C-LAS Relief-An Improved Feature Selection Technique in Data Mining

S.S.Baskar
Research Scholar
Dept. of Computer Science
St. Joseph's College (Autonomous)
Trichy, TN, India

L Arockiam
Associate professor
Dept. of Computer Science
St. Joseph's College (Autonomous)
Trichy, TN, India

ABSTRACT
Feature selection or Feature subset selection is a process of reducing the attribute space in the feature set. It is also stated that feature selection is a technique of identifying a subset of features. These subsets of features are selected by removing irrelevant or redundant features in the feature set. A good feature set is said to be that it contains highly correlated features with the class. Such feature set improves the efficiency of the classification algorithms and also the classification accuracy. The Chebyshev distance with median variance in the weight estimation of attributes in the Relief imparts the consistency and good accuracy. In this paper a novel algorithm called C LAS-Relief is used to improve the reliability and accuracy of classification. Here C LAS-Relief stands for Chebyshev distance LAS-Relief. The efficiency and effectiveness of proposed method is experimented using agriculture soil data sets, Soybean and Ozone data sets. Similarly the new approach is compared with LAS-Relief approach using Naive bayes and J48 classifiers. The classification accuracy of C-LAS-Relief is superior over LAS-Relief. C LAS-Relief algorithm increases the accuracy of classification compared to LAS-Relief algorithm.

Keywords: Relief, Chebyshev distance, Naive Bayes, J48, Data Mining

1. INTRODUCTION
Feature selection is one of the important tasks in data mining and machine learning. The feature selection technique diminishes the features from large sets. In real world scenario, it is not necessary to use all the features in datasets to derive the target class. The features with high relevancy are to be selected using feature selection techniques. In some cases, there are redundant and irrelevant features in the data sets. The main objective of the feature selection is to remove the redundant and irrelevant features from the data sets. This can be achieved by using appropriate feature selection techniques. Feature selection process reduces the dimensionality of the data sets. Thus the reduced feature set helps to allow the learning algorithms to operate faster and effectively. The redundant and irrelevant feature in the feature set diminishes the quality of the classification. At the same time, the feature selection technique should bring the minimum subset of features which is able to model the target most appropriately. Finding the minimum feature with more relevancies and less redundancy in the feature space would reduce system complexity and it will reduce the system processing time. This in turn saves the computation resources as well as processing time. In general, feature selection or feature reduction approaches are widely used in image processing, data mining and machine learning as well as artificial intelligence. Feature Selection plays a critical role in many domains for minimising the cost and computation time for finding the target concept. There is a serious challenge with limited training samples for selecting useful features by existing feature selection algorithms. In this paper, the novel feature selection algorithm C LAS-Relief has been proposed. This novel algorithm incorporates the Chebyshev distance for finding the near Hit [H] and near Miss [M]. This novel algorithm is named as C LAS-Relief. As the C LAS-Relief algorithm uses the Chebyshev distance; the features are correctly assessed based on the relevancy. The use of this algorithm in the feature selection selects appropriate feature in the feature sets. The redundant and irrelevant features are ignored by use of C LAS-Relief algorithm.

2. RELATED WORK ON FEATURE SELECTION
In real world data, the representation of data often uses too many features. But only a few features may be used to relate the target concept. In the bulk set of data collection, there is possibility of relevant and irrelevant features in the feature sets. The irrelevant or redundant features in the feature sets do not have a specific role on target class. The novel feature selection algorithm ignores the redundant and irrelevant features on feature selection process. The selection of potential and appropriate features from the feature space reduces the dimensionality of the feature space and allows the learning algorithms to work faster and efficiently.

Sun Yi Jun [5] reported feature selection by Principle component analysis and compression (information theory). This played major role for feature selection by way of eliminating the features with less information for prediction. Liu and his co-workers [4] adopted the feature selection technique for various domain areas to improve the model.

J. Hua et.al [1] worked on comparison study of few feature selection method in the area of bioinformatics. They used Information Gain, Gini Index, T-Test.

2.1 Related work on Relief algorithm
Relief algorithm was first proposed in [2]. After that a lot of variants came into usage for feature selection. Every variant of Relief algorithm has its own merits and demerits depending on the nature of data sets. The key idea of Relief is to iteratively estimate feature weights according to their ability to discriminate between neighbouring models.

Kononenko [3] described enhancements of Relief algorithm that enable to cope with multi-class, noisy and incomplete domains. This Relief algorithm was named as Relief F. Iterative-Relief is put forward in [5]. Adaptive Relief is termed as A-Relief. This A-Relief algorithm offers effective feature subset for the further identification [6].

