Proceedings of the 2014 Industrial and Systems Engineering Research Conference
Y. Guan and H. Liao, eds.

An Improvement on Association Rule Based Classification of Medical Data

Anas Hourani, Chun-An Chou and Sina Khanmohammadi
Department of Systems Science and Industrial Engineering
State University of New York at Binghamton
Binghamton, NY, USA

Abstract

Association rule based classification is one of the popular data mining techniques applied in medical domain. The major advantage is its interpretable results that medical doctors can easily adopt for diagnostic decision-making. The classification framework consists of data discretization, association rule generation, and classification. The discretization step is required to convert numerical features such as blood pressure into a categorical format, to make it suitable for association rules mining. Existing discretization methods such as Omega algorithm construct several non-adjacent intervals to represent new categorical variables. However, such algorithms are not generalizable because of failure to recognize new observations that lie between constructed intervals; this will impact the accuracy of association rules based classification. To overcome this problem, an associative classification framework based on an improved discretization algorithm is proposed. In the discretization step, a centroid of each constructed interval is identified to represent that interval. Using the identified centroids, numerical data is discretized and fed to an Apriori algorithm for rule induction. Consequently, a new observation is classified using a majority-voting scheme of the generated rules. The framework was tested on various medical datasets from the University of California Irvine repository and the Aneurisk dataset repository. The results show that the proposed framework gives a higher accuracy when compared to existing approaches.

Keywords
Medical decision-making, association rules, discretization, classification

1. Introduction
Data mining techniques are widely used for many applications [1], especially for medical decision-making where a large variety of clinical data needs to be analyzed. For example, in medical diagnosis, data mining is employed to extract meaningful information (i.e., rules or patterns) from historical data that helps to determine (or predict) possible condition of new patients, such as normal or abnormal. Recently, a supervised learning method called association rule based classification (or associative classification) has been significantly applied to medical diagnosis and prognosis. The principle concept of associative classification is to find informative association relationships (i.e., association rules) based on the frequency of training data items [2], which subsequently are used for determining the classification of new observation points. Prior to generating association rules, data discretization is a key preprocessing step to identify a finite number of states (categories) from continuous feature values. For example, the risk status of patients with heart disease can be categorized into several levels: high, middle, and low by applying thresholds (i.e., cut points) to numerical blood pressure. In addition to the advantage of interpretability, there are other advantages for using discrete (categorical) features instead of continuous features when solving large-scale complex problems. Discrete features are simpler, closer to knowledge representation, and easier to understand. Furthermore, the performance of machine learning algorithms in terms of computational complexity and accuracy can be improved by using discrete datasets [3].

Despite the success of association rules used for classification has been shown, classification accuracy is limited by the construction of categorical data from continuous data due to the loss of information during the discretization process [4, 5]. Hence, there is a tradeoff between the amount of discretization (number of intervals) and the classification accuracy that can be resulted from associative classification techniques. Considering this tradeoff, a lot of researchers attempted to develop discretization algorithms that can balance the information loss and number of identified discrete intervals. One of existing powerful discretization algorithms is Omega algorithm that constructs intervals to represent new categorical variables based on class information of training data [6, 7]. However, the constructed intervals are not always adjacent to each other. As a result, new observations points that lie between the intervals cannot be recognized; therefore, the overall classification accuracy is decreased. In this paper, we introduce a concept of centroids to replace intervals, where each interval is represented by a centroid. Then, each value in continuous feature is addressed to the nearest centroid. Hence, the entire space between centroid is included and there are no uncovered gaps between them.

We begin in Section 2 by reviewing related work. In Section 3, the proposed method is described in detail. In Section 4, the results for testing the proposed method on 10 benchmark datasets are discussed. Finally, in Section 5 the conclusion of this work is provided.

