A Discretization Algorithm Based on Extended Gini Criterion

Yasmin Mohd Yacob, Harsa Amylia Mat Sakim, and Nor Ashidi Mat Isa

***Abstract*—A supervised, global and static algorithm for the discretization of continuous features is presented in this paper. The proposed discretization algorithm extends Gini-criterion concept. The algorithm hybridizes binning, and entropy method based on the amendment from 1R and Ent-Minimum Description Length Principle (MDLP) method. The proposed algorithm is comparable due to its simplicity, yet need further improvement and testing using various data sets. The proposed algorithm can be a good alternative method in the entropy-based discretization family.**

***Index Terms*—Continuous, discretization, entropy, Gini criterion.**

I. Introduction

Data sets conversion of continuous features into discrete attributes is known as discretization [1]. Many algorithms in the machine learning studies are developed for nominal or categorical data. Nevertheless, in reality, a lot of the data often involve continuous values. Those features need to be discretized before utilizing such an algorithm. Data originate in a mixed form which is nominal, discrete and/or continuous. Discrete and continuous data are ordinal data types, which are ordered values. Meanwhile nominal features do not exhibit any order among values. Nominal data are considered categorical data without any sequencing [2]. Discretization transformed continuous features into a finite number of intervals. Discretized intervals can be treated as ordinal values during induction and deduction. In addition, discretization makes learning more accurate and faster [1].

The discretization methods can be classified based on three dichotomies, which are supervised versus unsupervised, global versus local and static versus dynamic. Supervised methods perform attribute discretization indicated from class information, whereas unsupervised methods do not rely on class information. Local methods discretize attributes in a subset of instances while global discretization methods utilized the entire instance space [1]. Therefore, local method is usually associated with dynamic discretization whereby only part of the region space is applied for discretization. Global methods partitioned each feature into regions independent of the other attributes [3]. A dynamic discretization method discretizes continuous values when a classifier is being built, while static approach discretizes before the classification task [1]. The static method, such as entropy-based MDLP, carried out discretization on each feature and discovers the maximum number of interval for each attribute independent of the other feature. On the other hand, the dynamic method exercises discretization by searching possible *n* values throughout the whole space for all features simultaneously.

According to Liu et al., discretization's framework can be classified into splitting and merging methods [1]. Splitting algorithm consists of four steps, which are sorting the feature value, search for suitable cut point, split the range of continuous value according to cut point and stop when a stopping criterion is met, otherwise repeat the second step. Splitting methods are categorized into four types of method, which are binning, entropy, dependency and accuracy. To name a few, some of the algorithms developed under binning are Equal Width, Equal Frequency, and 1R [4], and entropy-based methods, which are ID3 [5], D2 [6], Minimum Description Length Principle (MDLP) [7] and Mantaras distance [8]. Meanwhile, Zeta [9] and Adaptive Quantizer [10] are algorithms under Dependency and Accuracy methods respectively.

On the other hand, merging methods follow four steps, which are sorting the value, finding the best two neighbouring intervals, merging the pair into one interval and terminates when the chosen stopping criteria satisfies. Some of the algorithms under merging methods are χ^2, ChiMerge [11] and Chi2 [12].

This paper proposed a discretization method which is supervised, global and static. It is a hybrid method between binning and entropy whereby it combines 1R Holte and entropy based method named Gini. Many algorithms in machine learning studies, including feature selection are developed for discrete data. Therefore, discretization is essential in the area of machine learning. The proposed discretization measure considers class information by extending Gini-criterion studies and also performs a natural stopping. In addition, it is also based on splitting method. This study also proposed a threshold measure to choose the good splitting point based on extended Gini's criterion.

II. Algorithm

This section presents the discretization notations, measure, description of the discretization process and a complete discretization algorithm.

A. Notations and Definitions

Suppose there is a supervised classification task with k

Manuscript received May 1, 2012; revised May 23, 2012. This work was supported in part by the Universiti Sains Malaysia Research University grant No. 814082 and Universiti Sains Malaysia Postgraduate Research Scheme No. 1001/PELECT/8042009.

