ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Improved Discretization and Decision Tree Approach For Continues Attributes Dataset

Madda Sukumar
M.Tech Student ,
Computer Science Engineering Department,
Sir C.R.Reddy College of Engineering.

Mr.M.Ganesh Babu, M.Tech
Asst Professor,
Computer Science Engineering Department,
Sir C.R.Reddy College of Engineering.

Abstract:

Data mining, an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Discretization concerns the process of transferring continuous models and equations into discrete counterparts.

This process is usually carried out as a first step toward making them suitable for numerical evaluation and implementation on digital computers. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

It is one way to display an algorithm. However, data in solid world are sometimes continuous by nature. Even for algorithms that will directly encounter continuous features, learning is most often ineffective and effective. Hence discretization addresses this problem by finding the intervals of numbers which happen to be more concise to represent and specify.

Discretization of continuous attributes is one of the important data preprocessing steps of knowledge extraction. The proposed improved discretization approach significantly reduces the IO cost and alsorequires one time sorting for numerical attributes which leads to a better performance in time dimension on rule mining algorithms.

According to the experimental results, our algorithm acquires less execution time over the Entropy based algorithm and also adoptable for any attribute selection method by which the accuracy of rule mining is improved.

Keywords:

Decision Tree Analysis, Discretization, Preprocessing, Data Mining, Machine learning

Introduction:

Data mining involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records, unusual records and dependencies. This usually involves using database techniques such as spatial indices.

These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. It bridges the gap from applied statistics and artificial intelligence to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Pre-processing:

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

Data mining involves six common classes of tasks:

Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

Regression – attempts to find a function which models the data with the least error.

Summarization – providing a more compact representation of the data set, including visualization and report generation.

Decision Tree:

In decision analysis a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of 3 types of nodes:

Decision nodes - commonly represented by squares
Chance nodes - represented by circles
End nodes - represented by triangles

Advantages of decision trees:

- Are simple to understand and interpret. People are able to understand decision tree models after a brief explanation.
- Have value even with little hard data. Important insights can be generated based on experts describing a situation (its alternatives, probabilities, and costs) and their preferences for outcomes.
- Allow the addition of new possible scenarios.
- Help determine worst, best and expected values for different scenarios.
- Use a white box model. If a given result is provided by a model.
- Can be combined with other decision techniques. The following example uses Net Present Value calculations, PERT 3-point estimations (decision #1) and a linear distribution of expected outcomes (decision #2).

Disadvantages of decision trees:

- For data including categorical variables with different number of levels, information gain in decision trees are biased in favor of those attributes with more levels.
- Calculations can get very complex particularly if many values are uncertain and/or if many outcomes are linked.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Related Work And Background:

The discretization of a continuous-valued attribute consists of transforming it into a finite number of intervals and to re-encode, for all instances, each value on this attribute by associating it with its corresponding interval. There are many ways to realize this process.

One of these ways consists in realizing a discretization with a fixed number of intervals. In this situation, the user must a priori choose the appropriate number: too many intervals will be unsuited to the learning problem and too few intervals can risk losing some interesting information. A continuous attribute can be divided in intervals of equal width (figure 1) or equal frequency (figure 2). Other methods exist to constitute the intervals, for example based on the clustering principles –e.g., K-means clustering discretization (Monti & Cooper, 1999).

Figure 1 Equal Width Discretization

Figure 2 Equal Frequency Discretization

Nevertheless, for supervised learning, these discretization methods ignore an important source of information: the instance labels of the class attribute. By contrast, the supervised discretization methods handle the class label repartition to achieve the different cuts and find the more appropriate intervals.

The figure 3 shows a situation where it is more efficient to have only 2 intervals for the continuous attribute instead of 3: it is not relevant to separate two bordering intervals if they are composed of the same class data. Therefore, the supervised or unsupervised quality of a discretization method is an important criterion to take into consideration.