2. Related Work

There were many supervised or unsupervised discretization approaches proposed in the recent years [1]. Equal width interval and equal frequency interval binning methods are the simplest unsupervised discretization methods. The equal width interval method divides the feature range into a number of equal-sized bins (intervals), whereas the equal frequency interval method divides the feature range into *k* bins that contain the same number of observation points [8]. The disadvantage of these unsupervised methods is the fact that they do not take into account the class information during the binning process [3]. ChiMerge method is another discretization method proposed by Kerber [9], where the cut points are determined based on Chi-square statistic test. Later, Liu and Setiono presented an improved ChiMerge method called Chi2. Chi2 algorithm repeats the discretization process using Chi-square statistic test until finding inconsistencies in the data [10]. A more generalized version of ChiMerge method (StatDisc) proposed by Richeldi and Rossotto [11]. This bottom-up method uses statistical tests to create a hierarchy of discretized intervals. In StatDisc method, a predefined number of intervals are merged at a time, whereas only two intervals are merged at a time in ChiMerge.

For supervised methods, one of the simplest methods is the 1R method developed by Holte [12], which pure intervals are defined for every feature and each interval contains observation points in the same class. To avoid overfitting, the number of observation points for each interval should be larger than a predetermined number. Shehzad proposed a new supervised algorithm, called entropy-based discretization intervals using scope of classes (EDISC) [13]. This method was developed based on the traditional entropy-minimum description length principle (Entropy-MDLP) discretization method. In EDISC method, the cut points are identified for each continuous feature with taking into consideration the optimality of cut points for the present class. This method minimizes the occurrence of instances of other classes to ensure the discovery of class-tailored cut points for each attribute. Wei proposed a multivariate discretization method [14]. This method used clustering techniques to unhide the interesting patterns from data, and then the genetic algorithm is applied to discretize multi-attributes according to entropy criterion.

In addition to new methods, there has been a research going on to improve the performance of discretization methods. In this regards, a resampling based bagging technique was introduced by Qureshi and Zighed [15]. Bagging method produces a set of candidate points with the objective of reducing the variance. Thus, the quality of the discretized data set is improved, where the quality refers to the number of intervals and information loss in the discretized dataset. In other words, when comparing two discretized datasets, the dataset with lower number of intervals and less information loss is considered as having a higher quality.

One of the simple and most effective discretization algorithms is called Omega algorithm and was proposed by Ribeiro et al [6]. The advantage of Omega algorithm is that it can perform discretization and feature selection simultaneously. In Omega algorithm, first the values of each feature in the dataset are sorted, and the initial cut points are determined. Then, the number of intervals is reduced by merging the identified intervals using the inconsistence rate. Inconsistence rate is defined as percent of observations that do not belong to the majority class of

an interval, where the majority class is the most frequent class label in that interval. Finally, the continuous values are discretized based on the resulted intervals.

The discretized dataset has many applications in machine learning and data mining techniques. One of the main data mining methods that use discretized dataset is associations rule mining. Association rule mining is used to find the relationship between items in a dataset. Abundant of association rule algorithms were proposed in literature. Apriori algorithm is one of the common association rule algorithms developed by Agrawal and Srikant [16]. A frequent 1-itemsets (a set that consists of 1 item) can be found after first scan over the data, and then frequent 2-itemsets (a set that consists of 2 items) are generated from the frequent 1-itemsets. This process iterates until stop generating frequent *k*-itemsets. Han et al. proposed a new association rule algorithm called FP-growth [17]. FP-growth mines all frequent itemsets without generating candidate itemsets. A descending-order list of frequent items is found from first scan over the data. Thus, the data is compressed into a frequent-pattern tree (FP-tree). Then, a conditional pattern base for each frequent length-1 pattern is constructed to mine the FP-tree. Next, the conditional FP-tree of the pattern is created, and the mining process is recursively performed on this tree. By concatenation of the suffix pattern with the new pattern that generated from conditional FP-tree, the pattern growth is achieved. Equivalence Class Transformation (Eclat) algorithm is proposed by Zaki [18], where transactions are explored in a vertical data format (a list for each item consists of transactions indexes where this item frequent). This is opposite to Apriori and FP-growth algorithms, which the transactions are explored in a horizontal data format (i.e., a list of transactions and each transaction consists of set of items).

3. Associative Classification Framework

In this paper, an associative classification framework incorporating the proposed discretization method is presented. Figure 1 represents the overall framework. In Figure 1(a), the first step is to discretize all continues features in the data set. In the next step the discretized features are utilized to generate association rules, and finally, a majority voting scheme is employed to classify the new observation points.

