ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20170220135853
Original scientific paper

Improvement of the Accuracy of Prediction Using Unsupervised Discretization Method: Educational Data Set Case Study

Gabrijela DIMIĆ, Dejan RANČIĆ, Ivan MILENTIJEVIĆ, Petar SPALEVIĆ

Abstract: This paper presents a comparison of the efficacy of unsupervised and supervised discretization methods for educational data from blended learning environment. Naïve Bayes classifier was trained for each discretized data set and comparative analysis of prediction models was conducted. The research goal was to transform numeric features into maximum independent discrete values with minimum loss of information and reduction of classification error. Proposed unsupervised discretization method was based on the histogram distribution and implementation of oversampling technique. The main contribution of this research is improvement of accuracy prediction using the unsupervised discretization method which reduces the effect of ignoring class feature for educational data set.

Keywords: discretization; data mining; educational data set; entropy; equal width binning; histogram; machine learning; oversampling

1 INTRODUCTION

Effective prediction models demand a detailed approach to data discretization as a basic task in the pre-processing phase. In the field of machine learning and data mining there is a number of algorithms that are primarily oriented towards working with discrete values. In the real world, data are of mixed type, i.e. in most cases of continuous type. This is why it is necessary to integrate the application of machine learning and data mining algorithms with discretization methods in order to perform the transformation of continuous data. The main objective of the applied discretization method is certainly the reduction of value domains by dividing into discrete intervals while maximizing the independence of different discrete feature values and class marks [1]. Thus, continuous values are being transformed into an adequate set of discrete values more relevant for the interpretation [2, 3]. The advantages of using discrete values are related to the less memory requirements, intelligibility and simplicity by using meaningful marks, regulating discrepancy variations in the estimation of smaller fragmented data, reducing data quantity by identifying and removing redundant data, the algorithm accuracy and speed. A good discretization algorithm should balance the loss of information with the process of generating a reasonable number of split intervals for the adequate search space. The compromise must be found between the information quality (homogeneous intervals according to the prediction attribute) and statistic quality (sufficient size of instances in each interval for ensuring the generalization). Choosing an adequate discretization method implies obtaining a satisfactory compromise between these two objectives. Many studies indicate a positive effect of discretization on induction tasks: rules with discrete values are more intelligible, and higher accuracy is achieved in the cases of prediction and classification. The discretization effect can be measured in terms of accuracy, time necessary for performing learning algorithm and result intelligibility. There is a great number of different discretization methods described in the literature. Most methods apply iterative candidate space search using different rating functions for estimating results. The key question is not only whether one discretization method is more superior to another, but under what terms a certain method can achieve better performances for the given issue. Previous research of application of unsupervised and supervised discretization algorithms indicated that the supervised methods affect the achievement of better performance classification and prediction.

Discretization methods can be classified as supervised/unsupervised [4], hierarchical/non-hierarchical [5], top-down/bottom-up [6], static/dynamic [7], global/local [8], parametric/non-parametric and univariate/multivariate [9]. Unsupervised equal binning width, equal binning frequency, clustering algorithms like k-means methods imply splitting the continuous attribute values domains into sub–scopes, not taking the class information into account, or more accurately, on the basis of user-defined parameters. In supervised discretization methods, class information is used for finding adequate intervals by defining most optimal cut – points. Supervised discretization can use metrics based on the errors in the training data, the disparity measure, i.e. interval entropy or some statistic measures.

The research described in the paper is focused on determining the procedure for improving the efficacy of unsupervised discretization methods in the case of transformation of continuous educational dataset features.

Improved unsupervised discretization method is achieved applying the oversampling technique and randomize filter. Three experiments included in this survey were implemented:

1) Entropy – based discretization method with the minimal description length principle stopping criteria
2) Equal width binning unsupervised discretization method with dynamic search
3) Histogram graphs based on Sccot rule.

2 RELATED WORK

Researchers from the machine learning field have introduced a great number of discretization methods. The review of discretization algorithms can be found in [2]. In the paper [4], Dougherty et al. perform the comparative analysis of 5 discretization methods over 16 datasets from UC Irvine ML Database Repository [10]. Two

