Data-Mining by Probability-Based Patterns

M. Karegar[1], A. Isazadeh[2], F. Fartash[3], T. Saderi[4], A. Habibizad Navin [5]
[1, 4, 5] *Department of Computer Engineering Islamic Azad University of Tabriz*
[2]*Department of Computer Science, Tabriz University*
[3]*Department of Information Technology Engineering, International University of Chabahar*
{kargar,saderi,ah_habibi}@iaut.ac.ir, isazadeh@tabrizu.ac.ir, fartash@iuc.ac.ir

Abstract. *In this paper a new method is suggested for designing patterns in data-mining. These patterns are designed using probability rules in decision trees and are cared to be valid, novel, useful and understandable. By using the suggested patterns in data-mining, the system gets efficient information about the data stored in its data-bases and uses them in the best planning for special objectives.*

Keywords. Data-mining, Pattern, Relation, Relationship, Graph, Tree, Risk Management.

1. Introduction

By technology development and highly affect of computer usage in people's daily life, the way of gathering and saving information has differed a lot. Deep change in storing and retrieving data - raw information - from paper based to computer based has brought up a condition needing new techniques for data management necessarily. Increasing need for gathering information in different social systems and on the other hand, decreasing time periods from asking to get a reply, made computer engineers think about ways for speed process on stored data and get useful information as soon as possible. As it is said, "necessity is the mother of invention," from late 1980s data mining has been developed and become warmly accepted amongst software engineers and database designers [1, 2].

Representing local features of data in the forms called patterns is the main task of data mining [3]. In this paper, a new pattern in data mining is represented by using probability-trees to have a process on different type of data stored in databases and give familiar information to system users.

1.1. Terminology

Data mining is a popular technique in searching for interesting and unusual patterns in data, has also been enabled by the construction of data warehouses, and there are claims of enhanced sales through exploitation of patterns discovered in this way. In other words, data mining is extraction of interesting (non-trivial, implicit, previously unknown and unexpected and potentially useful) information or patterns from data sources, e.g. databases, texts, web, images, etc [4, 5, 6].

Data mining is not data warehousing, (deductive) query processing, SQL/reporting, software agents, expert systems, Online Analytical Processing (OLAP), statistical analysis tool, statistical programs or data visualization. Further definitions are available in [7, 8, 9, 10, 11].

Pattern is a local feature of the data, departure from general run of data, a group of records that always score the same on some variables, a pair of variables with very high correlation and unusual combination of products purchased together. Patterns must be valid, novel, potentially useful and understandable; validity is holding on new data with some certainty; novelty is to be non-obvious to the system; usefulness is to be possible to act on the item and understandability is being interpretable by humans [12].

In data-bases relation has the same meaning with table. In this paper relation is used instead of using the word table.

1.2. Problem

The data which are the result of system operation and the reflex of policies and strategies of the stakeholders are stored in data-bases during the system runtime. Some guesses and suppositions may be represented for the success

or failure of the system; but the problem is that, the weak and strong points of that guesses and suppositions can not be recognized so easily. What's more, there is no reassurance for their perfection.

1.3. Solution

Designing patterns and relating each of them by numbers, help us select our required queries from patterns so easily. New attributes turning up from the data-bases presents the fact that, the prime guesses were unable to recognize them. In this paper, Intra-nodes play this role. The weak and strong points of each cases are pointed by the weights related to the edges of the trees. While coming to the conclusion that some attributes would have affect on the pattern or they might be so important in some conditions or sometimes, we may change its range of affection by changing the amount of c_1 coefficient (described in subsection 3.4), in designing the pattern from the graph.

1.4. The claim

We claim that such a solution does exist. By drawing a graph, computing the values and priorities of each part, performing them in a decision tree and concluding a pattern from it, we reach to a point that we can do our deductive data-mining so clearly. The information we gather from the patterns will help stakeholders manage the system easily, with least amount of risk or errors.

1.5. Objective

As described above, expansion of data-mining is growing rapidly due to the large amount of data in data-bases and starvation for information in other parts. The objectives we want to achieve by suggesting a new method for designing patterns is using probability in data-mining to gain a simple, understandable and novel pattern. This pattern can give numerical information about the attributes in data-bases and help managers of a system in risk management and correct guidance.

