ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)
Vol. 4, Issue 3 (July - Sept. 2017)
International Journal of Advanced Research in Education & Technology (IJARET)

An Entropy Based Effective Algorithm for Data Discretization

[I]Priyanka Das, [II]Sarita Sharma
[I]M.Tech. Scholar, MATS University, Aarang, Raipur, Chhattisgarh, India
[II]Asst. Prof., MATS University, Aarang, Raipur, Chhattisgarh, India

Abstract
The task of extracting knowledge from databases is quite often performed by machine learning algorithms. Many algorithms can only process discrete attributes. Real-world databases often involve continuous features. Those features have to be discredited before using such algorithms. Discretization methods can transform continuous features into a finite number of intervals, where each interval is associated with a numerical discrete value. This paper analyzed existing data discretization techniques for data preprocessing. Firstly, the importance and process of discretization is studied. Furthermore, we conduct an experimental study of discretization methods involving the most representative and newest discretizers.
A new entropy based data discretization is also proposed in the proposed system. Experimental results show better accuracy in classification.

Keywords
Discretization, Continuous data, Data Mining, Classification.

I. Introduction

The process of automation of processing and extract ion of knowledge from data becomes an important task that is often performed by machine learning (ML) algorithms. One of the most common tasks performed by ML algorithms is generation of classification rules from class-labeled examples. T he examples are described by a set of numerical, nominal, or continuous attributes. Many existing inductive ML algorithms are designed expressly for handling numerical or nominal data, while some algorithms perform better with discrete-valued attributes despite the fact that they can also handle continuous attributes [1][9].This drawback can be overcome by using a discretiza tion algorithm as a front-end for the learning algo rithm. Discretization is a process of transforming a continuous attribute values into a finite number of intervals and associating with each interval a discrete, numerical value. The usual approach for learning tasks that use mixed-mode (continuous and discrete) data is to perform discretization prior to the learning process [1][7] [8][12]. The discretization process first finds the number o f discrete intervals, and then the width, or the boundaries for the intervals, given the range of values of a continuous attribute. Most often the user must specify the number of interval s, or provide some heuristic rule to be used [2]. The proposed CILA algorithm performs both tasks by automatically selecting the number of discrete intervals and finding width of every interval based on interdependency between class and attribute values. Discretization algorithms can be divided into two categories: 1. Unsupervised (class-blind) algorithms that discretize attributes without taking into account respective class labels. The representative algorithms are equal-width and equal -frequency discretizations [3]. 2. Supervised algorithms discretize attributes by taking into account the class-attribute interdependence. The representative algorithms are: Patterson-Nibett algorithm [11], which is built-in as a front en d into a decision trees algorithm [13] [19], statistics based algorithms like Chi Merge [9] or Chi2 [10], class-attribute interdependency algorithms like CADD algorithm [2] and clustering-based algorithms like K- means Discretization algorithm [15].

Discretization should significantly reduce the number of possible values of the continuous attribute since large number of possible attribute values contributes to slow and ineffective process of inductive machine learning [1]. Thus, a supervised discretization algorithm should seek possibly minimum number of discrete intervals, and at the same time it should no t weaken the interdependency between the attribute values and the class label. The proposed CILA discretization algorithm not only discretizes an attribute into the small number of intervals but also makes it easier for the subsequent machine learning task by maximizing the class-attribute interdependency. In this paper we discuss machine learning, especially inductive learning and its related algorithms, and a new information theoretic discretization method optimized for supervised and unsupervised learning methods is proposed and described, so we proposed improvement of ILA [1 4] which is called Continuous Inductive Learning Algorithm (CILA), and tested in inductive learning example to show how the discretization methods deal with continuous attributes. Inductive learning system can be effectively used t o acquire classification knowledge from examples; m any existing symbolic learning algorithms can be applied in domains with continuous attributes when integrated with discreti zation algorithms to transform the continuous attributes into ordered di scretization. Most of the existing machines learning algorithms are able to extract knowledge from databases that store discrete attributes (features). If the attributes are continuous, the algorithms can be integrated with a discretization algorithm that transforms them into discrete attributes. The attribute in learning problem may be nominal (categorical), or they may be continuous (numerical) . The term “continuous” is used in the literature to indicate both real and integer valued attributes . Continuous-valued attr ibute must , therefore, be discretized prior to attribute selection. There are many ways to discretize a continuous attribute , so to partition a continuous variable two important decisions must be made [2]: 1- The number of discrete intervals must be selected, the selection of the optimal number of intervals is seldom addressed by existing discretization methods, and in most cases the human user selects the appropriate number of intervals. 2- The width of the intervals must be determined, which mean the boundaries of the discretized intervals need to be defined. The simplest discretization procedure is to divide the range of a continuous variable into equal-width intervals, which determined the range of values from the minimum and maximum ob served attribute values, the range is then divided equally by a user- defined number of intervals, but the weakness of the is procedure is that in case where the outcome observations are not distributed evenly, a large amount of important information can be lost after the discretization

ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)

process, to solve this problem a method based on the concept of maximum marginal entropy has been developed, this method partitions a continuous attributes using a criterion to maximize shannon's entropy and this minimize the loss of information. Once the number of interval is selected.

II. Methodology

The Main objective of this research work is to enhance the quality of different datasets using a discretization algorithm known as "Effective Entropy based CAIM algorithm .Methodology is implemented by using MATLAB which is a Multi paradigm numerical computing Environment. It is a programming language that allows matrix manipulation, plotting of functions and various kind of data and also implements algorithm. It creates various user interface and interfacing with programs that are written in other languages. This chapter discusses the implementation of Effective Entropy based discretization algorithm and the various component of multi paradigm numerical computing environment.

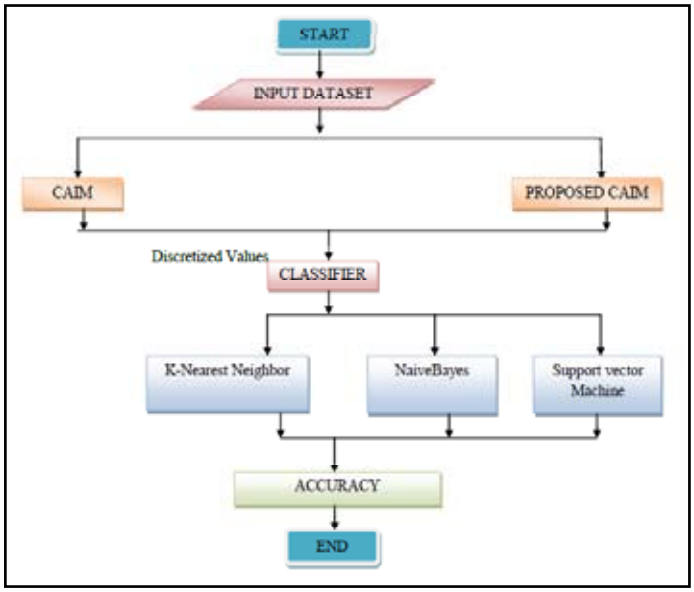

Fig. 2.1 Proposed Methodology

It also provides description of the various parameters and analysis used in this study. In Proposed CAIM, using different classifiers and various balanced datasets, calculate in different parameters like Accuracy, No. of intervals, Runtime etc. for discretization using class-attribute interdependency maximization based on Entropy

A. Algorithm Step

Step 1 Initially, the "Effective – Entropy based CAIM algorithm" starts scanning the different types of datasets given as input.
Step 2 Each input datasets are processed through "Effective – Entropy based CAIM algorithm" and calculates discretized values for each input dataset.
Step 3 Discretized values from Proposed Algorithm processed through classifiers i.e. KNN, NB &SVM.
Step 4 And accuracy is calculated by these classifiers.
Step 5 End

