World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:4, No:2, 2010

Open Science Index, Computer and Information Engineering Vol:4, No:2, 2010 publications.waset.org/7743.pdf

Distributed Data-Mining by Probability-Based Patterns

M. Kargar, and F. Gharbalchi

Abstract**—**In this paper a new method is suggested for distributed data-mining by the probability patterns. These patterns use decision trees and decision graphs. The patterns are cared to be valid, novel, useful, and understandable. Considering a set of functions, the system reaches to a good pattern or better objectives. By using the suggested method we will be able to extract the useful information from massive and multi-relational data bases.

Keywords**—**Data-mining, Decision tree, Decision graph, Pattern, Relationship.

I. Introduction

TECHNOLOGY development and the increasing prevalence of computer usage in people's daily life cause new techniques for data management and gathering information.

Increasing need for gathering information in different social systems and on the other hand, decreasing time periods from asking to get a reply, made every one think about ways for speed process on stored data and get useful information as soon as possible [1, 2]. Because of various activities in different systems distribution of data is undeniable. In recent years, database technology has advanced in stride. Vast amounts of data have been stored in the databases and business people have realized the wealth of information hidden in those data sets. Data-Mining then become the focus of attention as it promises to turn those raw data into valuable information that businesses can use to increase their profitability [3].

Representing local features of data in the forms called patterns is the main task of Data-Mining. In this paper, a new way of pattern extraction in distributed 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.

A. *Terminology*

Data-Mining is a step in the Knowledge Discovery in Data (KDD) process that consists of applying data analysis and discovery algorithms which, under acceptable computational limitations, produce a particular enumeration of patterns (or generate a model) over the data [4]. The key challenge in Data-Mining is the extraction of knowledge and insight from massive databases [5].

Data-Mining is not data warehousing, 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 [6, 7, 8, 9, 10].

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 [11].

In databases, relation has the same meaning with table. In this paper, relation is used instead of using the word table.

Distributed Data-Mining [12] [13] [14] is the subfield of Data-Mining which focuses on analyzing data remotely distributed in different, interconnected locations [15]. Indeed distributed Data-Mining separately works with data bases but at the end produces one pattern [5].

B. *Problem*

In distributed systems, when question becomes mooted, we need to answer to that question by extracting the relevant pattern. We should divide the pattern into sub-patterns and keep the principals of the pattern safe, as the components of the pattern are in the distributed system. Overall the problem is to find a way to divide this pattern without any break in logical relationship between sub-patterns.

C. *Solution*

As mentioned above, the distributed environment can cause the breaks in the relation of the sub-patterns. To prevent those breaks, the approach of probability-based patterns is employed. And through this, all of the relations in data based are considered to keep principals of the sub-patterns safe. Thus to answer the question and to find the related numbers we divide the presented pattern- mooted question- into sub-patterns considering the relations between all fields.

Mas'ood Kargar is with the Department of Computer Engineering, Islamic Azad University Tabriz Branch, Tabriz, Iran. He is also collaborating in Islamic Azad University of Tabriz Branch Computer Research Laboratories. IAUT-CRL (phone: +98-9141035597; email: kargar@iaut.ac.ir).

furugh.gharbalchi is with the Department of Computer Engineering, Islamic Azad University, Tabriz Branch, Tabriz, Iran (email: furugh.gharbalchi@gmail.com).

D. The Claim

We claim that extracting such pattern is possible. Our final aim in distributed Data-Mining is to reach the pattern. Such pattern is a matrix that contains sub-patterns that these sub-patterns themselves are matrixes and each of them formed in one data base. These sub-patterns will connect to each other and the result is the final pattern. Therefore if we can extract such sub-patterns then we connect them to each other we will have the final pattern. Thus we introduce the new approach in distributed Data-Mining in relational data bases that extract patterns, called probability based patterns.

E. 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. Most of the data-mining methods look for a pattern in one table or data base but many of the data bases we are dealing with are in multi-relational and distributed form. The objectives we want to achieve by suggesting a new method for designing patterns is using probability in distributed data-mining to gain relevant sub-patterns and connect them to each other to reach final desired pattern. This pattern can give numerical information about the attributes in databases.

