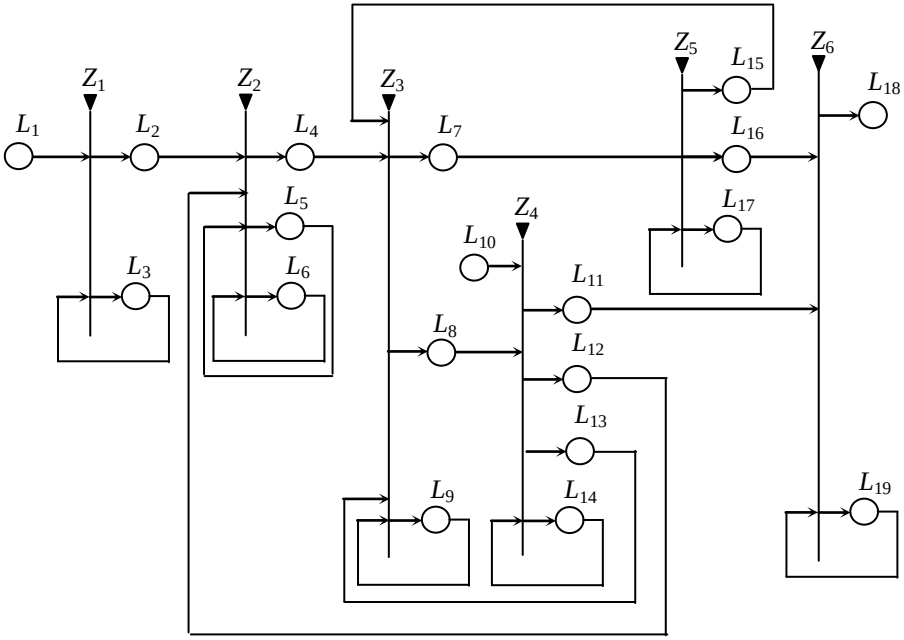
Конструиран е обобщеномрежов модел за определяне на вероятността за възникване на пожар в горите с помощта на

асоциативни правила. При моделирането на процеса са извлечени чести елементи чрез алгоритъма *Eclat.* Той използва вертикален формат за данните за генериране на чести зависимости. Представеният ОМ-модел трябва да подбере добре входните метеорологични наблюдения и коректно да предскаже предварително неизвестни параметри за времето. Той може да бъде използван за наблюдаване на вероятността за пожар чрез извличането на чести елементи в зависимост от различни метеорологични условия.

На Фиг. 4 е представен ОМ-модел на процеса за извличане на чести елементи чрез алгоритъм *Eclat*. Той съдържа следното множество от преходи A:

$$A = \{Z_1, Z_2, Z_3, Z_4, Z_5, Z_6\},$$

където преходите описват процесите:



Фиг. 4: ОМ-модел на процеса на генериране на асоциативни правила чрез алгоритъм Eclat

Z_1 – Работа с транзакционен склад от данни, съдържащ метеорологични данни

Z_2 – Трансформиране на транзакциите във вертикален и изчисляване на минималната подкрепа за всеки елемент

Z_3 – Сортиране на елементите и подмножествата от елементи в нарастващ ред по минимална подкрепа

Z_4 – Определяне на минималната подкрепа, зададена от потребителя и намиране на честите елементи

Z_5 – Намиране на честите подмножества чрез свойството *Apriori*

Z_6 – Записване на честите елементи и подмножества.

Дадено е подробно описание на модела.

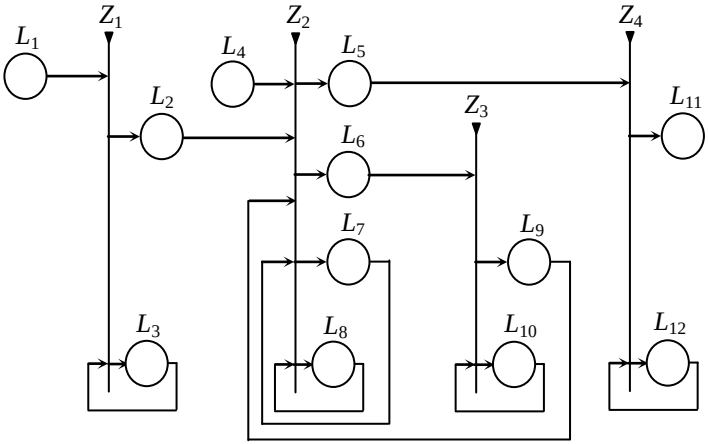
2.1.5. Моделиране на процеса на извличане на зависимости от последователности чрез Generalized Sequential Pattern Algorithm (GSP), използвайки апарата на обобщените мрежи

Зависимости от последователности са зависимости, които присъстват в данните над зададен минимален праг в определен времеви интервал. Откриването на зависимости от последователности се прилага за отриване на чести последователности в база данни. Зависимостите от последователности могат да бъдат използвани за създаване на асоциативни правила. Алгоритъмът *GSP* е реализиран в *RapidMiner* с данни за метеорологични наблюдения от база данни, съдържащи времеви данни, инфрачервена камера и детектор за дим, за да се определи опасността от горски пожар. Предложеният ОМ-модел може да бъде използван за контрол на процеса по извличане на зависимости от последователности, зависещ от метеорологичните параметри.

