International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

Comparison of the effect of unsupervised and supervised discretization methods on classification process

Mehmet HACIBEYOĞLU [1], Mohammed H. IBRAHIM*[2]

Accepted 3rd September 2016

Abstract: Most of the machine learning and data mining algorithms use discrete data for the classification process. But, most data in practice include continuous features. Therefore, a discretization pre-processing step is applied on these datasets before the classification. Discretization process converts continuous values to discrete values. In the literature, there are many methods used for discretization process. These methods are grouped as supervised and unsupervised methods according to whether a class information is used or not. In this paper, we used two unsupervised methods: Equal Width Interval (EW), Equal Frequency (EF) and one supervised method: Entropy Based (EB) discretization. In the experiments, a well-known 10 dataset from UCI (Machine Learning Repository) is used in order to compare the effect of the discretization methods on the classification. The results show that, Naive Bayes (NB), C4.5 and ID3 classification algorithms obtain higher accuracy with EB discretization method.

Keywords: *Discretization, Supervised and Unsupervised Discretization, Continuous Features, Discrete Feature, classification algorithms.*

1. Introduction

Many Machine Learning and Data Mining classification algorithms have application possibility only to the discrete data. But, most data in practice have continuous feature. The datasets usually contains mixed forms of nominal, discrete, and continuous data. Discrete values have intervals between a continuous series of values. The number of continuous values for an attribute can be endless, but the number of discrete values may be few or have an end value [1]. An example to continuous features is blood sugar content. Whereas an example of discrete features is gender. Process of converting continuous values to discrete values is called discretization. Discretization is generally used to sort and reshape continuous variables of attributes into categorized features. However, there are endless possibilities of discretization methods depending on the intervals which exist within domain. The idea of discretization is to divide the range of a numeric or ordinal attribute into intervals through user given or computed cut points. (Dougherty, J. et al; 1995) made a comparison between unsupervised discretization method (Equal width, Equal frequency) and supervised discretization method (Entropy-based, Purity-based) from classification accuracy point of view. They found that the classification accuracy of the classification algorithms (Naive-Bayes, C4.5) significantly improved when features of the datasets were discretise using an entropy-based discrete method. At (Hacibeyoglu, M. et al; 2011), the results of comparison between using discrete method and continuous method for six datasets showing that the performance of the classification accuracy is improved, when the features of datasets discretise. Studies for Cluster Algorithm also can be used as discretise values. (Gupta, A. et al; 2010) Proposed accounting the interdependencies among different attributes and discretise the data using minimum entropy with minimum description length as the stopping criteria, used K-mean clustering methods and shared nearest neighbour. While (Joiţa, D. 2010) used the K-means clustering algorithm for discretization.

In this study, we applied supervised EB and unsupervised EW and EF discretization methods on 10 UCI dataset with continuous values. These datasets are applied to classification algorithms that work on discrete features, and then the accuracy of classification methods are compared. At the next section, the discretization methods unsupervised EW & EF and supervised EB are explained. At section 3, the classification algorithms are given. The experimental results are shown at Section 4 and the paper is concluded at Section 5.

2. Discretization and classification of discretization methods

Data discretization process is a method aims to reduce the volume of distinct values of continuous variables through dividing its range into limited set of unrelated intervals and then relating these intervals by specific descriptive labels. Usually discretization steps are sorting continuous values, finding cut points and finally applying conversion process [1].

Discretization methods are categorized along different needs, discretization of continuous values to obtain higher accuracy rate on handling data with high cardinality. Main classification of discretization is as supervised and unsupervised. Generally the categorization of the supervised and unsupervised discretization techniques depends on class information, for example if the discretization process uses class information, then we say it is supervised discretization. Otherwise, it is called as unsupervised discretization [1], [2]–[3].

[1] *Computer Engineering Dep., Necmettin Erbakan Un., Konya, Turkey*
[2] *Computer Engineering Dep., Necmettin Erbakan Un., Konya, Turkey*
** Corresponding Author: Email: mibrahim@konya.edu.tr*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

2.1. Unsupervised Discretization Methods

The simple discretization (equal-width and equal-frequency interval binning) is among the unsupervised discretization methods and binning does not use class information. Continuous ranges are subdivided into smaller ranges through user specified width or frequency. Usually in the unsupervised discretization methods, the number of parts must be supplied by the user [1], [2] and [6].

