ISSN 2395-1621

Functional Association Rule Mining For Continuous Attributes

#1Ms. Pooja Kulkarni, #2Mrs.Vaishali Kolhe

1kulkarnipooja8080@gmail.com,
2vlkolhe@gmail.com

#1M.E. Student, Department of Computer Engineering,
#2Assistant Professor, Department of Computer Engineering

D.Y. Patil College of Engineering, Akurdi,
Savitribai Phule Pune University, India

ABSTRACT

Association Rule Mining (ARM) is an important branch of methods for extracting patterns from data sets. Earlier, ARM is concerned with categorical data sets. When it is used to process continuous variables, it converts the values of the variables into intervals. A continuous attribute is one which can take any value within the specified range. The continuous attribute is transformed into a finite number of intervals associated with a discrete value. The importance of discretization methods stems from interest in extending to continuous variable classification methods, such as decision trees or Bayesian networks, which were designed to work on discrete variables. The continuous variables can also be handled without discretization.

Keywords: Association rule mining (ARM), Continuous Attribute.

ARTICLE INFO

Article History

Received: 8th January 2017

Received in revised form :

8th January 2017

Accepted: 10th January 2017

Published online :

11th January 2017

I. INTRODUCTION

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is likely to also purchase milk." One of the reasons behind maintaining any database is to enable the user to find interesting patterns and trends in the data. For example, in a supermarket, the user can figure out which items are being sold most frequently. But this is not the only type of 'trend' which one can possibly think of. The goal of database mining is to automate this process of finding interesting patterns and trends. Once this information is available, there is a way get rid of the original database. The output of the data-mining process should be a summary of the database. This goal is difficult to achieve due to the vagueness associated with the term "interesting". The solution is to define various types of trends and to look for only those trends in the database. One such type constitutes the association rule.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout [1]. Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed. The abundance of data generates the appearance of a new field named data mining. Data collected in large databases become raw material for these knowledge discovery techniques and mining tools for gold were necessary. The current expert system technologies, which typically rely on users or domain experts to manually input knowledge into knowledge bases. This procedure contains errors, and it is extremely time consuming and costly.

Data mining tools which perform data analysis may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. Data mining represents the automatic process to discover patterns and relations between data stored in large databases called warehouses, the final product of this process being the knowledge, meaning the significant information provided by the unknown elements AR mining (ARM) is an important branch of methods for extracting patterns from data sets. Conventionally, ARM is concerned with categorical data sets. When it is used to process continuous variables, it converts the values of the variables into intervals. This discretization process determines the granularity of the ARs being generated. In contrast, the FAR proposed can handle nonlinearity in the relationship and can deal with continuous variables directly and without the need to convert them into intervals. Therefore, it does not require a discretization process, neither does it need predetermining granularity levels.

II. RELATED WORK

Many classification algorithms require that the training data contain only discrete attributes. To use such an algorithm when there are numeric at- tributes, all numeric values must first be converted into discrete values-a process called discretization. This paper describes ChiMerge, a general, robust algorithm that uses the x2 statistic to discretize numeric attributes [1].

USAMA M. FAYYAD present a result applicable to classification learning algorithms that generate decision trees or rules using the information entropy minimization heuristic for discretizing continuous-valued attributes. The result serves to give a better understanding of the entropy measure, to point out that the behavior of the information entropy heuristic possesses desirable properties that justify its usage in a formal sense, and to improve the efficiency of evaluating continuous-valued attributes for cut value selection. Along with the formal proof, we present empirical results that demonstrate the theoretically expected reduction in evaluation effort for training data sets from real-world domains.

Discretization can turn numeric attributes into discrete ones. Feature selection can eliminate some irrelevant and/or redundant attributes. Huan Liu invented Chi2 which is a simple and general algorithm that uses the c2 statistic to discretize numeric attributes repeatedly until some inconsistencies are found in the data. It achieves feature selection via discretization. It can handle mixed attributes, work with multiclass data, and remove irrelevant and redundant attributes [2].

Discretization should consider the effects on all variables in the analysis and that two regions X and Y should only be in the same cell after discretization if the instances in those regions have similar multivariate distributions (Fx , Fy) across all variables and combinations of variables. Stephen D. Bay presented a bottom up merging algorithm to discretize continuous variables based on this rule [10].