Разработеният тук модел (Фиг. 5) е разширение на модела от т. 1.2. Той съдържа 4 прехода и 12 позиции. Преходите представят следните процеси:

- Z_1 – настройка на мерките, критериите minsup, maxgap, mingap, window size от потребителя
- Z_2 – избор на транзакции от базата данни с метеорологични измервания и анализиране на единичните последователности
- Z_3 – генериране на k+1 кандидат-подпоследователности от единичните чести последователности

- Z_4 – записване на получените правила от последователностното извличане.



Фиг. 5: Обобщена мрежа на процеса на извличане на последователности чрез GSP алгоритъм

Дадено е подробно описание на модела.

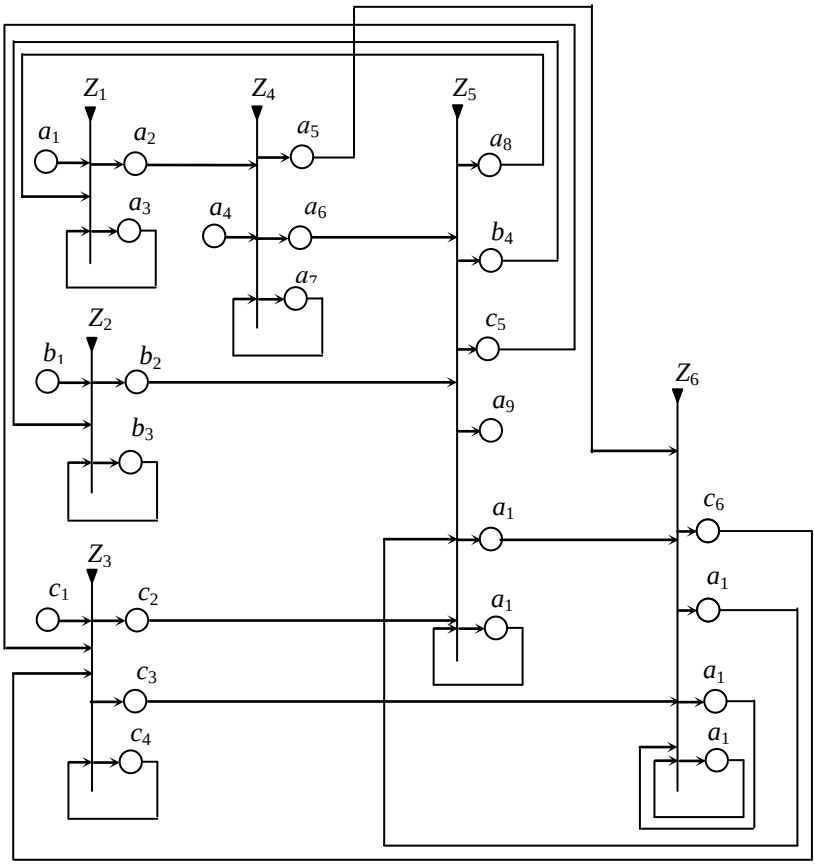
2.1.6. Обобщеномрежов модел на процеса на извличане на асоциативни правила

На фиг. 6 е представен обобщеномрежов модел на процеса на извличане на асоциативни правила. Преходите на модела представят стандартните стъпки при откриването на асоциативни правила без значение кой алгоритмите ще бъде използван.

Той съдържа 6 прехода и 24 позиции. Множеството на преходите A е следното:

$$A = \{Z_1, Z_2, Z_3, Z_4, Z_5, Z_6\},$$

където преходите описват следните процеси:



Фиг.6 Обобщеномрежов модел на процеса на извличане на асоциативни правила

Z_1 – дейности с хранилището с данни

Z_2 – избор на алгоритъм за генериране на асоциативни правила

Z_3 – предварителна обработка на избрани данни

Z_4 – избор на критерии

Z_5 – генериране на асоциативни правила

Z_6 – валидиранеи тестване на получените правила.

Позициите в модела са три групи и са свързани с три типа ядра, които постъпват в тях. Дадено е подробно описание на модела.

