ISSN:0975-9646

Feature Selection Algorithm with Discretization and PSO Search Methods for Continuous Attributes

Madhu.G[1], Rajinikanth.T.V[2], Govardhan.A[3]

[1]*Dept of Information Technology, VNRVJIET, Hyderabad-90, INDIA,*

[2]*Department of Information Technology, GRIET, Hyderabad-85, India.*

[3]*School of Information Technology, J.N.T.University, Hyderabad-85, India.*

Abstract— **Discretization plays a significant role during the transformation of continuous attributes into discrete values in the pre-processing step in data mining, significant attention in the machine learning community. Recently, many researchers have provided numerous discretization methods applied to classification problems. Moreover, these are potential to reduce the dimensionality of the data during the classification accuracy, due to this reason which may bias in classifier accuracies and results. In this paper, we propose the new discretization algorithm based on a popular statistical technique called a z-score and particle swarm optimization technique for feature selection to identify the highly influenced features using wrapper-based feature subset selection. Machine learning based algorithm (C4.5 decision tree) to generate highly accurate decision rules. Empirically, we conduct an experimental study on benchmark continuous data sets with different type of machine learning classifiers. The result shows high performance computed in terms of accuracy and inconsistency has been tested with a Wilcoxon signed-rank test.**

Keywords— **Classification, Continuous attributes, Discretization, Feature selection, PSO.**

I. INTRODUCTION

Discretization of continuous attributes is an important technique for the pre-processing task in the classification problems with simplification analysis, and it has played significant role in the machine learning algorithms [1][2][3][4][5]. However, learning process from continuous attributes to discrete values is often very less efficient and more classifier confusion [6][7]. Discretization first discussed about to qualitative data in classification learning algorithms [2] [8] [9] and the process can be performed either before learning or during the learning, is called as pre-processing. Recently, discretization techniques [2][10][11][12] has been presented into different dimensions, such as a) Supervised versus Unsupervised, b) Static versus Dynamic, c) Global versus Local, d) Parametric versus Non-parametric, e) Top-Down versus Bottom-up, f) Disjoint versus Non-disjoint g) Fuzzy versus Non-fuzzy, h) Ordinal versus Nominal, i) Eager versus Lazy. In this paper we propose a simplicity based discretization schema for dealing with consistency in continuous dataset, this is measured number of cut points without user specifications. Continuous attributes as a complex problem of searching a global minimal set of cut points which have shown that as NP-hard problems [4][13]. In this study, we propose a novel discretization method for continuous attributes using a standard deviation normalization technique (z-score) and particle swarm optimization (PSO) technique for feature selection for the discovery of high influenced features using wrapper-based feature subset selection. Machine learning based algorithm (C4.5 decision tree) to generate highly accurate decision rules. The classification results significantly achieved after discretization with PSO are much better than the classification results with simple PSO in terms of accuracy as well as a performance before discretization.

Kennedy and Eberhart [14] proposed a popular stochastic optimization technique, called Particle Swarm Optimization (PSO) motivated by social behavior of bird flocking or fish schooling. In the past decades PSO have attracted multidisciplinary researchers to solve the combinatorial optimization problems to continuous optimization problems, single and multi-objective problems, etc. In this study, we propose a new discretization method that is applied for continuous attributes to convert the discrete values after applied feature subset selection based on PSO technique which have been adapted to identify the influenced features in given dataset and to reduce the classifier confusion and improve the accuracy.

The present paper is organized as follows: Related work of the proposed discretization presented in Section II, while in Section III, our proposed discretization technique with wrapper based feature subset selection using particle swarm optimization technique step by step approach with C4.5 classifier on benchmark continuous datasets. In Section IV, experimental results and the description of the data sets are presented. Conclusions and discussion are deferred to Section V. Acknowledgements are in the VI.