unsupervised global methods, two supervised global methods (OneR and entropy minimization) and C4.5 algorithm representing supervised local method have been applied. Ismail et al. [11] used the decision tree classifier (C4.5) on the discretized data and an error measure to determine the relative value of discretization. Authors showed that the general effectiveness of discretization varies significantly depending on the shape of data distribution. Hacibeyoglu et al. [12] used comparison between discrete method and continuous method for six datasets and showed that the performance of the classification accuracy is improved, when the features of datasets discretise. Kurtcephe and Güvenir [13] presented a new method based on the receiver operating characteristics - maximum area under ROC curve-based discretization (MAD). Compared with alternative discretization methods, empirical results show that MAD is a strong candidate to be an effective supervised discretization method. In [14] authors conducted experiments and showed that the accuracy of the prediction model improves significantly when the discretization and over-sampling methods are applied.
H.-V. Nguyen et al. in [15] propose IPD, an information-theoretic method for unsupervised discretization that focuses on preserving multivariate interactions. In [16] author proposed discretization algorithms which significantly improves the results in terms of the accuracy. Dash et al. [17] analysed and compared supervised and unsupervised discretization techniques. Yang and Web [18] showed a new discretization method, combination of weighted proportional k-interval and non-disjoint discretization that helps Naive-Bayes classifiers to reduce average classification error. In [19] authors used 10 datasets from UCI (Machine Learning Repository) in order to compare the effect of the unsupervised discretization methods on the classification. The results showed that algorithms Naive Bayes, C4.5 and ID3 achieved higher accuracy with supervised discretization method based on entropy. Andre V. Carreiro et al. [20] analyse impact of different unsupervised and supervised discretization techniques on the classification accuracy. They used real clinical expression time series to predict the response of patients with multiple sclerosis to treatment with Interferon-ß. The experimental results show that using the discretization methods improves the classification accuracy and problem of a small number of instances and a large number of features is solved. Taijun et al. [21] propose a post-processing method for improving the quality of discretization adjusting the boundary points of interval in order to obtain a positive influence on the attribute. Supervised fuzzy discretization for classifying time series datasets is proposed in [22]. This method can be used without having expertise on data. Coefficients of discretization, equal time slicing, learning rate, and momentum are analysed.

3 DISCRETIZATION

Discretization implies the reduction of different continuous feature values by dividing the scope into the final set of disjoint intervals that are assigned with meaningful marks [23].

Let vector $A = (a_1, a_2, ..., a_k)$ denote values of numeric feature in the data set S and n be the number of instances. $Dom(A)$ is a set of all feature values and represents its active domain. Discretization of numeric feature A equals finding k interval of the active domain $Dom(A)$, which implies determining $k - 1$ cut points t_i.

The numeric feature A is transformed in the vector of discrete values defined with Eqs. (1) and (2).

$$A^{disc} = (a_1{}^{disc}, a_2{}^{disc}, ..., a_n{}^{disc}) \tag{1}$$

$$a_j{}^{disc} = i \tag{2}$$

if and only if $a_j \in P_i \;\; j = \{1, 2, ..., n\}$ and
$P = \{P_1, P_2, ..., P_i, ..., P_k\}$, $P_i = \{a \in Dom(A) : t_{i-1} \le a \le t_i\}$,
$P_k = \{a \in Dom(A) : t_{k-1} \le a \le t_k\}$ marks possible intervals.

Discretization process can be described with four basic steps [24]:

1) Sorting continuous attribute values that are being transformed
2) Choosing, defining estimation measures, testing the suitability of candidates for the cut-point, i.e. the number of k split interval domains
3) Splitting or merging continuous value intervals according to the appropriate criteria
4) Stopping the procedure based on the stopping criteria. The notion of a "cut - point" defines a real value belonging to the continuous attribute domain, with which the scope of the given attribute is divided into two intervals. One of the intervals includes values lower than or equal to the cut - point, and the other includes values higher than the cut - point. The number of the cut - point k – 1 is defined by the number of the split interval k which can be user – defined or defined on the basis of the set heuristic rule.

3.1 Entropy-Based Discretization

The entropy-based supervised method uses the information about the class candidate entropy during the definition of cut – points. The class entropy information is a purity measure that measures the quantity of information necessary for defining which class an instance belongs to. It observes the interval that contains all known feature values and cuts it with the recursive split into smaller subintervals until the set stopping criterium is satisfied. The cut – point will be chosen with estimating disparity measures, i.e. by defining the class entropy for partition candidates. In the entropy – based discretization method, the best point is defined on the basis of the potential cut – point candidates entropy.

Let the instances set be S, feature A and the split boundary point T. Entropy split T, marked with $E(A, T; S)$ is determined with the following formulas [25]:

$$E(A,T;S) = \frac{|S_1|}{|S|} Ent(S_1) - \frac{|S_2|}{|S|} Ent(S_2) \tag{3}$$

$$Ent(S_i) = -\sum_{i=1}^{k} p(C_j, S_i) \log_2 (p(C_j, S_i)) \tag{4}$$

The subset entropy S_1, S_2 is calculated according to the Eq. (4), where $p(C_j, S_i)$ represents the percent of instances in S_i which have class C_j, k representing the number of classes marked with C_1, C_2, ..., C_k. In the Eq. (3), the instance set S is split into two intervals S_1 and S_2 using the cut - point T for attribute A value. The entropy function *Ent* for the given set is calculated on the basis of class sample distribution in the set. The best candidate for the cut - point T among all candidates for $E(A, T; S)$ is the one that has the minimum entropy value. After choosing the cut - point, continuous feature values are split into two parts. This procedure is repeated recursively until the set stopping criterium is satisfied. In the entropy-based discretization method, the stopping criterium is defined with the following formulas:

$$Gain(A,T;S) < \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N}, \tag{5}$$