2.2. Йерархичен обобщеномрежов модел на процеса на клъстеризация

Тук са представени накратко основните техники за клъстеризация и е разработен йерархичен обобщеномрежов модел на процеса на клъстеризация (фиг. 7). Той описва стандартните стъпки, които се извършват при клъстерния анализ.

Моделът съдържа 8 прехода и 32 позиции. Множеството на преходите A е следното:

$$A = \{Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8\},$$

където преходите описват следните процеси:

Z_1 – избор на входни данни

Z_2 – избор на критерии

Z_3 – избор на метод за клъстеризация

Z_4 – разделяне на входните данни на обучаващо, валидиращо и тестово множества";

Z_5 – предварителна обработка на данните – почистване на данни, отстраняване на екстремни стойности, обработка на липсващи стойности, трансформиране

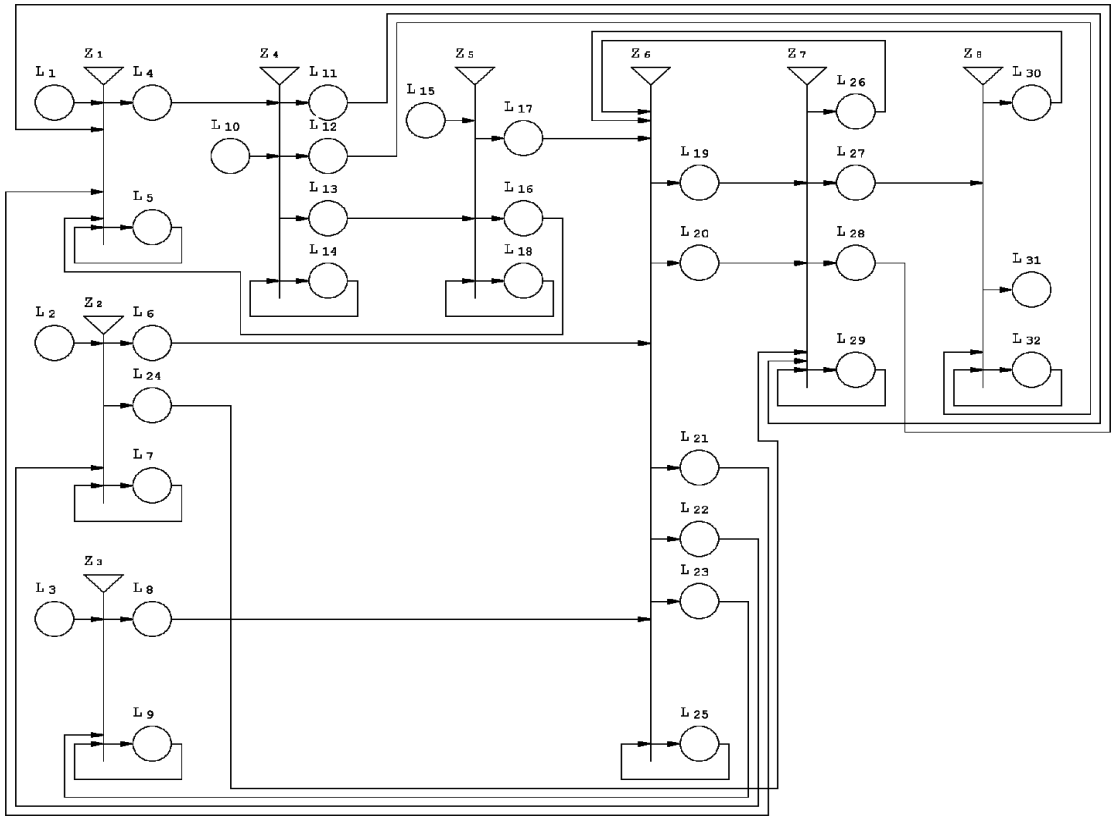
Z_6 – процес на клъстеризация -данните се групират или разделят в клъстери