Authors are with the School of Electrical and Electronic Engineering, Universiti Sains Malaysia, 14300 Penang, Malaysia (e-mail: yasmin160273@ yahoo.com; amylia@eng.usm.my; ashidi@eng.usm.my).

class labels and the training data set S consists of |S| instances. Let A be any of continuous attributes, there exists a discretization scheme D on the attribute A. Let [a,b] be the value domain of a continuous attribute A. Discretization of A is a process of searching for an optimal cut off points set $B_D = \{c_1, c_2, \ldots, c_n\}$, where $a < c_1, c_2, \ldots, c_{n-1} < b$, which divides [a,b] into n disjoint intervals, $I_D = [a, c_1], (c_1, c_2], \ldots, (c_n, b]$.

B. Measure

For the ith interval $(b_{i-1}, b_i]$, a class conditional probability $p(i) = (p_1^{(i)}, \ldots p_j^{(i)}, \ldots, p_k^{(i)})$ can be generated, where p_j is the jth class probability in ith interval and satisfies

$$\sum_{j=1}^{k} p_j^{(i)} = 1$$

The basic Gini equation is as follows:

$$G(\text{interval } i) = 1 - \sum_{j=1}^{k} (p_j^{(i)})^2$$

To decide the suitable cut off points for each attribute A, each cut off point is computed using Gini Gain formula:

$$\text{Gini Gain}(A;T,S) = \text{Gini}(S) - (\frac{|S1|}{S} * \text{Gini}(S1) + \frac{|S2|}{S} * \text{Gini}(S2)). \quad (1)$$

Results from [13] showed that in most experiments, Normalized Gain works better than Information Gain and Gain Ratio. Therefore, the study proposed the Gini Gain formula to be extended to Normalized Gini Gain.

$$\text{Normalized Gini Gain}(A;T,S) = \frac{\text{Gini Gain}(A;T,S)}{\log_2 n},$$

where n is the number of partition. (2)

In the proposed algorithm, the stopping criterion is used as a threshold measure to choose good cut off points or boundary points.

$$\text{Normalized Gini Gain} > \frac{\log_2(N-1)}{N} + \frac{\Delta(A;T,S)}{N}$$

where N is the number of instances and

$$\Delta(A;T,S) = \log_2(3^k - 2) - [k * \text{Gini}(S) - (k_1 * \text{Gini}(S_1)) - (k_2 * \text{Gini}(S_2))],$$

where k_i is number of class labels represented in set S_i. (3)

A similar idea as above, the study proposed normalized Gini in the threshold measure described as follows:

$$\text{Normalized Gini Threshold Measure} = \frac{[\frac{\log_2(N-1)}{N} + \frac{\Delta(A;T,S)}{N}]}{\log_2 n},$$

where n is the number of partition. (4)

Therefore, if a cutoff point computed from Normalized Gini Gain is greater than Normalized Gini Threshold Measure, then the point is a good cut off point and is chosen or selected.

C. Description of Discretization Process

The differences between traditional 1R and extended Gini's discretization algorithm are described in Fig. 1 and Fig. 2 below.

Cut off point

Attr. i	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9
Class i	1	1	1	1	1	1	2	2	2

Fig. 1. Selection of cut off point in traditional 1R and default for extended Gini discretization method

1R cut off point Extended Gini cut off point

Attr. i+1	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9
Class i	1	1	1	1	1	1	1	1	2

Fig. 2. Selection of cut off point for extended Gini discretization method

D. Algorithm

The proposed algorithm employs a hybrid method between binning and entropy. It applies 1R and entropy-based MDLP with some amendment in the 1R splitting method and use of Normalized Gini instead of traditional entropy formula. The algorithm followed 1R splitting method by separating the range of continuous values into a number of disjoint intervals and adjust the boundaries based on the class labels associated with the continuous values. Each interval should contain a minimum of six instances, and if it cuts the class interval, the algorithm will find the next nearest class separator. Initially, there are a lot of cut- off points and to ensure the validity and goodness of the points; the Normalized Gini Measure will determine the suitable cut-off points. This algorithm has a natural stopping and based on splitting method.

