Extracting Knowledge in Data Warehouses using Fuzzy AprioriTid

Somaieh Goudarzvand
Islamic Azad University North Tehran Branch

Ali Harounabadi
Islamic Azad University Central Tehran Branch

Mohammad Mansour Riahi Kashani
Islamic Azad University North Tehran Branch

ABSTRACT
Multidimensional databases and OLAP tools that provide an efficient framework for data mining have been pushing us to the OLAM architecture. OLAP is widely used to illustrate meaningful and interactive analysis of data on the complex structure. In contrast, detecting hidden patterns in the data and exploring them is for the data mining. OLAP and data mining are believed to complete each other for analyzing large data sets in decision support systems efficiently. Unlike previous work in this field, this method does not rely on the availability of knowledge in a particular field. Variables will be selected with the consideration of user to build cubes. Hierarchical clustering is used to obtain dynamic relationships between variables at different levels of data. Results of the Adult data set shows that the obtained Lift from Fuzzy AprioriTid compared with Apriori algorithm increased.

Keywords
Data warehouses, Extracting knowledge, Fuzzy AprioriTid

1. INTRODUCTION
The goal of Knowledge Discovery is mining new patterns, potentially useful, valid and understandable. Recently, a combination of OLAP and data mining because of its importance in knowledge discovery of large cube has been used in research. Several methods have been investigated in the literature that addresses the issue of extracting Rules from multidimensional database [9] [10]. Although, in the mentioned works, there has been done a lot to ease the process of extracting knowledge, some issues remained unresolved. In addition to being time-consuming and hard work, analysis of the problem can be studied from another aspect as well. It is that analysts do not have sufficient mastery of the domain so it is possible some useful rules be ignored by them because of lack of their information. These deficiencies push us to extract informative rules automatically.

Extracting knowledge from high volume data can be also an incredible task. In this case, using an algorithm like Apriori should be suggested. Apriori Algorithm predicts a lot of rules so that it is necessary to decrease number of rules by grouping nominal or numerical data. In in [10] the test has been done on nominal data and has used Multidimensional Scaling theory, however, this research is discussed on the numeric data and focus on triangular fuzzy membership.

Approach in this research have discussed in the following. Firstly, an agglomerative hierarchical clustering is used to obtain data abstraction at different levels of data, so that a dendrogram is created. Each level in that dendrogram contains a set of child clusters that were split off a parent. Each cluster is contained with numeric and nominal variables which Nominal variables are utilized as dimension table and numeric variables used as fact table. In fact, cube can be constructed in any level and every cluster. Useful rules are then extracted from the cubes in different clusters using Weighted Fuzzy AprioriTid and Fuzzy AprioriTid. Result has been compared with Apriori and illustrates that using Fuzzy for extracting rules lead to a higher lift in comparison with Apriori. Another test on the diversity of the rules shows that diversity in both Fuzzy AprioriTid and Apriori algorithm has not changed considerably.

Thus, the article follows: in the second part techniques used in the methodology are discussed. In the third section, the literature is reviewed in the relevant area. In the fourth section, the proposed methodology is presented. Then, in Section V, evaluation of the case studies is done, at the end of the paper conclusions and some discussions for future research are discussed.

2. BACKGROUND
2.1 Agglomerative Hierarchical Algorithm
The preference for Agglomerative Hierarchical Clustering is that it offers natural hierarchical clustering algorithms more than others.

2.2 Creating A Data Cube Containing Useful Information
Data cube is built from dimension and fact tables. An important issue is what dimension and fact tables can be apply to construct cubes. In this step, user can select the highly rated variables for making cube. User also can create various combinations of them such as highly ranked numeric variable with less ranked nominal variable.

