Probabilistic Multi Relation Classification Approach by Using Genetic Algorithms: A Review

Surbhi Pathak, Mr Arun Kumar Jhapate
Department of Computer Science and Engineering,
Sagar Institute of Research & Technology, Bhopal
pathaksurbhi28@gmail.com

***Abstract*— Classification is an important subject in data mining and machine learning, which has been studied extensively and has a wide range of applications. The traditional variable weighting methods suffer from unbalanced phenomenon: the view with more variables will play more important role than the view with less variables. In the two-level variable weighting method, the view weights will be only determined in the view level, while the variable weights will be only determined in a view. Therefore, the two levels of variable weights will eliminate the unbalanced phenomenon and compute more objective weights. Higher degree of classification rate leads better classification. This paper presents birds' eye over Bayesian probabilistic classification and GA approach over it and explain various step of Genetic algorithms and approach for classification.**

Keywords- *Bayesian Network, Genetic algorithms, probabilistic classification*

I INTRODUCTION

Data mining [1] an non-trivial extraction of novel, implicit, and actionable knowledge from large data sets is an evolving technology which is a direct result of the increasing use of computer databases in order to store and retrieve information effectively .It is also known as Knowledge Discovery in Databases (KDD) and enables data exploration, data analysis, and data visualization of huge databases at a high level of abstraction, without a specific hypothesis in mind. The working of data mining is understood by using a method called modeling with it to make predictions. Data mining techniques are results of long process of research and product development and include artificial neural networks, decision trees and genetic algorithms.

This retrieval of data as and when needed contributes the technology of data mining. Data mining can be viewed as a result of the natural evolution of information technology. This technology provides a wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Data mining is the extraction of interesting patterns or knowledge from huge amount of data. It can be known by different names like knowledge discovery (mining) in Databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence and others.

The term "data mining" [2] is nothing but analysis of data in a database using tools which look for trends or anomalies without the knowledge of meaning of the data and is primarily used by statisticians, database researchers and business communities. A data mining software does not just change the presentation, but discovers previously unknown relationships among the data.

The information on which the data mining process operates is contained in a historical database of previous interactions. In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. Some kinds of information that is collected are as follows: Some kinds of information that is collected are as follows:

Every transaction in the business industry is (often) "memorized" for perpetuity. Such transactions are usuallytime related and can be inter-business or intra-businessoperations effective use of the data in a reasonable timeframe for competitive decision making is definitely the most important problem to solve for businesses that struggle tosurvive in a highly competitive world.

Our society is amassing colossal amounts of scientific data that need to be analyzed. Unfortunately, wecan capture and store more new data faster than we cananalyze the old data already accumulated.

This Paper discusses about one of the application areas of partition based clustering algorithms k-Means and Fuzzy C-Means by means of an experimental approach choosing a real time telecommunication data. Many applications have been proposed by using different algorithms. Now, it is necessary to discuss some of the applications of related areas. This will be helpful to understand the related breakthrough in computations and engineering applications. They discuss that both theoretical and practical efforts in band images often neglect the characteristics having interactions and mutual influence among attributes or criteria, even in the stages of different brand life cycles. This study aims to create a hierarchical framework for brand image management.

The analytical network process and fuzzy sets theory have been applied to both mindshare in brand images and inherent interaction/interdependencies among diverse information resources. A real empirical application is demonstrated in the department store. Both the theoretical and practical background of this work have shown the fuzzy analytical network process can capture expert's knowledge existing in the form of incomplete and vague information for the mutual inspiration on attribute and criteria of brand image management.

In the two-level variable weighting method, the variable weights V are used to identify the important variables in each view, and the view weights W are used to identify compact cluster structures within these views. If the view contains compact cluster structures, a large view weight is assigned so as to enhance the effect of such view; on the contrary, if the view contains loose cluster structures, a small view weight is assigned to eliminate the effect of such view. Compared with the traditional variable weighting method, the new method can take both individual variables and multiple views into consideration and capture the differences among different views and variables.

