14th Int. Workshop on Generalized Nets
Burgas, 29–30 November 2013, 1–10

Generalized net of the process of association rules discovery by Eclat algorithm using weather databases

Veselina Bureva and Evdokia Sotirova

"Prof. Asen Zlatarov" University
1 "Prof. Yakimov" Blvd, Burgas–8010, Bulgaria
e-mails: vesito_ka@abv.bg, esotirova@btu.bg

Abstract: In the present paper, a Generated net model is constructed to determine the possibility of forest fire by association rules. To model the process, we use frequent pattern mining by the Eclat algorithm. A pattern is considered to be frequent when it occurs in the data more often than a predefined minimum support frequency. Frequent pattern mining is a step of the process of association rules discovery. Eclat algorithm uses vertical data format for generating frequent patterns, with associative rules having the *If A then B* form. The proposed Generated net model should both fit well the input metrological observations, and correctly predict previously unknown weather parameters. It can be used for monitoring of the possibility of fire via frequent pattern mining depending on metrological conditions.
Keywords: Generalized Net, Association rules, Weather databases, Frequent pattern mining, Data Mining, Knowledge Discovery.
AMS Classification: 68Q85, 62H30.

1 Introduction

The meteorological data from the World meteorological stations are collected in databases and some of them can be used for analyzing the danger from wildfires. Standard weather analysis and forecasts for the next few hours (12 or 24 hours) are typically quite accurate, but after several days, forecast accuracy falls off quickly. Also the quantity of meteorological data online is increasing, which makes it important to use specific techniques for analysis. So by using Data mining techniques such as classification, clustering, prediction, outlier analysis, we can find and extract usable hidden knowledge from largely available weather forecast databases [8, 9, 12]. This can help for understanding the climate variability and climate prediction.

Data mining includes a large number of fields of science like database and data warehouse technology, machine learning, high-performance computing, pattern recognition, image and signal processing, data visualization, neural networks, information retrieval, statistics, and spatial or temporal data analysis. The most common data mining tasks are description, estimation, prediction, classification, clustering and association [4, 10].

In the current paper is presented one of the most used techniques for data mining - association rules discovery. It is used for extracting of the unseen relationships, patterns between items in large datasets and data streams [1]. Association rules are a form of unsupervised learning. In much other analysis the result is often given like rule. The process of discovering association rules use two-step approach: frequent itemset generation and rule generation. Frequent patterns can be itemsets, subsequences with support greater than the minimum support . This threshold is defined by the user.

The associative analysis can be used for discovering and exploring hidden relationships between items in large amounts of weather forecast databases. The extracted patterns are presented in the form of association rules that are represent in the *If A then B* form, and can be used to predict some severe weather situations, such as thunderstorms. Most algorithms for mining association rules identify relationships among transactions using binary values. Transactions with quantitative values and items are, however, commonly seen in real-world applications.

In the paper was constructed a model for the extraction a frequent patterns by Eclat algorithm in weather databases using the apparatus of Generalized nets (GNs, see [2, 3]). The generated model should both fit well the input metrological observations and correctly predict previously unknown weather parameters.

Depending on the type of the data and kind of association rules, the different algorithms for knowledge discovery can be used. Each of them has the best effectiveness in practical examples. Apriori, FP-Growth and Eclat is the most used methods. In [11, 5] the algorithms for association rule discovery are described. The GN-models for the processes of frequent pattern mining by Apriori [6] and by Fp-Growth [7] algorithms are constructed. The difference between Apriori, FP-Growth and Eclat is the form of presenting the datasets. Apriori and FP-Growth use horizontal form for visualization while Eclat presents the data in vertical format. Eclat scans database once and prepares the data in vertical format. For each item in database it is created a *tidlist* (transactional list) which includes the numbers of transactions with the same item. The support for the item is equal of the length of his *tidlist*. Sometimes a bit matrix is used for presenting the algorithm. Each row corresponds to an item and each column confirms to a transaction. If the item is existed in transaction it is written "1", otherwise the item is deleted. The prefix tree is created. From parent-node to child-node is passed by new matrix which is created from intersecting of first row with the others last. This step is repeated. In the end, the support for the items is estimated. The frequent patterns mined can be used for generating candidate-itemsets.

2 Generalized net model

The GN-model for the process of frequent pattern mining by Eclat algorithm is presented in Fig. 1. It contains the following set of transitions A:

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

where the transitions describe these process:

- Z_1 - "Work of the transactional warehouse with weather data";
- Z_2 - "Transforming the transactions in vertical data layout and counting the minimum support for each item";
- Z_3 - "Sorting the itemsets and subsets in ascending order of minimum support";
- Z_4 - "Determining the user's minimum support and finding the frequent itemsets";
- Z_5 - "Finding the frequent subsets with Apriori property";
- Z_6 - "Recording the frequent itemsets and subsets".