Another important criterion to qualify a method is the fact that a discretization either processes on the different attributes one by one or takes into account the whole set of attributes for doing a overall cutting. The second case, called "multivariate discretization", is particularly interesting when some interactions exist between the different attributes.

On figure 4, a supervised discretization attempts to find the correct cuts by taking into account only one attribute independently of the others. This will fail: it is necessary to represent the data with the attributes X1 and X2 together to find the appropriate intervals on each attribute.

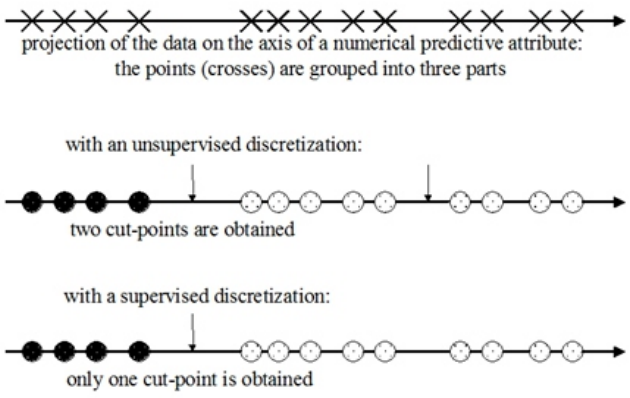


Figure 3 Supervised and Unsupervised Discretizations

Univariate Unsupervised Discretization:

The simplest discretization methods make no use of the instance labels of the class attribute. For example, the equal width interval binning consists of observing.

the values of the dataset, to identify the minimum and the maximum values observed, and to divide the continuous attribute into the number of intervals chosen by the user (figure 1).

Nevertheless, in this situation, if uncharacteric extreme values exist in the dataset ("outliers"), the range will be changed and the intervals will be misappropriate. To avoid this problem, it is possible to divide the continuous attribute in intervals containing the same number of instances (figure 2): this method is called equal frequency discretization method.

The unsupervised discretization can be grasped as a problem of sorting and separating intermingled probability laws (Potzelberger & Felsenstein, 1993).

The existence of an optimum analysis was studied by Teicher (1963) and Yakowitz and Spragins (1963).

Nevertheless, these methods are limited in their application in data mining due to too strong statistical hypotheses seldom checked with real data.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Univariate Supervised Discretization:

To improve the quality of a discretization in supervised data mining methods, it is important to take into account the instance labels of the class attribute. The figure 3 shows the problem of constituting intervals without the information of the class attribute. The intervals that are the better adapted to a discrete machine learning method are the "pure" intervals containing only instances of a given class. To obtain such intervals, the supervised discretization methods –such as the state-of-the-art method MDLPC– are based on statistical or information-theoretical criteria and heuristics (Fayyad & Irani, 1993).In a particular case, even if one supervised method can give better results than another (Kurgan & Krysztof, 2004) however, with real data, the improvements of one method compared to the others supervised methods are insignificant. Moreover, the performance of a discretization method is difficult to estimate without a learning algorithm. In addition, the final results can arise from the discretization processing, the learning processing or the combination of both.

Because the discretization is realized in an ad hoc way, independently of the learning algorithm characteristics, there is no guarantee that the interval cut will be optimal for the learning method. Only a little work showed the relevance and the optimality of the global discretization for very specific classifier such as naive Bayes (Hsu, Huang & Wong, 2003; Yang & Webb, 2003).

The supervised discretization methods can be distinguished depending on the way the algorithm proceeds: bottom-up (each value represents an interval and they are merged progressively to constitute the appropriate number of intervals) or top-down (the whole dataset represents an interval and it is progressively cut to constitute the appropriate number of intervals). However they are no significant performance differences between these two latest approaches (Zighed, Rakotomalala & Feschet, 1997).

Multivariate Unsupervised Discretization:

Association rules are an unsupervised learning method that needs discrete attributes. For such a method, the discretization of a continuous attribute can be realized in an univariate way but also in a multivariate way.