Z_7 – валидация на резултата от клъстеризацията

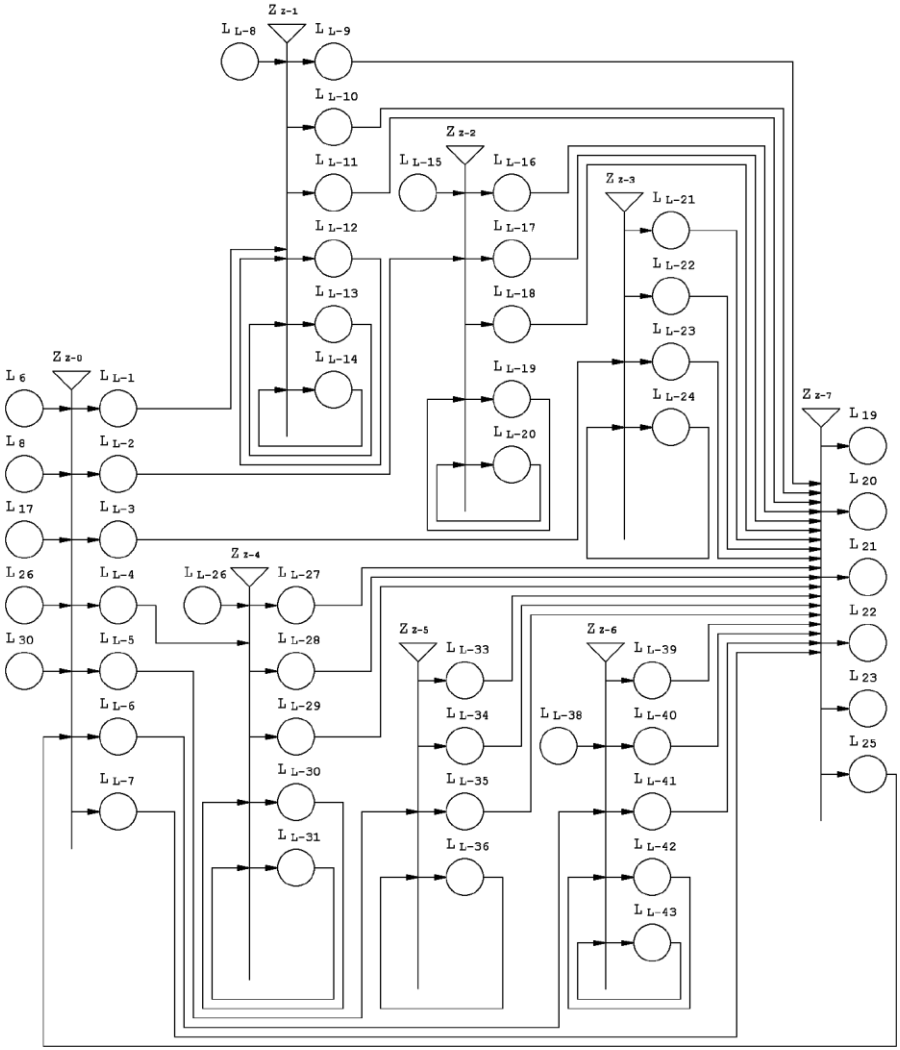
Z_8 – тестване на резултата от клъстеризацията.

Позициите в модела са три групи и са свързани с три типа ядра. Дадено е подробно описание на модела.

Преходът Z_6 от обобщеномрежовия модел на фиг. 7 може да се замени с подмрежа чрез прилагане на йерархичен оператор H_3 от [15, 17]. Подмрежата е представена на фиг. 8.



Фиг. 7 Обобщеномрежов модел на процеса на клъстеризация



Фиг. 8: Подмрежа на процеса на клъстеризация

Множеството на преходите A_{A-1} е следното:

$$A_{A-1} = \{ Z_{Z-0}, Z_{Z-1}, Z_{Z-2}, Z_{Z-3}, Z_{Z-4}, Z_{Z-5}, Z_{Z-6}, Z_{Z-7}\},$$

където преходите описват следните процеси:

Z_{Z-0} – входни параметри

Z_{Z-1} – йерархични методи за клъстеризация (hierarchical)

Z_{Z-2} – разделящи методи за клъстеризация (partitioning)

$Z_{Z\text{-}3}$ – методи за клъстеризация, базирани на плътност (density-based clustering)
$Z_{Z\text{-}4}$ – методи за клъстеризация, базирани на мрежа (grid-based clustering)
$Z_{Z\text{-}5}$ – методи за клъстеризация, базирани на модел (model-based clustering)
$Z_{Z\text{-}6}$ – други методи за клъстеризация - fuzzy clustering, soft computing
$Z_{Z\text{-}7}$ – изходни данни.

Позициите в модела са три групи и са свързани с три типа ядра, които постъпват в тях. Дадено е подробно описание на модела.

Всеки един от преходите $Z_{Z\text{-}1}$, $Z_{Z\text{-}2}$, $Z_{Z\text{-}3}$, $Z_{Z\text{-}4}$, $Z_{Z\text{-}5}$, $Z_{Z\text{-}6}$ (фиг. 8) може да бъде заменен с подмрежа отново чрез прилагане на йерархичен оператор H_3.

Изводи

Изграждането на обобщеномрежови модели на различни процеси от теорията на извличането на знания от данни е ефективен начин за тяхното визуализиране и разбиране за същността им. Конструираните в дисертационния труд обобщеномрежови модели продължават започнатите изследвания, свързани с представимостта на процеси от областта на Data mining чрез обобщени мрежи. При описание на различни Data mining техники с един и същи инструмент може да се направи сравнение между тях и те да се изследват от гледна точка на бързодействие, използвани ресурси за работата си, коректност на резултата и др. От друга страна чрез конструиране на ОМ-модели, реализиращи една и съща Data mining техника по различни алгоритми (в случая алгоритми за извличане на асоциативни правила), тези алгоритми може да се анализират и да се избере оптималният сред тях, в зависимост от обработваните данни.

