Journal of Informatics and Mathematical Sciences
Vol. 11, No. 1, pp. 31–40, 2019
ISSN 0975-5748 (online); 0974-875X (print)
Published by RGN Publications
DOI: 10.26713/jims.v11i1.890

http://www.rgnpublications.com

Research Article

Discretization of Unlabeled Data using RST & Clustering

Girish Kumar Singh* and Shrabanti Mandal

Department of Computer Science & Applications, Dr Harisingh Gour Central University, Sagar, India
***Corresponding author:** gkrsingh@gmail.com

Abstract. An algorithm can be applied on numerical or continuous attributes as well as on nominal or discrete value. If input to an algorithm required only attributes of nominal or discrete type then continuous attributes of the dataset need to be discretize before applying such algorithm. Discretization method can be of two types namely supervised and unsupervised. Supervised methods of dicretization utilize class labels of the dataset while in unsupervised method class labels are totally disregarded. In many literatures it has been shown that supervised methods gives good discretization result. Supervised algorithms cannot apply if dataset is unlabeled. In real life, many dataset do not have class (label) attribute and only unsupervised discretization methods are applicable in such cases. This paper presents discretization schemes for unlabeled data based on RST (Rough Set Theory) and clustering. The experiments have been performed to compare the proposed technique with other discretization methods for labeled data on two benchmark datasets. Two parameters Class-Attribute Interdependence Redundancy and the total number of intervals have been used to compare the proposed techniques with other existing techniques. The results display a satisfactory trade-off between the information loss and number of intervals for the proposed method.

Keywords. Discretization; Data mining; Rough set theory

MSC. 68P20 (Information storage and retrieval)

Received: April 13, 2018 **Accepted:** January 7, 2019

1. Introduction

Knowledge Discovery in Databases (KDD) is a sequence of processes, which includes data transformation, data selection, data preprocessing, data mining, interpretation and visualization of the mined patterns. Data preprocessing and data transformation produces the data in a suitable form on which data mining algorithm(s) can be applied. Data integration in which a single data matrix is created from data distributed over various resources is necessary tasks in transformation and in preprocessing if required, attributes are created. Exclusion of irrelevant attributes and discretization or grouping of the attributes values are also tasks of preprocessing. In discretization the values of continuous attribute are transformed into a finite number of intervals which are significantly less numerous than the continuous values. Discretization is a very necessary task in data mining [21], [22]. Discretization process is applied successfully in various areas such as soft computing bioinformatics, data mining [23].

Many researchers have worked in the field of discretization and many techniques have been developed. Generally discretization is a two step process. Number of discrete intervals is obtained in first step and the width or the boundaries of the intervals are obtained in the second step. There are some algorithms which can estimate the possible number of intervals instead of it is specified by the user. Ching et al. has proposed a heuristic rule to estimate the number of intervals [1]. In some techniques user need to specified the number of intervals.

Methods used for discretization can be classified on the basis of various criteria like local vs. global, supervised vs. unsupervised, univariate vs. multivariate and dynamic vs. static [2], [6], [21]. In global discretization technique all attributes are considered simultaneously and a conversion function is obtained for whole dataset while in local discretization conversion function is obtained only for single attribute [8], [4]. In supervised technique the class labels are taken into account [13], [8], [4], [15] while unsupervised technique does not consider the class labels. Multivariate techniques, consider all attributes concurrently to define the initial set of cut points or to decide the best cut point is also known as 2D discretization [18], univariate discretize only one attribute at a time. Multivariate techniques are more preferable as in complex problem attributes may highly interactive [19], [20]. In static technique the number of bins are presets, considering all attributes are independent attributes [13], [14]. In dynamic method an appropriate number of discrete intervals are anticipated based on the interdependencies of attributes.

Supervised techniques for discretization consider interdependence between class labels and the attribute values. Maximum Entropy [3], Patterson-Niblett [4], Information Entropy Maximization (IEM) [5], ChiMerge [8], Chi2 [9], Class Attribue Dependent Discretization (CADD) [1], Attribute Independence Maximization (CAIM) algorithm [2] are well known algorithms of this category.

Unsupervised discretization algorithm does not consider the attribute even if it present. Equal-width and Equal-frequency methods for discretization are example of this category [13],

[6], [8]. In equal-width, the range of attribute values divided into the pre-defined numbers of discrete intervals each of equal width. In equal-frequency method all values of the attribute is sort in ascending order and then the range divided into pre-defined number of intervals and each interval have same number of sorted attribute values.