Initially in the place L_3 there is one α-token. It will be in his own place during all the time of GN functioning. It has the following characteristic: "*transactional warehouse with weather data*".

The α-token in place L_3 generates new α-tokens at certain time moments which will can move to place L_2 with characteristic: "*selected transactions for frequent pattern mining*".

Token α_1 enters the net via place L_1 with initial characteristics: "*transactions with weather data*".

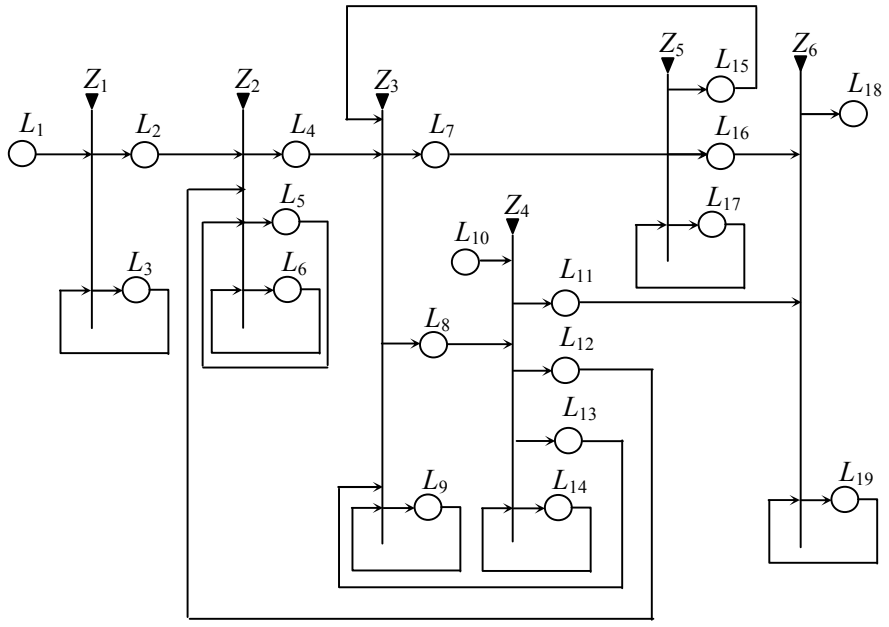

Figure 1. Generalized net of the process
of generating association rules by Eclat algorithm

The transition Z_1 has the form:

$$Z_1 = \langle \{L_1, L_3\}, \{L_2, L_3\}, R_1, \vee(L_1, L_3) \rangle,$$

where

$$R_1 = \begin{array}{c|cc} & L_2 & L_3 \\ \hline L_1 & false & true \\ L_3 & W_{3,2} & W_{3,3} \end{array},$$

and:

- $W_{3,2}$ ="There are selected transactions";
- $W_{3,3} = \neg W_{2,3}$.

The token, entering in place L_3 from place L_1 don't obtain new characteristic. The token in place L_3 generates new one, that enters in place L_2 with characteristic: *"selected transactions with weather data for frequent pattern mining"*.

The transition Z_2 has the form:

$$Z_2 = \langle \{L_2, L_{12}, L_5, L_6\}, \{L_4, L_5, L_6\}, R_2, \vee(L_2, L_{12}, L_5, L_6) \rangle,$$

where

$$R_2 = \begin{array}{c|ccc} & L_4 & L_5 & L_6 \\ \hline L_2 & false & false & true \\ L_{12} & false & false & true \\ L_5 & W_{5,4} & W_{5,5} & false \\ L_6 & false & W_{6,5} & W_{6,6} \end{array},$$

and:

- $W_{5,4}$ ="minimum support of itemsets from input transactions is counted";
- $W_{5,5} = \neg W_{5,4}$;
- $W_{6,5}$ ="the transactions with weather data are transformed on vertical data layout";
- $W_{6,6} = \neg W_{6,5}$.

The token, entering from place L_2 in place L_6 don't obtain new characteristic. The token in place L_6 generates new one that enters in place L_5 with characteristic: *"transformed transactions in vertical data layout"*.

At the second activation of the transition the token from place L_5 generates new token that enters in place L_4 with characteristic: *"itemsets with counted support"*.

