An Improved SVM Classifier for Discretization of Attributes using K-Means Clustering

Neelima Dixit
Department of Information Technology
Samrat Ashok Technological Institute
Vidisha (M.P.) India

ABSTRACT

Here in this broadside a novel approach for the Discretization of Nonstop Characteristics for the Classification of various datasets is proposed. The Planned Procedure implemented here works in Two Phases, in the first stage K-means Clustering is applied on the dataset to cluster the data on the basis of classes available in the dataset and second is to classify the Clustered Data using Support Vector Machine Classifier. The various Untried results achieved on different datasets proves that the planned procedure provides less mean number of cuts and reduced mean discretization time and also provides higher accuracy with better Scalability.

Keywords

Discretization, K-Means, Support Vector Machine, Clustering, Classifier.

1. INTRODUCTION

Rough Set Model was introduced by Z. Pawlak in 80's to assure they require for a prescribed structure to run in definite information expressed in terms of data obtained from experiments [1]. This theory was originally extended for a finite space of communication in which the information base is a separation, which is acquired by any correspondence relation described on the creation of conversation. In rough sets model, the information is categorized in a table baptized conclusion table. Rows of the conclusion table are in contact to objects and pilasters resemble to characteristics. In the statistics set, a class label to specify the class to which each row belongs. The class sticker is called as result characteristic the respite of the characteristics are the disorder characteristics. Here, C is used to signify the circumstance attributes, D for decision attributes, where $C \cap D = \Phi$, and tj denotes the j^{th} tuple of the data table. Rough sets theory describes three districts based on the corresponding classes encouraged by the characteristic standards: lower estimate, upper calculation, and boundary. Subordinate estimate contains all the objects which are classified definitely based scheduled the information composed and Upper calculation comprises all the substances, which can be classified almost certainly while the boundary is the variation between the upper estimate and the lower approximation. Hu et al., [2] presented the formal definitions of coarse set speculation. Imprecise refers to the fact that the granularity of knowledge causes indiscernibility. Uneven set philosophy outlines three districts based on the corresponding classes encouraged by the characteristic values these imprecise concepts can be defined approximately with available knowledge using three detailed notions called subordinate estimate ($\underline{R}X$) and upper guess (RX) and frontier. Lower estimate comprises all the substances, which are confidential confidently based on the information composed, and superior estimate comprises all the substances which can be classified probably, while the boundary is the difference between the upper estimate and the lower approximation. So, we can define a coarse set as any set distinct finished its subordinate and upper guesses. Alternatively, indiscernibility idea is essential to rough set theory. Informally, two substances in a conclusion table are imperceptible if one cannot differentiate amongst them on the derivation of a assumed set of characteristics. For this motive, indiscernibility is a meaning of the set of characteristics under concern. For each customary of characteristics they can consequently define a second indiscernibility relative, which is a gathering of couples of substances that are invisible to each other.

Let I = (U; A) be an material organization (attribute value system), anywhere U is a non-empty customary of determinate objects and A is a non-empty, determinate set of characteristics such that a: $U \rightarrow V_a$ for every $a \in A.V_a$is the set of values that attribute a may revenue. The evidence table allocates a charge a(x) from $V_a$to each characteristic a and article x in the cosmos U. With any $R \subseteq A$ there is an allied correspondence relative $IND(R) = \{(x; y) \in U^2 | \forall a \in R; a(x) = a(y)\}$. The relation IND(R)is called a R-indiscernibility relation. The partition of Uis a domestic of all correspondence courses of IND(R) and is denoted by U=IND(R).

Let $X \subseteq U$ be a target set that they aspiration to characterize using quality subsection P ;i.e., they are expressed that an chance set of substances X comprise a solitary class and they aspiration to communicate this class i.e., this subdivision using the correspondence courses encouraged by characteristic subsection R. In wide-ranging, X cannot be uttered accurately for the reason that the set may contain and prohibit objects which are impossible to differentiate on the beginning of attributes R.