2.3 Discovering Association Rules Using Fuzzy Aprioritid
Most old data mining algorithms were used nominal attribute to extract rules from databases. In this study the Rules that lie between numerical variables by Fuzzy AprioriTid approach is obtained. The algorithm focuses on the linguistic expression. The role of fuzzy sets is help to transmission of numerical data to nominal data, thus reduced the set of items in the search process. To compare the obtained results, the Diversity Measure and Lift are used.

$$Lift(A->B)=Confidence/Sup(B) \quad (1)$$

This is amount of independence between objects A and B that can be between 0 and infinity. Values close to 1 show A and B are independent of each other and hence do not show interesting Rules. If this value is less than 1 indicates that A and B are in negative relation and the value is greater than 1 indicates that A provides more information about B. Another studied measure is diversity of rules [11]. The importance of diversity is that if the triggering items in the antecedent rule for any given consequent rule be more than other rules, that rule is valuable in terms of diversity.

$$\text{Rae}= \sum_{i=1}^{m} \frac{ni(ni-1)}{N(N-1)} \quad (2)$$

Here m is the total number of rows in the summary table; ni the number of characters; N is the number of attributes.

3. REVIEW OF LITERATURE

Place Tables/Figures/Images in text as close to the reference

In the literature, methods of development of data warehouse and then extracting knowledge divided into four categories:

(1) Methods that rely on OLTP data model to create schema [2] [4] [5].

(2) Methods that decision makers have raised a series of analytical requirements to construct schema [7].

(3) Extracting knowledge from schema [3] [10].

In the first case, there are two advantages. That is, helping users to create schema from data model so it is guarantee that schema can be fed by the OLTP. In the second case it is assumed that the user has enough experience to express his analytic requirements. The second approach has also this advantage that all requirements of the users will be met, however, this can be considered as a negative point as well because some significant points may be neglected by them.

3.1 Constructing Schema Regarding Oltp Model

In [4] an automated tool using a relationship – Entity diagram is introduced. An example for this system is SAMSTAR. In [6] an approach is proposed to design data model that have not any dependency to a specific domain. In fact, in this approach the algorithm relies on the structural features of data source. It uses a couple of rules to extract concepts such as dimensions and facts from object-oriented databases and then creates a star schema.

3.2 Constructing Schema Regarding User's Need

In [7] the user asks a series of assumptions and a data mining system tries to find patterns that are consistent with the assumptions stated. The advantage of this approach is to discover rules that are consistent with the needs of the user. Although it would be another constraint so that useful numerous patterns may be removed. The reason is that user's awareness of the scope of domain is low and this makes a lot of patterns overlooked.

3.3 Extracting Knowledge From Schema

In [3] a method is provided to explore multi-dimensional data. In this way, the authors have proposed four different algorithms. First, pre-processing are done on the data. Preprocessing the data delete all the data in the fact table that are less than a predefined threshold value. Then the algorithms use this pre-processing table for the discovery of rules. The results of performance of these algorithms show that using pruning method on the extracted rules is appropriate. But no metrics for measuring diversity has been presented. A limitation of this method is to delete the rows of the fact table, uses just one threshold value for each row of table. In [1] proposed another approach to use values such as min, max, avg and sum to delete the data. Lift and Lovinger also been used to measure the popularity of the Rules. Knowledge discovery in multidimensional databases with fuzzy approach is discussed in [8].

4. PROCEDURE

In this section an overview is done. First, Agglomerative Hierarchical clustering algorithm is used to create different levels of abstraction of the data. This step is done in Orange software. Cube is made from nominal and numeric variables with the selection of user. Informative rules are then extracted using Fuzzy AprioriTid and Weighted Fuzzy AprioriTid. Figure 1 shows the process of methodology for extracting knowledge.

The rest of the methodology for extracting knowledge using Fuzzy AprioriTid and weighted Fuzzy AprioriTid has done in software that is proposed in [12].

4.1 Agglomerative Hierarchical Algorithm

In the first stage, Agglomerative hierarchical algorithms run on a given data set in orange software. Each level consists of a set of clusters in the generated . Cutoff point is done by user. Figure 2 shows the cutoff points, number of generated clusters. Here, cut off point is set to 3.24.

Dendrogram. Cutoff point is done by user. Figure 2 shows the cutoff points, number of generated clusters. Here, cut off point is set to 3.24.