1.6. Paper outline

This paper is organized in four sections. In Section 1 we started the problem, an idea solution to the problem and our claim regarding the solution. In Section 2 we present a brief survey of the previous work, their strength and weaknesses, and the areas, requiring further improvements. In Section 3 we present our work under the title of Probability-Based Patterns and describe how the job can be done. Finally we conclude the paper in Section 4 with an evaluation of our work and some future topics of research.

2. Previous work

Although the pattern suggested in this paper is new and there is no previous work done on it, but here some information in given about a number of previous models presented in articles.

2.1. Decision tree

Many data mining methods generate decision trees—trees whose leaves are classifications and whose branches are conjunctions of features that lead to those classifications. One way to learn a decision tree is to split the example set into subsets based on some attribute value test. The process then repeats recursively on the subsets, with each splitter value becoming a sub-tree root. Splitting stops when a subset gets so small that further splitting is superfluous or a subset contains examples with only one classification.

A good split decreases the percentage of different classifications in a subset, ensuring that subsequent learning generates smaller sub-trees by requiring less further splitting sorting out the subsets. Various schemes for finding good splits exist [13]. For further information about decision trees refer to [14].

2.2. Declarative networking

In Declarative Networking, the data and query model proposed for declarative networking are introduced. The language presented is Network Datalog (NDlog), a restricted variant of traditional Datalog intended to be computed in distributed fashion on physical network graphs. In describing the model, the NDlog query which performed distributed computation of shortest paths was used. In that model, one of the novelties of our setting, from a database perspective, is that data is distributed and relations may be partitioned across sites. To ease the generation of efficient query plans in such a system, NDlog gives the query writer explicit

control on data placement and movement. Specifically, NDlog uses a special data type and address, to specify a network location. For further information refer to [15].

2.3. Databases and logic

Logic can be used as a data model in databases. In frame of such a model we can distinguish:

Data Structures, Operators and Integrity Rules.

All these three elements are represented in the same unique way as axioms in the logic language. As a deductive database system we can treat such a system, which has the ability to define deductive rules and can deduce or infer additional information from the facts stored in a database. Deductive rules are often referred to as logic databases. A deductive database (DDB) can be defined as:

DDB= {I, R, F}, where: F – a fact set, R – a deductive rule set, I – an integrity constraints. For further information refer to [16].

In [17] a data-mining is surveyed in an artificial intelligence perspective that presents a new and interesting view to readers. Avoiding prolongation it is not noted here.

3. Probability-based patterns

In the suggested method, for representing a pattern in data-mining and risk management, a new step is put forward. Using probability in the pattern makes it so sensible, understandable and easy to be performed. The description of the pattern is kept on by representing an example:

Data-bases are considered as relations with logic relationships between them. The objective is to concentrate on the special data and to reach to a point that there exists desired amount of them or it grows with a special norm increasingly or decreasingly. In some cases such as profit, revenue, sale amount and etc the increasing growth is requested and in other cases such as withdrawal, warranty time, absolute produce cost and etc the decreasing growth is requested.

Because of the relationship between the relations, fluctuations of the amounts of a field may influence the amounts of other fields. So, by designing an appropriate pattern from the data-base, fluctuations of the amounts of a field could be managed in a way that it goes up and down, in the manner we want.

In the supposed data-base some relations exist as below:

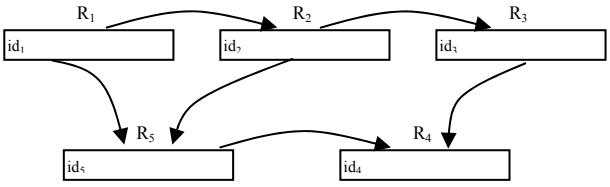

Figure 1.a. The subjective entity relationships

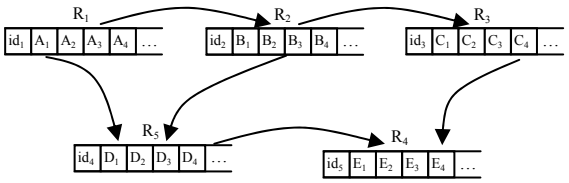

Figure 1.b. Relationships of the relations

The relation R_1 is supposed as the source and the relation R_4 is supposed as the destination. Other relations play the role of interfaces and are the connection terminals between the source and destination. However they can be supposed as independent destinations too.

As it is shown in figure 1, some relations are directly or indirectly related to R_1 and a change in their amounts can be affective in the change of the amounts of a field like A_1.