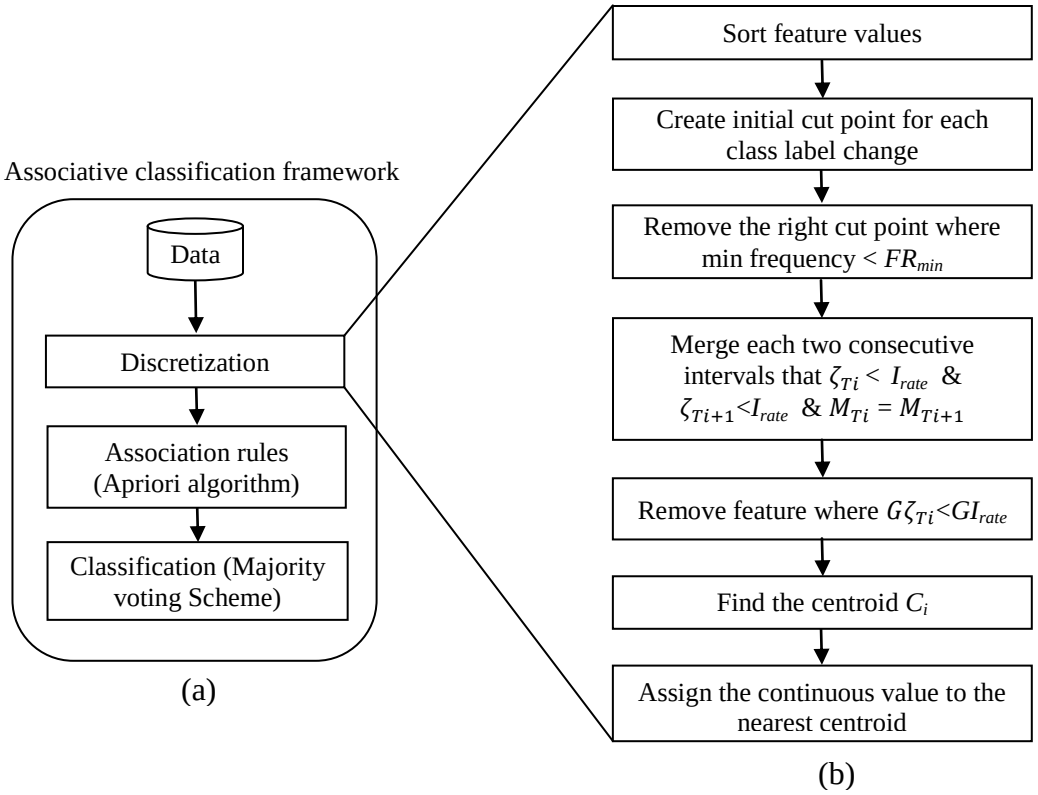

Figure 1: (a) The associative classification framework. (b) The diagram of the proposed discretization method.

For discretization step, a vector of discrete values is produced by the proposed method for each observation. These discretized values are called *centroids*. A centroid is a point used to represent an interval. An *interval* (also called *bin*) is a space or a distance between two points (or limits) with the property that any value lies between these two points belongs to that interval. The limit of an interval is called *cut point*. There are two cut points for each interval (lower-bound cut point and upper-bound cut point). In order considering the class information, the *majority class* is founded for each interval. Majority class refers to the most frequent class for an interval. The overall representation of the proposed discretization method is provided in Figure 1(b). Each of these steps will be explained in detail as follow:

Step 1: The continuous values are sorted for each feature *i*, where *i* represents the index of features. An example is provided in Figure 2.

Figure 2: Sorting the continuous values for feature *i*

Step 2: The initial cut points are defined for the sorted dataset. More specifically, each data point is compared with the proceeding data point and a cut point is set in case there is a change in the class label. A pure interval (an interval, which all data points belong to the same class label) is produced for each group of observations that ordered sequentially. These produced pure intervals have no inconsistence. Figure 3 represents this process.

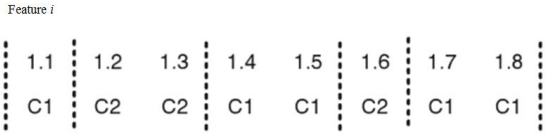

Figure 3: Defining the initial cut points

Step 3: In order to avoid a large number of intervals, the proposed method merges the intervals that don't meet the minimum frequency restriction FR_{min}, where FR_{min} is predefined by end users. For each of these identified intervals, the right cut point is removed to combine it with the proceeding interval as shown in Figure 4. The higher value of minimum frequency results in lower number of bins, and higher inconsistence within intervals.

