RELIEF-NCM: A Feature Selection Method for Enhancing Naive Bayesian Classifier

P. Kalpana[1] K.Mani[2]
[1]*Assistant Professor, Department of Computer Science, Nehru Memorial College, Puthanampatti, Tiruchirappalli-DT, Tamil Nadu, India*
[2]*Associate Professor, Department of Computer Science, Nehru Memorial College, Puthanampatti, Tiruchirappalli-DT, Tamil Nadu, India*
(E-mail:parasuramankalpana@gmail.com)

Abstract — Feature selection is a significant task in selecting the most informative and relevant features to the target from the feature space. RELIEF is one of the popular methods based on feature weight estimation. Many extensions of RELIEF have already been developed to improve the accuracy of the classifiers. RELIEF-DISC is one of the extended algorithms of RELIEF which is solely meant for continuous features but fails to deal with nominal features and multiclass problems. Also it considers the values next to the current instances for finding the nearest hit and miss ie., it looks the values towards the forward direction only. To overcome these, this paper focuses on generating a new feature filtering approach called RELIEF-NCM based on RELIEF-DISC. It finds the more appropriate hit and miss value by considering the values before and after the current instance. It converts the nominal features as numeric using target based encoding and also provides solution for the multiclass problem by splitting the problem into multiple binary class problems. The accuracy of the Naive Bayesian Classifier with the selected features by the proposed method outperforms better than RELIEF-DISC.

Keywords—*Feature Selection; Relevant features; RELIEF; RELIEF-DISC; RELIEF-NCM; Naive Bayesian Classifier.*

I. INTRODUCTION

Machine learning and data mining algorithms do not behave well when there are hundreds and thousands of potential features describing each input object because sometimes it may consist of many irrelevant, redundant and noisy features. The unnecessary features in the input will provide little information to the target and also they give more classification error. Thus a task of choosing small subset of features is ideally necessary and sufficiently required to describe the target and it can be performed using feature subset selection techniques [3]. The Feature Selection (FS) methods are useful for making supervised learning more effective and efficient. The feature weighting is a process of assigning weights to features, based on their relevance to a target. Once relevance weights have been assigned to features, the FS process discards the irrelevant features based on the threshold i.e., the features with weight below the threshold is removed.

Feature weighting algorithms are broadly classified into two categories viz., parametric and non-parametric approaches. The parametric approach fits the data to known distributions and then measures the relevance based on that model. The Linear Discriminant Analysis (LDA) is one of the parametric approaches which fits normal distributions to the samples in each class and finds the feature weights that maximize the inter-class distance while minimizing the intra-class distance. But the non-parametric approaches make no assumptions about the distribution of the data. RELIEF and its variants are the best example for non-parametric feature weighting algorithms. It measures the relevance of a feature at a sample in terms of the difference between the sample and other nearby samples of the same class and nearby samples of other classes. The relevance of a feature is the average relevance of the feature across all training samples (Draper, B., et al, 2003). Though the RELIEF-DISC is one of the non-parametric approaches which deals only with the continuous attributes, not supporting the nominal features and does not solve multiclass problems. To overcome these, a non-parametric feature weighting method called RELIEF-NCM has been proposed in this paper which is a variant of RELIEF-DISC and it uses sampling technique for sample selection where NCM is abbreviated as Nominal, Continuous and Multiclass.

The rest of the paper is structured as follows. Section 2 describes the review of literature. The proposed approach of this paper is presented in section 3. Section 4 discusses the results. Finally section 5 ends with conclusion.

II. REVIEW OF LITERATURE

Kira and Rendell [1] have introduced the original Relief algorithm which selects relevant features using statistical method. It does not depend on heuristics. It is an accurate method even if features interact and it is noise-tolerant. The algorithm does not produce a non-optimal feature set size. The original ReliefF algorithm decreases the relevance of some features and increases the relevance of others when irrelevant attributes are added to the dataset. To remove this drawback, the authors Draper, B., et al, [6] have developed an improved version of the algorithm called Iterative Relief and showed that

it removes the bias found in ReliefF and outperforms ReliefF for images. The authors Antonio Arauzo-Azofra et al, [5] have introduced a feature set measure for evaluating the feature sets in a search process. The measure helps in guiding the search process as well as selecting the most appropriate feature set. It was proved that the proposed measure performs better than the highly reputed wrapper approach and Liu's consistency measure.

In [9], the authors Wang, Y., & Makedon, F. have empirically evaluated the performance of Relief-F using the SVM and k-NN classifiers on three cancer classification datasets and proved that the performance of Relief-F is comparable with the other feature filtering methods such as Information Gain, Gain Ratio, and $\chi 2$-statistic using the leave-one-out cross validation method. Yuxuan SUN et al. [2] have developed a new model for RELIEF using mean-variance to eliminate the fluctuation in feature weight due to random selection of instances in the original RELIEF algorithm. It considers both the mean and the variances of the instances as the criterion for feature weight estimation and proved that the proposed model for RELIEF provides better performance for seismic signals of ground targets.