2.1.1. Equal Width Interval Discretization

The simplest discretization method is Equal-width interval discretization which divides the range of observed values for a feature into equal sized bins represented by k that is a parameter supplied by the user [1] and [2]. The process involves finding the minimum and maximum observed values through sorting of a continuous feature, A = {a_0, a_1, a_2, ... a_{n-1}, a_n}, a_{min} = a_0 and a_{max} = a_n. A is a continuous value array. Computing the interval may be done by dividing the range of the observed values for the variable into equally sized bins using the following formula: [1], [2] and [5].

$$Interval = \frac{(a_{max} - a_{min})}{k} \quad (1)$$

$$Boundaries = a_{min} + (i * interval) \quad (2)$$

The boundaries can be formed by i = 1...k-1 using the above formula. Equal Width Interval discretization steps as shown in Figure 1. And these steps are given as an example in Table 1 below. Where I is the instance, T is the temperature, C are the classes value, A is the sorted value of T and D which is the discretized value of *T*.

> ***Input:*** data is the array having continuous values of the attribute, ***A*** = {a_0, a_1, a_2, ….. a_{n-1}, a_n} and ***k*** is the number of parts, where ***k*** > 0;
> ***Output:*** data having discrete values.
> Step 1: Sort the value *A* in increasing order
> Step 2: Calculate the interval by equation 1
> Step 3: Binning the data by boundaries formula and determine the split point for *A*
> Step 4: The value of attribute in the array must be placed in the same boundaries

Figure 1. Equal Width Interval discretization algorithms

Table 1. Example for Equal Width discretization

I	1	2	3	4	5	6	7	8	9	10
T	85	90	86	96	80	70	65	95	75	91
C	no	no	yes	yes	yes	no	yes	no	yes	yes
A	65	0	75	80	85	86	90	91	95	96
Here a_{min}= 65, a_{max}= 96 and let k= 3 So Interval = (96-65)/3=10.3 then Boundary = 65+10.3 = 75.3 0 = [65, 75.3); 1 = [75.3, 85.6); 2 = [85.6, 96.2]										
D	1	2	2	2	1	0	0	2	0	2

2.1.2. Equal Frequency Interval Discretization

In the equal-frequency discretization algorithm the minimum and maximum values are determined for discretized attribute, and then all values are sorted in ascending order, and sorted continuous values divided into k intervals in a way that each interval contains approximately n/k data instances with adjacent values, A = {a_0, a_1, a_2, ... a_{n-1}, a_n} where n is a number of element in A. And A is set of data with array having continuous values [1], [2], [5] and [7]. In proportional K–interval discretization method, the data instances with identical value must be placed in the same interval, so it is not always possible to generate exactly K equal frequency intervals [7]. Equal Frequency discretization steps as shown in the Figure 2. And these steps are given as an example in Table 2 below. Where I is the instance, T is the temperature, C is the classes value, A sorted value of T and D is a discretised value of T.

> **Input:** data is an array having continuous values, *A = {a_0, a_1, a_2, ….. a_{n-1}, a_n}* and *k* is the number of parts. Where *k > 0*;
> **Output:** data having discrete values.
> Step 1: Sort the value *A* in increasing order
> Step 2: Determine the split point for *A* and calculate the data instances in each interval by dividing the number of elements in the array by number of parts.
> Step 3: The value of an attribute in the array must be placed in the same boundaries

Figure 2. Equal Frequency discretization algorithms

Table 2. Example for Equal Frequency discretization

I	1	2	3	4	5	6	7	8	9	10
T	85	90	86	96	80	70	65	95	75	91
C	no	no	yes	yes	yes	no	yes	no	yes	yes
A	65	0	75	80	85	86	90	91	95	96
Here a_{min}= 65, a_{max}= 96, *n* = 10, and let *k*= 3 So *Interval* = 10/3=3.3 and around (interval) Then *interval* = 3; each binning contained approximately 3 element, the rise is added to the last part. 0 = [65, 80); 1 = [80, 90); 2 = [90, 96].										
D	1	2	1	2	1	0	0	2	0	2

2.2. Supervised Discretization Methods

This discretization method converts numerical data to their categorical counterparts and use class information while choosing discretization nodes. Entropy based discretization is an example of this method [1], [2] and [5].

