ISSN:2320-0790

An International Journal of Advanced Computer Technology

E LAS-Relief-A Novel Feature Selection Algorithm In Data mining

S.S.Baskar[1], Dr. L Arockiam[2]

[1]Research Scholar, Dept. of Computer Science, St. Joseph's College (Autonomous), Trichirappalli, India

[2]Associate professor, Dept. of Computer Science St. Joseph's College (Autonomous), Trichirappalli, India

Abstract: Feature selection is vital task in any data mining pre processing procedure. This paper deals with the problem of estimating the quality features in the feature sets. The original feature estimating algorithm *LAS-Relief* algorithm can deal with discrete and continues attributes and it is limited to irrelevant feature removal. The new improved feature selection technique called *E LAS-Relief* deals with noisy and incomplete data sets. This paper shows that the novel algorithm *E LAS-Relief* outperforms on agriculture soil data sets for classification.

Keywords: Data mining, Relief algorithm, feature selection, Naive Bayes, J48 classifiers

1. INTRODUCTION

The key issue in the feature selection is finding the highly relevant features in the feature sets. Feature weighting is the process of assigning weights to features. Based on their relevance to a task, the weights are assigned to the features. These weights are considered to reflect the relative importance of the features and they are used for selection and discarding of features below a threshold which is defined by the user.

In current scenario, the feature sets are mixture of noisy, irrelevant and redundant features in the instances. Finding suitable and appropriate feature is more tedious task in the machine learning and data mining. There are many algorithms currently evolved in the context. Among the various feature selection quality estimation algorithm, Relief algorithm is considered to be efficient and simple. This algorithm deals with dependent and independent data sets. It uses the heuristic search techniques for finding the suitable attributes. LAS Relief algorithm gives importance for removing the irrelevant features but fails to deal with incomplete and noisy attributes in the data sets. The original Relief algorithms are a family of attribute weighting algorithms. They help identifying associations between features resulting in classification.

In this paper, the section 2 deals with related work with Relief based feature selection and its variants. The section 3 deals the fundamental concept of Relief, *LAS-Relief* and E *LAS-Relief* algorithm. Section 4 discusses the experiments with different data sets and its results. Section 5 gives the inference of the experiments.

2. REVIEW OF LITERATURE

2.1 Related work on feature estimation

Information Gain is one of the attribute quality estimation method proposed by Hunt et.al in 1966. Then this method is used by many authors [11]. The principle behind this Information gain is to estimate the difference between the prior entropy of class C and posterior entropy of given values V of attributes

$$\text{Gain} = \sum_C P(C) log2\ P(C) - \sum_V (-PV) \times \sum_C P(C|V)\ log2\ P(C|V) \quad (1)$$

Many methods are available for attribute quality estimation. They are Gini Index [2], distance measure [8] and J-measure [12]. Kira and Rendell [6] developed an algorithm called Relief. The Relief algorithm is simple and efficient in estimating quality attributes. The key idea of estimating the quality attributes in Relief is how well each attribute value is distinguished among instances and that is how near to each other. Thus, Relief searches for instances for two nearest neighbours. One is Nearest Hit. It means one from same class and other is nearest Miss. It means other from different class. The Relief's weight estimation W [A] of an attribute A is approximation of following difference in the probabilities.

W[A] = P (Different value of A | nearest instance of different class) – P (Different value of A| nearest instance of same class)
The logic is that the attributes having highly informative should differentiate between instances from different classes and should have the same value for instances from same class.

2.2 Related work on Relief and its variant

Kira and Rendell [6] proposed the Relief algorithm in 1992. It uses the statistical method instead of heuristic search for feature selection. Relief-F [7] was the extension of Relief algorithm. Relief-F has enabled to work with noisy and incomplete datasets and to deal with multi-class problems.
Robnik and Kononenko [10] reported that Relief was extended to handle noisy and missing data and solve multi classification issues which the original Relief algorithm cannot deal with. It statistically selects the relevant features instead of small size.

Heum Park and Hyuk-Chul Kwon [5] reported that new extended Relief algorithm showed better performances for all data sets than the Relief algorithm. Sun.Y [13] proposed an algorithm called Iterative RELIEF (I-RELIEF) algorithm. This algorithm is used to alleviate the deficiencies of RELIEF.

Blessie and Karthikayan [1] proposed the new algorithm as an extension of RELIEF based on Discretization. Discretization partitions features into finite set of adjacent intervals. Instead of using random sampling for selecting the instance, they suggested to take instance from each interval which reduces the computational complexity and maintains the quality of features.