N - number of set S instances

$$Gain(A,T;S) = Ent(S) - E(A,T;S) \tag{6}$$

$$\begin{aligned} &\Delta(A,T;S) = \log_2(3^k - 2) - \\ &-\left[k \cdot Ent(S) - k_1 \cdot Ent(S_1) - k_2 \cdot Ent(S_2)\right] \end{aligned} \tag{7}$$

k_i is the class mark number presented in the set S_i .

Taking into account the fact that each recursive discretization branch partition is assessed independently, some parts of the continuous value space will be finely partitioned, whereas the parts with the relatively low entropy will be roughly split. The aforementioned stopping criterium is known as Minimal Description Length Principle (MDL) and is described in the paper [26].

3.2 Equal-Width Binning

Equal-width binning (EWB) is one of the simplest direct unsupervised discretization methods [4]. The process includes sorting continuous features and then splitting the observed features domain into k interval (bins) of the same width (δ) with $k + 1$ cut – points.

Let active domain of feature A be marked with: $Dom(A) = (a_1, a_2, ..., a_n)$, $a_{\min} = \min\{a_1, a_2, ..., a_n\}$, $a_{\max} = \max\{a_1, a_2, ..., a_n\}$. Value δ for k equal intervals and cut – points $a_{\min}$, $a_{\min} + \delta$, ..., $a_{\max} = a_{\min} + k\delta$ is defined according to the formula:

$$\delta = \frac{(a_{\max} - a_{\min})}{k} \tag{8}$$

According to literature [27, 28, 29] the number of k intervals for the dataset of n instances with $a_{\min}$ and $a_{\max}$ minimum and maximum instance values respectively can be user defined or calculated on the basis of the Eqs. (9), (10), (11).

$$k = \log_2 n + 1 \tag{9}$$

$$k = \frac{(a_{\max} - a_{\min})}{h},\ h = \frac{2 \times IQR}{\sqrt[3]{n}} \tag{10}$$

IQR is an interquartile scope in the dataset

$$k = \frac{(a_{\max} - a_{\min})}{h},\ h = \frac{3.5 \times \sigma}{\sqrt[3]{n}} \tag{11}$$

σ is a standard deviation.

The number of k intervals is fixed and independent from the specific training data characteristics. This restriction can lead to some undesirable side effects. In case of a large dataset, a small number of split intervals can cause grouping of a wide instance spectre, which would definitely not have a positive effect on the applied learning algorithm. On the other hand, if the number of split intervals is too large, the intervals will have a small number of instances, and the importance and effects of the performed discretization could not be determined in that case.

3.3 Histogram Discretization

Histogram method belongs to unsupervised discretization techniques, since it does not use the class mark information [25]. The histogram represents geometric frequency table distribution which facilitates statistical data analysis. If X has the values $x_1, x_2, ..., x_n$ which appear in the N instance set $f_1, f_2, ..., f_n$ times. Values $f_1, f_2, ..., f_n$ satisfy the equation $f_1 + f_2 + ... + f_n = N$ and represent the frequencies. The intervals do not overlap each other and have certain boundary values. The same bin size width (bar width or class size) is defined and also the number of observed random variable instances for each interval represents the frequencies. The histogram implies the availability of all data, i.e. the lack of missing values in the analysed set. Taking into account extreme values and outliers, cut - points $a_1, a_2, ..., a_{k-1}$ and frequency instances $f_1, f_2, ..., f_k$ are defined while creating the histogram.

The k interval can be defined in the observed random variable value $(-\infty, a_1]$,$(a_1, a_2]$,..., $(a_{k-2}, a_{k-1}]$, $(a_{k-1}, +\infty)$. Using the visualized representation in terms of rectangular graphs between which there are no gaps, the histogram provides information about the distribution of the analysed dataset random variable. The algorithm for creating the histogram includes the following steps:

1) Sorting the random variable values in the ascending order
2) Defining minimum and maximum value (min, max)
3) Defining the number of split intervals (k)
4) Calculating the bin size according to the formula $binsize = \frac{\max - \min}{k}$
5) Calculating instance frequency for each interval (frequency table)
6) Creating rectangular graphs; x – axis represents split intervals; y – axis instance frequency for each interval.