In recent times, the discretization of uninterrupted attributes has gained significant attention in rough set theory. Many conventional discretization measures have been valuable to coarse sets [3]. Singh and Minz proposed a discretization approach implemented on grouping and coarse set conjecture [4]. Blajdo et al. compared the results of six promising discretization approaches from the standpoint of rough sets [5]. Tian et al. proposed a core-generating discretization method, which was used as the pre-processor of coarse set-based article assortment [6].

The main knowledge of coarse set philosophy is on an assumption of every instance or object is associated some information. The objects whose characteristics are defined as same, they are referred as indiscernible (or precise or similar) with respect to available information. The indiscernible objects in the set can be formed as a basic granule called elementary set. Since, available information has a granular structure; some objects can be framed as indiscernible whereas other objects can be vague which means these objects whose characteristics are not defined as same from available information. To resolve vagueness the concepts in uneven set philosophy are subordinate estimate and upper estimate. This concept works well if the data is a qualitative data, where each attribute can have limited number of distinct values. But if the data is quantitative, where attributes are continuous valued like length, age or speed etc., then the indiscernibility of occurrences can be unhurried based on familiarity of its values. By applying discretization [7-8] on continuous valued attributes they appear to grade discernibility between instances.

2. LITERATURE SURVEY

Here they propose [9] a overseen and multivariate discretization procedure — SMDNS in uneven sets, which is resulting from the conservative algorithm naive scaler i.e. Naive. Here in this algorithm use a decision table DT (U,C,D,V,f), in view of the fact that SMDNS uses both class information and the interdependence among different circumstance attributes in C to decide the discretization method the cuts acquired by SMDNS are much less than those acquired by Naive at the same time as the classification capability of DT remnants unaffected after discretization. To progress the computational presentation of SMDNS, they utilized the counting sort based technique to calculate the partitions of the universe and deleted the steps in SMD for scheming the consequence of each circumstance characteristic. Experimental results give you an idea about that the running time of SMDNS is much less than that of SMD. Particularly, they have confirmed that SMDNS can assurance that the classification capability of the given decision table continues unaffected after discretization.

Here in this research work is to evaluate the feature of these four discretization techniques using two criteria: an error rate evaluated by ten-fold cross-validation and the size of the decision tree generated by C4.5. Experimental results offered in [10] show that manifolds perusing is the most excellent discretization technique amongst these four discretization approaches. In [10], four discretization techniques were evaluated using a rule-based method. There is a opportunity that the results of depend on the selection of experimental set of connections. Consequently, to eliminate this preconception, they changed the unique setup and accomplished novel experiments using the typical C4.5 decision tree generation method. Our novel results entirely maintain the results of [10]. For 17 numerical datasets, four sets of experiments were performed: first, the C4.5 organization was used to analyze an mistake rate using ten-fold cross authentication; followed by, the similar techniques (equal intermission width and equivalent occurrence per interval) and multiple scanning, and for such discretized datasets, the similar C4.5 system was utilized to create an error rate.

Here author has to find the similar technique founded on calculating the C4.5 mistake rate was used in [11] to evaluate nine successful and recognized discretization techniques using 11 datasets. Seven of these 11 datasets (australian, bup a, glass, ionosphere, iris, pima and wine gratitude) were also utilized in our trials. For any of these seven datasets the most excellent result proficient using our techniques are enhanced than the equivalent most excellent result mentioned in [11]. Consequently, their choice for the four discretization techniques is well acceptable: they were used very proficient techniques. Their results give you an idea about that the multiple scanning discretization method is considerably enhanced than the internal discretization used in C4.5 and two globalized discretization techniques: equal intermission width and equivalent incidence per interval in expressions of the error rate computed by ten-fold cross-validation i.e. two-tailed test, 5% level of importance. In addition, decision trees produced from data discretized by multiple scanning are considerably humbler than choice trees produced directly by C4.5 and conclusion trees produced from datasets discretized and both globalized discretization techniques.