4.2 Creating A Data Cube Containing Useful Information

At this stage, data cube is produced. User can limit the search space by selecting dimensions or facts with highest rank, or by the user's chosen facts. Each user has particular analytical requirements. So, selection of dimensions and facts can be affected by the specific knowledge of user.

4.3 Extracting Rules By Fuzzy Aprioritid

Finally from the obtained cube in the previous step find the hiding rules in data by Fuzzy AprioriTid and weighted Fuzzy AprioriTid algorithms. The previous work in this issue has done on the nominal data and grouping them. In that research, grouping nominal data had led to more diversity and accuracy. The importance of numerical data in data warehouses and large data bases motivated us to do that on numerical data with the use of Fuzzy method.

5. CONCLUSION

To evaluate the above mentioned approach, Adult data set is used that have been extracted from UCI Machine Learning sites. Because of the limitation, testing was performed only on 30% of the data. For data clustering Orange software is used. For generation of fuzzy rules the software in (Coenen, 2008) has been used. The test has benefited from three algorithms for finding Association rules, and to compare the results. Sample cluster data given to the Weighted Fuzzy AprioriTid and Fuzzy AprioriTid and also Apriori algorithms. It is important to compare the difference between the creation of fuzzy sets or without fuzzy as association rules. Triangular fuzzy membership function is used in this research and fuzzy areas intended for the variables of age, time and education.

These tests have been used for two clusters. In this section, the rules resulting from the Fuzzy AprioriTid and Association Rule and Weighted Fuzzy AprioriTid algorithms in cluster 1 are compared. The amounts of lift in Fuzzy AprioriTid and also Weighted Fuzzy AprioriTid are much higher than Apriori that haven't used Fuzzy. In addition to that, the diversity of rules for two clusters is almost as the same and is not changed considerably.

As shown in Table (5) any change in diversity cannot be seen.

Here's a topic that is controversial that is despite the fact that these three algorithms are similar in diversity, the Lift Obtained from Fuzzy Algorithm are greater obviously. In general, the Lift of fuzzy algorithm is more than conventional algorithms used for extracting rules.

6. FUTURE WORK

In this paper, a method for extraction of knowledge from data is provided. With using that smaller numbers of variables is available. Rules extracted through Weighted Fuzzy AprioriTid and Fuzzy AprioriTid Compared to Association Rule have much higher lift measure, however, the amount of diversity has not changed. In this way, by the algorithm that is introduced, valuable rules are extracted at different levels of data. One thing for further research is that as in data warehouses a large volume of data is existed, finding the suitable range for using in linguistic variables can become a complicated issue. What should be done is to select the intervals without user or automatically done.

7. REFERENCES

[1] Ben Messaoud Riadh, Boussaid Omar, Loudcher Rabaséda Sabine, Missaoui Rokia. (2006). "Enhanced mining of association rules from data cubes," *In DOLAP '06 proceedings of the 9th ACM international workshop on data warehousing and OLAP* , . ACM, pp. 11–18.

[2] Carme,Jose-Norberto Mazon Andrea, Rizzi Stefano, "A model-driven heuristic approach for detecting multidimensional facts in relational data sources," 12th *Int. Conf. Data Warehousing and Knowledge Discovery*, Bilbao, Spain, pp. 13–24.

[3] Cokrowijoyo Tjioe Haorianto, Taniar David. (2005). "Mining association rules in data warehouses," *International Journal of Data Warehousing and Mining*, 1, 28–62.

[4] Cabibbo Lucca, Torlone Riccardo. (1998) "A logical approach to multidimensional databases," *Conf. Extended Database Technology*, Valencia, Spain , pp. 187–197.

[5] Feki Jamel, Hachachi Yasser.(2007) "Assisted data mart design: A method and a toolset," *Journal of Decision Systems* (in French), 16(3) , 303–333.

[6] Hachachi Yasser, Feki Jamel. (2013). "An automaticmethod for the design of multidimensional schemasfrom object oriented databases,".*InternationalJournal of Information Technology & Decision Making*. Vol. 12, No. 6 1223–1259.