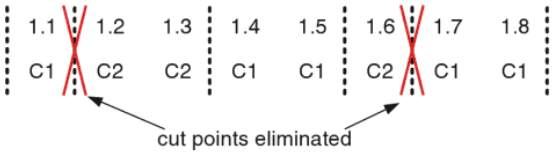

Figure 4: Merging intervals using FR_{min} =2

Step 4: A second merging process is performed. In this regards, the consecutive intervals whose inconsistence rate are less than a predefined value I_{rate}, and have the same majority class, are merged. The inconsistence rate for each interval T_i is calculated using Equation (1), where M_{Ti} denote the majority class for interval T_i .

$$\zeta_{Ti} = \frac{|T_i| - |M_{Ti}|}{|T_i|} \quad (1)$$

Step 5: A global inconsistence rate is calculated for each feature, and then the features whose global inconsistence rates are larger than a predefined value GI_{rate}, are eliminated. In other words, the most inconsistence features

are removed to avoid noise, and improve classification accuracy. The global inconsistence rate can be calculated using Equation (2).

$$\zeta_G = \frac{\sum_{T_i \in T} (|T_i| - |M_{Ti}|)}{\sum_{T_i \in T} |T_i|} \tag{2}$$

Step 6: A set of intervals is defined for each feature. Each interval is represented by the corresponding centroid, so intervals are replaced with centroids. The centroids are calculated using Equation (3). Let T_i denote the interval i, L_i denote the lower bound of interval i, U_i denote the upper bound of interval i, C_i denote the centroid of interval i, and n denote the number of intervals.

$$C_i = \begin{cases} L_i + \frac{U_i - L_i}{2}, & U_i - L_i \leq U_{i+1} - L_{i+1} \\ U_i - \frac{U_{i+1} - L_{i+1}}{2}, & U_i - L_i > U_{i+1} - L_{i+1} \\ L_i + (U_{i-1} - C_{i-1}), & i = n \end{cases} \tag{3}$$

Step 7: The original (continuous) dataset is converted to a discrete dataset using identified centroids. For this purpose, the distance between each value of each feature, and the corresponding centroids is calculated. Next, the continuous value is replaced with the nearest centroid label.

The Pseudo-code for the proposed method is presented as follows, where F_j denotes feature j, T_i denotes interval i, FR_{min} denotes the minimum frequency restriction, *IR* denotes the inconsistence rate, *MC* denotes the majority class, GI_{rate} denotes the global inconsistence rate, and C_i denotes centroid i.

```
FOR each feature F_j
    Sort(F_j)
    IF (class label i ≠ class label i+1) THEN
      Set(cut point)
    FOR each interval T_i
      IF (|T_i| < FR_min) THEN
      Merge (T_i , T_i+1 )
    END FOR
    FOR each interval T_i
      IF (IR (T_i ) < I_rate ) & (IR (T_i+1 ) < I_rate ) & (MC (T_i ) = MC (T_i+1 )) THEN
          Merge (T_i , T_i+1 )
    END FOR
    IF (Global Inconsistence < GI_rate ) THEN
      Remove (F_j )
    FOR each interval T_i
      Find (centroid C_i)
    END FOR
    FOR each continuous value
      AssignTo (centroid )
    END FOR
END FOR
```

The advantage of the proposed method is that it can recognize the points that lie between intervals, because it uses minimum distance from centroids instead of intervals for the discretization process. In other words, each value in continuous feature is addressed to the nearest centroid, therefore, the entire space between centroid is included and there are no uncovered gaps between them. Other discretization method like Omega do not have this capability and cannot recognize the points that lie between intervals as shown in Figure 5.