In the paper [12], a wide-ranging method has been proposed in developing IDS where RST and Q-learning algorithm are included to provide accommodation real time traffic data for sensing impositions with maximum classification correctness. Rough set theory is applied on discrete data only and so in the effort cut is applied on restricted attributes for discretization. Indiscernibility idea of RST is useful on discrete data for selecting set of most important attributes called reduct adequate to stand for the original data set. On the other hand, reduct is not exceptional and so the reduct which make available highest classification correctness is chosen to build the rule base classifier to classify the system traffic information either normal or anomaly. For the test data, concerning the similar cut value may produce unusual reduct, resulting fall of classification correctness. Consequently, discretization and article assortment are not to be take care of as independent events to classify network data precisely. In the proposed method, Q-learning algorithm has been modified to become skilled at different cut value for each restricted attribute and equivalent reduct and correctness are estimated to form the reward matrix. Modified Q matrix estimates optimum cut standards for each characteristic to accomplish highest classification correctness in detecting intrusions using network traffic data. The system is finished when two consecutive cut produces similar correctness or monotonically decreasing correctness.

3. PROPOSED METHODOLOGY

The Planned Procedure implemented here consists of following Steps:

1. Take an input Dataset.
2. Apply K-Means Clustering on the input dataset.
3. Discretization of Data on the basis of clustered data.
4. Apply Support Vector Machine based Classifier to Classify the data.
5. Generate Decision Tree from the SVM classifier.
6. Generate Rules from the Generated Decision Tree.

Datasets

For the Analysis and Comparison of the Planned Methodology with the Present Procedure Six Datasets are used.

(1) Ecoli data set (Ecoli),

(2) Iris Plants data set (Iris),

(3) Heart Disease data set (Heart),

(4) KDD Cup 1999 data set (KDD-99).

(5) Pima Indians Diabetes data set (Pima), and

(6) Glass Identification data set (Glass),

The Datasets used here for the analysis and comparison of Algorithms.

Table 1. Analysis of Six Datasets used

Property	Dataset					
	Ecoli	Glass	Heart	Pima	Iris	10% KDD
No. of Classes	8	7	5	2	3	4
No. of Objects	336	214	303	768	150	50
No. of Attributes	8	10	14	9	5	42

3.1 K-Means Clustering

K-means is unique of the meekest unverified education procedures that resolve the well identified bunching problematic. The technique shadows a modest and easy way to organize a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main impression is to outline k centroids, one for each group. These centroids would be positioned in a astute way since of diverse position grounds different result. So, the superior preference is to position them as much as promising far away from each other. A round has been introduced.

Finally, this procedure aims at reducing an *objective function*, in this case a squared error meaning. The objective function

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\| x_i^{(j)} - c_j \right\|^2$$

Where,

$\left\| \mathrm{x}_\mathrm{i}^{(\mathrm{j})} - \mathrm{c}_\mathrm{j} \right\|^2$ is a selected aloofness degree among a statistics opinion $\mathrm{x}_\mathrm{i}^{(\mathrm{j})}$and the group centre c_j, is an needle of the aloofness of the *n* information opinions from their individual cluster middles.

The procedure is collected of the subsequent stepladders:

1. Dwelling K arguments into the planetary characterized by the matters that are actuality grouped. These arguments epitomize initial assembly centroids.
2. Disperse each article to the assemblage that has the bordering centroid.
3. When all matters have been allotted, recalculate the sites of the K centroids.
4. Duplication Steps 2 and 3 pending the centroids no lengthier move. This harvests a departure of the substances into assemblages from which the metric to be curtailed can be considered.

3.2 SVM Classifier

Consider training sample$\{(\mathrm{x_i}, \mathrm{d_i})\}$, where $\mathrm{x_i}$ is the input pattern, d_i is the desired output:

$$aW_0^T X_i + cb_0 \geq +1, for d_i = +1$$

$$aW_0^T X_i + cb_0 \leq -1, for d_i = -1$$

Also the weight vector w should minimize the cost function

$$\varphi(W) = \frac{1}{2}W^T W$$

The data point which is very near is called the margin of separation ρ

The foremost purpose of using the SVM is to treasure the precise hyper plane of which the margin ρ is subjugated most favorable hyper plane a

$$W_0^T X + cb_0 = 0$$

For example, if we are choosing our model from the set of hyperplanes in *Rn*, then we have:

f(*x; {w; b}*) = sign(*w* . *x* + *b*)