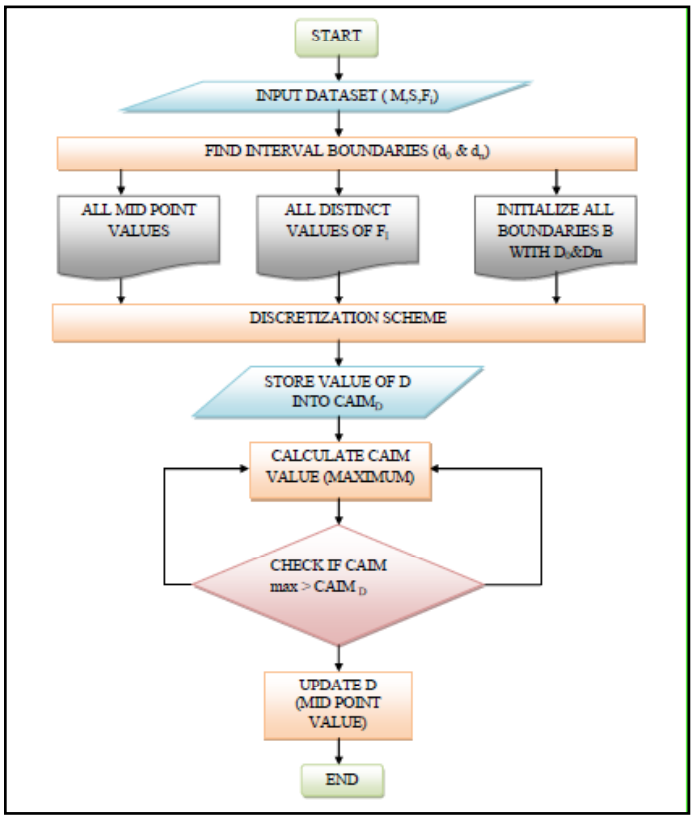

Fig. 2.2 : Basic Frame work of the proposed methodology

B. Description of the CAIM

The Class-Attribute Interdependency Maximization (CAIM) criterion measures the dependency between the class variable C and the discretization variable D for attribute F, for a given quanta matrix, and is defined as:

$$CAIM(C,D|F) = \frac{\sum_{r=1}^{n} \frac{\max_r{}^2}{M_{+r}}}{n},$$

where n is the number of intervals, r iterates through all intervals, i.e., r =1,2,….n, maxr is the maximum value among all qir values (maximum value within the rth column of the quanta matrix), i=1,2,3…….S, M+r is the total number of continuous values of attribute F that are within the interval (dr-1,dr].

C. Algorithm steps for CAIM

The algorithm consists of these two steps: .

- Initialization of the candidate interval boundaries and the initial discretization scheme and
- Consecutive additions of a new boundary that results in the locally highest value of the CAIM criterion.

The pseudocode of the CAIM algorithm follows.
Given: Data consisting of M examples, S classes, and continuous attributes Fi
For every Fi do:
Step 1.
1.1 Find maximum (dn) and minimum (do) values of Fi.
1.2 Form a set of all distinct values of Fi in ascending order, and initialize all possible interval boundaries B with minimum, maximum and all the midpoints of all the adjacent pairs in the set.

1.3 Set the initial discretization scheme as D :{[d0 dn]} , set GlobalCAIM=0.
Step 2.
2.1 Initialize k ¼ 1.
2.2 Tentatively add an inner boundary, which is not already in D, from B, and calculate corresponding CAIM value.
2.3 After all the tentative additions have been tried accept the one with the highest value of CAIM.
2.4 If (CAIM > GlobalCAIM or k < S) then update D with the accepted in Step 2.3 boundary and set GlobalCAIM=CAIM, else terminate. 2.5 Set k ¼ k þ 1 and go to 2.2.
Output: Discretization scheme D

D. Entropy Based Discretization

A conditional entropy of the decision d given an attribute a is

$$H(d|a) = -\sum_{j=1}^{m}(pj).\sum_{j=1}^{n} p(d1|d2).\log p(d1|d2)$$

Where a= a1, a2,.... am are all values of a and d1, d2,.... dn are all values of d. There are two fundamental criteria of quality based on Entropy.