The *Class Attribute Independence Maximization* (CAIM) algorithm [2] discretizes an attribute using class-interdependency [1]. CAIM algorithm uses a heuristic formula to calculate the number of intervals NFi for an attribute Fi and is given by $NFi = M/3C$, where M is the number of samples, and C is the number of classes.

The objective of discretization process is to minimum number of discrete intervals of a continuous attribute because large number of attribute values slows down the processing and hence inductive learning becomes ineffective [10]. Moreover, interdependence between discretized attribute's values and class labels should be maximize to ensure minimum loss of information due to process of discretization. A satisfactory tradeoff between these two objectives needs to be achieved [2].

Supervised discretization algorithms discretized the attribute values without much loss of information and number of intervals as compared to unsupervised algorithms. Also unsupervised techniques like equal width, equal frequency does not provide natural intervals. A method to discretization for labeled data using Clustering and *Rough Set Theory* (RST) is also proposed [17]. This technique explores clustering and RST to obtain the natural intervals. DBSCAN clustering has been employed to get the natural intervals and then approximations of the class attribute dependent concepts in RST are used to refine the intervals.

In this paper a method for discretization using RST and clustering for unlabeled dataset presented. To apply the RST labeled data is required and for this purpose PAM (*Partition Around Mean*) clustering algorithm has been used to create label of objects. The proposed method use the concepts explored in [17] to obtained natural intervals and the RST is used to refine these intervals.

The organization of this paper is as follows: the basic concepts of clustering, RST and method of descretization for labeled dataset using clustering RST (DUCRST) has been presented in Section 2. Section 3 described the proposed methods followed by the experimental detail in Section 4. Section 5 presents the results and analysis of the experiment and finally Section 6 conclude the paper.

2. Basic Concepts

2.1 Rough Set Theory

Rough Set theory (RST) introduced by Pawlak [11] is a technique which deals uncertainty in dataset. It can also be used as learning method to identify cause-effect relationship in databases.

A pair $S = (U, A)$, where U is a non-empty finite set called the universe and A is a finite set (non-empty) of attributes which describe the objects of U is called information system. For each

attribute α there is value set V_α that is $\forall\ \alpha \in A$, $\alpha : U \to V_\alpha$. An information system of the form $S = (U, A \cup D)$ is called decision System, where A is called the set of conditional attributes and D is called decision attribute and $A \cap D = \emptyset$.

For any $B \subseteq A$ an equivalence relation associated to B known as B-indiscernibility relation is defined by:

$$IND_s(B) = \{(x, x') \in U \times U \mid \alpha \in B\alpha(x) = \alpha(x')\}. \tag{2.1}$$

The equivalence classes of this relation are denoted by $[x]_B$. This indiscernibility relation is the basis of rough sets. Universal set U can be partitioned into equivalence classes using indiscernibility relation.

For any $X \subseteq U$, X can be described by the information contained in some $B \subseteq A$ using the B-lower approximation defined by $\underline{B}X = \{x : [x]_B \subseteq X\}$ and B-upper approximation given by $\overline{B}X = \{x : [x]_B \cap X \neq \emptyset$.

$\underline{B}X$ contained those objects which are certainty classified on the basis of the information in B as members of X, while $\overline{B}X$ contained those objects which can be possible members of X based on the information in B. The objects of boundary region defined by $BN_B = \overline{B}X - \underline{B}X$, cannot decisively classify as a member of X on the basis of information in B.

The B-outside region of X given by $U - \overline{B}X$ contained those objects, which are certainly not a member of X on the basis of information in B.

A set X is called *rough set* if the boundary region of the set is non-empty and if the boundary region is empty then set is called *crisp set*.

Let B be a set of conditional attributes then B-positive region $POS_B(D)$ using the relation $IND(D)$ is define as

$$POS_B(D) = \cup\{\underline{B}X : X \in D^*\} \tag{2.2}$$

where D^* is the partition corresponding to relation $IND(D)$.

The objects of positive $POS_B(D)$ region classified into distinct classes defined by $IND(D)$. An attribute set B is said to be highly significant if cardinality of $POS_B(D)$ is high.

Based on the information expressed by attribute set B, Rough membership function is used to measure how strongly an object x is a member of rough set X. Thus the significance of an attribute can be measure by Rough membership function which is given by,

$$\mu_X^B(x) = \frac{card(X \cap [x]_{IND(B)})}{card([x]_{IND(B)})}. \tag{2.3}$$

2.2 Clustering

Clustering is a process which groups data into groups or clusters. The objects of a cluster are highly similar, but are very dissimilar to object of other clusters. There are number of clustering algorithms has been proposed. Various approaches like partitioning, density-based, hierarchical, grid-based, nearest-neighbor, fuzzy, etc has been used in clustering techniques.