Разработените обобщеномрежови модели представят основните техники за извличане на асоциативни правила по различни алгоритми. Моделирани са алгоритмите Apriori, FP-Growth, GSP и Eclat при зададени подкрепа, доверие и лифт. Моделиран е и алгоритъм за конструиране на дърво на решението, чрез което се извличат асоциативни правила.

Конструираните обобщеномрежови модели позволяват:

1. Да се анализират основни алгоритми, свързани с извличане на знания от данни;
2. Да се симулират процесите на функциониране на алгоритмите с цел подобряване на начина на тяхното протичане;
3. Да се анализират реални данни от големи масиви с данни на базата на различни типове алгоритми за създаване на асоциативни правила.

Съставени са йерархични ОМ-модели за:

1. Откриване на асоциативни правила без значение кой от алгоритмите ще бъде използван;
2. описание на процеса на клъстеризация;
3. описание на процеса на избор на клъстеризиращ алгоритъм.

Част от преходите в тези йерархични модели могат да се заменят с подмрежи, които да детайлизират моделираните процеси.

При изготвянето на дисертационния труд се появиха редица идеи за бъдещи изследвания, например процесът по извличане на асоциативни правила да бъде разширен с вариант за създаване на интуиционистки размити асоциативни правила. Чрез тяхната помощ например изследването на метеорологичните данни би било по-точно. Интуиционистки размити оценки могат да се използват също и при процеса на клъстеризация. Така ще се направи по-точна категоризация на данните, попадащи между клъстерите.

Глава 3. Реализации на процеси от областта на извличане на знания от данни

В тази глава са представени програмни реализации на алгоритмите и техниките, за които са представени обобщеномрежови модели в Глава 2. За целта са използвани софтуери със свободен код за обработка на данни чрез техники за извличане на знания - *RapidMiner* и статистическият език за програмиране *R*, който включва в себе си и библиотеки за прилагане на методи за извличане на знания.

3.1 Реализация на алгориъм за извличане на чести елементи и генериране на асоциативни правила Apriori чрез чрез средствата на статистическият език R използвайки метеорологични данни

За реализацията на извличането на асоциативни правила чрез езика *R* са използвани метеорологочните записи от база данни "*weather*". За тестването е създадена виртуална таблица с данни за времето – влажност, наличие на вятър, температура, детектор за наличие на дим и инфра-червена камера. Въведени са 102 записа, които са обработени за откриването на скрити зависимости в данните. Преди да се пристъпи към анализа се осъществява връзка между системата за управление на база данни (СУБД), реализирана на *MySQL* (*MySQL Connector- ODBC*) и интегрираната среда за разработка на *R*, в която ще се извърши извличането на знания.

След сканирането на данните е получен резултат от 6 асоциативни правила. За по-голяма прегледност е зададено те да бъдат сортирани по тяхната стойност за подкрепа.

Следва изпълнение на процедура по окастряне на асоциативните правила, за да се определи кои от тях могат да бъдат надеждни критерии за преценка и предвиждане на бъдещи температурни зависимости (фиг. 9).

```
> subset.matrix <- is.subset(rules.sorted, rules.sorted)
> subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA
> redundant <- colSums(subset.matrix, na.rm=T) >= 1
> which(redundant)
[1] 2 5
> rules.pruned <- rules.sorted[!redundant]
> inspect(rules.pruned)
  lhs                          rhs                     support confidence     lift
1 {fire_detector_color=no} => {smoke_detector=no} 0.7352941  0.9493671 1.166692
2 {Humidity=high}          => {smoke_detector=no} 0.4313725  0.9777778 1.201606
3 {Humidity=normal}        => {Temperature=hot}   0.4313725  0.7719298 1.290768
4 {Temperature=cool}       => {smoke_detector=no} 0.4019608  1.0000000 1.228916
```