F. Paper Outline

This paper includes 7 sections. In section I we mention the problem and the idea of solution and our claim regarding the solution. In section II we have an overview to the previous works, the strong and weaknesses, and requiring their further improvements. In section III we explain our work and in section IV the problem of our work is figured out. In section V the solution of our method is propounded. And finally we conclude the paper in section VI.

II. Previous Work

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

A. Neural Network

Artificial neural networks, as a parallel, fine-grained implementation of non-linear static or dynamic systems, were originally developed as a parallel computational model. A very important advantage of these networks is their adaptive capability, where “learning by example” replaces the traditional “programming” in problem solving. Another important advantage is the intrinsic parallelism that allows fast computations. Artificial neural networks are a viable computational model for a wide variety of problems, including pattern classification, speech synthesis and recognition, approximation, associative memory, and modeling and control of non-linear unknown systems, in addition to the application of multimedia Data-Mining. The third advantage of artificial neural networks is the generalization capability, which allows correct classification of new patterns. A significant disadvantage of artificial neural networks is their poor interpretability [16].

B. On-Line Analytical Processing (OLAP)

OLAP is part of the spectrum of decision support tools and it uses multidimensional array representation. Data analysts use OLAP tools to test the relevance of a theory, whereas they apply data mining tools to a problem in hopes that the findings will suggest an answer. Furthermore, OLAP is also complementary in the early stages of the knowledge discovery process because it can help exploring data, for instance by focusing attention on important variables, identifying exceptions, or finding interactions. [17, 18]

C. Decision Tree

Decision trees are very popular for Data-Mining applications since they obtain reasonable accuracy and relatively less computationally expensive to contrast and use. A decision tree consists of internal nodes and leaves. Each of the internal nodes has splitting decision and splitting attribute associated with it. The leaves have a class label assigned to them. Once a decision tree is built, a prediction process is relatively straightforward: the classification process begins at the root, and a path to a leaf is traced by using the splitting decision at each internal node. The class label attached to the leaf is then to the incoming record [19]. Given the growth in distributed databases at geographically dispersed locations, the methods for decision tree induction in distributed settings are gaining importance [20].

D. Probability-Based Patterns

Using probability in the pattern makes it so sensible, understandable and easy to be performed. Totally in probability-based patterns data-bases are considered as relations with logic relationships between them. Thus to attain the pattern via this approach, there are steps to be considered:

1. Presuppositions in designing a pattern, is the first step that the relations are analyzed and the important attributes are perceived.

2. Drawing the directed graph, is the next step. The directed graph- data base model- is drawn based on the relationships of data base.

3. Giving name to the nameless vertexes;

4. Computing the importance and priority is done based on the out put and input degrees.

5. Designing the tree in which the weights of edges are reckoned referring the statistics of the data base.

6. Designing the pattern is the final step that is performed where the maximum number of nodes and maximum degree of the tree are impressive factors.

For further information about the steps refer to [1, 2, 3].

III. Distributed Data-Mining by Probability-Based Patterns

In the suggested method, for representing pattern in distributed data mining, a new step is put forward.

As mentioned, here the data bases are considered as relations with logical relationships between them. The objective is to answer the question that is asked by coordinator by presenting the related numbers of each attribute which are mentioned in the question. The description of the pattern is kept by representing an example:

In Distributed Data Mining, that is typically not feasible to share or communicate data at all; local models are built at each site, and are then merged /combined via various methods [5]. Therefore here we are dealing with the distributed environment which causes the distribution of pattern components too. Totally this job is completed in 5 main steps.

A. Presenting the Effective Attributes

According to the goal of the pattern extraction, coordinator asks a question that presents the effective attributes of pattern, and this will be a start point operation.

The sample of the question is:

How the students' educational status affects the course types they choose?