2.2.1. Entropy Based Discretization

This method uses division approach. Entropy (or information content) is calculated based on class labels. Intuitively works towards purifying each data group by finding the best cut points, so each data in the pure groups will have the same class label as much as possible. It is characterized by finding intervals that gives the maximum information gained [1] and [8]. The formula to calculate Entropy is as follows;

$$E(S, A) = -\sum_{i=1}^{n} P_i \log_2(P_i) \quad (3)$$

$$P_i = {}^{v_i}/_{v} \quad (4)$$

Entropy means “impurity”. But the impurity here means: diversity, having too much data with different specifications in one group. When calculating entropy, we will first calculate entropy for estimated class value E(S). Then we will calculate entropy for specified cut points E(S, A). Finally, the best cut point by calculating information gained by the flowing equation is determined [1] and [8].

$$Information\ Gain = E(S) - E(S, A) \quad (5)$$

Entropy based discretization pseudo code is shown in Figure 3 below. All these steps are given as an example in Table 3 below.

Input: Data is an array having continuous values, $A = \{a_0, a_1, a_2, \ldots\ldots a_{n-1}, a_n\}$
Output: Data has discrete values.
Step 1: sort the value *A* in increasing order
Step 2: calculate entropy for class data
Step 3: determine the best split point for *A* by
- Calculate Entropy for each split point
- Calculate information gain for each split point

Step 4: select the split point with highest information gain
Step 5: recursively the partition on each split point when terminate once information gain falls below a certain threshold or once you reach a specified number of bins.

Figure 3. Entropy-Based discretization algorithms

Table 3. Example for Entropy-Based discretization

I	1	2	3	4	5	6	7	8	9		10
T	85	90	86	96	80	70	65	95	75		91
C	no	no	yes	yes	yes	no	yes	no	yes		yes
A	65	0	75	80	85	86	90	91	95		96
	No = 4, Yes = 6 probability No P_{no} = 4/10 = 0.4 probability Yes, P_{yes} = 6/10 = 0.6 Entropy (Play) = E (4, 6) = E (0.4, 0.6) = -0.4 * $\log_2$ (0.4) – 0.6 * $\log_2$ (0.6) = 0.970. Let k = 2 (80 and 86) then. For 1. Split <=80 = {65, 70, 75, 80} and > 80 = {85, 86, 90, 91, 95, 96} For 2. Split <=86 = {65, 70, 75, 80, 85, 86} and > 86 = {90, 91, 95, 96} Calculate entropy and information gain for each split then chose the split with the highest information gain. Entropy E = - Σ pi log2 (pi) Entropy for split 1 = 0.954 and Entropy for split 2 = 0.888. **Information Gain = E(S) – E(S, A)** Info for split 1 = 0.970 - 0.954 = 0.016 and Info for split 2 = 0.970 - 0.888 = 0.082 then 0 = [65, 86); 1 = [86, 96];										
D	1	2	1	2	1	0	0	2	0		2

3. Classification Algorithms

Classification concept is basically the distribution of the data among predefined various classes on a dataset [1]. Classification algorithms learn this distribution type from the given education cluster and they try to distribute data when unclassified test data are received. The values indicating these classes on the dataset are named as labels and they are used in order to denote the classes either for education and test. Some classification algorithms process according to categorical values (Decision Tree, Naive Bayes) while others process with numerical values (ANN) [1].