II. RELATED WORKS

In this section, we demonstrate our discretization method for the purpose of pre-processing step in the data mining process. A complete description of discretization technique is as follows:

A. Discretization Strategy

Discretization is a method for quantifying the numerical attributes into nominal or categorical attributes. In this context the term continuous is used for integer or real (numeric) [9]. Let $S = \{x_1, x_2, \ldots \ldots x_n\}$ be the set of real-valued features or continuous values, real-world data sets are generally distinct. Now we consider a discretization

process for all the continuous values in the dataset need to be a standardized statistical technique z-score (given below).

$$z = \frac{S-\mu}{\sigma} \quad \ldots\ldots (1)$$

where S is an original score obtained from a sample (a population), σ is the standard deviation of the population and μ is the mean of the population. We assume that the minimum value of z-score is 'a' and maximum value of z-score is 'b' form the dataset S.

$$a \leq x_i < b \quad for\ i = 1,2,3,\ldots..n \ldots\ldots\ldots (2)$$

To define the interval of all possible values for random variables X.

$$X = [a,b) = \{x/\ a \leq x_i < b\} \quad \ldots\ldots\ldots (3)$$

after that partition the interval $X = [a,b)$ into a k-equal-width bins as follows:

$$[a,b) = \bigcup_{i=1}^{k-1} B_i = B_0 \cup B_1 \cup B_2 \ldots\ldots\ldots\ldots \cup B_{k-1} \ldots (4)$$

Let us define Sturges' formula [10], which derived from a binomial distribution and implicitly assumes an approximately normal distribution.

$$k = \lceil \log_2 N + 1 \rceil \ldots\ldots\ldots\ldots (5)$$

where 'N' number of row size in the dataset.

Now define a width of the interval is $\delta = \frac{b-a}{k}$ … (6)

where b=maximum value (z), a=minimum value (z) and k= number of bins (equal width). Therefore, the bins are given below:

$$B_0 = [a, a+\delta),\ B_1 = [a+\delta, a+2\delta) \ldots$$
$$\ldots\ldots B_{k-1} = [a+(k-1)\delta, a+k\delta) \ldots\ldots\ldots\ldots (7)$$

Algorithm –I: Discretization Algorithm

Input: Dataset 'S' consisting of the number of rows and column observations, with continuous attributes ' Ai' and class attribute C in the set 'S'.

Output: all attributes are Discretized format in dataset S.

1. Collect all the records for each of 'A_i' in the data set S, not those in the decision column attributes (i.e. $A_i \subseteq S$).
2. let s=S;
3. for each A_i // Select features A_i of training dataset s.
   // D_i=data_for_discretization.
4. D_i = individual columns of the dataset
5. Z=(D_i -mean(D_i))/std(D_i); // apply z-score on dataset s.
6. B_{i0}= min (Z);
7. B_{in}= max (Z);
8. B_i =[B_{i0} , B_{in}];
9. if ($s \neq \emptyset$) then
10. k=ceil(log2(rows)+1);
11. width_Interval=(B_{i0}-B_{in})/k;
12. No_of Bins = k;
13. For rowdis=1 to rows // row discretization
14. for i=1 to bin
15. if(Z(rowdis)>= B_{i0}+((i-1)*width_Interval) && Z(rowdis)<= B_{i0}+i*width_Interval) then
16. update D_i with maximum number of bins.
17. s= reduction_dataset (s, D_i);
18. D= union of { D, D_i }
19. Stop.

III. DISCRETIZATION AND PSO ALGORITHM

All discrete values of the particle swarm optimization technique is used as a 'particle' in the search space, which adjusts the position in the search space based on its own flying experience and flying with another particle. PSO is a heuristic technique inspired by the sequence of steps of a bird flock [14]. This heuristic approach is used to improve the optimization efficiency. In this study we used PSO after discretization to particle search strategy, which improve the machine learning classifier accuracy and performance of the algorithm.

The PSO technique consists of 'N' particles in swarm and i^{th} particle in a swarm. In order, each particle p_i can be viewed as a point in k dimensional space $p_i = (p_{i1}, p_{i2}, \ldots, p_{ik})$ for $i = 1,2,\ldots..N$. To measure the index of the particle, which has the best fitness is denoted as '*gbest*'. The fitness function of the best positions of the particles is given by $f_i = (f_{i1}, f_{i2}, \ldots, f_{ik})$ for $i = 1,2,\ldots..N$. The velocity of the particle moving in the k-dimensional search space is $V_i = (v_{i1}, v_{i2}, \ldots, v_{ik})$ for $i = 1,2,\ldots..N$ [15][16]. PSO technique combines the local search method, called self experience and global search methods, called neighboring experience. The velocity and particles are updated with the following equations.

$$v_{il} = \omega * v_{il} + \lambda_1 * rand1() * (f_{il} - x_{il}) + \lambda_1 * rand2() * (f_{gl} - x_{il}) \quad \ldots. (8)$$

$$x_{il} = x_{il} + v_{il} \quad \ldots\ldots\ldots\ldots (9)$$

where $l = 1,2,3,..K$, and ω is the inertia weight and it is a +ve linear function for time changing under the iteration process. The λ_1, λ_2 are be the acceleration constants that pull the particles towards 'pbest' and 'gbest' has shown in (8). The random numbers rand1 and rand2 generate the functions.