We can try to acquire *f*(*x; _*) by indicating a meaning that completes well on exercise information:

$$J(w, b, \alpha) = \frac{1}{2}W^T W - \sum_{i=1}^{N} \alpha_i \left[d_i\left(W^T . x^i + b\right) - 1\right]$$

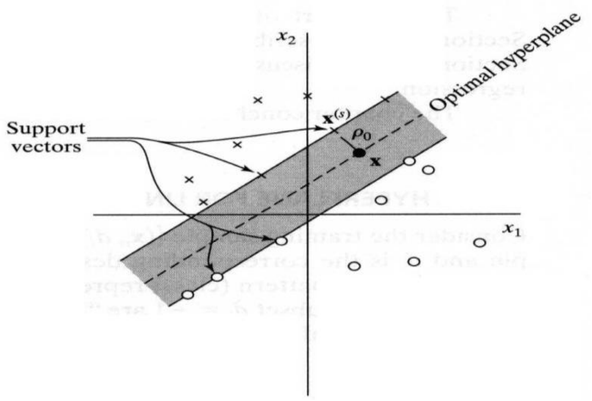

Fig 1:Basic Architecture of SVM

4. RESULT ANALYSIS

The Table shown below is the examination and judgment of mean number of cuts and the mean discretization time between the existing MSDNS algorithm and proposed algorithm. The planned procedure shows better presentation in comparison with SMDNS algorithm.

$$\begin{aligned} Discretization\ Time \\ &= Current\ Time \\ &- Time\ before\ Discretization \end{aligned}$$

Table 2.Assessment of discretization schemes on 10% KDD

Algorithm	Mean No. of Cuts	Mean Discretization Time (s)
SMDNS	70	69.8
Proposed	55	58.3

The Table shown below is the examination and judgment of Detection Rate between existing SMDNS Algorithm and the proposed algorithm.

The Comparison done here is on the basis of each type of attack and Detection rate of all Categories. The planned method has better Detection rate in Assessment to the prevailing algorithm.

$$DR = \frac{No. of\ Attacks\ Detected}{Total\ No. of\ Available\ Attacks}$$

Table 3. Comparison of Classification results on 10% KDD

	DR for each attack category (%)				DR for all attack category (%)
Algorithm	DoS	R2L	U2R	Probe	
SMDNS	99.9688	96.1962	67.3077	98.2621	99.8113
Proposed	99.9723	97.453	70.17	99.64	99.912

The Table shown below is the examination and judgment of Average Cataloguing Accurateness between existing SMDNS Algorithm and the proposed algorithm.

The Comparison done here is on the basis of each type of Datasets. The projected tactic has better Average Cataloging Accurateness in Comparison to the existing algorithm.

$$Accuracy = \frac{Correctly\ Classified\ Values}{Total\ Data\ in\ Dataset}$$

Table 4. Comparison of Average Classification Accuracy

	Average Classification accuracy (%)				
Algorithm	**Ecoli**	**Glass**	**Heart**	**Pima**	**Iris**
SMDNS	78.3	72.9	80	77.5	96
Proposed	83.46	79.72	86.12	84.37	98.53

The Table shown below is the investigation and association of Scalability of the Execution time on the basis of number of attributes available in the datasets.

Table 5. Comparison of Execution Time

	Execution Time (S)	
No. of Attributes	SMDNS	Proposed
10	19	16
15	20	17
20	22	19
25	26	22
30	40	27
35	60	40
40	75	55
45	80	65

The Figure shown below is the investigation and judgment of Scalability of the Execution time on the basis of number of attributes available in the datasets.