B. Determining Final Pattern

The final pattern is determined by the attributes which are presented in the question of coordinator. As it is illustrated in Fig. 1 in our example the effective subjects which are mentioned by coordinator are student and course.

Attribute 1	Attribute 2	…
Value 1	Value 2	…

sub-pattern 1(student)

Attribute 1	Attribute 2	…
Value 1	Value 2	…

sub-pattern 2(course)

Fig. 1 The form of final pattern

C. Determining the Groups

As presented attributes in coordinator's question we determine the groups which these attributes are members of them. As it is shown in Fig. 2 according to the existing relations in data base, the effective groups forming the related sub-pattern are chosen.

Attribute 1	Attribute 2	Attribute 3
Value 1	Value 2	Value 3

sub-pattern 1(student)

Attribute 1	Attribute 2	…
Value 1	Value 2	…

sub-pattern 2(course)

Fig. 2 The groups of final pattern

D. Pattern Production in Probability-Based Decision Tree

As mentioned in previous works, to produce the probability-based patterns firstly we should have the data base model. Then each table of data base is presented in form of decision tree. The decision trees whose leaves are classifications, the branches are clusters and the priorities of their levels are computed. The sample of decision tree is presented in Fig. 3.

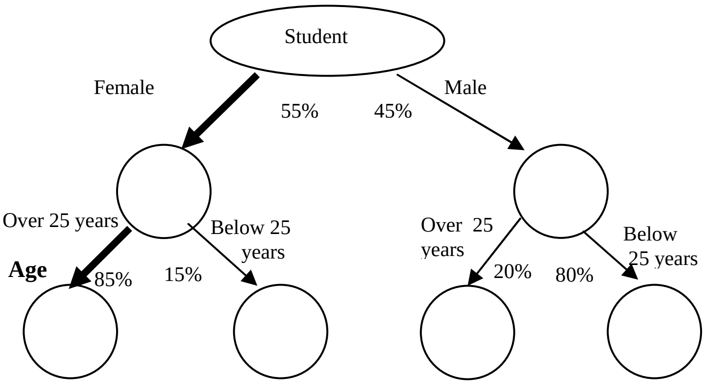

Fig. 3 Sample of decision tree

E. Pattern Extraction

The final step is to produce sub-patterns and find the numbers of the step 2 according to available statistics. In fact in every group the best path is selected. And this job is done referring to their weight in each edge of decision tree.

Now every sub-pattern is placed beside others in order to reach the desired final pattern. The sample of extracted sub-pattern is shown in Fig. 4.

Female	Less than25
55%	85%

sub-pattern 1(student)

College	M.S	art
55%	70%	65%

Major sub-pattern

Mediocre	Less than 4 terms
65%	60%

Edu. Status sub-pattern

Fig. 4 The sample of extracted sub-pattern

After applying all steps above, we will have few sub-patterns which contain the answer of the primary question then these sub-patterns join together to form the final desired pattern. And it is the end of process.

F. Toward a Reliable Model

After reaching the pattern, we notice that if the tree levels change, the pattern will change too. And this problem is the result of intricate clusters.

The major two advantages of decision graph that can be listed are:

- Adaptation of designed model with decision model,
- No effect of number of different values in clustering localization.

So it is beneficial to use decision graph instead of decision tree.

G. Pattern Production in Probability-Based Decision Graph

As the pattern production in probability-based decision trees, all the above steps will be taken here but instead of decision tree we draw decision graph. After computing the priority by formula1 we will be able to draw the graph. In this approach the unnecessary attributes may be generated. According to the input and the output degrees we decide about the path.

$$H(A_i) = C_1(DI_{RAi}) + C_2(DO_{RAi}) + C_3(H_{RAi}) \qquad (1)$$

At the formula above, DI is internal degree, DO is external degree and H is number of hops to the target relation from which the priority of Ai is computed. C1, C2, C3 are constants that show the importance of internal degree, external degree and hops. Obviously hop has an important role in every case and the external degree is much important and effective than the internal degree.