II LITERATURE SURVEY

This section gives an extensive literature survey on the multiple relational classification using genetic algorithms. We study various research paper and journal and know about data classification. All methodology and process are not described here. But some related

work in the field of association classification discuss by the name of authors and their respective title.

Zhen- Hui Song &Yi Li, [6]. Describe in the field of data classification as Associative classification (AC) has shown great promise over many other classification techniques on static dataset. However, the increasing prominence of data streams arising in a wide range of advanced application has posed a new challenge for it. The author describes and evaluates AC-DS, a new associative classification algorithm for data streams which is based on the estimation mechanism of the Lossy Counting (LC) and landmark window model. They apply AC-DS to mining several datasets obtained from the UCI Machine Learning Repository and the result shows that the algorithm is effective and efficient. An associative classification approach based on association rules for mining data streams. Empirical studies show its effectiveness in taking advantage of massive numbers of examples. AC-DS's application to a high-speed stream is under way.

S.P. Syed Ibrahim, K. R. Chandran, M. S. Abinaya[7] Describe in the field of data classification as weighted association rule mining reflects semantic significance of item by considering its weight. Classification extracts set of rules and constructs a classifier to predict the new data instance. The author proposes compact weighted associative classification method, which integrates weighted association rule mining and classification for constructing an efficient weighted associative classifier.

Pei-Yi Hao, Yu-De Chen [8] Describe in the field of data classification as Association Classification not only has widely adopted but also has performed well in data mining. The literatures have been argued that the small disjunction and using multiple class-association rules have significant effect on classification accuracy. The author proposed a CMAR (Classification based on Multiple Class-Association Rules) and Adriano Veloso proposed Lazy Associative Classifier algorithm for Small Disjunction mining. In addition, we collocate with a new weight calculation method in our algorithm to solve weight bias problem of CMAR.

You Wan, Chenghu Zhou [9] Describe in the field of data classification as Spatial Co-location patterns are similar to association rules but explore more relying spatial auto-correlation. They represent subsets of Boolean spatial features whose instances are often located in close geographic proximity. Existing co-location patterns mining researches only concern the spatial attributes, and few of them can handle the huge amount of non-spatial attributes in spatial datasets. Also, they use distance threshold to define spatial neighborhood. However, it is hard to decide the distance threshold for each spatial dataset without specific prior knowledge.

Achilleas Tziatzios and Jianhua Shao, Grigorios Loukides[10]Describe in the field of data classification as the ability to learn classification rules from data is important and useful in a range of applications. While many methods to facilitate this task have been proposed, few can derive classification rules that involve ranges (numerical intervals).

Rupali haldulakar, Prof. Jitendra Agrawal[11]Describe in the field of data classification as Strong rule generation is an important area of data mining. The author proposed a design a novel method for generation of strong rule. In which a general Apriori algorithm is used to generate the rules after that they use the optimization techniques. Genetic algorithm is one of the best ways to optimize the rules .In this direction for the optimization of the rule set we design a new fitness function that uses the concept of supervised learning then the GA will be able to generate the stronger rule set. In this direction we optimize association rule mining using new fitness function. To make genetic algorithm more effective and efficient it can be incorporated with other techniques so it can provide a best result.

XING Xue, CHEN Yao. WANG Yan-en[12] Describe in the field of data classification as association rules which applied in data mining that aims to analyze large source data and reveal knowledge hidden in the database and proposed a association rules mining to the software of examination paper evaluation system, obtaining the useful information which is hidden in the database. It's concluded that the algorithm provides a valuable analysis of information to the examination paper evaluation system Keywords-association rules. Association rules is the one most important theory in data mining, which have a wide range of applications in the various fields, but, applied to the Evaluation of reliable, it can be said that has just begun, with the mining association rules theoretical the constant further research and using, the rational, efficient and objective analysis of The examination, all from the Association Rules theoretical support. And, judging from the current access to a large number of information, Association rules, applied to the analysis of the test research, has aroused the expert's wide attention.