Histograms that are most often used are the equal width, where the scope of observed values is split into k intervals of the same length, or the equal frequency, where the scope of observed values is split into k intervals containing equal number of instances. For both algorithms it is necessary to define the parameter k, i.e. the number of split intervals, which is also the main issue. In the research data analysis, the histogram application implies recursive application to each partition in order to automatically generate the multilevel hierarchy concept until a predefined number of levels is achieved. For the recursive procedure control, the minimum interval value can be used or the minimum value number per interval.

Creating the histogram for different k parameter values enables choosing the most suitable one, depending on its final purpose.

4 CASE STUDY

The research described in the paper includes extraction process, preparation of data and experimental application of the three methods of discretization. EWB method has been implemented with the dynamic search for optimal value *k*.

Determination of split number in the case of histogram discretization is done by applying the Scott rule for calculating interval width. By using the supervised entropy method, the domains of numeric features have been discretized with different number of discrete values.

The dataset for the analysis contained 276 instances selected from Computer Graphics Moodle course. The course was held during the summer term in the academic year 2015/2016 at the High School of Electrical Engineering and Computer Science of Applied studies. Activities for every student at the Moodle course were represented using the set of instances with appropriate features. Based on the analysis of the domain value, numerical and categorical features of dataset were identified. Numeric features had uneven value domains, whereas categorical features were binominal with two possible values and polynomial with multiple possible values.

Table 1 Numeric features for extracted dataset

Feature	Domain	Description of the feature
LAB	[0,000 ; 14,900]	Points won at the lab.exercises
DZ1	[0,000 ; 2,000]	Points won in solving five homeworks
DZ2		
DZ3		
DZ4		
DZ5		
P1	[0,000 ; 20,000]	Average points of all attempts in solving preparatory tests
P2		
P3	[0,000 ; 30,000]	
T1	[0,000 ; 20,000]	Points won in first and second test
T2		
FT	[0,000 ; 30,000]	Points won on the final exam
PDF	[0,000 ; 8,000]	Usage of PDF materials
LVT	[0,000 ; 12,000]	Usage of e-tutorials
LESS		Usage of Moodle lessons
MARK	[3,000 ; 10,000]	Final grade

For the analysed dataset global values were used for completing the missing values. In the case of input features, missing values were completed with value 0 which stated that a student had not realized the activity. Numeric feature *MARK* was defined as a class feature and missing value completion was done with the value 3, which stated that a student had not taken the exam.

The training set created from two thirds, test set from one third of each discretized dataset of instances. Naïve Bayes (NB) classifier was trained and tested on each discretized dataset. Measures accuracy (*Acc*%) and relative absolute error of classification (*RAE*%) have been considered. Domain values and descriptions of numeric features are given in Tab. 1.

4.1 First Experiment

The first experiment represented the implementation of entropy – based discretization method with the MDL stopping criteria. The supervised entropy method takes into consideration information about the class of a candidate in order to choose the discretization boundaries. This method observes one large interval of all observed feature values, and then performs the recursive split into subintervals until the stopping criteria is reached. The numeric feature *MARK* was transformed into the nominal type so that the values of 3, 5, 6, 7, 8, 9, 10 correspond with the class labels{exam_not_taken, exam_failed, five, six, seven, eight, nine, ten} respectively. The input numeric feature values were classified in the descending order. The value domain was split into the points from the sorting list where the class mark value was being changed. For each cut – point, the entropy value of induced partitions, i.e. subsets left and right from the cut – point was calculated. The candidate with the minimum entropy value was chosen as a candidate for the cut – point T among all candidates for $E(A,T;S)$. The process was repeated recursively until both subsets contained only the same - class instances and the stopping criteria were reached. The domains of numeric feature values of the analysed dataset were discretized with different discrete value numbers. The discrete value numbers of input numeric features are given in Tab. 2.

Table 2 Discrete value numeric features

Features	Discrete Values
LAB	6
DZ1, DZ2, DZ3, DZ4	3
DZ5	4
T1	8
T2	6
FT	7
PDF, LVT, LESS	1

Cut – points were not determined for *PDF*, *LVT* and *LESS* features which led to the conclusion that those features did not affect the class feature *MARK*. This case could be explained with the fact that the mentioned features represent optional activities of the Moodle course which students could use in the learning process, but there were no scores. PDF, LVT and LESS features were excluded from further analysis. The NB classifier generated the model with the accuracy of $Acc = 86{,}23\%$ and classification error of $RAE = 20{,}03\%$.

4.2 Second Experiment

In the second experiment EWB unsupervised discretization method was applied. Dynamic search for optimal split interval number was carried out by

simultaneous discretization of numeric features for values k = 2, 3, ..., 10 [30]. Each discretized set was tested with NB classifier. Accuracy and relative absolute error of classification for generated prediction models are given in Tab. 3.