In this study, we compute the discrete values of continuous attributes using a proposed discretization algorithm discussed in section. 2, after discretization dataset splits into training and testing folds based on k-fold cross validation procedure. For each training dataset, initialize the swarm size and then select the features based on PSO for identifying feature subsets with C4.5 classifier. Then we test the test accuracy obtained influenced features the previous step with C4.5 classifier have shown in Algorithm-II.

Algorithm-II: Discretization based PSO feature selection

Input: Dataset '$S = (n,m)$' consisting of number of rows and column observations, with continuous attributes.

Output: Discrete values dataset of the feature subsets and accuracy of the dataset S.

Step 1: Collect all the records with continuous values in the data attribute set S, not those in the decision attributes column (i.e. $\subseteq S$).

Step 2: Apply the new discretization method on the dataset S using the procedure discussed in section II.A

Step 3: Divide the discredited dataset S into training (T_1) and testing (T_2) sets using a stratified a k- fold cross validation test.

Step 4: For each k compute the following

(i) Initialize the number of particles in PSO as $N = 5$, the fitness value as $V_0 = 0$ for $i = 0$.

a) Apply the wrapper method on training dataset with PSO search for identifying influenced features using number of particles as N with C4.5 classifier for its evolution

b) Select the influenced features obtained in Step 4(a) to generate new training $(F_1)$and testing (F_2) datasets.

(ii) Build the C4.5 classifier using the records obtained from F_1 and obtain the test accuracy using testing F_2. Denote this accuracy E_1.

(iii) Compute the feature values of datasets F_1 and T_1respecetively. Denoted these values are N_1 & N_2

(iv) Compute fitness values as $V_i = 0.8 * A + 0.2 * (N_1/N_2)$.

Step 5: Repeat the Steps (4)-(i) to Step (4)-(iv) for each fold.

Step 6: Compute the accuracy of the dataset S.

Step 8: RETURN S

Step 9: STOP

The algorithm terminates for the fold if the feature score falls below the two previous scores and population giving maximum accuracy. We train with C4.5 classifier using features extracted in the previous steps and compute the test accuracies based on test dataset. This process continues for all the ten folds in the data set and we get the average accuracy, the process has been presented in Algorithm-II.

IV. EXPERIMENTS AND RESULTS

The experimental study was conducted with proposed algorithm is implemented in MATLAB R2010a© with a personal computer having an Intel (R) core (TM) 2 Duo, CPU E8400 @ 3.00GHz processor with 4GB RAM and Windows XP operating system. In order to ensure that the accuracy of the predictive model, we used C4.5, algorithms in Weka© [17], KEEL(Knowledge Extraction based on Evolutionary Learning) software tool Open Source - V2012-02-16 [18] are considered for evaluating the performance with our algorithm. The evaluation and the performance proposed algorithms used the number of decision rules, classifier accuracy and identifying influenced features in the data sets.

A. Datasets

In this section we represent the characteristics of the benchmark data sets, which from the KEEL repository [18] shown in Table.1. The selected datasets range between 4 to 34 features. These nine datasets are continuous as well as mixed type and each dataset is followed by continuous and mixed type attributes respectively.

In this experiment, we use a stratified 10-fold cross validation (k=10) statistical measure. In k-fold procedure the data splits into 10 fold datasets and perform k tests, each training on k-1 sets and validating on continuous data set (here k=10).

Table 1. Summary of benchmark datasets used in experiments

Name	#Attributes (R/I/N)	# of continuous Attributes	# Instances	# of Classes
Bupa	6 (1/5/0)	6	345	2
Cleveland	3(13/0/0)	13	303	5
German	20 (0/7/13)	20	1000	2
Glass	9(9/0/0)	9	214	7
Heart	13 (1/12/0)	13	270	2
Ionosphere	33 (32/1/0)	34	351	2
Iris	4(4/0/0)	4	150	3
Wine	13 (13/0/0)	13	178	3

B. Results

The popular algorithms like, C4.5, Logistic Regression (LOR) and Naïve Bayes classifiers are implemented in Weka© [17] with our proposed algorithm. To compare the performance of our algorithms used a non-parametric statistical test, called Wilcoxon [19]. The performance of our algorithm and state-of-the-art classifiers has shown in Table.2. From this table we conclude that our algorithm is better than other classifiers. Also, we performed a Wilcoxon test for matching pairs between the accuracies of our algorithm and other state-of-the-art classifiers clearly we have presented Wilcoxon test measures in Table.3.