3.1. Decision Tree Algorithm

Decision trees are frequently used data mining approaches for classification and estimation. Despite the fact that other methods such as ANN can be used for classification, decision trees provide advantage of easy interpretation and comprehensibility to the decision makers [3]. Classification of the data using decision tree technique is a two-step process comprising learning and classification. A previously known education data is analyzed by the classification algorithm in order to constitute a model during the learning phase. The constituted model is shown as classification rules or a decision tree. During the classification phase, on the other hand, test data is utilized in order to determine the accuracy of the classification rules or a decision tree. If the accuracy is in an acceptable range then the rules are used for classification of new data. It should be determined that which domains will be used in what type of order to constitute the tree. Entropy metric is the most widely used measurement for this purpose. The results obtained using the domain are uncertain and instable proportionally to the entropy measurement in that domain. Thus the minimum entropy or maximum information gain measurement of the domains are used in the roots of the decision tree. Entropy can be mathematically defined as; if $(p_1, p_2, \ldots\ p_n) = \sum_{i=1}^{n} p_i \log_2 p_i$. $(p_1, p_2, \ldots\ p_n)$ Which states the probabilities, then the sum of all these probabilities should be exactly 1. Finally the information gain is calculated for each *p* via the following equation [9] and [10].

$$infgain\ (D,S) = H(D) - \sum_{i=1}^{n} P(D_i)H(D_i) \quad 6$$

In a decision tree, all paths from the root node to the leaf node proceed by way of AND [11]. There are multiple modules of the decision tree algorithms [10], ID3 and C4.5 modules are used in this paper.

3.2. Naive Bayes algorithm

Naive Bayes Classifier is an approach of probability, which can be used with a proposition that first seems to be very limiting in classification problems [10]. This proposition is the necessity of independency of each defining quality or parameter to be used in pattern identification in respect of statistics. No matter how this proposition is limiting the application field of Naïve Bayes Classifier, this approach yields results comparable to more complex methods such as ANN by stretching statistical independency condition.
Naive Bayes Classifier is the simplified state of Bayes theorem by means of the independency proposition. Bayes theorem is expressed with the following equation:

$$P(A \mid B) = \ (P(B \mid A)\ P(A))/P(B) \quad 7$$

$P(A|B)$; The probability of incident A when incident B is realized
$P(B|A)$; The probability of incident B when incident A is realized
$P(A)$ And $P(B)$; Prior probabilities of incidents A and B.
Here, prior probabilities add subjectivity to the Bayes theorem. In other words, for instance, $P(A)$ is the information that is obtained prior to any obtained data about incident A. On the other hand $P(B|A)$ is post probability because it gives information about the realization probability of incident B when incident A is realized after data collection [1] and [10].

4. The experimental results

In this study, using discretization methods of unsupervised (EW and EF) and supervised (EB) to convert an attribute's continuous value of UCI dataset of Table 4 into discrete values. Where *NS* is The Number of continuous Attributes, *ND* is The Number of discrete Attributes, *NI* is The Number of Instance and *NC* is The Number of Classes. Using Machine learning classification algorithms (NB, ID3, C4.5), we carried out the classification process on discrete datasets of converted values. Number of parts used in EW and EF discretization methods are (*k*= 3 To 22; *k*+= 2), the means (3, 5, 7, ... 21) are used, and at each *k* value classification performance is calculated, as a result, we have

taken the average of overall part numbers according to the high performance value of each classification algorithm, in this case k = 9, in Table 5 and 6 we have given classification accuracy according to the value of k. In EB discretization method, as we mentioned above, the number of parts is calculated according to the data information gain, each attribute of data set is divided into different number of parts. In Table 5, 6 and 7 classification results are given according to discretization methods used.