Hisao Ishibuchi first fuzzify the concept of association rules. That is, He showed fuzzy versions of two measures (i.e., confidence and support) that are used for evaluating each association rule in the field of data mining. Then he illustrated these two measures of fuzzy rules for function approximation and patten classification problems. Finally he examined the relation between the classification performance of fuzzy rules and the definition of their certainty grades through computer simulations. Simulation results show that the direct use of confidence as a certainty grade is not always appropriate from the viewpoint of classification performance [9].

Marco Vannucci presents the problem of the unsupervised discretization of continuous attributes for association rules mining. It shows commonly used techniques for this aim and highlights their principal limitations. To overcome such limitations a method based on the use of a SOM is presented and tested over various real world datasets.

Marco Vannucci presented the problem of the unsupervised discretization of continuous attributes for association rules mining. It shows commonly used techniques for this aim and highlights their principal limitations. To overcome such limitations a method based on the use of a SOM is presented and tested over various real world datasets [7]. Wai-Ho Au presented a new method to determine the membership functions of fuzzy sets directly from data to maximize the class-attribute interdependence and, hence, improve the classification results. In other words, it forms a fuzzy partition of the input space automatically, using an information theoretic measure to evaluate the interdependence between the class membership and an attribute as the objective function for fuzzy partitioning. To find the optimum of the measure, it employs fractional programming [3].

Karla Taboada presented an association rule mining algorithm that is suited for continuous valued attributes commonly found in scientific and statistical databases. We propose a method using a new graph-based evolutionary algorithm named ‘genetic network programming (GNP)’ that can deal with continuous values directly, that is, without using any discretization method as a preprocessing step. GNP represents its individuals using graph structures and evolves them in order to find a solution; this feature contributes to creating very compact programs and implicitly memorizing past action sequences. Then he added fuzzy membership to it [5][13].

Francisco J. Ruiz introduced a new method for supervised discretization based on interval distances by using a novel concept of neighborhood in the target’s space. The proposed method takes into consideration the order of the class attribute, when this exists, so that it can be used with ordinal discrete classes as well as continuous classes, in the case of regression problems[11].

Bilal Alatas presented Pareto-based multi-objective differential evolution (DE) algorithm is proposed as a search strategy for mining accurate and comprehensible numeric association rules (ARs) which are optimal in the wider sense that no other rules are superior to them when all objectives are simultaneously considered. The proposed DE guided the search of ARs toward the global Pareto-optimal set while maintaining adequate population diversity to capture as many high-quality ARs as possible [14]. Rong Cong introduced a new method for discretization of continuous attributes is put forward to overcome the limitation of the traditional rough sets, which cannot deal with continuous attributes. The method is based on an improved algorithm to produce candidate cut points and an algorithm of reduction based on variable precision rough information entropy [4].

Bing Wang introduced a novel form of association rules (ARs) that do not require discretization of continuous variables or the use of intervals in either sides of the rule. This rule form captures nonlinear relationships among variables, and provides an alternative pattern representation for mining essential relations hidden in a given data set.

III. METHODS OF HANDLING CONTINUOUS ATTRIBUTES

A. *Dealing with continuous attributes via discritization:*

Discretization, also named quantization, is the process by which a continuous attribute is transformed into a finite number of intervals associated with a discrete value. The importance of discretization methods stems from interest in extending to continuous variable classification methods, such as decision trees or Bayesian networks, which were designed to work on discrete variables. The use of discrete variables, besides diminishing the computational cost of some automatic learning algorithms, also facilitates the interpretation of the obtained results [10],[3]. Discretization can be considered as a previous stage in the global process of inductive learning. In decision trees, discretization as a pre-processing step is preferable to a local discretization process as part of the decision tree building algorithm [4]. This stage can be carried out directly by an expert or automatically by means of a suitable methodology. In any case, the discretization process entails implicit knowledge of the data. This knowledge is introduced explicitly into the learning process by an expert or extracted implicitly from the data as a prior step to the global learning process, if discretization is carried out automatically [11].