3. BASICS BEHIND RELIEF ALGORITHMS

Relief algorithm is a very simple, fast, and effective approach to attribute weighting. The output of the Relief algorithm is a weight between −1 and 1 for each attribute, with more positive weights indicating more predictive attributes. It has many variants depending on the nature of data and attributes characteristics. The Relief Algorithm works as the following principles. The weight of an attribute is updated iteratively as below procedure. A sample is selected from the data, and the nearest neighbouring sample that belongs to the same class (nearest hit) and the nearest neighbouring sample that belongs to the opposite class (nearest miss) are identified. Nearest Hit and Nearest Miss are main portions of this algorithm. The nearest Hit and nearest Miss is calculated based on the Manhattan distance between two points. The change in feature weights is considered for feature selection in the classification of target class. Such features are given more weight for classification process. Thus weight of feature plays vital role for finding accurate class.

On the other hand, a change in feature weight value does not provide any change in the class is considered as down weighting of the feature. This procedure of updating the weight of the attribute is performed for a random set of samples in the data or for every sample in the data. The weight updates are then averaged so that the final weight is in the range [0, 1]. The feature weight estimated by Relief has a probabilistic interpretation. It is proportional to the difference between two conditional probabilities, namely, the probability of the attribute's value being different conditioned on the given nearest Hit and nearest Miss respectively.

4. MANHATTAN DISTANCE IN LAS-RELIEF

The function diff() in Relief algorithm as well as LAS-Relief is used for calculating the distance between instances to find the nearest neighbours. The total distance is simply the sum of distances over all features. In LAS-Relief algorithm, Manhattan distance is used to weigh the distance between every two instances. The Manhattan distance between any two instances Ri, Rj is calculated by the Eq.1.

Dis (Manhattan) = | Ri-Rj| --------> [Eq.1]

4.1 *C LAS-Relief* Algorithm Basics

The key idea of the C LAS-Relief algorithm is to estimate the quality of attributes according to how well their values distinguish between instances that are near to each other. First C LAS-Relief selects the instances randomly. The random selected instance is Ri.

The C LAS-Relief searches for its two nearest neighbours: one from the same class, called nearest hit (H), and the other from the different class, called nearest miss (M). Then the quality estimation of weights for all features depending on the instances values for Ri, M, and H. If instances Ri and H have different values of the attribute A, then the attribute A separates two instances with the same class which is not desirable to decrease the quality estimation of feature weight W [A]. On the other hand if instances Ri and M have different values of the attribute A then the attribute A separates two instances with different class values which is desirable to increase the quality estimation W [A]. The entire process is repeated for m times, where m is a user-defined parameter.

C LAS-Relief algorithm uses the Chebyshev distance. The Chebyshev distance in the C LAS-Relief algorithm is used for calculating the distance between attributes in two instances. As Chebyshev distance takes the maximum value between points, the two neighbourhood values H and M are determined by using Chbyshev distance. Then the diff () function correctly estimates the features quality of attributes. This improves the classification accuracy.

C LAS-Relief algorithm uses the Chebyshev distance for calculating the near Hit value and near Miss value of the instances. Since Chebyshev distances take the maximum of absolute value of two points, the relevant features in the feature space is correctly selected. The Chebyshev distance in C LAS-Relief algorithm is better than LAS-Relief algorithm. The Eq.1 shows the Chebyshev distance formula. The original Relief uses the Manhattan distance in the diff () function

Dis (Chebyshev) =Max (|X1-X2|,|Y1-Y2|)

Proposed algorithm of *C LAS- Relief*

1. *Set all weights W[A] := 0:0;*
2. *for i := 1 to m do begin*
3. *Randomly select an instance Ri;*
4. *Find nearest hit H and nearest miss M;*
5. *for A: = 1 to a do*
6. *W[A] := W[A] - diff(A, Ri, H)* ^{2}*/m + diff(A, Ri, M)*^{2}*/m;*
7. *end;*

**********Weight Updation*******

8. $W[F_i]$ = *median Value W[A]*
9. *If* $W[F_i] \geq \delta$ *put it into the selected feature subset* T;

The proposed algorithm uses the Chebyshev distance instead of Manhattan distance. The median variance with Chebyshev distance in the C LAS-Relief algorithm enhances the accuracy of attribute weight estimation.

5. EXPERIMENTS AND RESULTS

Agriculture soil, Ozone and Soybean data sets are used in the experiments. The number of features and number of instances of three data sets are tabulated in the Table.1. This experiment selects the appropriate features based on the high relevancy for classification.