Table 4. Name and Properties of Datasets

Datasets name	*NS*	*ND*	*NI*	*NC*
Breast Cancer Wisconsin	9	0	699	2
Pima Indians Diabetes	8	0	768	2
Glass Identification	9	0	214	6
Iris	4	0	150	3
Stat log Heart	5	8	270	2
Australian Credit	6	8	690	2
German Credit	7	13	1000	2
E.coli	7	1	336	8

Table 5. Classification Accuracy of EW

Datasets name	*NB*	*ID3*	*J4.5*
Breast Cancer Wisconsin	97,42	93,82	95,42
Pima Indians Diabetes	75,91	51,69	75,13
Glass Identification	68,69	55,61	59,81
Iris	95,33	89,33	96,00
Stat log Heart	82,22	58,15	79,63
Australian Credit	85,80	67,54	86,52
German Credit	74,50	59,10	72,60
E.coli	85,71	58,63	74,11

Table 6. Classification Accuracy of EF

Datasets name	**NB**	**ID3**	**J4.5**
Breast Cancer Wisconsin	97,42	93,82	95,42
Pima Indians Diabetes	75,91	51,69	75,13
Glass Identification	68,69	55,61	59,81
Iris	95,33	89,33	96,00
Stat log Heart	82,22	58,15	79,63
Australian Credit	85,80	67,54	86,52
German Credit	74,50	59,10	72,60
E.coli	85,71	58,63	74,11

Table 7. Classification Accuracy of EB

Datasets name	*NB*	*ID3*	*J4.5*
Breast Cancer Wisconsin	97,00	84,39	94,99
Pima Indians Diabetes	77,86	77,08	78,26
Glass Identification	74,30	73,36	73,83
Iris	94,00	94,00	94,00
Stat log Heart	83,33	80,37	81,85
Australian Credit	85,51	75,36	87,65
German Credit	75,80	63,20	72,10
E.coli	87,46	85,02	86,24

The analysis of the results given in Tables V, VI and VII, from 8 datasets show that in EW method provides high classification accuracy for only Breast Cancer Wisconsin and Iris datasets and in EF method provides high classification accuracy Glass Identification and German Credit datasets while in EB method provides high classification accuracy for the remaining four datasets. In general supervised EB provides better accuracy then unsupervised EW and EF.

5. Conclusion

Discretization algorithms are known to have advantages and disadvantages relative to one another. Several new methods have been developed keeping these advantages and disadvantages in mind. As it is mentioned before; such as, in equal width discretization method, even if this bare discretization method is attractive among the others, there are reported opinions about major loss of data after discretization process, because of determining the gap width is not done correctly during the division into equal intervals. Other methods, such as the EW and EF require the user to enter the parameters while EB method determines the parameters itself. It is possible to say that, in this case, if user entered values are wrong parameters it could pose a problem in terms of results obtained. Data discretization process plays an important role in the data classification process, since it is the process of entropy based discretization data depends on the class of data in discretization data the discretization is relatively true, and it increases Probability of data classification is relatively. As a result of this study in categorization algorithms (NB, ID3, J4.5) which discretization method gives better performance than the others.

References

[1] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[2] Dougherty, J., Kohavi, R., & Sahami, M. (1995, July). Supervised and unsupervised discretization of continuous features. In Machine learning: proceedings of the twelfth international conference (Vol. 12, pp. 194-202).

[3] Hacibeyoglu, M., Arslan, A., & Kahramanli, S. (2011). Improving Classification Accuracy with Discretization on Data Sets Including Continuous Valued Features. Ionosphere, 34(351), 2.

[4] Gupta, A., Mehrotra, K. G., & Mohan, C. (2010). A clustering-based discretization for supervised learning. Statistics & probability letters, 80(9), 816-824.

[5] Joiţa, D. (2010). Unsupervised static discretization methods in data mining. Titu Maiorescu University, Bucharest, Romania.

[6] Gama, J., & Pinto, C. (2006, April). Discretization from data streams: applications to histograms and data mining. In Proceedings of the 2006 ACM symposium on Applied computing (pp. 662-667). ACM.

[7] Jiang, S. Y., Li, X., Zheng, Q., & Wang, L. X. (2009, May). Approximate equal frequency discretization method. In 2009 WRI Global Congress on Intelligent Systems (Vol. 3, pp. 514-518). IEEE.

[8] Agre, G., & Peev, S. (2002). On supervised and unsupervised discretization. Cybernetics and information technologies, 2(2), 43-57.

[9] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[10] HSSINA, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. Int. J. Adv. Comput. Sci. Appl, 4(2).