In the latter case, each attribute is cut in relation to the other attributes of the database, this approach can then provide some interesting improvements when unsupervised univariate discretization methods do not yield satisfactory results.

The multivariate unsupervised discretizations can be performed by clustering techniques using all attributes globally. It is also possible to consider each cluster obtained as a class and improve the discretization quality by using (univariate) supervised discretization methods (Chmielewski & Grzymala-Busse, 1996).

An approach called multi-supervised discretization (Ludl & Widmer, 2000a) can be seen as a particular unsupervised multivariate discretization. This method starts with the temporary univariate discretization of all attributes. Then, the final cutting of a given attribute is based on the univariate supervised discretization of all others attributes previously and temporarily discretized. These attributes play the role of a class attribute one after another. Finally, the smallest intervals are merged.

For supervised learning problems, a paving of the representation space can be done by cutting each continuous attribute into intervals. The discretization process consists in merging the bordering intervals in which the data distribution is the same (Bay, 2001). Nevertheless, even if this strategy can introduce the class attribute in the discretization process, it can not give a particular role to the class attribute and can induce the discretization to non-impressive results in the predictive model.

Multivariate Supervised Discretization:

When the learning problem is supervised and the instance labels are scattered in the representation space with interactions between the continuous predictive attributes (as presented on figure 4), the methods previously seen will not give satisfactory results. HyperCluster Finder is a method that will fix this problem by combining the advantages of the supervised and multivariate approaches (Muhlenbach & Rakotomalala, 2002). This method is based on clusters constituted as sets of same class instances that are closed in the representation space. The clusters are identified on a multivariate and supervised way: First, a neighborhood graph is built by using all predictive attributes to

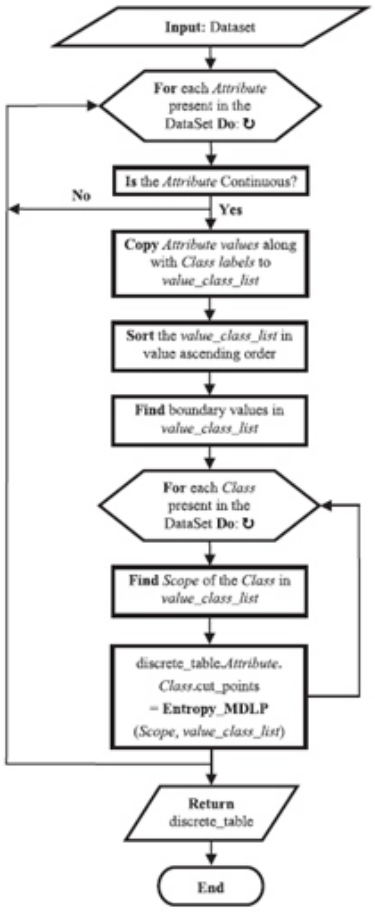
ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

determine which instances are close to others; Second, the edges connecting two instances belonging to different classes are cut on the graph to constitute the clusters; Third, the minimal and maximal values of each relevant cluster are used as cut-points on each predictive attribute.

The intervals found by this method have the characteristic to be "pure" on a pavement of the whole representation space even if the purity is not guaranteed for an independent attribute; It is the combination of all predictive attribute intervals that will provide pure areas in the representation space.

Proposed Method:

The technique is essentially preprocessing since all the cut points are discovered and stored in a discretization table prior to the start of the learning phase. The technique is compared with four other state-of-the-art discretization methods used as preprocessing discretization procedures.