Table-1 Data set Description

No.	Name of the data set	No of features	No of Instances
1	Agriculture soil data sets	13	200
2	Ozone	72	2534
3	Soybean	35	683

The efficiency of the proposed C LAS-Relief algorithm is evaluated on the basis of number of selected features and also by accuracy classification. This C LAS-Relief feature selection algorithm is compared with LAS-Relief.

5.1 Feature Selection

The feature selection efficiency between C LAS-Relief and LAS-Relief are studied and their results are shown in the Table-2. C LAS-Relief has resulted comparatively superior over than the LAS-Relief in feature weight estimation. In C LAS-Relief, the numbers of selected features above the threshold value are comparatively lesser than the LAS-Relief. From the Table -2, it has been clearly noted that C LAS-Relief estimates the relevant features along with redundant and irrelevant feature sets.

Table-2 Comparison of *LAS-Relief* and *C LAS-Relief* methods on Agriculture data set

Feature selection Method	Results										
	Features	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**
LAS Relief Algorithms	Selected Times	4	17	3	2	19	19	18	4	17	3
	Selected Probability (%)	23	89	*19*	*7*	*87*	*95*	*90*	*34*	*80*	*83*
	Selected Features	F2,F3,F5,F6,F7,F9									
C LAS - Relief Algorithms	Selected Times	19	18	8	3	17	16	4	18	15	17
	Selected Probability (%)	*95*	*91*	*4*	*15*	*85*	*80*	*20*	*90*	*80*	*84*
	Selected Features	F1,F2,F5,F6,F8,F9,F10									

The C LAS-Relief algorithm selects the appropriate relevant features compared to LAS Relief algorithm. Thus C LAS-Relief algorithm increases the accuracy of the classification. From the table-2, it is observed that the result reveals that C LAS- Relief algorithm estimates the quality feature than LAS-Relief. Some time LAS-Relief ignores the features with high relevancy. These missing relevant features are correctly selected in C LAS-Relief. This is because of the Chebyshev distance measure used in C LAS-Relief.

The Table-3 shows the comparison result of two methods with different data sets. The C LAS Relief algorithm selects the less number of features when compared to LAS-Relief. This is due to distance function for calculating the two neighbourhoods. The C LAS-Relief uses the Chebyshev distance.

Table.3 Selected features from data sets

S.No	Method	Data sets name	No of Features	Selected Features
1	C LAS-Relief	Agriculture Soil	13	6
		Soybean	35	24
		Ozone	72	51
2	LAS-Relief	Agriculture Soil	13	8
		Soybean	35	31
		Ozone	72	56

The selected features after feature estimation by from the total number of features in agriculture soil, Soybean and Ozone data sets. C LAS-Relief selects the less no of features than the LAS-Relief algorithm. The proposed method C-LAS-Relief is found to be best in reducing the size of feature sets compared to LAS-Relief feature selection methods. In case of Agriculture data set, C LAS-Relief reduces the number of features to 6. But in LAS-Relief it has been reduced to 8. The C LAS-Relief reduces 72 features of original dataset to 51 in Ozone data sets. This is better reduction than LAS-Relief. In the LAS-Relief method, the features are reduced to 56. The figure.1 shows the reduced data set selection from the original data sets.

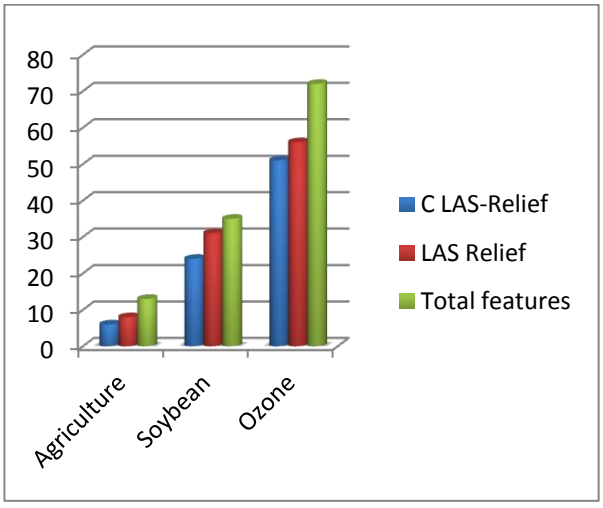

Figure.1 Selected features from total feature sets.