- The first is an information gain associated with an attribute a and defined by

$$I(a) = Hd(U) - H(d|a)$$

- The second is information gain ratio, for simplicity called gain ratio, defined by

$$G(a) = \frac{I(a)}{Ha(U)}$$

E. Algorithm of Entropy based Discretization

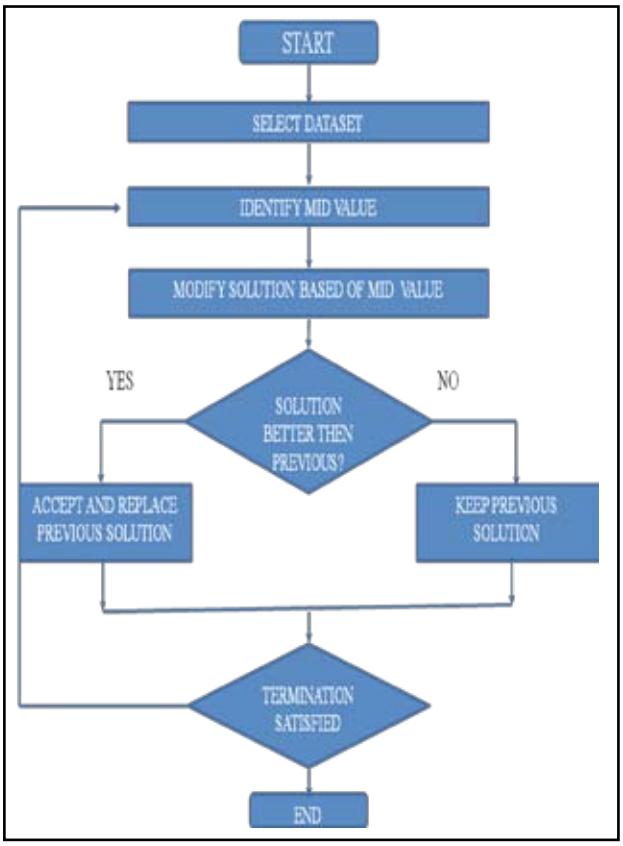

Fig. 2.3 : Flowchart of Entropy based Discretization

F. Algorithm Steps for entropy based discretization algorithm

Step1 Input: Dataset with *i* continuous attribute, *M* examples and *S* target classes;
Step2 Begin
Step 3 For each continuous attribute *Ai*
Step 4 Find the maximum *dn* and the minimum *d0* values of *Ai*;
Step 5 Form a set of all distinct values of *A* in ascending order;
Step 6 Initialize all possible interval boundaries *B* with the minimum and maximum
Step 7 Calculate the midpoints of all the adjacent pairs in the set;
Step 8 Set the initial discretization scheme as *D*: {[*d0*,*dn*]}and *Globalentropy*= 0;
Step 9 Initialize *k* = 1;
Step 10 For each inner boundary *B* which is not already in scheme *D*,
Step 11 Add it into *D*;
Step 12 Calculate the corresponding *entropy* value;
Step 13 Pick up the scheme *D'* with the highest *entropy* value;
Step 14 If *entropy* > *Globalentropy* or *k* < *S* then
Step 15 Replace *D* with D';
Step 16 *Globalentropy* = *entropy*;
Step 17 *k* = *k* + 1;
Step 18 Goto Line 10;
Step 18 Else
Step 19 *D'* = *D*;
Step 20 End If
Step 21 Output the Discretization scheme *D'* with *k* intervals for continuous attribute *Ai*;
Step 22 End