III CLASSIFICATION

Classification [13] is an important subject in data mining and machine learning, which has been studied extensively and has a wide range of applications. Classification based on association rules, also called associative classification, is a technique that uses association rules to build classifier. Generally it contains two steps: first it finds all the class association rules (CARs) whose right-hand side is a class label, and then selects strong rules from the CARs to build a classifier. In this way, associative classification can generate rules with higher confidence and better understandability comparing with traditional approaches. Thus associative classification has been studied widely in both academic world and industrial world, and several effective algorithms [14, 15] have been proposed successively. However, all the above algorithms only focus on processing data organized in a single relational table.

In practical application, data is often stored dispersedly in multiple tables in a relational database. Simply converting multi-relational data into a single flat table may lead to the high time and space cost, moreover, some essential semantic information carried by the multi-relational data may be lost. Thus the existing associative classification algorithms cannot be applied in a relational database directly.

Table 1: Training set

T1	T2	T3	T4	CLASS
1	2	3	4	Yes
1	2	3	4	Yes
1	2	3	4	Yes
1	2	3	4	No
1	2	3	4	Yes
1	2	3	4	No
1	2	3	4	No
1	2	3	4	No

IV DECISION TREE

Decision trees [5] are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. An example of a decision tree for the training set.Using the decision tree as an example, the instance At1 = a1, At2 = b2, At3 = a3, At4 = b4 would sort to the nodes: At1, At2, and finally At3, which would classify the instance as being positive (represented by the values "Yes"). The problem of constructing optimal binary decision trees is an NP complete problem and thus theoreticians have searched for efficient heuristics for constructing near-optimal decision trees. The feature that best divides the training data would be the root node of the tree. There are numerous methods for finding the feature that best divides the training data such as information gain. While myopic measures estimate each attribute independently.

V BAYESIAN NETWORKS

A Bayesian Network (BN) [4] is a graphical model for probability relationships among a set of variables features.

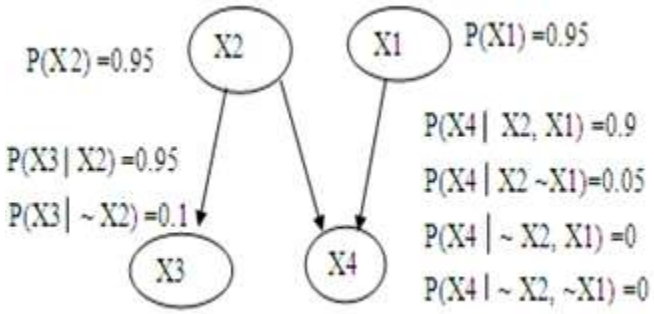

Figure 1: Bayesian Network

The Bayesian network structure S is a directed acyclic graph DAG) and the nodes in S are in one-to-one correspondence with the features X. The arcs represent casual influences among the features while the lack of possible arcs in S encodes conditional independencies. Moreover, a feature (node) is conditionally independent from its non-descendants given its parents (X1 is conditionally independent from X2 given X3 if P (X1|X2, X3) =P (X1|X3) for all possible values of (X1, X2, X3) [5].

VI K-NEAREST NEIGHBOR CLASSIFIERS

Nearest neighbor classifiers [4] are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points,

$X=(x1, x2, \ldots\ldots\ldots\ldots,x_n)$ and $Y= (y1,y2,\ldots\ldots y_n)$ is

$$d(X, Y)= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

The unknown sample is assigned the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space [5].