Table.2. Test classifiers of with other classification algorithms.

Datasets	**Algorithm-I**	**C4.5**	**LOR**	**Naïve Bayes**
Bupa	62.02	57.97	60.86	60.28
Cleveland	57.09	51.12	44.55	56.76
German	72.50	70.00	68.20	72.50
Glass	67.75	37.38	43.92	55.14
Heart	79.25	79.25	69.60	79.25
Ionosphere	91.16	51.15	44.55	56.76
Iris	98.00	96.00	96.00	96.00
Wine	89.88	39.88	75.28	80.89

Table.3. Results obtained by the Wilcoxon test for algorithm-I and state-of-the-art classifiers.

Vs	R+	R-	Test Statistics	P-value	Asymptotic P-value
C4.5	28.0	0.0	0.0	0.015	0.014
LOR	36.0	0.0	0.0	0.007	0.009
Naïve Bayes	34.5	1.5	1.5	0.019	0.017

From the Table.3. We conclude that the comparison with C4.5 has shown in a positive rank sum of 28.00 with a $p < 0.05$ at $\alpha = 0.05$ significant level. That will show the superiority in the form of performance of our algorithm over C4.5 classifier. Similarly LOR has resulted in a positive rank sum of 36 with a $p < 0.05$ at $\alpha = 0.05$ significant level shows the improved performance of our algorithm over LOR. Naïve Bayes the $p > 0.05$ the

positive rank sum of 34.5, it is so far greater than the negative rank sum of 1.5. The test classifier accuracy in percentage has shown in Fig.1.

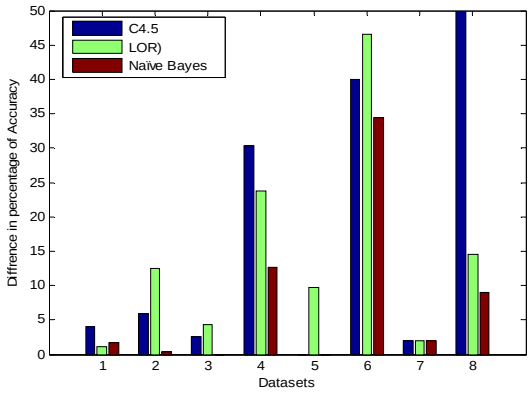

Fig.1. Test classifiers for Algorithm-I vs. other classifiers

To compare the performance of Algorithm-II with and without PSO feature selection with a wrapper based approach on benchmark datasets presented in Table.4.

Table.4. Comparison of Our Algorithm-II with and without PSO feature selection

Datasets	Without Feature Selection (PSO) #Original Features	Accuracy (%)	With Feature Selection (PSO) #Influenced Features	Accuracy (%)
Bupa	6	60.98	2	65.21
Cleveland	13	57.09	3	59.40
German	20	72.30	4	72.50
Glass	9	63.55	6	67.75
Heart	13	79.25	3	83.33
Ionosphere	34	91.16	7	94.58
Iris	4	97.33	2	98.00
Wine	14	89.88	9	92.13

From the Table.4 we clearly demonstrated that our Algorithm-II has performed better performance than with traditional algorithms. The PSO feature subset selection to identify the swarm size and C4.5 classifier as a wrapper based feature selection. Clearly we presented that our proposed PSO feature selection algorithm is to identify the highly influenced futures in the given datasets. The percentage difference in the accuracies with and without PSO feature section of our algorithm is shown in Fig.2.