3.1. Presuppositions in designing a pattern

In order to design a pattern, these presuppositions should be considered:

1- The relations that are related to a relation like R1 directly or with less number of hops can have more affect on it. So the dependency amount of near relations is more than the far ones (from the perspective of hop amounts of relationships).

Although the presupposition 1 is a basic rule in designing the pattern, but it never causes the far relations to be worthless. The exit of a far relation from the set of relations which are affective in the pattern depends on its dominant and related data.

2- The relation which includes more relationships is an important relation and should be set in higher priority in designing the pattern.

3- The R set includes dominant and affective attributes in the pattern. These attributes are selected through the prime and experienced guess which becomes perfect during pattern

designing period and possibly some attributes would be less important (not affective).

4- An attribute is considered as a head or a key attribute and pattern is designed according to it.

Objective Attribute/Vertex	Attribute 1	Attribute 2	…
-	Coefficient 1	Coefficient 2	…

Figure 2. The primarily pattern

As it is shown in figure 2, the pattern is the objective attribute or vertex of the attribute that the pattern is designed for. In this pattern, it is declared that by which attributes the objective attribute is affected. Now by trading off the attributes, a pattern would be suggested in which the compatibility of attributes 1 to n is considered. The following steps explain how the pattern is designed:

1- Firstly a graph is drawn based on the R set;

2- Then importance and priority of the nodes should be computed;

3- The decision tree is drawn based on the graph and the statistics gathered from the data-base, for designing the pattern;

4- Deduction of superior patterns is done from the graph.

3.1. Drawing the directed graph

1- In this graph, there is an edge from A_1 to B1 if A_1 and B_1 are members of R set and R_1 and R_2 are directly related. The direction of the edge is the direction of relation between R_1 and R_2.

2- There is a route from A1 to B_1 if A_1 and B_1 are members of R set and they don't have any direct relationship between each other. The nameless nodes are used at the route and the names of connections are noted on the edges. This job is done to name those nodes in production process.

3- The direction of the edge while drawing is from A_1 to B_1 if the common field in both, is the main key in A1 and a secondary key in B_1.

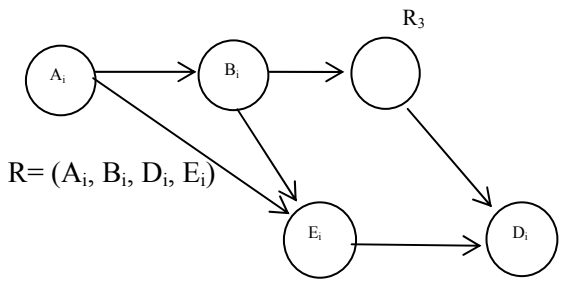

Figure 3. A sample designed graph

4- As there is no loop used in the relational model to cause the relational integrity, so the direction of the edges are exclusive.

5- Depending is more important than becoming dependant; therefore the input degree is less important than the output degree. The figure 3 is given as an example.

3.2. Giving name to the nameless vertexes

If there is a vertex that has no name, this means that no attribute or component of it has been used in designing the pattern. Therefore the related relation would be filtered according to the recognized major amounts or the father relations or the ones which there is a dependency to them. For example, R_2 is named as the father relation of R_3; according to the clustering in R_2 the selected data are the ones which are in connection with at least one of the clusters.

3.3. Computing the importance and priority

The formula below is used in computing the value of a node in which, Pr is the priority computed by the formula; c1 and c2 are two constants; d_i and d_o are respectively input and output degree; h is the hop number or the distance from the vertex A. While computing h the direction of edges is not considered.

$$Pr = c_1 * d_o + d_i + c_2(n-h)$$
$$n = h_{max}+1$$

The result of computing the priorities produces a list which includes the members of R and some that are not members. In that list A is the head and other members are tails. There, fields may be put together with the relations. Non of the fields from the intermediate relations are selected this time, but while designing tree from the existing data in the data-base and extracting operation, some fields of them might be chosen.