The transition Z_3 has the form:

$$Z_3 = \langle \{L_{15}, L_4, L_{13}, L_9\}, \{L_7, L_8, L_9\}, R_3, \vee(L_{15}, L_4, L_{13}, L_9) \rangle,$$

where

$$R_3 = \begin{array}{c|ccc} & L_7 & L_8 & L_9 \\ \hline L_{15} & false & false & true \\ L_4 & false & false & true \\ L_{13} & false & false & true \\ L_9 & W_{9,7} & W_{9,8} & W_{9,9} \end{array},$$

and

- $W_{9,7}$ ="there are sorted itemsets for finding frequent subsets";
- $W_{9,8}$ ="there are sorted itemsets for finding frequent itemsets";
- $W_{9,9} = \neg(W_{9,7} \wedge W_{9,8})$

The tokens entering in place L_9 (from L_4, L_{13} and L_{15}) do not obtain new characteristics. The token in place L_9 generates two new tokens that enter in places L_7 and L_8 with characteristics respectively: *"sorted subsets"* in place L_7, and *"sorted itemsets"* in place L_8.

The token enters the net via place L_{10} and has initial characteristic: *"minimum support"*.

The transition Z_4 has the form:

$$Z_4 = \langle \{L_{10}, L_8, L_{14}\}, \{L_{11}, L_{12}, L_{13}, L_{14}\}, R_4, \vee(\wedge(L_{10}, L_8), L_{14}) \rangle,$$

where

$$R_4 = \begin{array}{c|cccc} & L_{11} & L_{12} & L_{13} & L_{14} \\ \hline L_{10} & false & fakse & false & true \\ L_8 & false & false & false & true \\ L_{14} & W_{14,11} & W_{14,12} & W_{14,13} & W_{14,14} \end{array},$$

and

- $W_{14,11}$ ="there are frequent itemsets for generating association rules";
- $W_{14,12}$ ="there are no frequent itemsets";
- $W_{14,13}$ ="the frequent itemsets need to be sorted";
- $W_{14,14}= \neg(W_{14,11} \wedge W_{14,12} \wedge W_{14,13})$.

The tokens entering in place L_{14} (from L_{10} and L_8) do not obtain new characteristics. The token in place L_{14} generates three new tokens that enter in places L_{11}, L_{12} and L_{13} with characteristics respectively:

"frequent itemsets for generating association rules" in place L_{11},

"not frequent itemsets" in place L_{12}, and

"frequent itemsets for sorting" in place L_{13}.

The transition Z_5 has the form:

$$Z_5 = \langle \{L_7, L_{17}\}, \{L_{15}, L_{16}, L_{17}\}, R_5, \vee(L_7, L_{17}) \rangle,$$

where

$$R_5 = \begin{array}{c|ccc} & L_{15} & L_{16} & L_{17} \\ \hline L_7 & false & false & true \\ L_{17} & W_{17,15} & W_{17,16} & W_{17,17} \end{array},$$

and

- $W_{17,15}$ ="the frequent subsets need to be sorted";
- $W_{17,16}$ ="there are frequent subsets for generating association rules";
- $W_{17,17}= \neg(W_{17,15} \wedge W_{17,16})$.

The token entering in place L_{17} (from L_7) do not obtain new characteristic. The token in place L_{17} generating two new tokens that enter in places L_{15} and L_{16} with characteristics respectively:

"*frequent subsets for sorting*" in place L_{15},
and "*frequent subsets for generating association rules*" in place L_{16}.

The transition Z_6 has the form:

$$Z_6 = \langle \{L_{16}, L_{11}, L_{19}\}, \{L_{18}, L_{19}\}, R_6, \vee(L_{16}, L_{11}, L_{19}) \rangle,$$

where

$$R_6 = \begin{array}{c|cc} & L_{18} & L_{19} \\ \hline L_{16} & false & true \\ L_{11} & false & true \\ L_{19} & W_{19,18} & W_{19,19} \end{array},$$

and:

- $W_{19,18}$ = "association rule is created";
- $W_{19,19}= \neg W_{19,18}$.

3 Realization of the algorithm

Frequent pattern mining by Eclat algorithm is presented in the following example. For datasets exploration is used statistical language *R*. The library "arules" needs to be installed. The weather datasets is stored in "weather.csv" file. It has attributes-wind (calm, breeze, gale), temperature (cool, mild, hot), outlook (sunny, overcast, rainy), humidity (normal, hide) and fire (yes, no). Part of the weather data is described on Fig. 2. The records are visualized in the form of transactions. The most important step is to write the minimum support. In the example there are 84 transactions and minsup of 0.3 (30%) defined by the user. The steps of the process for frequent pattern mining by Eclat algorithm in *R* is presented in Fig. 3. The analysis for frequent pattern mining describes the possibility for fire depending of the weather.