Partition Around Medoid (PAM) [16] which is proposed by Kaufman and Housseeuw has been used to label the data. PAM (k-medoid method) algorithm is robust in the presence of outliers. Clusters found by this method are insensitive to the order of input of objects and resultant clusters are invariant in case of orthogonal transformations and translations of data points. Medoid (also known as cluster representative) is the most centrally located object within the cluster. PAM's algorithm determines one medoid, for each cluster and then each non-selected object is grouped with the medoid to which it is most similar.

Density Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based approach which can discover clusters of arbitrary shape and noise in a dataset [12]. Natural intervals of attribute values are obtained using this method. DBSCAN based on the notion of neighborhood. If the distance between two objects is less than or equal to some fixed distance say Ep then these two objects are *neighbour* of each other. The neighborhood of an object o is the set of objects which are neighbour to o. If neighborhood of o contains at least minimum say *MinPt* number of objects then the object o is called *core object*.

An object p is *directly density-reachable* from an object q if q is a core object and p is in *neighborhood* of q. An object p is *density-reachable* from a point q if there is a chain of objects $q = p_1, p_2, \ldots, p_n = p$ such that p_{i+1} is directly density-reachable from p_i.

DBSCAN randomly select an object o and if this object is a core object with respect to two parameters Ep and *MinPt* then a cluster with o as a core object is created. The clusters are growing by adding all objects which are density reachable from a core object of the clusters.

2.3 Discretization method for labeled data using RST & Clustering

This method of discretization for unlabeled data is a two phase approach. Values of continuous attribute are partitioned using DBSCAN into natural intervals. Intervals obtained in first phase are refined in second phase using RST tools to have good discretization results.

Three threshold values namely *MaxPt*, *MinPt*, and *MaxLen*, are used to categorize intervals obtained in first phase into three categories namely small, large and normal. If the number of attribute values in an interval is less than *MinPt* then such interval is called small. For large interval I following criteria must be satisfied:

$(Card(I) > MaxPt)$ OR $(Range(I) > MaxLen)$ OR both.

An interval I will be called as normal if

$(MinPt \leq Card(I) \leq MaxPt)$ AND $(Range(I) \leq MaxLen)$.

Second phase optimize the number of partitions and loss of information while maintaining the significance of the attribute. In this phase neighboring small intervals are merged and large intervals are split into two or more intervals.

3. Proposed Discretization Method Unlabeled data

The proposed method of discretization uses the concepts of algorithm for discrtization of unlabeled data. To apply RST, the data must be labeled. The labeling data is required in refinement of intervals. To have labele of objects PAM clustering method is used to make clusters and for each cluster an cluster id is assigned. The assigned cluster id is used as label for each objects belonging to corresponding cluster.

Given an unlabeled dataset of N objects which are described by n attributes $A = \{a_1, a_2, \ldots, a_n\}$ with no class attribute. Let three parameters $MaxPt$, $MinPt$ and $MaxLen$ be the maximum and minimum number of points in any of the intervals, and the maximum range of values in the intervals respectively. Following are the steps of the proposed discretization method:

1: Store the distinct values of attribute a_i in $D[\ ]$.

2: Arrange the values in $D[\ ]$ in ascending order.

3: Run DBSCAN $(D, Eps, MinPt)$ to get natural intervals $In_1, In_2, \ldots, In_m$.

4: Refine_ul $(In_1, In_2, \ldots, In_m, MinPt, MaxPt, MaxLen)$ to get the optimal intervals.

Refine_ul$(In_1, In_2, \ldots, In_r, MinPt, MaxPt, MaxLen)$

do

For each interval In_j

If $(|In_j| \geq MinPt$ OR $Range(In_j) > MaxLen)$ then

Select the objects from the dataset for which the attribute value of a_i is in the interval I_j.

Make two clusters of selected objects using PAM.

Assign class label 1 to the objects of one cluster and 2 to the objects of the other cluster.