Table 3 Accuracy and classification error of NB models

	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Acc (%)	58.51	68.09	74.6	78.72	80.85	81.91	79.79	80.85	77.66
RAE (%)	56.16	49.14	37.73	31.56	28.12	30.46	29.30	26.99	29.87

For values k = {2, 3, 4, 5, 6} linear improvement performances of classification models were observed. For k = 6, NB classifier created model with accuracy from *Acc* = 80,85% and classification error from *RAE* = 28,12%. Uneven changes were noticed with further interval number increase. In case k = 7 it was observed increasing accuracy to *Acc* = 81,91% but also increasing classification error to *RAE* = 30,46%. However, for k = 8 accuracy was decreased to *Acc*=79,79% and classification error to *RAE* = 29,30%. In case k = 9 accuracy was *Acc* = 80.9%, and classification error *RAE* = 26,99%.

In the last examined case, k = 10, performance of classification model was decreased again, i.e. the accuracy decreased to *Acc* = 77,66% and classification error increased to *RAE* = 29.8%.

4.3 Third Experiment

The third experiment was related to histogram discretization method. Sorting, defining minimum and maximum values and histogram analysis were carried out. The split interval number was determined on the basis of the Scott rule (Eq. (11)). According to the fact that in the preparatory phase of the training dataset missing values were replaced with the value of 0 which marked that student had not realized the particular activity and did not achieve the scores, the same minimum value was determineted, min = 0, for all numeric features. The h and k values of the training dataset with n = 276 instances are given in Tab. 4.

Table 4 Defining the split interval number

Feature	min	max	*Mean*	*StDev*	h	(max−min)/h	k
LAB	0	14,9	10,926	4,331	2,33	6,40	7
DZ1	0	2	1,171	0,795	0,43	4,68	5
DZ2	0	2	1,150	0,879	0,47	4,23	5
DZ3	0	2	1,132	0,877	0,47	4,24	5
DZ4	0	2	0,962	0,790	0,42	4,71	5
DZ5	0	2	1,005	0,869	0,47	4,28	5
T1	0	20	11,422	6,175	3,32	6,02	7
T2	0	20	11,221	6,169	3,32	6,03	7
FT	0	30	16,872	10,059	5,41	5,55	6

The interval number k was obtained by rounding to the higher value so that all set instances belong to the appropriate interval. The Upper Bound was determined and the frequency was calculated for each interval, as well as the instance number that has values within the boundaries of the particular interval. Instances with missing values transformed in the value of 0, in the case of features DZ1, DZ2, DZ3, DZ4, DZ5, T1, T2, FT were placed in the first interval, and in the case of *lab* in the first three intervals. Based on calculation of frequency tables, the histogram distribution graphs were created. NB classifier was trained on the histogram discretized dataset. Prediction model has achieved accuracy of *Acc* = 79.59%, with classification error of *RAE* = 32.39%.

The performances of NB models for the dataset discretized by applying unsupervised and supervised methods are given in Tab. 5.

Table 5 Performances of NB classifier models

	EWB		Histogram	Entropy
	k=6	k=9		
Acc %	80,85	80,85	79,59	86,23
RAE %	28,12	26,99	32,39	20,03

As could be assumed, the greatest accuracy was achieved by applying the supervised entropy discretization method. EWB discretization with k = 6 and k = 9 has achieved equal accuracy but with different classification errors and it was not possible to determine optimal split interval number. It can be considered that the fact about the different distribution of the features value in the active domain has not been taken into account. Simultaneous discretization with the same value of split interval number for all numeric features can result in suppressing positive discretization effects. The histogram method achieved the lowest classification accuracy of *Acc* = 79.59% and the greatest error value *RAE* = 32.39%. Considering the results of conducted experiments, the question is how to improve unsupervised discretization methods and achieve as precise as possible prediction model for the educational training dataset.

5 PROPOSED APPROACH

The proposed approach for improving the efficiency of unsupervised discretization methods modifies the dataset disbalance by applying the Synthetic Minority Oversampling Technique (SMOTE) [31] on the discretized sets. As the result of the applied technique, distribution of instances with minor class feature values was carried out. By creating minor class synthetic instances, predictive accuracy of the analysed set was increased. The SMOTE algorithm was implemented on the sets discretized with histogram discretization and EWB method for the split interval number k = 6 and k = 9.