	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

Figure 2. Part of the weather data

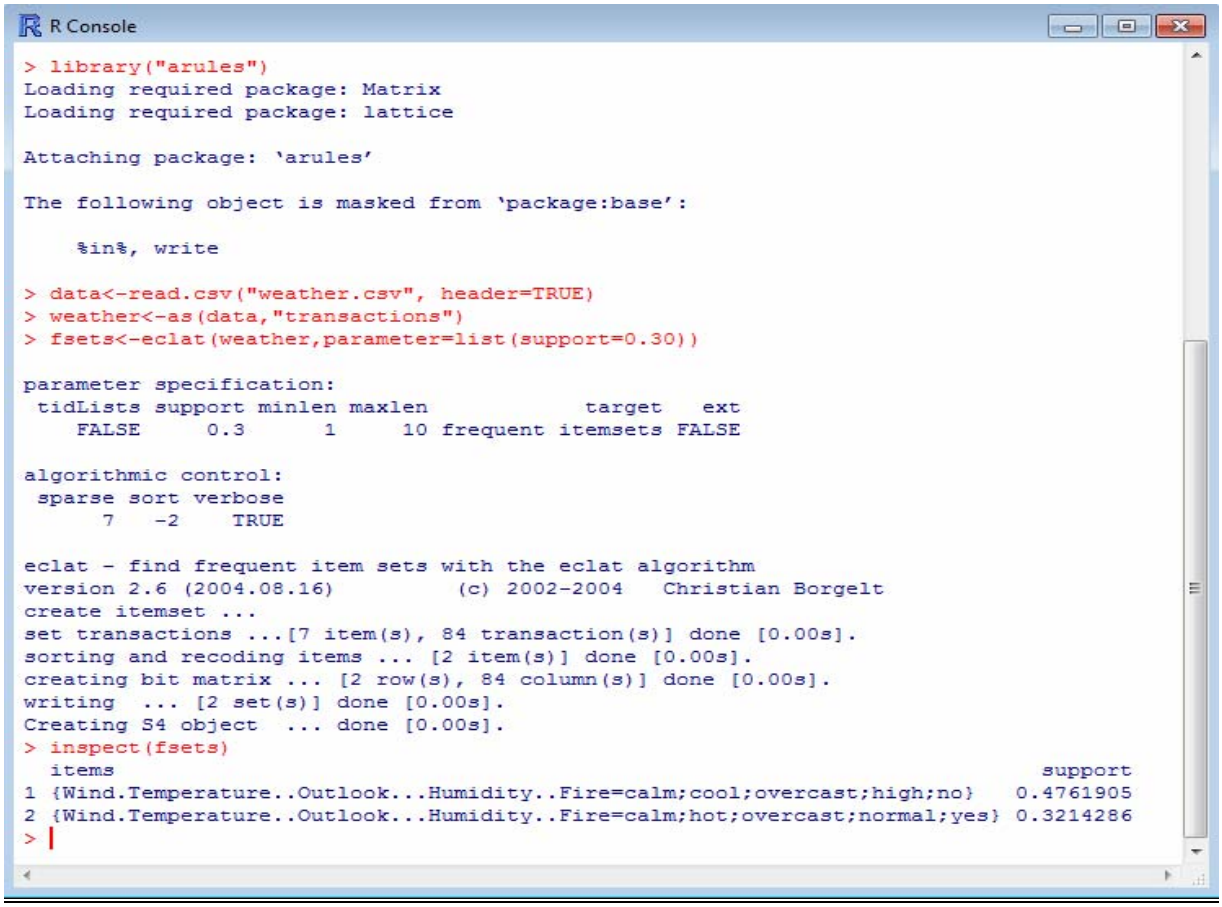

Figure 3. Loading the weather data in *R* and giving the minimum support threshold of 0.3 for extracting frequent patterns by Eclat

The results have a form:

```
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)
```