A_1: country of manufacture:
Iran, Malaysia, Singapore
B_1: TV type:
LCD, Plasma, CRT …
D_3: TV brand:
LG, Sony, Samsung

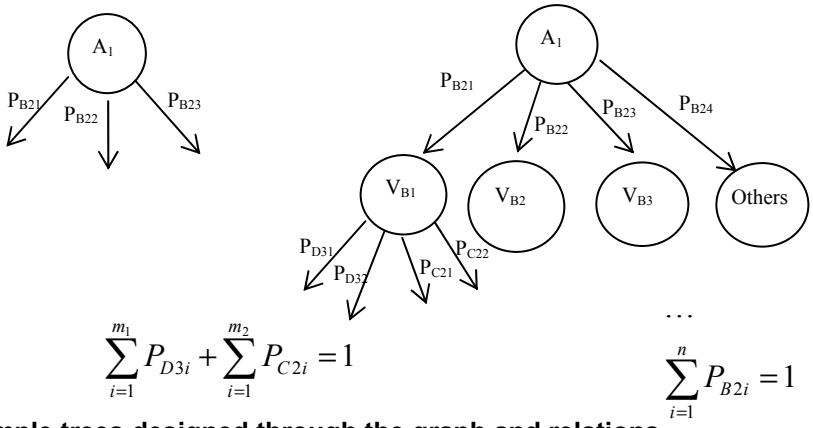

Figure 4. Sample trees designed through the graph and relations

3.4. Designing the tree

1- In the supposed directed graph, if there is an edge from A to B, A is called father and B is called child. This naming is done by other neighbors either. The superior attribute of the pattern may be in root or other vertexes.

2- The weight of the edge is recognized by the frequency number relating to that relationship.

3- If the intermediate vertex of a graph is a relation instead of a node, then the non-key attributes are in two kinds: scalar and non-scalar. The scalar amounts can be classified by superior intervals and the non-scalar ones by the superior amounts.

4- As the weight sum of each node is computed relatively, it should equal with 1. While adding a father or a child to the tree, we should take care that it didn't exist before till the probable loops in graph can't get way to the tree. This condition could be the last one to produce the tree.

Figure 4 shows a sample of designed tree through the graph and relations. It is supposed in the figure that dominant amount in R_3 gained by statistics is the attribute C_2 which is called the license.

As it is shown in figure 5, there is no loop at the graph so the infinity loop won't occur while producing the tree. Not mentioning the node E_i the node D_i would have no affect because it is not on the route from A_i; however if E_i is mentioned, then there would be three routes from A_i to C_i.

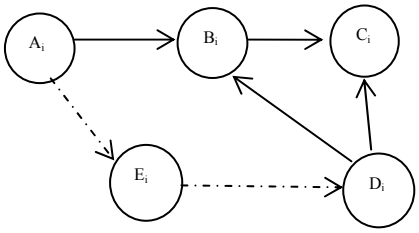

Figure 5. A sample designed graph

3.5. Designing the pattern

After producing the tree, if M_1 is named as the maximum degree of a tree and M_2 as the maximum number of the nodes in the pattern, then uttermost we will have M_1*M_2 trees. For each tree all the routes to the leaves are gathered in a list. Each list performs a route. For each list a number, resulted by the multiply of the edges of that route, will be considered as the value of the pattern and each list itself becomes a pattern.

Some of the superior patterns are selected from the patterns created by the variety of the trees, either considering the weight or the requested attributes. Selection of the patterns can be a new approach in the data-base survey. Subscription between the patterns can also become a new step for surveying till the operation of new pattern producing gets start with a new set of attributes. These attributes are the ones called R' and are gathered or produced from a subset from R and a subset from superior attributes in patterns other than R. This job will be affective in maintaining the heuristic (by guess) attributes of R_4.

Pattern1

A_1	B_{22}	D_{31}	…
1	1/6	1/2	…

$pv_1 = 1 * 1/6 * 1/2 * \ldots$

Pattern2

A_1	B_{22}	C_{31}	…
1	1/3	1/8	…

$pv_2 = 1*1/3*1/8* \ldots$

Figure 6. Sample patterns designed through the tree.

As it is shown above in figure 6, the value of the head A is always 1 in presupposition. It is better to order the data according to the coefficients.

3.6. Pattern Evaluation

This part consists two parts:

- Computing the weight of patterns,
- Maintaining the weight of patterns.

The weight computation is shown in figure 6 which is the product of weights. As the number of attributes may increase or decrease during pattern designing, the comparison of products would not be correct if the numbers of cases differ. So maintaining before data-mining is very important. In maintaining the a product like P_1, it is multiplied with a number which equals with the number of P_1 attributes divided to the product of numbers of all other attributes of patterns. The result will be the maintained value of pattern 1 which can be compared with other maintained values of patterns.