ChiMerge, a general, robust algorithm that uses the x^2 statistic to discretize (quantize) numeric attributes. The ChiMerge algorithm consists of an initialization step and a bottom-up merging process, where intervals are continuously merged until a termination condition is met [1]. Chi2 is a simple and general algorithm that uses the x^2 statistic to discretize numeric attributes repeatedly until some inconsistencies are found in the data. It achieves feature selection via discretization. It can handle mixed attributes, work with multiclass data, and remove irrelevant and redundant attributes. The Chi2 algorithm applies the x^2 statistic which conducts a significance test on the relationship between the values of an attribute and the categories [2].

Hisao Ishibuchi first fuzzify the concept of association rules. That is, He showed fuzzy versions of two measures (i.e., confidence and support) that are used for evaluating each association rule in the field of data mining. Then he illustrated these two measures of fuzzy rules for function approximation and pattem classification problems. Finally he examined the relation between the classification performance of fuzzy rules and the definition of their certainty grades through computer simulations. Simulation results show that the direct use of confidence as a certainty grade is not always appropriate from the viewpoint of classification performance [9].

SOM–based discretization method tries to preserve the original sample distribution. Supervised discretization: In association rule mining and, more generally, in data mining the emphasis is not on predictive accuracy but rather in discovering unknown and useful patterns. When coping with classification problems, each record in the dataset contains a label-attribute so as to indicate the class it belongs to and this information is widely used during the supervised discretization of the other attributes. On the contrary, in association rule mining there is almost never a class-attribute and records are not labelled, thus supervised discretization is not applicable. Unsupervised discretization: Unsupervised discretization methods are generally based on the distribution of attribute values. The simplest and most used discretization method divides the range of observed attribute values into k equal sized intervals. In [2] the optimal number of intervals to create is established, given an arbitrary measure of the maximum information lost due to the discretization, the so–called partial completeness. It is also demonstrated that, for any given number of intervals, Equal Width (EW) partitioning minimizes the partial completeness. A related method, the equal frequency (EF) interval, divides the continuous attribute into k intervals where, given m instances, each interval contains m/k values [7].

Classification is an important topic in data mining research. To better handle continuous data, fuzzy sets are used to represent interval events in the domains of continuous attributes, allowing continuous data lying on the interval boundaries to partially belong to multiple intervals. Since the membership functions of fuzzy sets can profoundly affect the performance of the models or rules discovered, the determination of membership functions or fuzzy partitioning is crucial. In this paper, we present a new method to determine the membership functions of fuzzy sets directly from data to maximize the class-attribute interdependence and, hence, improve the classification results[3].

Interval Distance-based discretization method considers the order of the output variable and can work with ordinal output variables with a large number of different values as well as with continuous variables. The IDD is neither a bottom-up nor a top-down method, but one which, unlike the usual supervised discretization techniques, finds the cutpoints in a single step, dramatically improving computational speed with respect to other techniques[11].
RST method can reduce the number of cut points and improve efficiency of reduction. Adopting variable precision rough information entropy as measure criterion, it has a good tolerance to noise. Experiments show that the algorithm yields satisfying reduction results[4].

B. *Dealing with continuous attributes without discritization:*

A method using a new graph-based evolutionary algorithm named ‘genetic network programming (GNP)’ deals with continuous values directly, that is, without using any discretization method as a preprocessing step. GNP represents its individuals using graph structures and evolves them in order to find a solution; this feature contributes to creating very compact programs and implicitly memorizing past action sequences. A method is proposed that can deal with continuous attributes, where attributes in databases correspond to judgment nodes in GNP, each continuous attribute is checked for whether its value is greater than a threshold value and association rules are represented as the connections of the judgment nodes. In addition, the threshold value is evolved by mutation in order to obtain as many association rules as possible. Then fuzzy membership is added to it [5][13].

A novel form of association rules (ARs) is introduced that do not require discretization of continuous variables or the use of intervals in either sides of the rule. This rule form captures nonlinear relationships among variables, and provides an alternative pattern representation for mining essential relations hidden in a given data set. This new rule form is referred as a functional AR (FAR). A new neural network based, co-operative, coevolutionary algorithm is presented for FAR mining.