Blessie, E. C., and Karthikeyan, E. [10] have proposed a RELIEF-DISC algorithm, an extension of RELIEF based on discretization for handling continuous features. The proposed algorithm helps to avoid random sampling for selecting the instances and also eliminates the need to specify the sample size. They proved that the RELIEF-DISC maintains the quality of features as that of RELIEF and also produces shorter decision tree than RELIEF. In [4] Fan Wenbing et al, have developed an adaptive relief algorithm to alleviate the deficiencies of Relief by dividing the instance set adaptively. They compared the adaptive relief with the existing algorithms name Relief, Relief-F and I-Relief using the image datasets and proved that the proposed method drastically enhances the classification accuracy and also resolves the blind selection problem in the original relief algorithm.

The authors Sarrafzadeh, A., et al, [7] have studied the effect of reducing and selecting the most effective subset of features by using Legendre moments to extract features. They used ReliefF algorithm to select the most relevant and non-redundant features and support vector machine to classify images and illustrated that ReliefF has improved the retrieval speed and accuracy. S.S.Baskar and L Arockiam [8] have developed a new Relief algorithm called LAS-Relief algorithm, which is evolved from mean variance Relief algorithm. The Manhattan distance has been used for selecting the nearest hit (*NH*) and nearest miss (*NM*) instances. The novel LAS-Relief algorithm uses median as a measure for feature weight estimation.

The LAS-Relief algorithm enhances the classification accuracy of Naive Bayes Classifier (NBC) and J48 than Mean-Variance Relief algorithm. A feature selection method called E LAS-Relief has been initiated by S.S.Baskar and L Arockiam, [3] for improving the LAS-Relief. The LAS-Relief algorithm concentrates on the discrete and continues attributes and it is limited to irrelevant feature removal whereas the E LAS-Relief focused on removing noisy and incomplete data sets. It has been proved that the novel algorithm E LAS-Relief outperforms on agriculture soil data sets for classification. S.S.Baskar and L.Arockiam [14] have presented a novel algorithm for FS called C-LAS Relief based on Chebyshev distance to improve the reliability and accuracy of classifiers and they proved that the classification of NBC and J48 is superior over LAS-Relief.

From the literature it has been found that the RELIEF algorithm has been alleviated to several levels, which pawed way for enhancing the accuracy of classifiers and also proved that the RELIEF algorithm is one of the most successful FS algorithms due to its simplicity and effectiveness. To extend further the RELIEF-NCM algorithm has been proposed in this paper.

III. RELIEF-NCM: A MODIFIED APPROACH TO RELIEF-DISC

In the original RELIEF, the instances are selected from the given dataset randomly and the number of samples selected from the training set is based on sample size. It uses squared Euclidean distance measure for finding the nearest hit and miss [13]. If the sample size is small, then most of the features weight may be less than the threshold value which ultimately leads the features to be irrelevant and hence they are not to be included in the relevant features. On the other hand, if the size of the sample is large, then most of the features weight may be greater than the threshold value and hence the irrelevant features may be considered as relevant ones.

To eradicate these issues the RELIEF-DISC has been introduced only for continuous features [10]. In that the entire dataset is sorted first and then the unique values of each feature are divided into finite number of intervals with different classes. Also it uses the first instance in each interval as the current instance, the next instance in the current interval and the first instance in the next interval as the nearest hit and miss respectively. In this way the need for specifying the sample size is eliminated and the sample size is equal to *m,* the number of intervals. But the current instance is not selected randomly which results in reduction of computational complexity and maintains the quality of features.

It is noted that in the RELIEF-DISC, the first instance in each interval is always selected as the current instance which may sometimes be biased. Also it always considers the second element in each interval as the *NH*, it is not possible to find such *NH* if an interval contains only one value. But for the last interval there is no possibility of finding the *NM*. Since the RELIEF-DISC always selects the second element as the *NH* for all intervals, it may not be more appropriate hit when the last element in the previous interval of the same class has lower distance with the current instance than the distance between the current instance and the second element. Similarly

selecting the *NM* is also inappropriate as it selects the first element in the next interval. Further the original RELIEF-DISC focuses only on continuous features and does not solve the multiclass problems.

To overcome the limitations of RELIEF-DISC, this paper proposes a modified approach called RELIEF-NCM which consists of three phases viz., selecting the current instance, finding the NH and NM and weight estimation and finding the relevance features.