5.2 Classification Accuracy Analysis.

The classification accuracy is studied by using two classifier called Naive Bayes (NB) and J48. The study revealed that the accuracy of C LAS-Relief is better than the LAS-Relief in both Naive Bayes and J48 classifier. Precision, Recall and F Measure have been taken for studying the accuracy measure in Naive Bayes and J48. The performance of C LAS-Relief is better than LAS-Relief method depending on the average of

Recall, Precision and F Measure. The results are shown in the table-4.

The classification accuracy is higher in C LAS-Relief when compared to LAS-Relief algorithm. The improvement in the accuracy in classification is due to selection of appropriate features with high relevant. The C LAS-Relief selects the quality features with higher relevancy than the LAS-Relief. The more accurate features having high relevancy is taken in C LAS-Relief than the LAS-Relief.

In C-LAS-Relief algorithm, all the relevant features are selected without missing it. The possibility of relevant features becoming irrelevant is minimum in C LAS-Relief compared to LAS-Relief.

Table-4 Accuracy analysis of two methods

S.No	Method	Data sets	Precision	Recall	F measure
1	**C LAS-Relief**	Agriculture Soil	0.860	0.859	0.859
		Soybean	0.917	0.869	0.880
		Ozone	0.893	0.871	0.870
		Average	**0.890**	**0.866**	**0.870**
2	**LAS- Relief**	Agriculture Soil	0.855	0.839	0.829
		Soybean	0.910	0.849	0.840
		Ozone	0.843	0.831	0.830
		Average	**0.869**	**0.839**	**0.833**

The precision, Recall and F measure have been taken for evaluation measure for classification accuracy. Based on the results, C LAS-Relief outperforms than LAS-Relief. The comparative results with average value of Precision, Recall and F Measure have been shown in figure.2.

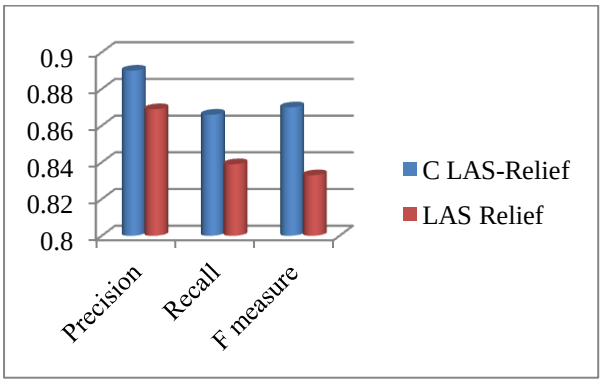

Figure.2 Comparison of C LAS-Relief and LAS Relief methods

This result indicates that C LAS-Relief performs better than LAS-Relief. The average values of precision, Recall and F measure are 0.890, 0.866, and 0.870 for C LAS-Relief algorithm. These values are higher than the LAS-Relief algorithm. The result shows that the new C LAS-Relief algorithm is better than LAS-Relief in accuracy of classification.

6. CONCLUSION

The C LAS-Relief algorithm selects the more relevant features in the feature set. This algorithm estimates the quality features in the feature sets by finding the higher relevancy features. The accuracy of classification is higher than LAS-Relief on Soil, Soybean and Ozone datasets. The classification accuracy of C LAS-Relief is measured by precision, recall and F-measure. Soil, Soybean and Ozone datasets are used to evaluate the C LAS-Relief algorithm. The accuracy of the classification is higher on C LAS-Relief than LAS-Relief. From this experiment, it is concluded that the C LAS-Relief algorithm shows better accuracy than LAS-Relief. The C LAS-Relief algorithm outperforms than the LAS-Relief.

7. REFERENCES

[1] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data", Pattern Recognition, Vol. 42, No. 3, 2009, pp. 409–424.

[2] Kira K, Rendell L A, "Practical approach to feature selection", In ML92: Proceedings of the ninth international workshop on Machine learning. Morgan Kaufmann Publishers Inc, 1992, pp.249-256.

[3] I. Kononenko. Estimating attributes: Analysis and extensions of relief. In Proceedings of the European Conference on Machine Learning, 1994.

[4] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information", Pattern Recognition, Vol. 42, No. 7, 2009, pp. 1330 –1339.

[5] Sun Yi Jun, "Iterative Relief for feature weighting algorithms, theories and applications", IEEE Trans on Pattern Analysis and Machine Intelligence, Vol.29, No.6, 2007, pp 1035-1051.

[6] Fan Wenbing, Wang Quanquan and Zhu Hui, "Feature Selection Method Based on Adaptive Relief Algorithm" 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010) IPCSIT Vol. 53 (2012) © (2012) IACSIT Press, Singapore, 2012