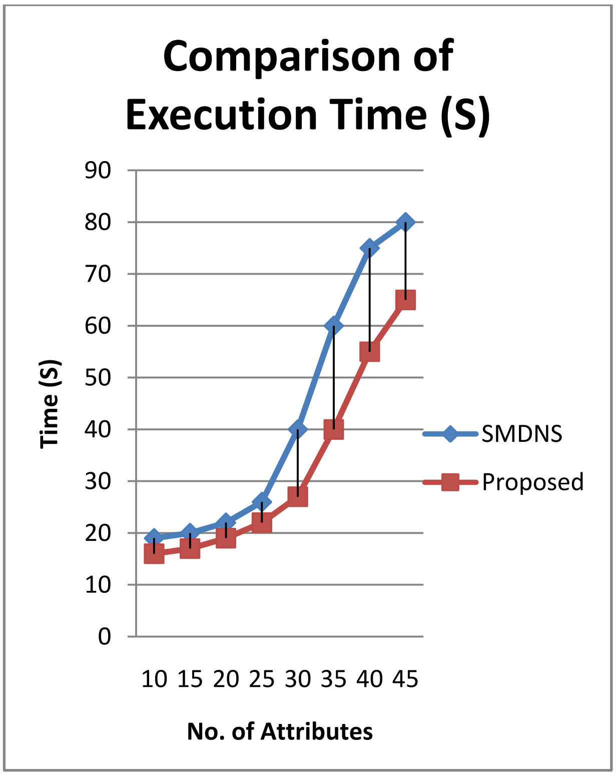

Fig 2:Comparison of Execution Time

5. CONCLUSION

The Planned Procedure implemented here for the Discretization of values on different datasets using Mixture Combinatorial technique of K-means bunching and SVM based Classifier provides efficient results in comparison with the existing methodology implemented on Fuzzy Theory. The Analysis and Comparison done on various Datasets proves that the planned procedure implemented provides low mean Discretization time and high standard cataloging Accuracy as compared to the existing Algorithm.

6. REFERENCES

[1] Z. Pawlak. Rough sets. International Journal of Computer and Information Sciences, 11(5):341–356, 1982.

[2] R.Jensen and Q.Shen. A Rough Set – Aided system for Sorting WWW Book-marks. In N.Zhong et al.(Eds.), Web Intelligence: Research and Development, 95-105, 2001.

[3] F. Min, L.J. Xie, Q.H. Liu, H.B. Cai, A divide-and-conquer discretization algorithm, in: Proc. of the 2nd Int. Conf. on Fuzzy Systems and Knowledge Discovery(FSKD 2005), LNAI, vol. 3613, Springer, 2005, p. 1277C1286.

[4] G.K. Singh, S. Minz, Discretization using clustering and rough set theory, in: Proc. of the 2007 Int. Conf. on Computing: Theory and Applications (ICCTA 2007), 2007, pp.330–336.

[5] P. Blajdo, J.W. Grzymala-Busse, Z.S. Hippe, M. Knap, T. Mroczek, L. Piatek, A comparison of six approaches to discretization — a rough set perspective, in: Proc. of the 3rd Int. Conf. on Rough Sets and Knowledge Technology (RSKT 2008), LNAI, vol. 5009, Springer, 2008, p. 31C38.

[6] D. Tian, X.J. Zeng, J. Keane, Core-generating discretization for rough set feature selection, in: Transactions on Rough Sets XIII, LNCS, vol. 6499, Springer, 2011, p. 135C158.

[7] H S Nguyen. Approximate Boolean Reasoning: Foundations and Applications in Data Mining, Transactions on Rough Sets, Lecture Notes in Computer Science, Springer, pages 334–506, 2006.

[8] Z Pawlak, A Skowron. Rough Sets and Boolean Reasoning, Information Sciences 177:41–73, 2007.

[9] Feng Jiang, Yuefei Sui, “A novel approach for discretization of continuous attributes in rough set theory”, 0950-7051, Elsevier, 2014.

[10] Grzymala-Busse, J.W. Discretization based on entropy and multiple scanning. Entropy 2013, 15, 1486–1502.

[11] Liu, H.; Hussain, F.; Tan, C.L.; Dash, M. Discretization: An enabling technique. Data Min. Knowl. Discov. 2002, 6, 393–423.

[12] Nandita Sengupta, Jaydeep Sen, Jaya Sil, Moumita Saha, “Designing of on line intrusion detection system using rough set theory and Q-learning algorithm” 0925-2312, Elsevier, 2013.