Yuxuan SUN et.al [14] proposed the Mean Variance based Relief algorithm to derive a small subset features from datasets with stability. This algorithm stabilized the feature weight estimation from the variation by way of mean value in the weight.

Matthew E Stokes and Shyam Visweswaran [9] developed a new feature selection algorithm by introducing the spatially weighted variation of Relief. This algorithm is called Sigmoid Weighted ReliefF Star (SWRF*). This algorithm is applied to synthetic SNP data sets. They reported that SURF* performed well on SNP data than Relief-F.

Fan Wenbing et.al [4] proposed a new approach from Relief algorithm This is called as an adaptive Relief (A-Relief) algorithm. This algorithm mitigates the issue of Relief algorithm by dividing the instance set adaptively. In A-Relief algorithm, each feature inspected deeply to detect the bogus feature, before the feature was trained through *A-Relief***.**

3. FUNDAMENTAL OF RELIEF ALGORITHM BASED FEATURE ESTIMATION

The Relief algorithm is a simple and efficient and uses the heuristic search techniques. The Pseudo code of Relief algorithm is given as follows.
Set all weights W[A] =0
for i =1 to m do
begin
randomly select an instance R;
find the nearest hit H and nearest miss M ;
for A= 1 to all_ attributes do
W[A] = W[A – diff(A,R,H / m+ diff(A,R, H) / m;
End;
In the pseudo code, the function *diff ()* calculates the difference between values of attributes [A] for two instances. The difference between two values is either 1 for discrete values or 0 for continuous values. Value 1 means that the attribute values are different between two instances. If the difference is 0 , values are equal. For continuous attributes, the difference is actual difference. The difference value is normalized to the interval of 0 and 1. The normalization with m values assures that weights are in the interval of [-1, 1]
The Function difference *diff()* in the Relief algorithm is used for calculating the distance between instances to find the nearest neighbours. The total distance is sum of distance of overall attributes. The default distance function for measuring the distance in the Relief algorithm is Manhattan distance. This distance function helps to find the nearest neighbour of instances.
The difference function for nominal and numerical attribute value are as follows
Nominal Attribute

$$Diff(A,I_1,I_2) = \begin{cases} 0; \text{ Value } (A, I_1) = \text{Value } (A,I_2) \\ 1; \text{ Otherwise} \end{cases} \quad (2)$$

Numerical Attribute

$$Diff(A,I_1,I_2) = \frac{|value\ (A,I1) - value\ (A,I2)}{\max(A) - \min(A)} \quad (3)$$

3.1 Limitation in the *LAS Relief* algorithm

In the *LAS-Relief* algorithm, noisy values of attributes or features are not taken into account in the feature estimation. The noisy features in the feature set strongly affect the selection of nearest neighbours. This in turn affects the feature weight estimation as well as accuracy of classification. This issue is addressed in the paper by introducing the novel Relief algorithm. This algorithm is called

E *LAS-Relief*. This novel algorithm adopts the squared Euclidean distance function for distance measure instead of Manhattan distance. This algorithm searches neighbour with *k* times rather than one time.

3.2 *E LAS-Relief* Algorithm's novelty

This novel algorithm introduces the new distance function for calculating the distance between the two instances. This new distance function is squared Euclidean distance. This measures the distance between two instances [3]. Squared Euclidean Distance is not a metric. This distance is frequently used in optimization problems in which distances have to be compared.

Formula for squared Euclidean distance is

$$d^2(p, q) = (p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2 \ldots\ldots\ldots + (p_n - q_n)^2 \quad (4)$$

4. EXPERIMENTS AND RESULTS

To estimate the performance of *E LAS-Relief* algorithm, the agriculture soil data sets are taken for analysing accuracy of soil classification. This algorithm is compared with LAS-Relief algorithm for accuracy measure. Here Precision, Recall and F measure are taken for accuracy estimation. E LAS-Relief fetches appropriate features by estimating informative feature. The mechanism of k times calculation of nearest miss and nearest hit handles the noisy and incomplete instances more appropriately.

4.1 Characteristics of data sets

The agriculture soil data set contains 200 instances with 13 attributes. Nearly 3 percentages of instances are found to be incomplete with noise. The target classes of agriculture soil data set are two. E LAS-Relief algorithm is experimented with agriculture data set. Then it is again tested in two artificial data sets. They are Ozone and Soybean data sets. Number of instances and number of features of the data sets are given in the Table-1.

Table-1 Data sets 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 two feature selection methods such as *E LAS-Relief* and *LAS-Relief* are used for the experiments in this paper. Twenty times the classification was performed for m = 20 to m = 39 respectively, for example Wm=20[A]; Wm=21[A]; ....Wm=39[A]. Afterwards W [A] is calculated. Then they are sorted by size, and the top 5 features are chosen as the selected features for later target recognition.