Algorithm: Improved Discretization method.
Attributes:Ai
Input:
N, number of examples.
Ai, continuous attributes.
Cj, class values in training set.
Global Threshold value
Output: Interval borders in Ai
Procedure:
1. **for** each continuous attribute Ai in training dataset **do**
2. Do normalize the attribute within 0-1 range
3. Sorting the values of continuous attribute Ai in ascending order.
4. **for** each class Cj in training dataset **do**
5. Find the minimum (Minvalue) using StdDev attribute value of Ai for Cj
6. Find the maximum (Max) attribute value of Ai for Cj.
7. **endfor**
8. Find the cut points in the continuous attributes values based on the Min and Max values of each class Cj.
Best Cutpoint range measure:
9. Find the conditional probability P(Cj/A) on each cut point and select the cut point with maximum probability value.
Stopping criteria:
10. **If** the cut point using the maximum probability value is exist and satisfies the global threshold value then it can be taken as an interval border else consider the next cut point, where information gain value and global threshold value satisfy the same point.
12. **endfor**

Improved Decision tree measure:

Modified Information or entropy is given as

$$\text{ModInfo}(D) = -S_i \sum_{i=1}^{m} l \log \sqrt[3]{S_i} \quad \text{,m different classes}$$

$$= -S_1 \log \sqrt[3]{S_1} + S_2 \log \sqrt[3]{S_2}$$

Where S_1 indicates set of samples which belongs to target class 'anamoly', S_2 indicates set of samples which belongs to target class 'normal'.

Information or Entropy to each attribute is calculated using

$$Info_A(D) = \sum_{i=1}^{v} |D_i| / |D| \times ModInfo(D_i)$$

The term Di /D acts as the weight of the jth partition. ModInfo(D) is the expected information required to classify a tuple from D based on the partitioning by A.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

CONCLUSION AND FUTURE SCOPE:

Discretization of continuous features plays an important role in data pre-processing. This paper briefly introduces that the generation of the problem of discretization brings many benefits including improving the algorithms' efficiency and expanding their application scope. There have been drawbacks in the existing literature to classify discretization methods. The idea and drawbacks of some typical methods are expressed in details by supervised or unsupervised category.

Proposed Improved discretization approach significantly reduces the IO cost and also requires one time sorting for numerical attributes which leads to a better performance in time dimension on rule mining algorithms. According to the experimental results, our algorithm acquires less execution time over the Entropy based algorithm and also adoptable for any attribute selection method by which the accuracy of rule mining is improved.

REFERENCES:

[1] Khurram Shehzad, EDISC: A Class-Tailored Discretization Technique for Rule-Based Classification IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 8, AUGUST 2012.

[2] A DISCRETIZATION ALGORITHM BASED ON GINI CRITERION XIAO-HANG ZHANG, JUN WU, TING-JIE LU, YUAN JIANG, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007.

[3] A Novel Multivariate Discretization Method for Mining Association Rules Hantian Wei, 2009 AsiaPacific Conference on Information Processing.

[4] A Rule-Based Classification Algorithm for Uncertain Data, IEEE International Conference on Data Engineering.

[5] M. C. Ludl, G. Widmer. Relative unsupervised discretization for association rule mining. In Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, Springer, 2000.

[6] Stephen D. Bay. Multivariate discretization for set mining. Knowledge and Information Systems, 2001, 3(4): 491-512.

[7] Stephen D. Bay and Michael J. Pazzani. Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery, 2001, 5(3): 213-246.

[8] CAIM Discretization Algorithm Lukasz A. Kurgan.

[9] Effective Supervised Discretization for Classification based on Correlation Maximization Qiusha Zhu, Lin Lin, Mei-Ling Shyu.

[10] X.S.Li, D.Y.Li. A New Method Based on Density Clustering for Discretization of Continuous Attributes, Journal of System Simulation, 15(6):804-806,813,2005.

[11]: R.Kass, L.Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwa z discriterion, Journal of the American Statistical Association, Vol.90:928-935, 1995.

[12]: Comparative Analysis of Supervised and Unsupervised Discretization Techniques Rajashree Dash.

[13]Bay, S.D. (2001). Multivariate Discretization for Set Mining. Knowledge and Information Systems, 3(4), 491-512.

[14]Chmielewski M.R., & Grzymala-Busse J.W. (1994). Global Discretization of Continuous Attributes as Preprocessing for Machine Learning. Proceedings of the 3rd International Workshop on Rough Sets and Soft Computing, 294-301.