The SMOTE algorithm implementation implied automatic definition of a minor class, setting parameter values to 5 for choosing the nearest neighbour, and the percent of SMOTE instances that would be created was set to 100%. Two minor classes were noticed, one with 24 and the other with 27 instances. Two iterations of SMOTE algorithm application were carried out. After the second iteration, the overall number of instances was n = 327. Since the synthetically created minor class instances were concentratedly generated at the end of the set, the Randomize filter was implemented, and thus a random instance order in the training set was created. After that, NB classifier was trained on training set with 218 instances and tested on testing set with 109 instances. The generated model performances are given in Tab. 6.

Table 6 NB model performances after the SMOTE algorithm implementation

	Histogram	EWB		Entropy
		(k=6)	(k=9)	
Acc %	88,28	84,68	81,98	89,19
RAE %	20,39	24,20	23,53	16,44

As given in Tab. 6, applying the SMOTE technique has affected the improvement of the applied models efficacy in terms of increasing the accuracy of prediction models. For unsupervised methods, the best result was achieved with the histogram discretization. In that case, accuracy was increased to *Acc* = 88.28% and classification error was decreased to *RAE* = 20.39%. For EWB method, it is evident that accuracy for k = 6 has been better than the case of k = 9. However, classification error has been lower in the case of k = 9. Determining the split interval number by the EWB method was excluded due to the differences of the numeric feature domain values and suppression of positive discretization effects. The comparison of prediction accuracy for the dataset discretized by applying unsupervised methods before and after SMOTE algorithm implementation is given in Fig. 1.

Before the proposed approach, accuracy of unsupervised discretization histogram method has been lower by even 6.64% compared to the accuracy achieved with supervised discretization entropy method. Using improved unsupervised histogram discretization method, prediction accuracy has been lower by only 0.91% compared to the accuracy achieved with supervised discretization entropy method. Minimum loss of information was achieved by calculating bin size of the intervals using the Scotts rule based on the values of standard deviation. Disbalance of discretized dataset was modified with synthetically created instances of minority class by applying the SMOTE Oversampling Technique. The comparison of prediction accuracy for the dataset discretized by unsupervised histogram discretization method and supervised discretization entropy method is given in Fig. 2. For the proposed improved unsupervised discretization method the pseudo-code is given below.

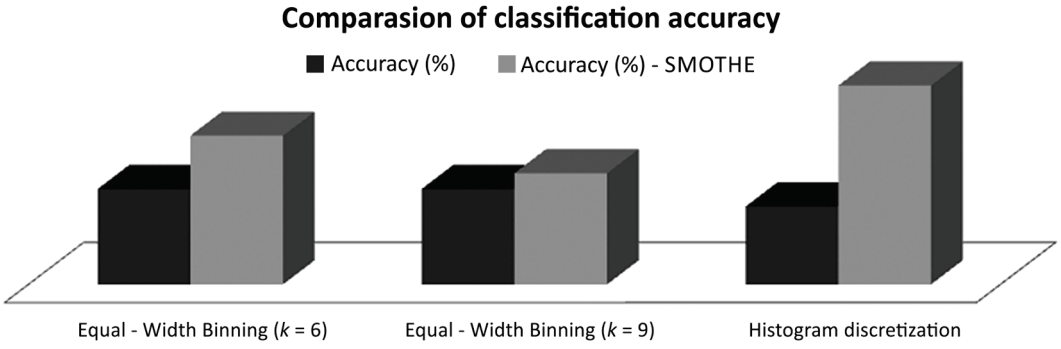

Figure 1 Comparison of classification accuracy for unsupervised discretization methods

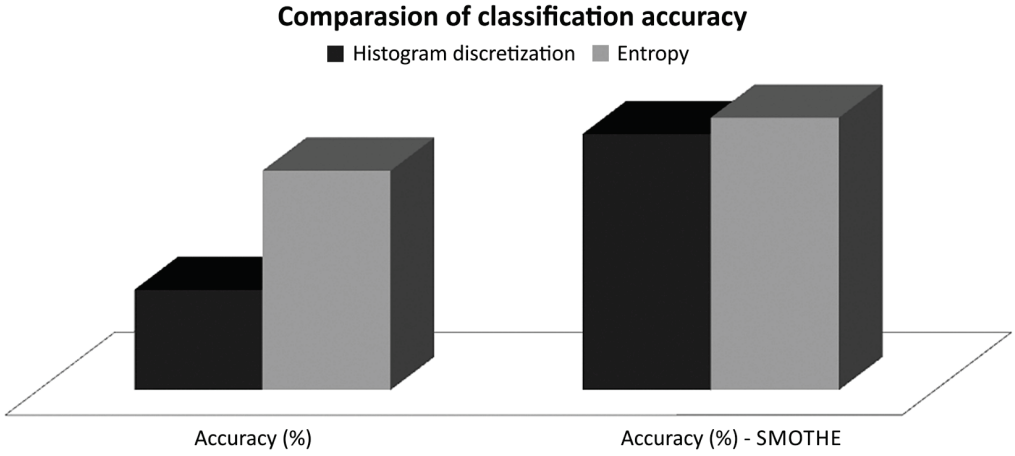

Figure 2 Comparison of classification accuracy for histogram and entropy discretization methods