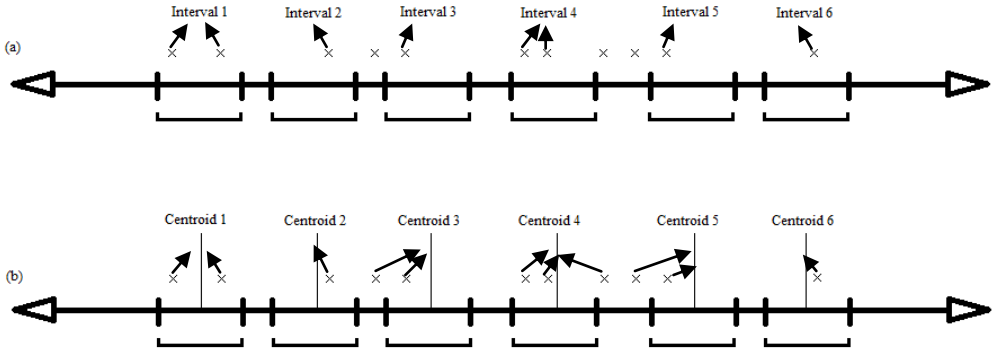

Figure 5: Discretizing the points that lie between intervals using the proposed method.

4. Experimental Results

In order to evaluate the performance of the proposed discretization method, 10 benchmark medical datasets are chosen from the University of California Irvine (UCI) repository [19], and the Aneurisk data repository [20]. The properties of these datasets are shown in Table 1.

Table 1: The properties of used datasets

Dataset	Number of Features	Number of Observations
WDBC	32	569
Breast	11	699
Bupa	7	345
Parkinsons	23	195
Pima	8	768
WPDC	34	198
Mammographic	6	961
Hepatitis	20	155
Heart Disease	14	303
Aneurysm*	39	102

* Retrieved from Aneurisk data repository

For each of these datasets, several discretization methods are applied such as Omega, equal width, equal frequency, and the proposed improved method. The Apriori algorithm is used to generate rules based on the discretized data, and these rules are used for the classification purpose by applying a majority voting scheme. For the validation purpose, a *k-fold* cross-validation method is used, which involves partitioning the data into k folds (subsets). A classification model is trained based on the k-1 folds (training dataset), and tested on the remaining one fold (testing dataset). This process is repeated k times for each fold as a testing set. In our experiment, the dataset is divided into 5 folds (i.e., $k = 5$).

Table 2: The accuracy of association rule classification

Dataset	Proposed Method	Omega Method	Equal Width	Equal Frequency
WDBC	0.93	0.92	0.64	0.63
Breast	0.89	0.88	0.81	0.82
Bupa	0.63	0.64	0.44	0.37
Parkinsons	0.75	0.72	0.73	0.74
Pima	0.72	0.71	0.66	0.48
WPDC	0.74	0.73	0.76	0.76
Mammographic	0.87	0.85	0.60	0.58
Hepatitis	0.78	0.78	0.81	0.81
Heart Disease	0.83	0.82	0.73	0.83
Aneurysm	0.78	0.75	0.60	0.52

The classification accuracy based on the proposed discretization method is evaluated, and compared to the performance of classification based on the other discretization methods. The results indicate the classification accuracy based on the proposed discretization method outperforms the other discretization methods. Table 2 shows the comparison of association rule classification based on different discretization methods.

Figure 6 illustrates a graphical representation for comparing the accuracy of association rule classification based on several discretization methods.