IV. RESULT ANALYSIS

A series of comparison experiments is conducted with the two state-of-the-art continuous variable ARM approaches without discretization: GAR mining algorithm [4] and the MODENAR [5]. These two algorithms are also based on evolutionary computation; the ARs are derived from the optimization process without the preprocessing of discretization. Compared with CCFARM, which works on FARs, these two algorithms aim at mining interval-based AR forms. Since the basic rule forms are different, we only carried out the comparison using three metrics: accuracy, data set coverage, and rule size.

CCFARM is again tested with three different ANN structures. In addition, we conduct one-tailed paired t-tests between the results derived from CCFARM and the results from MODENAR with significance level 0.05. It is shown that the FARs extracted by CCFARM have higher average accuracy value compared with those mined by MODENAR. In Table VII, the best performances are highlighted with bold text.

V. CONCLUSION

Discretization, also named quantization, is the process by which a continuous attribute is transformed into a finite number of intervals associated with a discrete value. The importance of discretization methods stems from interest in extending to continuous variable classification methods, such as decision trees or Bayesian networks, which were designed to work on discrete variables. The continuous variables can also be handled without discretization.

VI. ACKNOWLEDGMENT

The authors would like to thank the publishers and researchers for making their resources available. We also thank the college authority for providing the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

REFERENCES

[1]Randy Kerber,"ChiMerge: Discretization of numeric attributes,"

[2] Liu, Huan, and Rudy Setiono. "Feature selection via discretization." *IEEE Transactions on knowledge and Data Engineering* 9.4 (1997): 642-645.

[3] Au, W-H., Keith CC Chan, and Andrew KC Wong. "A fuzzy approach to partitioning continuous attributes for classification." *IEEE Transactions on Knowledge and Data Engineering* 18.5 (2006): 715-719.

[4] Cang, Rong, et al. "New method for discretization of continuous attributes in rough set theory." *Journal of Systems Engineering and Electronics* 21.2 (2010): 250-253.

[5] Taboada, Karla, et al. "Association rules mining for handling continuous attributes using genetic network programming and fuzzy membership functions." *SICE, 2007 Annual Conference*. IEEE, 2007.

[6]Fayyad, Usama M., and Keki B. Irani. "On the handling in decision tree of continuous-valued attributes generation." *Machine Learning* 8.1 (1992): 87-102.

[7] Vannucci, Marco, and Valentina Colla. "Meaningful discretization of continuous features for association rules mining by means of a SOM." *ESANN*. 2004.

[8] Ishibuchi, Hisao, Tomoharu Nakashima, and Takashi Yamamoto. "Fuzzy association rules for handling continuous attributes." *Industrial Electronics, 2001. Proceedings. ISIE 2001. IEEE International Symposium on*. Vol. 1. IEEE, 2001.

[9] H. Ishibuchi, T. Nakashima, and T. Yamamoto, "Fuzzy association rules for handling continuous attributes," in Proc. IEEE Int. Symp. Ind. Electron., vol a, 2001, pp. 118–121.

[10] Bay, Stephen D. "Multivariate discretization of continuous variables for set mining." *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000.

[11] Ruiz, Francisco J., Cecilio Angulo, and Núria Agell. "IDD: a supervised interval distance-based method for discretization." *IEEE Transactions on knowledge and data engineering* 20.9 (2008): 1230-1238.

[12] Taboada, Karla, et al. "Association rule mining for continuous attributes using genetic network programming." *IEEJ Transactions on Electrical and Electronic Engineering* 3.2 (2008): 199-211.

[13] K. Taboada, E. Gonzales, K. Shimada, S. Mabu, K. Hiragana, and J. Hu, "Association rule mining for continuous attributes using genetic network programming," IEEJ Trans. Elect. Electron. Eng., vol. 3, no. 2, pp. 199–211, Mar. 2008.

[14] B. Alatas, E. Akin, and A. Karci, "MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules," Appl. Soft Comput., vol. 8, no. 1, pp. 646–656, 2008.

[15] A. H. Sung, "Ranking importance of input parameters of neural networks," Expert Syst. Appl., vol. 15, pp. 405–411, Oct./Nov. 1998.

[16] D. K. Y. Chiu and T. W. H. Lui, "NHOP: A nested associative pattern for analysis of consensus sequence ensembles," IEEE Trans. Knowl. Data Eng., vol. 25, no. 10, pp. 2314–2324, Oct. 2013.