Input
n, number of instances
Ai, continuous features
min - minimum value
max – maximum value
σ - standard deviation
Discretization with Scott rule
For each continuous attribute Ai in training dataset do

Sorting the random variable values in the ascending order
Defining minimum and maximum value (min, max)
Defining value of standard deviation σ
Calculating range measure
binsize -h

$$h = \frac{3.5 \times \sigma}{\sqrt[3]{n}}$$

number of split intervals- k

$$k = \frac{(a_{\max} - a_{\min})}{h}$$

Upper Bound, frequency table
and for
Output: *Unsupervised discretization based on histogram distribution*
Improved histogram discretization with SMOTHE
Defining class variable
Transform numeric type class variable into nominal
Setting number of minor classes
Setting parameter values to 5
Percent of SMOTE instances = was set to 100%
Number of implemented SMOTHE = Number of minor classes
Randomize instances
Output: *Synthetically created minor class instances*

Based on differences between discrete values achieved with supervised entropy method and improved unsupervised discretization method it can be concluded that the discrete values obtained by the histogram method is generally approximate to the number accomplished by the entropy method.

6 CONCLUSION

The primary objective of the research was improving the efficacy of the classification prediction model by determining the extended process of numeric features discretization in the phase of pre-processing the educational training dataset. A small number of instances in the dataset and determining the multiclass feature led to the oversampling issue. As expected, the entropy model transformed the numeric features into discrete values for which the generated NB classification model achieved the highest accuracy, both for the original dataset of 276 instances, and for the set with synthetically created instances. It has been concluded that the accuracy of the prediction can be improved in the case of the application of unsupervised discretization methods when the SMOTE algorithm and Randomize filter are applied in the pre-processing phase. Since the discrete value number is as approximate to the achieved number as in the case of entropy method, it has been determined that classification errors are reduced by applying the unsupervised histogram method.

The main contribution of this research was the improvement of the efficacy of unsupervised discretization methods, which allows greater accuracy of the prediction model and reduces the effects of ignoring class features. In the case of educational dataset, a precise classification of students was realized based on the activities carried out during the semester without information of the final mark.

The continuation of the research will be directed towards determining the procedure in the pre-processing phase which would achieve better performances of the prediction system in the case of instance subset separation and with missing values from the analysed educational training dataset.

7 REFERENCES

[1] Cios, K., Pedrycz, W., Swiniarski, R., & Kurgan, L. (2007). *Data Mining a Knowledge Discovery Approach*, Springer.

[2] Liu, H., Hussain, F., Tan, C. L. et al. (2002). Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery, 6*(4), 393-423. https://doi.org/10.1023/A:1016304305535

[3] Xu, E., Liangshan, S., Yongchang, R., Hao, W., & Feng, Q. (2010). A new discretization approach of continuous attributes. *Asia-Pacific Conference on Wearable Computing Systems, 5*(2), 136-138. https://doi.org/10.1109/APWCS.2010.40

[4] Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization discretization of continuous features. *Proc. 12th International Conference on Machine Learning*, Los Altos, CA : Morgan Kaufman, 194-202. https://doi.org/10.1016/B978-1-55860-377-6.50032-3

[5] Quinlan, J. (1993). *C4.5: Programs for Machine Learning*, M. Kaufmann, San Mateo, CA.

[6] Chmielevski, M. R. & Grzymala-Busse, J. W. (1994). Global discretization of continuous attributes on preprocessing for machine learning. *3rd International Workshop on Rough Sets and Soft Computing*, 294-301.

[7] Bay, S. (2000). Multivariate discretization of continuous variables for set mining. *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 315-319. https://doi.org/10.1145/347090.347159

[8] Sheng-yi, J., Li, X., Zheng, Q., & Wang, L. X. (2009). An approximate equal frequency discretization method. *WRI Global Congress Intelligent System, 3*(4), 514-518.

[9] Gama, J. & Pinto, C. (2006). Discretization from data streams: Applications to histograms and data mining. *Symposium on Applied Computing*, 662-667. https://doi.org/10.1145/1141277.1141429

[10] Asuncion, A. & Newman, D. J. UCI Machine Learning Repository, University of California, Irvine, CA, School of Information and Computer Science, http://archive.ics.uci.edu/ml/

[11] Ismail, M. K. & Ciesielski, V. (2003). An empirical investigation of the impact of discretization on common data distributions. *Proc. of the 3rd Int'l Conference on Hybrid Intelligent System*, 692-701.