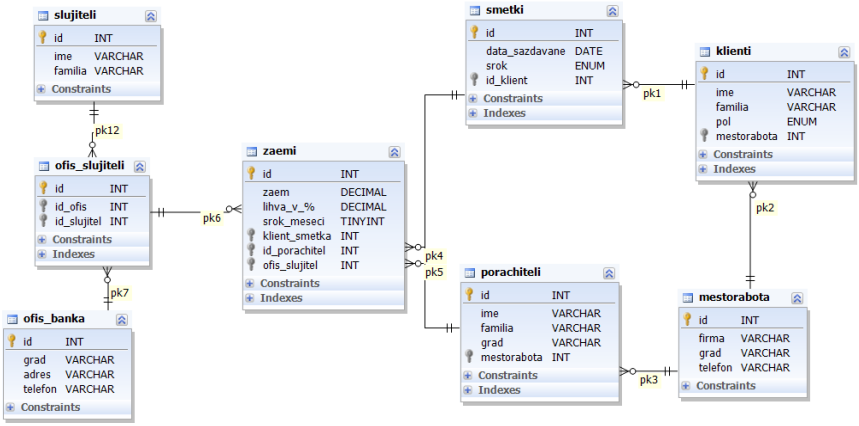
Фиг. 9 Стъпка по окастряне

Като резултат са получени четири асоциативни правила, представящи зависимостите в наличните данни. Необходимо е да се отбележи факта, че при по-големи количества входна информация има по-голяма вероятност за извличане на силни асоциативни правила. При натрупване на информация в базата данни в определен момент от време ще бъде възможно предвиждането на вероятността от пожар

при конкретни температурни зависимости. В представения анализ беше представен и описан приложно процеса на извличане на асициативни правила чрез *Apriori* алгоритъм. За допълнително онагледяване и разбиране на процесите в областта на извличането на знания от данни активно се използват средствата за визуализация.

3.2 Реализация на алгориъм за извличане на чести елементи и генериране на асоциативни правила FP-Growth чрез чрез средствата на софтуерния продукт RapidMiner използвайки данни от банкова база данни

Процесът по извличане на асоциативни правила е илюстриран със софтуера *RapidMiner*, който използва като основен алгоритъм *Fp-Growth*. За целта на анализа е съставена примерна релационна база от данни за малка част от дейността на банка. За реализацията и е използвана системата за управление на бази от данни *MySQL*, чието администриране е осъществено чрез инструмента *dbForge Studio Express for MySQL*. Схемата на базата данни е показана на фиг. 10. Въведената информация е с обем от 102 записа.



Фиг. 10: Схема на база данни "bank"

При стартиране на процеса данните се извеждат в режим *Result Overview*, изглед *Data View*. В *Meta Data View* се преглеждат типовете на данните и някои налични статистики за тях. Таблицата със заемите съдържа данни от тип *real* и тип *integer*. Операторът *GP-Growth* се

нуждае от биноминални (binominal) данни, за да извлече честите елементи. За елиминиране на различията на данните се извършва предварителна обработка. С цел да се получи по-структуриран, прегледен и разбираем процес, етапите от трансформацията на информацията са проведени в подпроцес, реализиран чрез оператора *Subprocess*. Подпроцесът представлява вътрешно ниво на обработка.

Асоциативният анализ се извършва с част от полетата на таблицата - lihva_v_%, srok_meseci и zaem. Записите са селектирани чрез оператор *Select Attributes*. Предварителната обработка на данните започва с дискретизация по честота, осъществена чрез оператора Discretize by Friquency. Като резултат се получават данни от номинален тип.

При налични, предварително обработени данни започва процесът на откриване на асоциативни правила. Първоначално се откриват честите елементи. На тяхна база се създават асоциативни правила, които потенциално се намират в базата данни. Задава се доверие за правилата, което може да приема стойности от 0 до 1. За анализа е записан минимален праг от 0,6. *RapidMiner* предлага възможност за оценяване на асоциативно правило чрез мерките *lift, convinction, ps, gain* и *laplace*. Асоциативните правила, удовлетворяват минимален праг на подкрепа от 0,3 и минимален праг на доверие от 0,6.

Извлечените асоциативни правила от базата данни са следните:

Правило 1: Ако лихвата на заема е до 3,3%, *то* той ще бъде изплатен в срок до 73 месеца;

Правило 2: Ако лихвата на заема е до 3,3 %, *то* той ще бъде с размер до 23,500 лв;

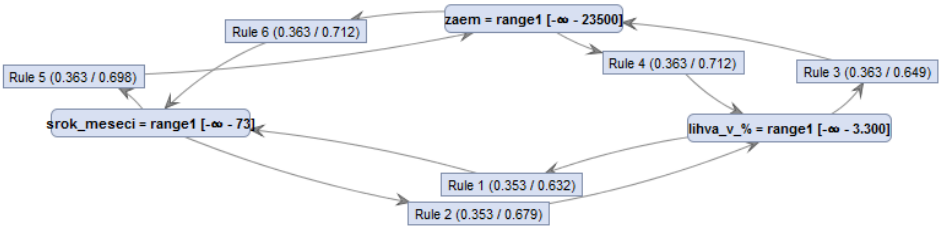
Правило 3: Ако заемът е в срок до 73 месеца, *то* лихвата му ще бъде до 3,3%;

Правило 4: Ако заемът е в срок до 73 месеца, *то* неговият размер ще бъде до 23,500 лв;

Правило 5: Ако заемът е с размер до 23,500 лв, *то* лихвата му ще бъде до 3,3%;

Правило 6: Ако заемът е с размер до 23,500 лв, *то* той ще бъде в срок до 73 месеца.