$IPart = Seed\_Point(a_i, \{v_{j_1}, v_{j2}, \ldots, v_{jk}\})$

Create two intervals $In_{j1} = [v_{j1}, IPart]$ and $In_{j2} = [IPart, v_{jk}]$ in place the In_j

elseif $|In_j| < MinPt$ then

If In_k is neighbour of In_j AND (number of points in both In_j and $In_k \leq MaxPt$) AND $(Range(In_j, In_k) \leq MaxLen)$ then merge intervals In_j and In_k

Endif

Endif

Endfor

while (no change in number of intervals) // End of do – while

Seed_Point$(a_j, \{v_{j1}, v_{j2}, \ldots, v_{jk}\})$ {

$I = [v_{j1}, v_{jk/2}]$ // $v_{ik/2}$ is the middle value of $\{v_{j1}, v_{j2}, v_{j3}, \ldots, v_{jk}\}$
$MaxRV = Max(\{f(a_i, c_p, I)\}) \ \forall \ c_p$,
for (each v_{jl}, $l = k/2$ down to 2) {
  $I = [v_{j1}, v_{jl}]$
    $TempRV = Max(\{f(a_j, c_p, I)\}) \ \forall \ c_p$;
    if $TempRV > MaxRV$ then
        $MaxRV = TempRV$;
      else
        break;
}
if ($l < k/2$) then return v_{jl} as seed point for the interval
  else
  for (each v_{jl}, $l = k/2$ to $k-1$) {
    $I = [v_{jl}, v_{jk}]$
    $TempRV = Max(\{f(a_j, c_p, I)\}) \ \forall \ c_p$;
    if $TempRV > MaxRV$ then
      $MaxRV = TempRV$;
    else
    v_{jl} as seed point for the interval
  }
} // end of function

The function $f(a_i, c_p, I)$ is rough membership function. For an interval $I = (v_1, v_j]$ of the values of attribute a for a class c_p is defined as

$$f(a, c_p, I) = (Card(\underline{a}_I), X_{C_p}) \setminus Card(X_{a,I}) \tag{3.1}$$

where $X_{a,I} = \{x \mid x \in U, a(x) \to I\}$ and $a_I, X_{C_p} = \{x \mid a(x) \to I, D(x) = C_p\}$.

4. Experimental Detail

The proposed method of discretization for unlabeld has been compared with other seven supervised and unsupervised methods for discretization. The seven algorithms are: equal-width, equal frequency, Patterson-Niblett, IEM, Maximum Entropy, CADD, CAIM and DUCRST. The first two algorithms equal-width, equal frequency are unsupervised algorithms for which the number of intervals has been estimated using formula $n_{Fi} = M/3C$ [3].

To apply both type of algorithm, two labeled datasets are used namely Iris Plants dataset (IRIS) and Pima Indians Diabetes dataset (PID) obtained from the UC Irvine ML repository. IRIS dataset has 150 samples and three class labels with four conditional attributes. PID is a

two class dataset having eight attributes and 768 samples. To apply unsupervised algorithm label/class attribute has not been used.

5. Results and Analysis

CAIR value [7] and total number of intervals are the two parameters used to evaluation of the different discretization algorithms. CAIR value is calculated using the concepts of class attribute mutual information and class-attribute joint entropy. Higher CAIR value shows higher interdependence between the class labels and the discrete intervals. CAIR value is also independent of number of unique values of the continuous attribute and number of class labels. Table 1 shows the CAIR value of different algorithms and Table 2 shows the total number of discrete intervals.

Table 1. CAIR Value based comparison of Discretization methods

Discretization Method	Datasets	
	Iris	Pid
Equal-width	0.40	0.058
Equal-frequency	0.41	0.052
Patterson-Niblett	0.35	0.052
IEM	0.52	0.079
Max.-Entropy	0.30	0.048
CADD	0.51	0.057
CAIM	0.54	0.084
DUCRST	0.56	0.107
Proposed Unlabeled Method	0.53	0.105

Table 2. Number of intervals based comparison of Discretization Methods

Discretization Method	Datasets	
	Iris	Pid
Equal-width	16	106
Equal-frequency	16	106
Patterson-Niblett	48	62
IEM	12	17
Max.-Entropy	16	97
CADD	16	96
CAIM	12	16
DUCRST	12	33
Proposed Unlabeled Method	12	37

The proposed discretization algorithm for unlabeled dataset achieved the high class-attribute interdependency and close to CAIM and DUCRST for both dataset. Regarding the total number of intervals the proposed algorithm generates equal number of intervals as generated by CAIM, IEM and DUCRST methods for the IRIS dataset which is lowest number of intervals. The number of intervals significantly higher than the number of intervals obtained by CAIM and IEM for PID dataset, but it is significantly less than the other discretization method.