[12] Hacibeyoglu, M., Arslan, A., & Kahramanli, S. (2011). Improving classification accuracy with discretization on datasets including continuous valued features. *Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering, 5*(6), 623-626.

[13] Kurtcephe, M. & Güvenir, H. A. (2013). A discretization method based on maximizing the area under receiver operating characteristic curve. *International Journal of Pattern Recognition and Artificial Intelligence, 27*(1), 1350002 (26 pages)

[14] Jishan S. T. et al. (2015). Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics 2:1, Springer Open Journal*.

https://doi.org/10.1186/s40165-014-0010-2
[15] Nguyen, H. V., Müller, E., Vreeken, J. et al. (2014). Unsupervised interaction-preserving discretization of multivariate data. *Data Mining and Knowledge Discovery, 28*(5), 1366-1397. https://doi.org/10.1007/s10618-014-0350-5
[16] Al-Ibrahim, A. (2011). Discretization of continuous attributes in supervised learning algorithms. *The Research Bulletin of Jordan ACM, II*(IV), 158-165.
[17] Dash, R., Paramguru, R. L., & Dash, R. (2011). Comparative analysis of supervised and unsupervised discretization techniques. *International Journal of Advances in Science and Technology, 2*(3).
[18] Yang, Y. & Webb, G. (2002). A comparative study of discretization methods for Naive-Bayes classifiers. *Proceedings of Pacific Rim Knowledge Acquisition Workshop*.
[19] Ibrahim, M. & Hacibeyoglu, M. (2016). Comparison of the effect of unsupervised and supervised discretization methods on classification process. *International Journal of Intelligent Systems and Applications in Engineering, 4*(Special Issue 1), 105-108. https://doi.org/10.18201/ijisae.267490
[20] Andre, C., Artur, J., Ferreira, F., & Sara, M. (2012). Towards a classification approach using Meta-Biclustering: Impact of discretization in the analysis of expression time series. *Journal of Integrative Bioinformatics, 9*(3), 207.
[21] Taijun, H. et al. (2015). Improving discretization by post-processing procedure. *International Journal of Engineering and Technology, 7*(2), 414-421.
[22] Umut, O. (2016). Time series adapted supervised fuzzy discretization: an application to ECG signals. *Turk J Elec Eng & Comp Sci, 24*, 3987-3998. https://doi.org/10.3906/elk-1411-36
[23] Kerber, R. (1992). Discretization of numeric attributes. *Proceedings of the 10th National Conference on Artificial Intelligence*, MIT Press, Cambridge, 123-128.
[24] Hussain, F., Liu, H., Tan, C. L., & Dash, M. (1999). Discretization: an enabling technique. *Technical Report–School of Computing*, Singapore.
[25] Witten, I. H. & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, CA: Morgan Kaufman, San Mateo.
[26] Fayyad, U. M. & Irani, K. B. (1993). Multi-interval discretisation of continuous-valued attributes for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1022-1027.
[27] Sturges, H. (1926). The choice of a class interval. *Journal of the American Statistical Association, 21*(153), 65-66. https://doi.org/10.1080/01621459.1926.10502161
[28] Freedman, D. & Diaconis, P. (1981). On the histogram as a density estimator: L2 Theory. *Probability Theory and Related Fields, 57*(4), 453-476.
[29] Scott, D. (1979). On Optimal and Data-based Histograms. *Biometrika, 66*(3), 605-610. https://doi.org/10.1093/biomet/66.3.605
[30] Kayah, F. (2008). *Discretizing Continuous Features for Naive Bayes and C4.5 Classifiers*, University of Maryland publications: College Park, MD, USA.
[31] Nitesh, V. C. et al. (2002). Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Contact information:

Gabrijela DIMIĆ, PhD candidate
High School of Electrical Engineering and Computer Science
Vojvode Stepe 283, 11000 Belgrade, Serbia
Phone: +381 69 2005 134
E-mail: gdimic@gmail.com, gabrijela.dimic@viser.edu.rs

Dejan RANČIĆ, PhD
Faculty of Electronic Engineering
Aleksandra Medvedeva 18, 18000 Niš, Serbia
E-mail: dejan.rancic@elfak.ni.ac.rs

Ivan MILENTIJEVIĆ, PhD
Faculty of Electronic Engineering
Aleksandra Medvedeva 18, 18000 Niš, Serbia
E-mail: ivan.milentijevic@elfak.ni.ac.rs

Petar SPALEVIĆ, PhD
University of Prishtina, Faculty of Technical Sciences
Kneza Miloša 7, 38220 Kosovska Mitrovica, Serbia
E-mail: petar.spalevic@pr.ac.rs