Всички открити асоциативни правила, удовлетворили минималния праг на доверие и подкрепа, са изобразени под формата на граф на фиг. 11.



Фиг. 11 Асоциативните правила, изобразени под формата на граф

Тестването на откритите асоциативни правила се осъществява чрез прилагането на извлечените закономерности върху нова база от данни. Реализирана е връзка към втора база от данни, съдържаща 37 записа. Извършена е предварителна обработка на новите данни във втори подпроцес. Стойностите са преобразувани в бинарен вид. Крайният резултат е записан във файл на Excel.

3.3 Реализация на алгориъм за извличане на чести елементи и генериране на асоциативни правила Eclat чрез чрез средствата на статистическият език R използвайки метеорологични данни

За извличането на чести елементи чрез алгоритъм Eclat е използван статистическият език *R*. Множествата с метеорологични данни са съхранени във файла "weather.csv". Те имат следните атрибути - *wind* (calm, breeze, gale), *temperature* (cool, mild, hot), *outlook* (sunny, overcast, rainy), *humidity* (normal, hide) and *fire* (yes, no). Част от метеорологичните данни са представени на Фиг.12. Записите са визуализирани под формата на транзакции. Най-важната стъпка в процеса е да се запише минимален праг на подкрепа. В примера са въведени 84 транзакции и минимален праг на подкрепа от 0.3 (30%), който се определя от потребителя. Анализът за извличане на чести

елементи описва вероятността от пожар в зависимост от метеорологичните условия.

	A	B	C	D	E
1	**Wind**	**Temperature**	**Outlook**	**Humidity**	**Fire**
2	calm	hot	overcast	normal	yes
3	gale	cool	rainy	high	no
4	calm	hot	overcast	normal	yes
5	calm	cool	overcast	high	no
6	gale	cool	rainy	high	no
7	calm	cool	overcast	high	no
8	calm	cool	overcast	high	no
9	breeze	hot	sunny	normal	yes
10	calm	cool	overcast	high	no
11	calm	cool	overcast	high	no
12	breeze	hot	sunny	normal	yes
13	calm	hot	overcast	normal	yes

Фиг. 12: Част от метеорологичните данни

Резултатите имат формата:

If Wind=calm,Temperature=hot, Outlook=overcast and Humidity=normal Then Fire=yes (minsup=0.32)

If Wind=calm,Temperature=cool,Outlook=overcast and Humidity=high Then Fire=no (minsup=0.48)

3.4 Реализация на алгориъм за извличане на последователни зависимости gsp(generalized a sequential patterns) чрез средствата на софтуерния продукт Rapidminer използвайки метеорологични данни

Съставеният ОМ модел на процеса на извличане на последователни зависимости чрез алгоритъм *GSP*, представен в т. 2.5, може да бъде използван за откриване на последователни закономерности в метеорологични данни от база данин, съдържащи времеви данни, инфрачервена камера и детектор за дим за да се определи възможността за горски пожар. Алгоритъмът *GSP* е реализиран в *RapidMiner*. *GSP* използва минимално времево ограничение, максимално времево ограничение и свойството на плъзгащият се прозорец. Последователното извличане на зависимости е стъпка към разширяването на извличането на знания от данни към времевия анализ. ОМ модели на алгоритмите *FP-Growth* и Eclat са използвани в

разширените алгоритми за последователно извличане на закономерности *PrefixSpan* и *SPADE*.

3.5 Реализация на техника за конструиране на дърво на решението чрез средствата на софтуерния продукт Rapidminer използвайки метеорологични данни

Процесът по конструиране на дърво на решението се състои от предварителна обработка на данните, задаване на атрибут за етикетите на класовете, генериране на модел и последващото му тестване. Първоначално е осъществена връзка към *MySQL* база данни, съдържаща метеорологични данни и след това са избрани необходимите данни за анализа. За текущият анализ е избрано поле "humidity".

Конструирано е дърво на решението, което класицифира данните, намирайки температурните зависимости във времето. При добавяне на записи за настъпили пожари в минало моделът може да предскаже температурните условия, при които има възможност за пожар. От конструираното дърво на решението са извлечени следните правила:

If Outlook=a few showers *then* Humidity=high*;*

If Outlook=overcast and Wind=breeze and Temperature=cool *then* Humidity=high;

If Outlook=overcast and Wind=breeze and Temperature=hot *then* Humidity=normal;

If Outlook=overcast and Wind=calm *then* Humidity=normal;

If Outlook=overcast and Wind=gale *then* Humidity=normal;

If Outlook=overcast and Wind=light breeze *then* Humidity=normal;

If Outlook=partly cloudy *then* Humidity=normal;

If Outlook=partly sunny *then* Humidity=normal;

If Outlook=periods of rain *then* Humidity=high;

If Outlook=storm of hail *then* Humidity=high;

If Outlook=sunny *then* Humidity=normal.