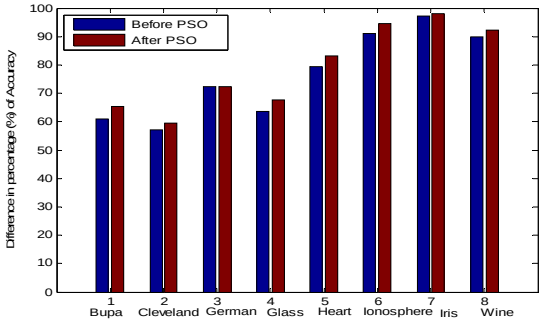

Fig.2. Test classifiers with and without PSO feature selection

V. CONCLUSIONS

In this paper, we have discussed a new discretization algorithm for continuous attributes. The proposed discretization algorithm has outperformed the state-of –the-art classifiers in mixed type attribute datasets. This algorithm significantly performs on continuous attributes for machine learning algorithms. We proposed a feature selection algorithm based on particle swarm optimization technique with our discretization algorithm. The feature selection method is applied with wrapper based approach. We have used evolutionary algorithm such as PSO method after discretization for continuous attributes. The comparison of the results stated that the superiority of our algorithms with other popular classification algorithms with feature selection (PSO) and without feature selection. Finally, we conclude that our new discretization based particle swarm optimization technique is superior to other feature selection algorithms.

ACKNOWLEDGMENT

The authors would like to thank Prof. V. Sree Hari Rao & Dr.M. Naresh Kumar for their valuable suggestions

REFERENCES

[1] Chmielewski, M.R. & Grzymala-Busse, J.W., Global discretization of attributes aspreprocessing for machine learning, 3rd International Workshop on Rough Set and Soft Computing, pp. 294-301, 1994.

[2] Dougherty J.; Kohavi R. & Sahami M., Supervised and unsupervised discretization ofcontinuous features. Proceedings of 12th International Conference on Machine Learning, pp. 194-202, 1995.

[3] Nguyen, S.H. & Skowron, A., Quantization of real value attributes: rough set and boolean reasoning approach, Second Joint Annual Conference on Information Sciences, pp. 34-37, 1995.

[4] Nguyen, H.S., Discretization problem for rough sets methods, First International Conference on Rough Sets and Current Trends in Computing, Springer-Verlag, pp. 545-552, 1998.

[5] Liu, H.; Hussain, F.; Tan, C.L. & Dash, M., Discretization: an enabling technique. Data Mining and Knowledge Discovery, vol. 6 (4), pp.393-423, 2002.

[6] An. A, Cercone.N, Discretization of Continuous Attributes for Learning Classification Rules, 3rd Pacific-Asia Conference, Methodologies for Knowledge Discovery and Data Mining, 1999, pp.509-514.

[7] Ying Yang, Geoffrey I. Webb, and Xindong Wu, Discretization Methods, Data Mining and Knowledge Discovery Handbook, Second Edition, O. Maimon, L. Rokach, Eds, 2010, pp. 101-116.

[8] Kerber R, Chimerge:Discretization for numeric attributes, National Conference on Artificial Intelligence, pp.123-128, 1992.

[9] Kohavi R, Sahami M, "Error-based and entropy-based discretization of continuous features", In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996, pp.114-119.

[10] X. Liu and H. Wang., A discretization algorithm based on a heterogeneity criterion, IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 9, pp. 1166–1173, 2005.

[11] C. T. Su and J. H. Hsu., An extended Chi2 algorithm for discretization of real value attributes, IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 3, pp. 437–441, 2005.

[12] F. E. H. Tay and L. Shen., A modified Chi2 algorithm for discretization, IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 3, pp. 666–670, 2002.

[13] Amitava Roy, Sankar K.Pal., Fuzzy Discretization of Feature Space for a Rough Set Classifier, Pattern Recognition Letters, vol.24, 2003, pp.895-902.

[14] Kennedy, J. & Eberhart, R.C., Particle Swarm Optimization, Proceedings of IEEE International Conference on Neural Networks, IV: pp. 1942-1948, 1995.

[15] X. Wang, J. Yang, X. Teng, W. Xia,, R. Jensen., Feature Selection based on Rough Sets and Particle Swarm Optimization. Computer Methods and Programs in Biomedicine. 83(2), pp.459-471, 2007.

[16] X. Wang, J. Yang, R. Jensen, X. Liu., Rough Set Feature Selection and Rule Induction for Prediction of Malignancy Degree in Brain Glioma, 2006.

[17] I. Witten, E. Frank., Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, San Francisco, 2005.

[18] J.Alcala-Fdez et al., KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework, Journal of Multiple-Valued Logic and Soft Computing, vol.17, no.2-3, pp.255-287,2011.[Online]. Available: http:// http://www.keel.es/

[19] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin, vol. 1(6), pp.80-83,1945.