Table-2 Comparison of *E LAS-Relief* and *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	17	4	3	2	19	19	18	4	6	3
	Selected Probability (%)	89	23	*19*	*7*	*87*	*95*	*90*	*34*	*36*	*83*
	Selected Features	F1,F3,F5,F6,F7									
E 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									

From the Table-2, it is inferred that important features are taken from the feature sets. In *E LAS-Relief*, number of highly relevant feature selected is more than the *LAS-Relief* algorithm. The

novel algorithm *E LAS-Relief* selects the highly relevant feature from noisy as well as incomplete datasets.

4.2 Classification Accuracy Analysis

The accuracy of classification is tested in the two classifiers. Here J48 and Naive Bayes classifiers are used for analysing classification accuracy of *LAS-Relief* and *E LAS-Relief* algorithms. The evaluation measures are precision, Recall and F measure.

Table-3 Accuracy analysis of *E LAS-Relief* and *LAS-Relief* methods for three data sets In Naive Bayes

S.No	Method	Data sets	Precision	Recall	F measure
1	E LAS-Relief	Agriculture Soil	0.863	0.862	0.868
		Soybean	0.918	0.871	0.882
		Ozone	0.895	0.873	0.872
		Average	**0.892**	**0.868**	**0.874**
2	LAS- Relief	Agriculture Soil	0.852	0.837	0.825
		Soybean	0.909	0.844	0.832
		Ozone	0.841	0.830	0.829
		Average	**0.867**	**0.837**	**0.828**

The classification accuracy is analysed by using Naive Bayes (NB) classifier. The results show that the accuracy of *E LAS-Relief* is higher than the *LAS-Relief* in all the three data sets on Naive Bayes classifier. The results are shown in the Table-3.The improvement in accuracy is because of the appropriate feature selection in the feature space.

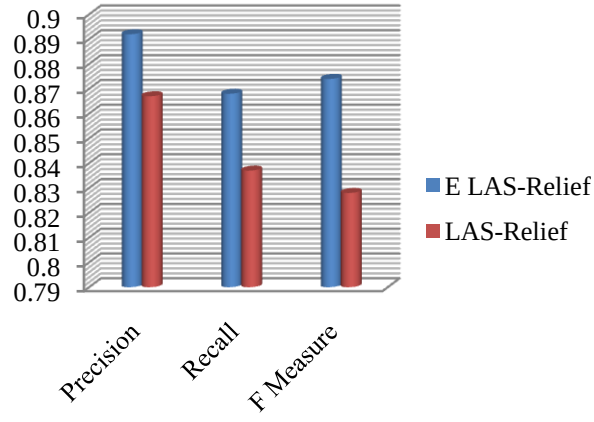

Figure-1 Comparison of *E LAS-Relief* and *LAS- Relief* methods in Naive Bayes classifier

E LAS-Relief algorithm shows higher efficiency in the classification accuracy than *LAS-Relief* algorithm. The Table-3 and Figure-1 depict that *E LAS-Relief* is higher in efficiency than *LAS-Relief*. In figure-1 the average value of Precision, Recall and F Measure are taken from three data sets.

Table-4 Accuracy analysis of *E LAS-Relief* and *LAS-Relief* methods for three data sets in J48

S. No	Method	Data sets	Precision	Recall	F measure
1	E LAS-Relief	Agriculture Soil	0.872	0.868	0.874
		Soybean	0.922	0.875	0.886
		Ozone	0.898	0.878	0.877
		Average	**0.897**	**0.874**	**0.879**
2	LAS- Relief	Agriculture Soil	0.848	0.843	0.838
		Soybean	0.917	0.842	0.822
		Ozone	0.831	0.821	0.819
		Average	**0.865**	**0.835**	**0.826**

Agriculture soil, Soybean and Ozone data sets are analysed using J48 classifier for classification accuracy. The results are depicted in the Table-4. From the Table-4, it is inferred that *E LAS-Relief* algorithm performs well in classification accuracy than *LAS-Relief* algorithm. *E LAS-Relief* algorithm handles the noisy as well as incomplete data sets. This results the higher value of Precison, Recall and F measure in J48 classifier.