[7] KayaMehmet,Alhajj Reda .(2005)."FuzzyOLAPassociation rules mining-based modular reinforcementlearning approach for multiagent systems," *Part B: Cybernetics, IEEE Transactions*, 35, 326–338.

[8] Khare Neelu, Adlakha Neeru, Pardasani K. R..(2009) "An Algorithm for Mining Multidimensional Fuzzy Assoiation Rules," *(IJCSIS) International Journal of Computer Science and Information Security, Vol. 5, No. 1, 2009* .

[9] M. Chung Soon & Mangamuri Murali (2005). "Miningassociation rules from the star schema on a parallelNCR teradata database system,". *In International conference on information technology: coding and computing (ITCC'05)*. Nevada.

[10] Usman Muhammad, Pears Russel (2013)." Discovering diverse association rules from multidimensional schema." *Expert Systems with Applications*. 40 , 5975–5996.

[11] Zbidi Naim , Faiz Sami & Limam Mohamed. (2006)." On mining summaries by objective measures of interestingness,". *Machine Learning*, 62, 175–198.

[12] Coenen, F. (2008), *The LUCS-KDD Weighted Fuzzy Apriori-TSoftware*, http://www.csc.liv.ac.uk/~frans/KDD/Software/WFaprio ri_TFP/weightedFuzzyAprioriTidFP.html, Department of Computer Science, The University of Liverpool, UK.