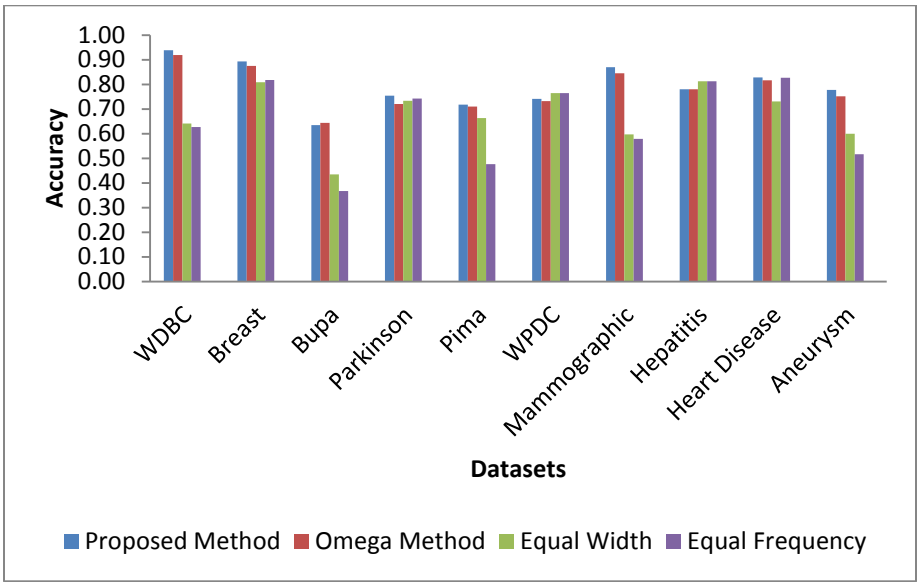

Figure 6: The accuracy of association rule classification

In order to determine the highest classification accuracy for the proposed method, additional experiments are conducted using different support values. A support value *S* is the percentage of the frequency of a rule in the dataset. Table 3 shows the classification accuracy on WDBC dataset (as example) for different *S* values. The training and testing accuracy are shown in this table, as well as the number of rules for the association rule classification algorithm. It can be seen in Table 3 that the best accuracy is achieved when $S = 0.05$, however, the number of identified rules is very high. Therefore, considering the tradeoff between number of rules and the accuracy, $S = 0.1$ is desirable as it can achieve a similar accuracy to $S = 0.05$ with less number of rules. This illustrated in Figure 7 as well.

Table 3: The classification accuracy and the number of rules change with *S* values (WDBC dataset as example)

		Training accuracy	Testing accuracy	Number of rules
Support	0.05	0.97	0.94	161209
	0.10	0.95	0.93	25513
	0.15	0.95	0.93	6144
	0.20	0.94	0.91	1972
	0.25	0.94	0.90	703
	0.30	0.94	0.89	452
	0.35	0.92	0.86	224
	0.40	0.92	0.86	124
	0.45	0.92	0.85	34
	0.50	0.92	0.85	10
	0.55	0.91	0.76	2
	0.60 – 1.0	0	0	0

When the *S* equals to 0.6 or above, it can be noticed that there are no rules identified. In other words, there are no rules with a frequency of 60% or more in the dataset. As a result, the associative classification accuracy for $S \geq 0.6$ equals to 0. The associative classification accuracy of the training dataset is considered as well, and the results indicate that there is no overfitting in this classification model.

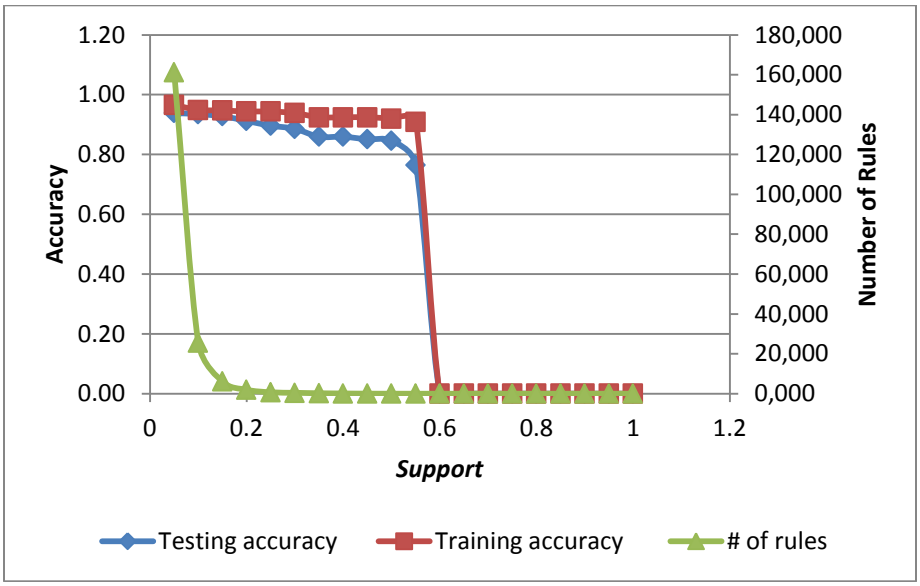

Figure 7: The classification accuracy and the number of rules among different *S* values