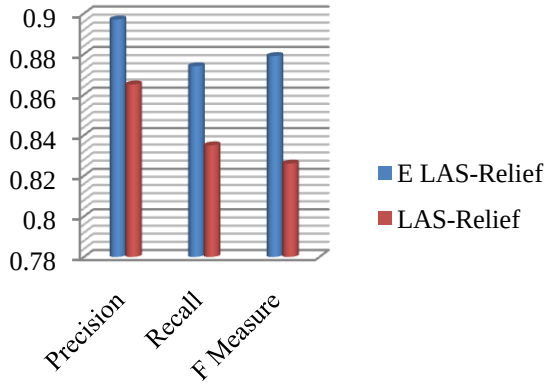

Figure-2 Comparison of *E LAS-Relief* and *LAS Relief* methods in J48 classifier

Figure-2 depicts that *E LAS-Relief* algorithm's classification accuracy which is higher than the LAS-Relief algorithm in the three datasets. Here the average value of Precision, Recall and F Measure are taken for plotting the chart. The results prove that the use of squared Euclidian distance measure in the *E LAS-Relief* algorithm selects the more appropriate features than the *LAS-Relief* algorithm This philosophy behind in *E LAS-Relief* algorithm enhances the classification accuracy over *LAS-Relief* algorithm.

5. CONCLUSION

E LAS-Relief algorithm is simple and efficient. It is an extension version of *LAS-Relief* algorithm. This algorithm is partially inspired by instance based learning techniques. This algorithm relies on squared Euclidean distance for searching k nearest neighbours in the same class and opposite class. *E LAS-Relief* employs the heuristic approach for appropriate feature estimation in the larger feature space. The efficiency of algorithm is measured in terms of accuracy of classification. The accuracy of classification is measured in terms of Precision, Recall and F measures. By employing the heuristic search approach, the noisy and incomplete portion of features in the feature space is managed efficiently in agriculture soil data sets and other two artificial data sets. From the experiment, it has been inferred that E *LAS-Relief* algorithm outperforms *LAS-Relief* algorithm. The experiment with *E LAS-Relief* algorithm is confined to agriculture soil data sets with two class classification. In future, E LAS-Relief algorithm may be extended to multi class concept as well as experimentation with domain area feature space. In *E LAS-Relief* algorithm, the number of nearest neighbours used is a user specified parameter that is constant from iteration to iteration. This plays a significant role in enhancement of this algorithm.

REFERENCES

[1] Blessie E.C and Karthikeyan E, "RELIEF-DISC: An Extended RELIEF Algorithm Using iscretization Approach for Continuous Features", Emerging Applications of Information Technology (EAIT), 2011 Second International Conference ,Feb 19-20,2011, pp 161 – 164.

[2] Breiman L, Friedman J.H, Olshen R.A, Stone C.J,"Classification and Regression Trees", Wadsforth International Group 1984.

[3] Elena Deza & Michel Marie **"**Deza, *Encyclopedia of Distances"*, page 94, Springer, 2009.

[4] 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, 2012, Vol. 53, No. 2

[5] Heum Park, Hyuk-Chul Kwon, "Extended Relief Algorithms in Instance-Based Feature Filtering", Advanced Language Processing and Web Information Technology,2007. ALPIT 2007. Sixth International Conference, August 22-24, 2007, pp 123-128.

[6] Kira K, Rendell L, "A practical approach to feature selection", Proc 9th International Workshop on Machine Learning, 1992, pp 249-256.

[7] Kononenko, I, "Estimating attributes: analysis and extensions of Relief", In: L. De Raedt and F. Bergadano (eds.): Springer Verlag. Machine Learning: ECML-94, 1994, pp 171–182.

[8] Mantaras R.L.: ID3 Revised: "A distance based creterian for attribute selection", In: Proc.Int.Symp. Methodologies for Intelligent Systems. Charlotte,North Carolina. USA., Oct 1989.

[9] Matthew E Stokes and Shyam Visweswaran, "Application of a spatially-weighted Relief algorithm for ranking genetic predictors of disease", Bio Data Mining 2012, Vol 5, No 20.

[10] Robnik Sikonjam, Kononenko I, "Theoretical and Empirical analysis of ReliefF and RReliefF", Machine Learning, Vol 53, No 1, 2003, pp 23-69.

[11] Quinlan R: Induction of decision trees. Machine learning 1: 81 106, 1986

[12] Smyth P, and Goodman R.N,"Rule induction using information theory", In. G.Piatetsky Shapiro & W. Frawley (eds.): Knowledge Discovery in Databases. MIT Press 1990

[13] 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.

[14] Yuxuan SUN, Xiaojun LOU, Bisai BAO,"A Novel Relief Feature Selection Algorithm Based on Mean Variance Model", Journal of Information & computational Science Vol.8, No 16, 2011, pp 3921-3929.