4. Conclusion

There are so many problems for great systems by the huge amount of data in data-bases. Data-mining by using patterns from data-bases can solve the problem very much, as describe in the first section. So many kinds of patterns are suggested, designed and used in systems that we have introduced some in section 2. The new pattern designing suggested in this paper, is done using probability and decision trees. It is valid, novel, useful and very understandable because of giving mathematical information. The relationship between the relations (tables) is done according to system analysis and the designing and gathering of data is based on that analysis. Various analysis give various designing; therefore a different pattern should be presented so that this supposition is mentioned in that pattern producing. The amount of data is also important in designing of patterns and is supposed so. As described in section 3 it can be used in risk management operations and guide great systems to increase their profitable attributes and reduce the harmful ones.

4.1. Future Work

The method we presented in this paper sets the stage ready for two interesting topics of research:

- Risk Management by the Probability Trees.
- Policy Planning for Guiding toward the Objective Pattern.

A New Approach:

A method can be suggested, in which the changing way of patterns, show the fact that the system is running on the risk, non-profitable or non-affective way; or the strategies and polices chosen are correct. Data entrance and their growth and weakness are performed by the patterns. So it can be understood if the policies should change rapidly or other time interval is needed to go on the research.

5. References

[1] Larose D T. Discovering Knowledge in Data: An Introduction to Data Mining. Copyright C 2005 John Wiley & Sons, Inc. Ch 1. pp.2-4.

[2] Pei J, Upadhyaya S J, Farooq F, Govindaraju V. Data Mining for Intrusion Detection: Techniques, Applications and Systems. Proceedings of the 20th International Conference on Data Engineering (ICDE'04) © 2004 IEEE.

[3] Bloedorn E. Mining Aviation Safety Data: A Hybrid Approach. The MITRE Corporation, 2000.

[4] Kuonen D. A Statistical Perspective of Data Mining. Published by CRM Today, December 2004, in CRM Zine (Vol. 48).

[5] Karasova V. Spatial data mining as a tool for improving geographical models. Master's Thesis, Helsinki University of Technology, 2005. p. 6-7.

[6] Laxman S, Sastry P S. A survey of temporal data mining. Sadhana Vol. 31, Part 2, April 2006, © Printed in India. p.173.

[7] Aflori C, Leon F. Efficient distributed data mining using intelligent agents. Supported in part by the National University Research Council under Grant AT no 66 / 2004.

[8] Piatetsky-Shapiro G, Djeraba C, Getoor L. What are the grand challenges for data mining? KDD-2006 Panel Report, SIGKDD Explorations, Volume 8, Issue 2.

[9] Alvarez J L, Mata J, Riquelme J C. Data mining for the management of software development process. International Journal of Software Engineering and Knowledge Engineering, (1994) World Scientific Publishing Company. p.3.

[10] McGrail A J, Gulski E, Groot E R S. Data mining techniques to access the condition of high voltage electrical plant. School of Electrical Engineering, University of New South Wales, SYDNEY, NSW 2052, AUSTRALIA, On behalf of WG 15.11 of Study Committee 15, 2002.

[11] Ordieres Meré J B, and Castej Limas M. Data mining in industrial processes. Actas del III Taller Nacional de Miner a de Datos y Aprendizaje, TAMIDA2005. P. 60.

[12] Hand D J, Mannila H, Smyth P. Principles of Data Mining (Adaptive Computation and Machine Learning). The MIT Press (August 1, 2001); Ch 6: models and patterns.

[13] Menzies T, Hu Y. Data mining for very busy people. Published by the IEEE Computer Society, 0018-9162/03/$17.00 © 2003 IEEE; P.19.

[14] Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. Copyright © 1996, American Association for Artificial Intelligence. p. 37-49.

[15] Loo B T, Condie T, Garofalakis M, Gay D E, Hellerstein J M. Declarative networking: language, execution and optimization. SIGMOD 2006, Chicago, Illinois, USA, Copyright 2006 ACM.

[16] Nycz M, Smok B. Intelligent support for decision-making: a conceptual model. Informing Science InSITE - "Where Parallels Intersect", June 2003, P. 916-917.

[17] Wu1 X. Data Mining: An AI Perspective. IEEE Computational Intelligence Bulletin, December 2004, Vol.4 No.2.