5. Conclusions

In this paper, an improved discretization method was proposed to enhance the discretization power of the Omega algorithm. The proposed discretization method is based on representing each identified interval with a centroid, and discretizing the dataset using the minimum distance of each data point from centroids. The strength of this new method is its capability to discretize the points that lie between constructed intervals, which the previous methods were unable to do. Different discretization methods such as Omega, equal width, and equal frequency were compared with the proposed method in terms of classification accuracy. The classification framework consists of rule induction from the discretized data using the Apriori algorithm, and applying the majority voting scheme to classify the data using identified rules. The results on 10 medical datasets from the UCI repository and the Aneurisk data repository show that the associative classification framework based on the proposed discretization method achieved higher accuracy when compared to other discretization methods.

Acknowledgements

The work is supported by in part the Grant (I920247) at Binghamton University and the research collaboration fund (66508) from the SUNY Research Foundation.

References

[1] Padhy, N., Mishra, D., and R. Panigrahi, 2012, "The Survey of Data Mining Applications And Feature Scope," arXiv preprint arXiv:1211.5723.

[2] Bouzouita, I. and Elloumi, S., 2011, "Generic Associative Classification Rules: A Comparative Study," International Journal of Advanced Science and Technology, (33), 69–84.

[3] Catlett, J., 1991, "On changing continuous attributes into ordered discrete attributes," in Machine learning—EWSL-91, 164–178.

[4] Vannucci, M. and Colla, V., 2004, "Meaningful discretization of continuous features for association rules mining by means of a SOM.," in ESANN, 489–494.

[5] Srikant, R. and Agrawal, R., 1996, "Mining quantitative association rules in large relational tables," in ACM SIGMOD Record, (25), 1–12.

[6] Ribeiro, M. X., Traina, A. J., and Traina, C. Jr, 2008, "A new algorithm for data discretization and feature selection," in Proceedings of the 2008 ACM symposium on Applied computing, 953–954.
[7] Ribeiro, M. X., Traina, A., Traina, Rosa, C., N. A., and Marques, P., 2008, "How to improve medical image diagnosis through association rules: The IDEA method," in Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on, 266–271.
[8] Dougherty, J., Kohavi, R., and Sahami, M., 1995, "Supervised and Unsupervised Discretization of Continuous Features," in Machine learning: proceedings of the Twelfth International Conference on Machine Learning, July 9-12, Tahoe City, California, 194-202.
[9] Kerber, R., July 1992, "ChiMerge: Discretization of Numeric Attributes," in Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, California, 123–128.
[10] Liu, H. and Setiono, R., 1997, "Feature selection via discretization," IEEE Transactions on Knowledge and Data Engineering, 9(4), 642–645.
[11] Richeldi, M. and Rossotto, M., 1995, "Class-driven statistical discretization of continuous attributes," in Machine Learning: ECML-95, Springer, 335–338.
[12] Holte, R. C., 1993, "Very simple classification rules perform well on most commonly used datasets," Machine learning, 11(1), 63–90.
[13] Shehzad, K., 2012, "EDISC: A Class-Tailored Discretization Technique for Rule-Based Classification," IEEE Transactions on Knowledge and Data Engineering, 24(8), 1435–1447.
[14] Wei, H., 2009, "A Novel Multivariate Discretization Method for Mining Association Rules," in Asia-Pacific Conference on Information Processing, 2009. APCIP 2009, 1, 378–381.
[15] Qureshi, T. and Zighed, D. A., 2009, "On improving discretization quality by a bagging technique," in Natural Computation, 2009. ICNC'09. Fifth International Conference on, 1, 226–231.
[16] Agarwal, R. and Srikant, R., 1994, "Fast algorithms for mining association rules," in Proc. of the 20th VLDB Conference, 487–499.
[17] Han, J., Pei, J., and Yin, Y., 2000, "Mining frequent patterns without candidate generation," in ACM SIGMOD Record, 29, 1–12.
[18] Zaki, M. J., 2000, "Scalable algorithms for association mining," IEEE Transactions on Knowledge and Data Engineering, 12(3), 372–390.
[19] Bache, K. and Lichman, M., 2013, "UCI Machine Learning Repository," [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[20] AneuriskWeb project website, 2012, [http://ecm2.mathcs.emory.edu/aneuriskweb]. Emory University, Department of Math&CS.