VII CLUSTERING

As we mentioned before, classification can be taken as supervised learning process, clustering is another mining technique similar to classification. However clustering is a unsupervised learning process. Clustering[4] is the process of grouping a set of physical or abstract objects into classes of similar objects, so that objects within the same cluster must be similar to some extent, also they should be dissimilar to those objects in other clusters. In classification which record belongs which class is predefined, while in clustering there is no predefined classes. In clustering, objects are grouped together based on their similarities. Similarity between objects is defined by similarity functions; usually similarities are quantitatively specified as distance or other measures by corresponding domain experts.Most clustering applications are used in market segmentation. By clustering their customers into different groups, business organizations can provide different personalized services to different group of markets. For example, based on the expense, deposit and draw patterns of the customers, a bank can clustering the market into different groups of people. For different groups of market, the bank can provide different kinds of loans for houses or cars with different budget plans. In this case the bank can provide a better service, and also make sure that all the loans can be reclaimed [5].

VIII GENETIC ALGORITHM

The genetic algorithm (GA) is an optimization and search technique [16] based on the principles of genetics and natural selection. A GA allows a population composed of many individuals (basically the candidates) to evolve under specified selection rules to a state that maximizes the fitness. A genetic algorithm mainly composed of three operators: selection, crossover, and mutation. In selection, a good string (on the basis of fitness) is selected to breed a new generation;crossover combines good strings to generate better offspring; mutation alters a string locally to maintain genetic diversity from one generation of a population of chromosomes to the next. In each generation, the population is evaluated and tested for termination of the algorithm. If the termination criterion is not satisfied, the population is operated upon by the three GA operators and then re-evaluated. The GA cycle continues until the termination criterion is reached. In feature selection, Genetic Algorithm (GA) is used as a random selection algorithm, Capable of effectively exploring large search spaces, which is usually required in case of attribute selection. For instance; if the original feature set contains N number of features, the total number of competing candidate subsets to be generated is 2N, which is a huge number even for medium-sized N. Further, unlike many search algorithms, which perform a local, greedy search, Gas performs a global search. This proposes a method for attribute subset selection based of correlation using GA. Correlation between the attributes will decide the fitness of individual to take part in mating. Fitness function for GA is a simple function, which assigns a rank to individual attribute on the basis of correlation coefficients. Since strongly correlated attributes cannot be the part of DW together, only those attributes shall be fit to take part in the crossover operations that

are having lower correlation coefficients. In other words we can say lower the correlation is higher the fitness value will be [16]. Specialty of the proposed method can more clearly be stated as follows:

A. It performs either equally good or better than many of the existing methods.
B. Its accuracy is more when applied on real and large dataset.
C. It is very simple and light because GA is used to search the optimal subset of attributes besides being used for searching the optimal techniques for attribute selections amongst the available ones.

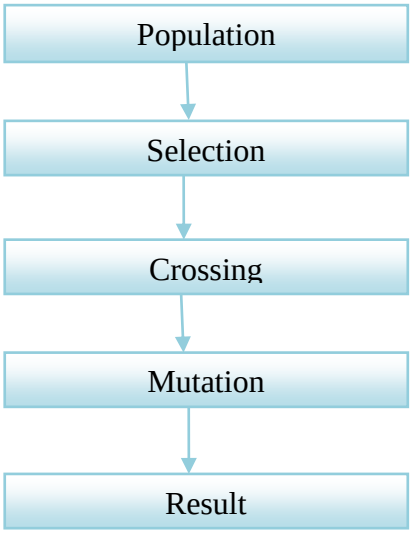

Figure 2: Stages of GA

A genetic algorithm [17] is a type of searching algorithm. It searches a solution space for an optimal solution to a problem. The algorithm creates a "population" of possible solutions to the problem and lets them "evolve" over multiple generations to find better and better solution. Algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. Cycle of the Algorithm: The algorithm operates through a simple cycle.

A. Creation of a population of strings.
B. Evolution of each string.
C. Selection of the best string.
D. Genetic manipulation to create a new population of strings.