To perform discretization of the attributes, the pseudo code of the extended Gini-based discretization algorithm is described in Fig. 3.

```
Compute Parent Gini
For each attribute
    Sort instances in ascending order
    While not end of instances
        By default the algorithm follows 1R
        Splitting method by having minimum 6 instances
        If cut off class then extend to the next instance
        until find next class separator
        EndIf
        Compute Equation (1) and Equation (2)
        Compute Equation (3)
        Compute Equation (4)
```

```
        If Normalized Gini Gain >
                Normalized Gini Threshold Measure
            Choose cut off points
        EndIf
    EndWhile
EndFor
```

Fig. 3. The pseudo code of Extended Gini discretization algorithm

III. Experiments

This section presents and analyses the experimental results. Table I show the information of data sets chosen from the UCI repository [14]. The number of instances of these data sets ranges from 150 to 768, purposed to test the algorithm's ability to adapt to various conditions. In fact, the number of attributes tested in the experiment ranges from small, medium and large sizes. The comparison of results for continuous, Ent-MDLP and Gini's experiments were taken from [1] and [3], which run using classifier C4.5. The experiments also compared with Ent-MDLP because it has been accepted as one of the best discretization methods. All the experiments are based on 10-fold cross validation. In the experiment for the Gini's method, both Diabetes and Wine data set are executed using classifier C5.0. This is because the C4.5 classifier is obsolete [15]. The proposed extended Gini's discretization method is executed using classifier C5.0.

TABLE I: Data Sets Information

Data sets	Instances	Attribute	Class
Iris	150	4	3
Thyroid	215	5	3
Diabetes	768	8	2
Breast Cancer	699	9	2
Wine	178	13	3
WDBC	569	30	2

Referring to Table II, comparison of results between the extended Gini's discretization and continuous data set showed comparable error rate for Iris, Thyroid, Diabetes and WDBC data set. Nevertheless, the error rate is twice as much as Breast Cancer and Wine data set when compared with continuous data set. Meanwhile, the extended Gini's discretization algorithm is comparable with Ent-MDLP for Iris, Thyroid, Diabetes and Wine data set. On the other hand, the results from Breast Cancer and WDBC data set are not comparable between discretization scheme of extended Gini and Ent-MDLP.

TABLE II: Comparisons of Classification Error Rate

Data sets	Continuous	Ent-MDLP	Gini	Ext-Gini
Iris	4.63 ± 6.33	5.99 ± 6.61	5.35 ± 4.22	6.00 ± 1.60
Thyroid	6.39 ± 4.42	4.24 ± 4.18	3.21 ± 3.81	7.00 ± 2.00
Diabetes	26.22 ± 2.65	25.21 ± 4.23	23.80 ± 2.34	29.00 ± 0.40
Breast Cancer	4.74 ± 3.82	4.56 ± 3.36	4.03 ± 2.83	9.30 ± 0.80
Wine	6.22 ± 6.84	7.95 ± 7.27	7.31 ± 1.70	13.50 ± 3.00
WDBC	5.98 ± 3.10	4.76 ± 3.10	4.41 ± 3.63	8.30 ± 1.20

Almost all the results from the proposed extended Gini algorithm do not show significant improvement from the Gini-based discretization algorithm [3]. Nevertheless, it is neither a failure. The extended Gini's discretization method resulted in a 6.00 error rate compared to 5.35 error rate for Iris's data set. Meanwhile for Diabetes data set, the error rate for the extended Gini's discretization method is 29.00 compared to 23.80 from the Gini's method. The Gini method resulted in 7.31 error rate compared to 13.50 in the extended Gini's study for the Wine data set. The Gini's error rate is nearly twice much lesser than the extended Gini's error rate in the WDBC data set. Regardless of the insignificant result; it is still considered comparable except for the Thyroid and Breast Cancer data set. In order to improve the algorithm performance, the extended Gini's threshold measure needs to be amended, and more data sets need to be run to get the overall result of the extended Gini's discretization algorithm. This paper captures the first run of the algorithm which is performed on six data sets.

IV. Conclusion

This paper proposes a discretization algorithm of continuous attributes based on extended Gini's criterion in a supervised, static and global manner. The algorithm hybridizes binning, and entropy method employed from 1R and entropy-based MDLP method. Normalized Gini Gain is applied as the threshold measure and is a class interdependent. Regardless of the insignificant preliminary results, the algorithm is still effective because of its simplicity feature to generate cut-off points. The proposed algorithm can be a better method once some amendment and further test run are conducted. To sum up, the proposed discretization algorithm can be a good alternative to the entropy-based discretization methods.