If we don't know the priorities the start point of the sub-patterns will be uncertain, however with considering priorities the first attribute and the others are determined.

The sample of decision graph is illustrated in Fig. 5.

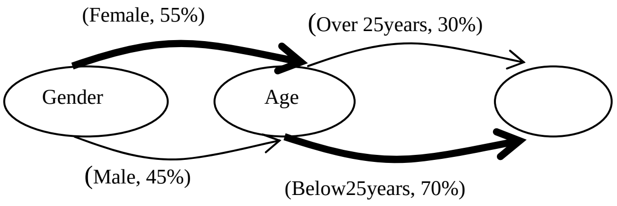

Fig. 5 The sample of decision graph

So by using the probability decision graph, we solve the local intricacy problem.

IV. Conclusion

Data-mining by using patterns from data-bases can lead us to profitable information. Distributed Data-Mining deals with loosely-coupled systems and it is essential in cases where there are multiple distributed Data-Bases as explained in introduction part. The extracted pattern through Distributed Data-Mining process is not necessarily the superior pattern since the presented attributes possibly are not the impressive attributed of the tables. If the relations are numerous it is better to use decision tree instead of decision graph due to their specifications.

A mentioned method can apply to distributed environment in order to extract the desired pattern due to discover the knowledge hidden in the data bases. Using this method can be helpful to increase the profitability of knowledge discovery process.

References

[1] M. Karegar, A. Isazadeh, F. Fartash, T. Saderi, A. Habibizad Navin. "Data-mining by the probability-based patterns," Published in the proceeding of the 30th International Conference on Information Technology Integrity, ITI 2008 IEEE. June 2008.

[2] M. Karegar, R. Mirmiran, F. Fartash, T. Saderi. "Risk-management by probability-based patterns in data-mining," Published in the proceeding of the International Conference on Information Technology Symposium 2008, ITSim 2008 IEEE. August 2008.

[3] Supatcharee Sirikulvadhana (2002), "Data Mining as A Financial Auditing Tool," unpublished thesis (M.Sc) The Swedish School of Economics and Business Administration.

[4] Sankar K. Pal and Pabitra Mitra (2004): "Pattern Recognition Algorithms for Data Mining," Calcutta, CHAPMAN & HALL/CRC

[5] Zaki Mohammed J., Ching-Tien Ho (2000): "Large-Scale Parallel Data Mining," Berlin, Springer.ch1,pp 1-2.

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] Jennings, N., Sycara, K., Wooldridge, M. "A Roadmap of Agent Research and Development, Autonomous Agents and Multi-Agent Systems," 1:7-38, 1998.

[13] Park, B., Kargupta, H. "Distributed Data Mining: Algorithms, Systems, and Applications," In the Handbook of Data Mining, N. Ye (ed.), Lawrence Erlbaum Associates, pp: 341-358, 2003.

[14] Freitas, A.; Lavington, S. H. "Mining very large data bases with parallel processing," Kluwer Academic Publishers The Netherlands, 1998.

[15] Danish Khan. "CAKE – Classifying, Associating & Knowledge DiscovEry An Approach for Distributed Data Mining (DDM) Using Parallel Data Mining Agents (PADMAs)," Published in the proceeding of International Conference on Information Technology Integrity, ITI 2008 IEEE. 2008.

[16] Zhongfei Zhang, Ruofei Zhang (2009): "Multimedia Data Mining A Systematic Introduction to Concepts and Theory," Boca Raton, CRC Press.

[17] Tan. P, Steinbach M., and Kumar V (2005): Introduction to Data Mining, Addison-Wesley, ch3.

[18] Herbert A.Edelstein (1999): Introduction to Data Mining and Knowledge Discovery,Third Edition.U.S.A:Two Crows Corporation, pp:8-9, 2005.

[19] Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar (2008): "Next Generation of Data Mining," CRC Press, ch8.pp: 155.

[20] Jie Ouyang Patel, N.Sethi, I.K. Chi-Square. "Test Based Decision Trees Induction in Distributed Environment," Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on Dec. 2008.