IX CONCLUSIONS

In the two-level variable weighting method, the variable weights V are used to identify the important variables in each view, and the view weights W are used to identify compact cluster structures within these views. If the view contains compact cluster structures, a large view weight is assigned so as to enhance the effect of such view; on the contrary, if the view contains loose cluster structures, a small view weight is assigned to eliminate the effect of such view. Compared with the traditional variable weighting method, the new method can take both individual variables and multiple views into consideration and capture the differences among different views and variables. Moreover, the traditional variable weighting methods suffer from unbalanced phenomenon: the view with more variables will play more important role than the view with less variables. In the two-level variable weighting method, the view weights will be only determined in the view level, while the variable weights will be only determined in a view. Therefore, the two levels of variable weights will eliminate the unbalanced phenomenon and compute more objective weights.TW k-means can be considered as the two-level variable weighting clustering algorithm, but it only computes automated two-level variable weighting clustering algorithm but do not apply any Fuzzy and time saving approach.

REFERENCES

[1] B.N. Lakshmi., G.H. Raghunandhan. "A conceptual Overview of Data Mining" Proceedings of the National Conference on Innovations in Emerging Technology-2011 Kongu Engineering College, Perundurai, Erode, Tamilnadu, pp.27-32. India.17 & 18 February, 2011.

[2] Han J. and M. Kamber (2000), Data Mining: Concepts and Techniques, Academic Press, San Diego, CA.

[3] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth "From Data Mining to KDD in Databases" pp. 0738-4602 1996.

[4] Xun Zhu1, Hongtao Deng2, Zheng Chen3 "A Brief Review On Frequent Pattern Mining" PP-4-11 2011 IEEE.

[5] Thair Nu Phyu "Survey of Classification Techniques in Data Mining" Vol I Imecs 2009, March 18 - 20, 2009, Hong Kong.

[6] Zhen- Hui Song &Yi Li, "Associative classification over Data Streams", IEEE, PP.2-10, 2010.

[7] S.P.Syed Ibrahim1 K. R. Chandran2 M. S. Abinaya3 "Compact Weighted Associative Classification" IEEE pp.8-11, 2011.

[8] Pei-yi hao, yu-de Chen "a novel associative classification algorithm: a combination of LAC and CMAR with new measure of Weighted effect of each rule group" IEEE pp.9-11, 2011.

[9] You Wan, Chenghu Zhou" QuCOM: k nearest features neighborhood based qualitative spatial co-location patterns mining algorithm" IEEE pp.8-11, 2011.

[10] Achilleas Tziatzios and Jianhua Shao, Grigorios Loukides" A Heuristic Method for Deriving Range-Based Classification Rules" IEEE pp.6-11, 2011.

[11] Rupali haldulakar, prof. Jitendra agrawal" Optimization of Association Rule Mining through Genetic Algorithm" (IJCSE) Vol. 3 No. 3 Mar 2011.

[12] XING Xue, CHEN Yao. WANG Yan-en" Study on Mining Theories of Association Rules and Its Application"IEEE PP.2-10, 2010.

[13] Yingqin Gu1,2, Hongyan Liu3, Jun He1,2, Bo Hu1,2 and Xiaoyong Du1,2 "A Multi-relational Classification Algorithm based on Association Rules" pp.4-9 2009 IEEE.

[14] W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient Classification Based on Multiple Class-Association Rules", Proceedings of the ICDM, IEEE Computer Society, San Jose California, 2001, pp. 369-376.

[15] X. Yin, and J. Han, "CPAR: Classification based on Predictive Association Rules", Proceedings of the SDM, SIAM, Francisco California, 2003.

[16] Rajdev Tiwari, Manu Pratap Singh "Correlation-based Attribute Selection using Genetic Algorithm" International Journal of Computer Applications (0975 – 8887) Volume 4– No.8, August 2010.

[17] Kalyanmoy Deb, "Introduction to Genetic Algorithms", Kanpur Genetic Laboratory (Kangal), Depart of Mechanical Engineering, IIIT Kanpur 2005.