G. Training and Testing of Datasets Using Classifier

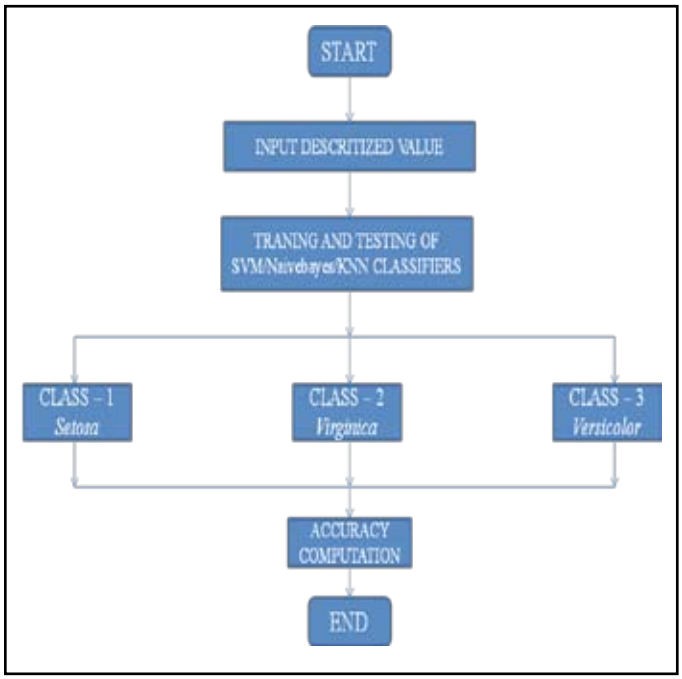

Fig 2.4 : Flowchart of Training and Testing of Datasets Using Classifier

For these classifiers calculate parameter (Accuracy) for all three datasets used.

ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)

III. Experimental Results

The goal of this research work is to study the improved quality of various datasets after applying Effective Entropy based CAIM algorithm which is a discretization based approach in data mining. Pre-generated scenario files are used to subject each algorithm to the same set of scenario in an identical fashion to perform a fair comparison. This chapter presents in details about the performance metrics, Input and Output and its parametric evaluation and comparatative analysis.

A. Performance Metrics

The performance metrics are used to quantitatively evaluate the Effective Entropy based CAIM algorithm. In this research after discretization and classification, evaluation is done the basis of following parameter.

Accuracy

The general term accuracy is used to describe the closeness of a measurement to the true value. When the term is applied to sets of measurements of the same measure and, it involves a component of random error and a component of systematic error. In this case trueness is the closeness of the mean of a set of measurement results to the actual true value and precision is the closeness of agreement among a set of results.

$$Accuracy = \frac{Number\ of\ correct\ classification}{Total\ number\ of\ test\ cases}$$

Taking those data which full fill the termination criteria will increase the probability of discretization.

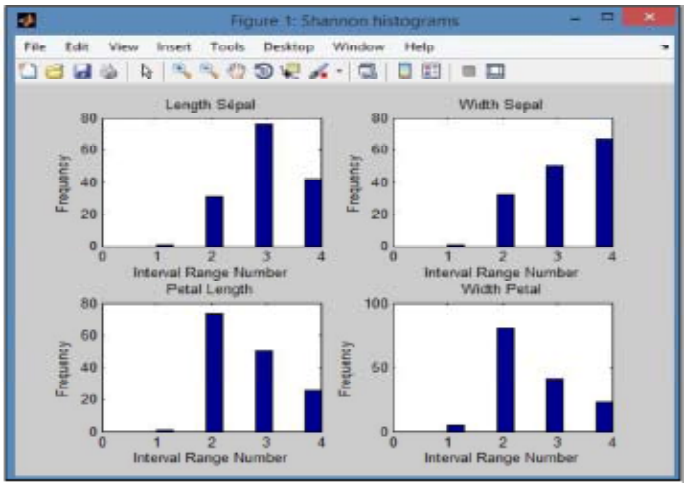

Fig. 3.1 : Discretized value of the iris dataset.

More over this paper includes the goal to generate an environment where the Entropy based Discretization algorithm, can be implemented and tested.