References

[1] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: An enabling technique," *Data Mining and Knowledge Discovery*, vol. 6, pp. 393-423, 2002.

[2] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *12th International Conference on Machine Learning*, Los Altos, California, 1995, pp. 194-202.

[3] Z. Xiao-Hang, W. Jun, L. Ting-Jie, and J. Yuan, "A Discretization algorithm based on Gini criterion," in *International Conference on Machine Learning and Cybernetics*, 2007, pp. 2557-2561.

[4] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63-90, 1993.

[5] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.

[6] J. Catlett, "On changing continuous attributes into ordered discrete attributes," in *5th European Working Session on Learning*, Berlin, 1991, pp. 164-177.

[7] U. Fayyad and K. Irani, "Multi-interval discretization of continuous attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1993, pp. 1022–1027.

[8] J. Cerquides and R.L. Mantaras, "Proposal and empirical comparison of a parallelizable distance-based discretization method," in *3rd International Conference on Knowledge Discovery and Data Mining (KDD1997)*, 1997, pp. 139-142.

[9] K. M. Ho and P. D. Scott, "Zeta: A global method for discretization of continuous variables," in *3rd International Conference of Knowledge*

Discovery and Data Mining (KDD 1997), Newport Beach, CA, 1997, pp. 191-194.
[10] C. C. Chan, C. Batur, and A. Srinivasan, "Determination of quantization in rule based model for dynamic systems," in Proceedings *of IEEE Conference on Systems, Man, and Cybernetics*, Charlottesvile, Virginia, 1991, pp. 1719-1723.
[11] R. Kerber, "Chimerge:Discretization of numeric attributes," in *9th* National *Conference Artificial Intelligence (AAAI92)*, 1992, pp. 123-128.
[12] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *7th IEEE International Conference on Tools with Artificial Intelligence,* Herndon, Virginia, 1995, pp. 388-391.
[13] B. H. Jun, C.S. Kim, H-Y. Song, and J. Kim, "A New criterion in selection and discretization of attributes for the generation of Decision Trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence, v*ol. 19, No. 12, 1997, pp. 1371-1375.
[14] C. L. Blake, and C.J. Mertz, UCI Repository of Machine Learning [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA:University of California, Department of Information and Computer Science, 1998.
[15] RuleQuest Research Data Mining Tools [http://www.rulequest.com], Updated: June 2010.

Yasmin Mohd Yacob received the bachelor's degree from University of Michigan, USA, and the MS degree from Universiti Putra Malaysia. She is currently a PhD candidate from School of Electrical and Electronic Engineering, Universiti Sains Malaysia. Her research interests include data mining, pattern recognition, image processing, and bio informatics. She is a student member of IEEE.

Associate Professor Dr. Harsa Amylia Mat Sakim received the Bachelor Engineering degree from University of Liverpool, the MS from University of Newcastle Upon Tyne, and PhD degree from Universiti Sains Malaysia. Dr. Harsa Amylia works in School of Electrical and Electronic Engineering at Universiti Sains Malaysia where she researches and teaches Signals and Systems, Control Systems, Instrumentation and Circuit Theory. She has published many papers in international journals including a book chapter related to Intelligent Systems especially in breast cancer studies. Her research interests include Artificial Intelligent, Biomedical Engineering and Medical Electronics and Instrumentation. She is a member of IEEE.

Associate Professor Dr. Nor Ashidi Mat Isa, B. Eng (USM), PhD (USM) received the B. Eng in Electrical and Electronic Engineering with First Class Honors from Universiti Sains Malaysia (USM) in 1999. In 2003, he received his PhD degree in Electronic Engineering (majoring in Artificial Neural Network) from the same university. He is currently lecturing at the School of Electrical and Electronics Engineering, USM. His research interests include intelligent system, image processing, neural network, Biomedical Engineering (i.e. intelligent diagnostic system) and algorithms. As of now, he and his research team (Imaging and Intelligent System Research Team, ISRT) managed to publish their works national and internationally. Their contribution can be found in numerous international and national journals, chapters in books, international and national proceedings.