6. Conclusion

This paper proposed method of discretization unlabeled data. In the proposed method the natural intervals are obtained with maximum mutual class-attribute interdependency and generates the possibly minimum number of intervals. The proposed method of also satisfies the different criteria to compare the discretization methods.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] J.Y. Ching, A.K.C. Wong and K.C.C. Chang, Class-dependent discretization for inductive learning from continuous and mixed mode data, *IEEE Trans. Pattern Analysis and Machine Intelligence* **17** 7 (1995), 641 – 651.

[2] L.A. Kurgan and K.J. Cios, CAIM discretization algorithm, *IEEE Trans. Knowledge and Data Engineering* **16**(2) (2004), 145 – 153.

[3] A.K.C. Wong and D.K.Y. Chiu, Synthesizing statistical knowledge from incomplete mixed mode-data, *IEEE Trans. Pattern Analysis and Machine Intelligence* **9** (7) (1987), 796 – 805.

[4] J.R. Quinlan, *C4.5 Programs for Machine Learning*, Morgan-Kaufmann (1993).

[5] U.M. Fayyad and K.B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, *Proc. 13th Int'l Joint Conf. Artificial Intelligence* (1993), 1022 – 1027.

[6] J. Dougherty, R. Kohavi and M. Sahami, Supervised and unsupervised discretization of continuous features, *Proc. 12th Int'l Conf. Machine Learning* (1995), 194 – 202.

[7] X. Wu, A bayesian discretizer for real-valued attributes, *The Computer Journal* **39**(1) (1996), 688 – 691, DOI: 10.1093/comjnl/39.8.688.

[8] R. Kerber, ChiMerge: discretization of numeric attributes, *Proc. Ninth Int'l Conf. Artificial Intelligence* (AAAI-91) (1992), 123 – 128.

[9] H. Liu and R. Setiono, Feature selection via discretization, *IEEE Trans. Knowledge and Data Eng.* **9**(4) (1997), 642 – 645.

[10] A.K.C. Wong and T.S. Liu, Typicality, diversity and feature pattern of an ensemble, *IEEE Trans. Computers* **24** (1975), 158 – 181.

[11] Z. Pawlak, Rough sets, *International Journal of Computer and Information Sciences* **11** (1982), 341 – 356.

[12] M. Ester, H.P. Kriegel, J. Sander and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (KDD'96) Portland, Oregon, 226 – 231 (1996).

[13] J. Catlett, On changing continuous attributes into ordered discrete attributes, in *Proceedings of the European Working Session on Learning*, Berlin, Germany, 164 – 178 (1991).

[14] R.C. Holte, Very simple classification rules perform well on most commonly used datasets, *Machine Learning* **11** (1993), 63 – 90.

[15] X. Wu and D. Urpani, Induction by attribute elimination, *IEEE Transactions on Knowledge and Data Engineering* **11**(5) (1999), 808 – 812.

[16] L. Kaufman and P.J. Rousueeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley & Sons (1990).

[17] G.K. Singh and S. Minz, Discretization using clustering and rough set theory, *International Conference on Computing: Theory and Applications* (ICCTA'07), 330 – 336 (2007).

[18] S. Mehta, S. Parthasarathy and H. Yang, Toward unsupervised correlation preserving discretization, *IEEE Trans. Knowledge and Data Eng.* **7**(9) (2005), 1174 – 1185.

[19] S. Ferrandiz and M. Boullé, Multivariate discretization by recursive supervised bipartition of graph, *Proc. Fourth Conf. Machine Learning and Data Mining (MLDM)* (2005), 253 – 264.

[20] P. Yang, J.-S. Li and Y.-X. Huang, HDD: A hypercube divisionbased algorithm for discretisation, *Int. J. Systems Science* **42**(4) (2011), 557 – 566.

[21] S. Garcıa, J. Luengo, J. A. Sáez, V. López and F. Herrera, A survey of discretization techniques: taxonomy and empirical analysis in supervised learning, *IEEE Transactions on Knowledge and Data Engineering* **25**(4) (2013), 734 – 750, DOI: 10.1109/TKDE.2012.35.

[22] J. Bai, K. Xia, Y. Chi and L. Liu, Continuous attribute discretization based on inflection point, *Journal of Information & Computational Science* **11**(4) (2014), 1327 – 1333, DOI: 10.12733/jics20103079.

[23] S. Ramirez Gallego, B. Krawczyk, S. Garcia, M. Wozniak and F. Herrera, A survey on data preprocessing for data stream mining: Current status and future directions, *Neurocomputing* **239** (2017), 39 – 57, DOI: 10.1016/j.neucom.2017.01.078.