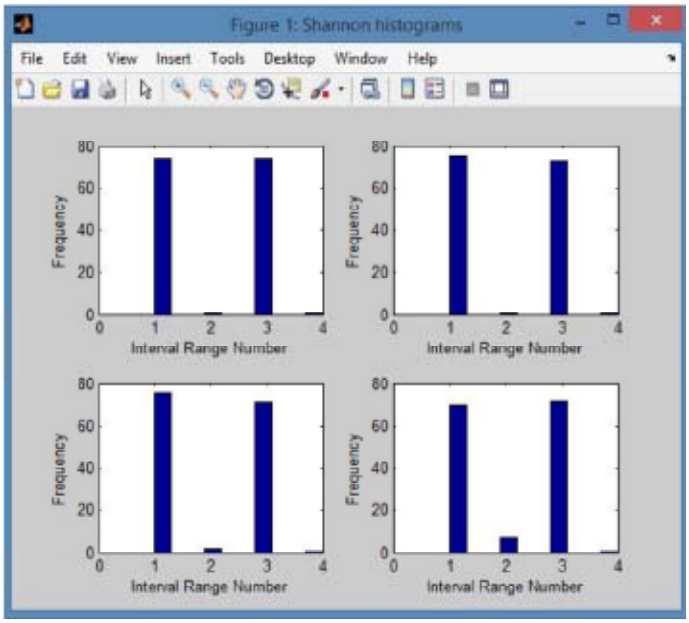

Fig. 3.2 : Discretized value of the wine dataset

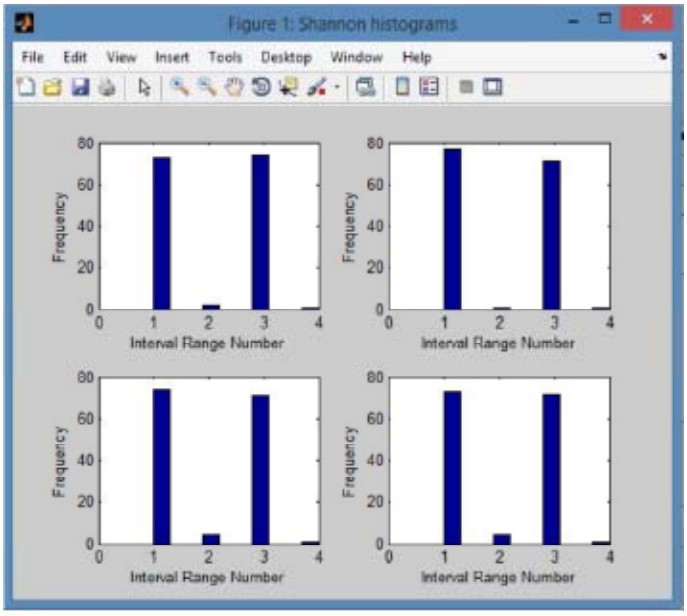

Fig. 3.3 : Discretized value of the Glass Dataset

The discretized values of Proposed caim algorithm are scanned by three different types of classifier i.e knn, svm, naive bayes. And the criteria are compared with the previous work. Also the below graphs shows values of different datasets after discretization using Effective Entropy based CAIM algorithm.

Table 3.1 : Comparison of Effective CAIM algorithm from different discretization algorithm

CAIM vs	Accuracy		
	KNN	NaiveBayes	SVM
EW	0.22	0.003	0.028
EF	0.0002	0.0010	0.028
CAIM	0.02	0.021	0.02
CACC	0.0039	0.002	0.021
Proposed CAIM	0.31	0.26	0.31

The Table 3.1 shows the evaluation results obtained for the said metrics after performing the discretized values of all three dataset i.e Iris, Wine and Glass using Entropy based caim algorithm. The table shows the classification performance for each of the classifier defined in the proposed algorithm. The last row represents the metric values with weighted average taken over all the classes using KNN, SVM and Naive Bayes classifier. The evaluated values shows better accuracy rate in comparison to other different discretization methods that is Equal Width, Equal Frequency and Class-Attribute contingency coefficient.

Average accuracy is computed by combining the accuracy obtained by all the dataset and applying the proposed entropy based discretization algorithm and then divided the total accuracy by the total number of different dataset.

Here , the below graphs shows average accuracy rate of three classifiers used for testing of unseen data from three different datasets that is Iris, Wine and Glass.

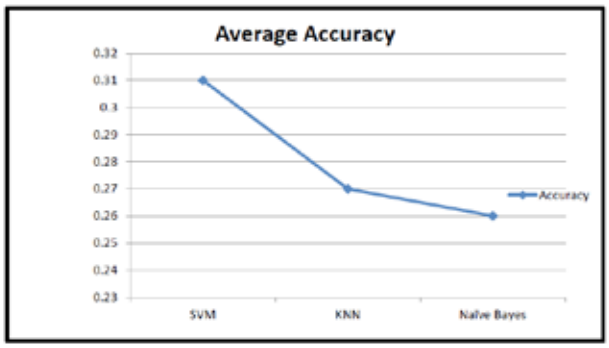

Fig. 3.4 : Average accuracy graph of svm, knn and naïve bayes classifier.