Заключение – основни резултати

Дисертационният труд включва кратък обзор на теориите на обобщените мрежи и извличане на знания от данни, обобщеномрежово моделиране на техники за извличане на знания от данни и тяхната реализация.

Приносите в настоящия дисертационен труд са от научно-приложен и приложен характер.

Научно-приложните приноси са следните:

Създадени са шест обобщеномрежови модела

- обобщеномрежов модел, отразяващ паралелната работа на процеса за конструиране на дърво на решението. За стъпките по създаване на дървото на решението е използван алгоритъма на Хънт. Дървото на решението се съставя чрез подход "отгоре-надолу". Като резултат се получава модел-дърво на решението, който илюстрира класификацията на отделните групи от данни. Възможно е записване на правила (класификационни, асоциативни), които да представят получените резултатни стойности от класификацията.
- три ОМ-модела, отразяващи процесите по извличане на асоциативни правила чрез прилагането на най-използваните алгоритми за извличане на чести елементи Apriori, FP-Growth и Eclat.
- обобщеномрежов модел, представящ процеса на откриване на зависимости от последователности (чести елементи, които се повтарят последователно във времето), който представлява разширение на алгоритъма Apriori.
- Йерархичен обобеномрежов модел, представящ дейностите по процеса на извличане на асоциативни правила. ОМ-моделът проследява всички стъпки, които е необходимо да бъдат извършени по време на извличане на асоциативни правила без значение избрания алгоритъм. За избор на конкретен алгоритъм на обработка се използва ОМ-модел за извличане на асоциативни

правила по Apriori, FP-Growth или Eclat алгоритми, който ще бъде подмрежа на йерархичния ОМ-модел.

- Конструирани са йерархичен обобщеномрежов модел на отделните стъпки на процеса на клъстеризация и йерархичен подмодел на избора на клъстеризиращ метод. Първият ОМ-модел представя паралелната работа по основните стъпки от обработката на данни с клъстерна техника. Вторият ОМ-модел, отразява работата по избора на тип на клъстеризиращ метод. Тук има възможност за разработване на следващи ОМ-модели, представящи процеса на клъстеризация по конкретен алгоритъм. Те ще са подмрежи на преходите на втория ОМ-модел.

Приложният принос се състои в тествания на основни техники от областта на извличане на знания. За по-голямата част от конструираните модели е илюстрирана реализация с актуални данни с цел онагледяване и изследване на практическата част от работата с посочените методи. За тестванията са използвани продукти със свободен лиценз на разпространение. Първата използвана среда за разработка е статистическият език *R*, който освен разширение за работа по извличане на знания от данни осигурява и голямо разнообразие от средства за визуализация. Вторият продукт е софтуерът *RapidMiner*, притежаващ изключително дружелюбен интерфейс.

Списък на публикациите по дисертационния труд

1. Бурева, В. Обобщеномрежов модел на процеса на създаване на асоциативни правила. Годишник на секция “Информатика” при Съюза на учените в България, Том 5, 2012, 73-83.
2. Бурева, В. Методи за извличане на закономерности от бази от данни. Академично списание "Управление и образование", Университет "Проф. д-р Асен Златаров" , кн. 4, том 8, 2012, 255-258.
3. Bureva, V., P. Chountas, K. Atanassov. A generalized net model of the process of decision tree construction. Proc. of 13th Int. Workshop on Generalized Nets, London, 29 October 2012, 1–7.
4. Bureva, V., E. Sotirova. Generalized net of the process of association rules discovery by Eclat algorithm using weather databases, 14-th Int. Workshop on Generalized Nets Burgas, 29–30 November 2013, 1–10.
5. Бурева, В. Алгоритми за извличане на асоциативни правила, Академично списание "Управление и образование", Университет "Проф. д-р Асен Златаров", Vol. 9 (6) 2013, 121-128.
6. Bureva, V., E. Sotirova, P. Chountas, Generalized Net of the Process of Sequential Pattern Mining by Generalized Sequential Pattern Algorithm (GSP), IEEE IS'14, Warsaw, (in press)
7. Бурева, В. Обобщеномрежов модел на процеса на извличане на асоциативни правила използвайки Frequent Pattern-Growth Method, Annual of “Informatics” Section Union of Scientists in Bulgaria, Том 6, 2013, 46-53.