The result gives interesting rules (Fig. 3). It is preferred rules which count is smaller. If the analysis gives many rules after the exploration there are lost their interestingness. The minimum support threshold is greater at the same time. For example then the user is written minsup of 0.01 it is received seven rules that are shown on Fig. 4.

```
> fsets<-eclat(weather,parameter=list(support=0.01))

parameter specification:
 tidLists support minlen maxlen            target   ext
    FALSE    0.01      1     10 frequent itemsets FALSE

algorithmic control:
 sparse sort verbose
      7   -2    TRUE

Warning in eclat(weather, parameter = list(support = 0.01)) :
  You chose a very low absolute support count of 0. You might run out of memory! Increase minimum support.

eclat - find frequent item sets with the eclat algorithm
version 2.6 (2004.08.16)         (c) 2002-2004   Christian Borgelt
create itemset ...
set transactions ...[7 item(s), 84 transaction(s)] done [0.00s].
sorting and recoding items ... [7 item(s)] done [0.00s].
creating bit matrix ... [7 row(s), 84 column(s)] done [0.00s].
writing  ... [7 set(s)] done [0.00s].
Creating S4 object  ... done [0.00s].
> inspect(rules)
Error in inspect(rules) :
  error in evaluating the argument 'x' in selecting a method for function 'inspect': Error: object 'rules' not found
> inspect(fsets)
  items                                                                  support
1 {Wind.Temperature..Outlook...Humidity..Fire=calm;cool;overcast;high;no}     0.47619048
2 {Wind.Temperature..Outlook...Humidity..Fire=calm;hot;overcast;normal;yes}   0.32142857
3 {Wind.Temperature..Outlook...Humidity..Fire=calm;hot;overcast;normal; yes}  0.08333333
4 {Wind.Temperature..Outlook...Humidity..Fire=calm;mild;rainy;high;no}        0.04761905
5 {Wind.Temperature..Outlook...Humidity..Fire=gale;cool;rainy;high;no}        0.03571429
6 {Wind.Temperature..Outlook...Humidity..Fire=breeze;hot;sunny;normal;yes}    0.02380952
7 {Wind.Temperature..Outlook...Humidity..Fire=calm;mild;overcast;normal; yes} 0.01190476
>
```

Figure 4. Giving the minimum support threshold of 0.01 for extracting frequent patterns by Eclat (with smaller interestingness)

The possibility for predicting the fire with many rules and smaller minimum support threshold is decreased. In conclusion, the user will prefer the associations with greater interestingness (minsup=0.3).

4 Conclusion

The constructed Generated net model can be used for determination of the possibility of forest fire by association rules depending on metrological conditions. For modelling we use frequent pattern mining by Eclat algorithm. To explore the weather conditions (wind, temperature, outlook, humidity and fire) we use statistical language *R*.

Acknowledgements

The authors are grateful for the support provided by the project DFNI-I-01/0006 “Simulating the behaviour of forest and field fires”, funded by the National Science Fund, Bulgarian Ministry of Education, Youth and Science.

References

[1] Agrawal, R., Imielinski T., And Swami A., Mining Association Rules Between Sets Of Items In Large Databases, in *Proceedings of ACM-SIGMOD Conference*, Washington, DC, 1993

[2] Atanassov, K. *Generalized Nets*. World Scientific, Singapore, 1991.

[3] Atanassov, K. *On Generalized Nets Theory*. Prof. M. Drinov Academic Publishing House, Sofia, 2007.

[4] Bureva, V. Methods for extracting patterns from databases, *Management and Education*, University "Prof. Asen Zlatarov", Burgas, Vol. 8 (4), 2012, 255–258 (in Bulgarian).

[5] Bureva, V. Algorithms for associative rule mining, *Management and Education*, University "Prof. Asen Zlatarov", Burgas, Vol. 9 (6) 2013, 121–128 (in Bulgarian).

[6] Bureva, V. Generalized model of the process of the creating the association rules using Apriori algorithm, *Annual of “Informatics” Section Union of Scientists in Bulgaria*, Vo. 5, 2012, 73–83 (in Bulgarian).

[7] Bureva, V. Generalized model of the process of the creating the association rules using Frequent Pattern-Growth Method, *Annual of “Informatics” Section Union of Scientists in Bulgaria*, 2013 (in bulgarian, in press).

[8] Ghosh, S., Nag, A., Biswas, D., Singh, J.P., Biswas, S., Sarkar, D., Sarkar, P.P., Weather Data Mining using Artificial Neural Network, *Recent Advances in Intelligent Computational Systems (RAICS)*, IEEE, 2011, 192–195.

[9] Kaur, G., Meteorological Data Mining Techniques: A Survey, *International Journal of Emerging Technology and Advanced Engineering*, Volume 2, Issue 8, August 2012, 325–327. http://www.ijetae.com/files/Volume2Issue8/IJETAE_0812_56.pdf

[10] Larose, D., *Discovering Knowledge In Data. An Introduction To Data Mining*, John Wiley & Sons, 2005.

[11] Tan, P., M. Steinbach, V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2006.

[12] Kalyankar, M., S. Alaspurkar, Data Mining Technique to Analyse the Metrological Data, *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 2, February 2013, 114–118. http://www.ijarcsse.com