Figure 3.4 shows the evaluation results obtained for the said metrics after performing the discretized values of all three dataset i.e Iris, Wine and Glass using Entropy based caim algorithm. The graph shows the classification performance for each of the classifier defined in the proposed algorithm. It represents the metric values with weighted average taken over all the classes using KNN, SVM and Naive Bayes classifier. On comparing above three classifiers, SVM gives the highest accuracy rate as compare to KNN and Naive Bayes classifier. The accuracy rate of SVM is 31% , KNN is 27% and Naves Bayes is 26% which is lowest among three.

IV. Conclusion

To conclusion of this research work can be encapsulated as follows:

- The Effective Entropy based CAIM algorithm is compared with Ur-CAIM algorithm.
- Both algorithms are tested and analyzed in terms of Accuracy. The results of the experiment shows that the Effective Entropy based CAIM algorithm gives better results than CAIM algorithm.
- On comparing the above two experiments, it is found that Accuracy Rate parameter is improved.

References

[1] **Salvador Garc´ıa, Juli ´an Luengo, Jos´e A. S´aez, Victoria L´opez, and Francisco Herrera** "A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning" *IEEE Transactions On Knowledge And Data Engineering*, Oct 2011.

[2] **A.K.C. Wong and D.K.Y. Chiu**, "Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 6, pp. 796-805, Nov. 1987.

[3] **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*, pp. 1022-1027, 1993.

[4] **Xiao-Hang Zhang, Jun Wu, Ting-Jie Lu, Yuan Jiang** "A discretization algorithm based on gini criterion" , *Proceedings Of The Sixth International Conference On Machine Learning And Cybernetics, Hong Kong,* 19-22 august 2007.

[5] **Lukasz A. Kurgan, and Krzysztof J. Cios**, "CAIM Discretization Algorithm" IEEE *Transactions On Knowledge And Data Engineering, Vol*. 16, *No*. 2, *February* 2004.

[6] **Cheng-Jung Tsai , Chien-I. Lee , Wei-Pang Yang** "A discretization algorithm based on Class-Attribute Contingency Coefficient" Information Sciences 178 (2008) 714–731

[7] **Alberto Cano · Dat T. Nguyen · Sebastián Ventura · Krzysztof J. Cios "**ur-CAIM: improved CAIM discretization for unbalanced and balanced data**"** Springer-Verlag Berlin Heidelberg 2014

[8] **Nor Liyana Mohd Shuib1, Azuraliza Abu Bakar2 and Zulaiha Ali Othman** "Performance Study on Data Discretization Techniques Using Nutrition Dataset" 2009 International Symposium on Computing, Communication, and Control (ISCCC 2009) Proc .of CSIT vol.1 (2011) © (2011) IACSIT Press, Singapore.

[9] **Z. Marzuki, F. Ahmad "**Data Mining Discretization Methods and Performances" Proceedings of the International Conference on Electrical Engineering and Informatics Institut Teknologi Bandung, Indonesia June 17-19, 2007

[10] **Lukasz Kurgan 1, and Krzysztof Cios** "Fast Class-Attribute Interdependence Maximization (CAIM) Discretization Algorithm**"** Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada,

[11] **Shivani V .Vora R.G.Mehta** "MCAIM: Modified CAIM Discretization Algorithm for Classification" International Journal of Applied Information Systems (IJAIS) – Volume 3– No.5, July 2012 – www.ijais.org.

[12] **Lukasz A. Kurgan, Member, IEEE, and Krzysztof J. Cios, Senior Member, IEEE** "Caim discretization Algorithm" Ieee Transactions On Knowledge And Data Engineering, Vol. 16, No. 2, February 2004.

[13] **B.Hemada, K.S.Vijaya Lakshmi "**Discretization Technique Using Maximum Frequent Values and Entropy Criterion"

International Journal of Advanced Research in Computer Science and software Engineering Volume 3, Issue 11, November 2013.

[14] **S. García, 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