8. APPENDIX

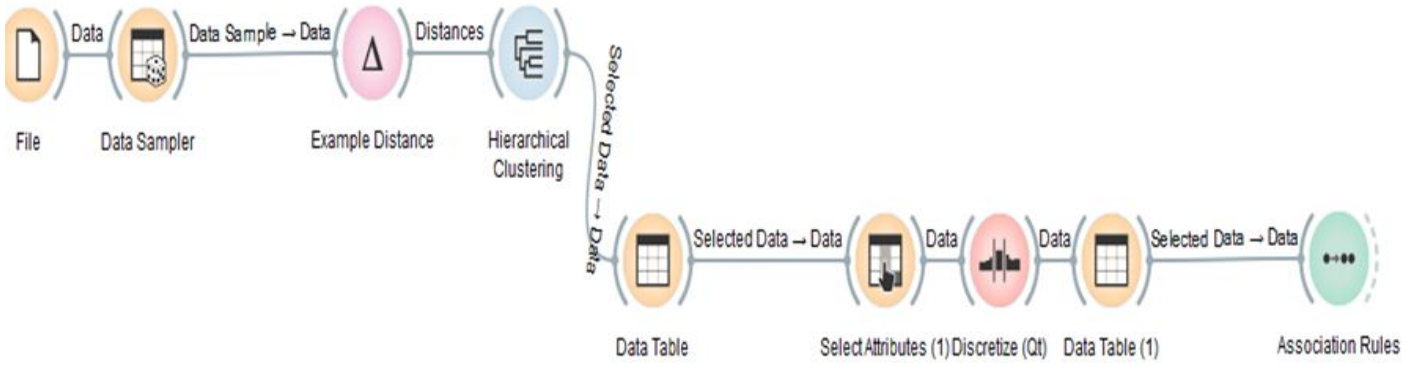

Fig 1: process of extracting knowledge using association Rule in orange software

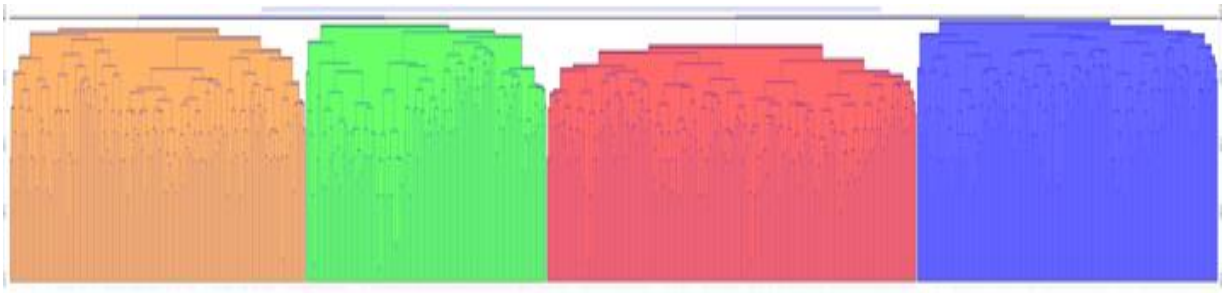

Figure 2: Clusters and Cut-off point

Table 1: Fuzzy Regions

Fuzzy Region	HourPerWeek	FuzzyRegion	Education	FuzzyRegion	Age
0-35	LOW	0-10	LOW	12-45	Young
25-65	Average	8-14	Average	35-65	MiddleAged
50-100	High	12-16	High	55-100	Old

Table 2: Fuzzy AprioriTid, Cluster1

Fuzzy AprioriTid , Cluster1 , Min Support=0.001 , Min Confidence=20 , 3516 data		
No	Rules	Lift
R1	Education=Low , Age=MiddleAged→Average Hour	29.874
R2	Education=Low , Age=Young → Average Hour	28.854
R3	Education=Average , Age=Young → Average Hour	28.563
R4	Education=Average , Age=MiddleAged → Average Hour	28.371
R5	Age=MiddleAged → Average Hour	27.976
R6	Age=Young → Average Hour	27.918
R7	Education=Average → Average Hour	27.669
R8	Education=Low → Average Hour	27.164
R9	Education=High , Age=MiddleAged→Average Hour	26.868
R10	Education=High→ Average Hour	25.318
Average Lift		**27.857**

Table 3: Weighted Fuzzy AprioriTid, Cluster 1

Weighted Fuzzy AprioriT, Cluster C1, Min Support=0.01,MinConfidence=10, 3516		
No	Rules	Lift
R1	Education=Low , Age=MiddleAged→Average Hour	6.204
R2	Education=Low , Age=Young → Average Hour	5.992
R3	Education=Average , Age=Young → Average Hour	5.932
R4	Education=Average , Age=MiddleAged → Average Hour	5.892
R5	Age=MiddleAged → Average Hour	5.812
R6	Age=Young → Average Hour	5.798
R7	Education=Average → Average Hour	5.745
R8	Education=Low → Average Hour	5.642
R9	Education=High , Age=MiddleAged→Average Hour	5.580
R10	Education=Low , Age=MiddleAged→ Average Hour	5.258
Average Lift		**5.785**

Table 4: Association Rule, Cluster 1

Association Rule , Cluster1 , Min Confidence=20, 3516 data		
No	Rules	Lift
R1	Age(37.5,48.5]→hour(39.5,40.5]	1.030
R2	Age(37.5,48.5], Education<=9.5→hour(39.5,40.5]	1.149
R3	Age(37.5,48.5],Education(9.5,11.5]→hour(39.5,40.5]	1.093
R4	Age(37.5,48.5],Education>11.5→hour(39.5,40.5]	0.782
R5	Education>11.5→hour(39.5,40.5]	0.734
R6	Age<=37.5, Education>11.5→hour(39.5,40.5]	0.649
R7	Age>48.5, Education>11.5→hour(39.5,40.5]	0.746
R8	Age<=37.5, Education>9.5→hour(39.5,40.5]	1.126
R9	Age>48.5, Education<=9.5→hour(39.5,40.5]	1.130
R10	Education<=9.5→hour(39.5,40.5]	1.103
Average Lift		**0.954**

Table5: Diversity Measure, Cluster1 and Cluster 2

ClusterLabel	**Rule Set**	**WeightedFuzzyApriori**	**FuzzyApriori**	**AssociationRule**
C1	R1-R10	0.05	0.05	0.051
C2	R1-R10	0.053	0.053	0.051