A. Selecting the current instance

Instead of selecting the first instance as in RELIEF-DISC as current instance in each interval, this phase gives equal preference in selecting the current instance using the simulation based deterministic procedure which uses the random numbers. The steps involved are

Step 1: Find the number of instances in each interval say 'n'

Step 2: Calculate the probability of selecting each instance I_i i.e., $p(I_i) = \frac{1}{n}$

Step 3: Compute the cumulative probability for each instance I_i, as $\sum_{i=1}^{x} p(I_i)$ where $1 \le x \le n$.

Also $\sum_{i=1}^{n} p(I_i) = 1$

Step 4: Assign the range of random numbers for each instance based on the cumulative probability

Step 5: Generate a random number and find the range in which the selected random number falls

Step 6: The instance which falls in the corresponding range of the random number generated is selected as the current instance.

B. Finding the NH and NM

Let I_i, I_{i-1}, I_{i+1} are the current, previous and next intervals of feature F_i respectively. Also $2 \le i \le n-1$ and $n(I_i)$ be the number of values in I_i. The different possible cases in finding *NM* and *NH* are

$$d1 = dist\_cal(I_i, last\_value(I_{i-1}))$$

$$d2 = dist\_cal(I_i, first\_value(I_{i+1}))$$

$$d3 = dist\_cal(I_i, last\_value(I_{i-2}))$$

$$d4 = dist\_cal(I_i, next\_value(I_i))$$

$$d5 = dist\_cal(I_i, previous\_value(I_i))$$

$$d6 = dist\_cal(I_i, first\_value(I_{i+2}))$$

case 1: if n(I_i)=1then

$$\min\_dist\_1 = \min(d1, d2)$$

$$NH = first\_value(I_{i+2})$$

$$NM = \begin{cases} last\_value(I_{i-1}), \text{ if } \min\_dist\_1 = d1 \\ first\_value(I_{i+1}), \text{ if } \min\_dist\_1 = d2 \end{cases}$$

case 2: if n(I_i)=2 then

$$\min\_dist\_2 = \min(d3, d4)$$

$$\min\_dist\_3 = \min(d5, d6)$$

$$NH1 = NH = \begin{cases} \begin{cases} last\_value(I_{i-2}), \text{ if min\_dist\_2=d3} \\ next\_value(I_i), \text{ if min\_dist\_2=d4} \end{cases} \\ \begin{cases} previous\_value(I_i), \text{ if min\_dist\_3=d5} \\ first\_value(I_{i+2}), \text{ if min\_dist\_3=d6} \end{cases} \end{cases}$$

NM = same as case 1

case 3: If n(I_i)>2 then

$$\min\_dist\_4 = dist\_cal(d5, d4)$$

$$NH = \begin{cases} NH1 \\ NH2 = \begin{cases} previous\_value(I_i), \text{ if min\_dist\_4=d5} \\ next\_value(I_i), \text{ if min\_dist\_4=d4} \end{cases} \end{cases}$$

NM = same as case 1

For the interval I_1, the *NH* and *NM* is considered as

case 1: if n(I_i)=1then

$$NH = first\_value(I_{i+2})$$

case 2: if n(I_i)=2 then

$$\min\_dist\_5 = \min(d5, d6)$$

$$NH = \begin{cases} NH3 = next\_value(I_i) \\ NH4 = \begin{cases} previous\_value(I_i), \text{ if min\_dist\_5=d5} \\ first\_value(I_{i+2}), \text{ if min\_dist\_5=d6} \end{cases} \end{cases}$$

case 3: if n(I_i)>2 then

$$NH = \begin{cases} NH3 \\ \begin{cases} NH4 \\ NH2 \end{cases} \end{cases}$$

For all the above cases,

$$NM = first\_value(I_{i+1})$$

For the interval I_n, the *NH* and *NM* is considered as follows

case 1: if n(I_i)=1then

$$NH = last\_value(I_i - 2)$$

case 2: if n(I_i)=2 then

$$NH = \begin{cases} NH5 = \begin{cases} last\_value(I_{i-2}), \text{ if min\_dist\_2=d3} \\ next\_value(I_i), \text{ if min\_dist\_2=d4} \end{cases} \\ previous\_value(I_i) \end{cases}$$

case 3: if $n(I_i)>2$ then

$$NH = \begin{cases} NH5 \\ previous\_value(I_i) \\ NH2 \end{cases}$$

For all the above cases,

$$NM = last\_value(I_{i-1})$$

It is noted that this phase of the proposed work uses the Manhattan distance to find *NH* and *NM* because it require less time to find distance between any two points than the squared Euclidean distance and it is computed as

$$d(X,Y) = \sum_{i=1}^{n} | x_i - y_i | \quad (1)$$

Where X and Y are two vectors $X=(x_1,x_2,x_3,...,x_n)$ and $Y=(y_1,y_2,y_3,...,y_n)$ [12]. It is shown in procedure dist_cal().

```
procedure dist_cal (Instance_1, Instance_2)
begin
   dist ← abs(Instance_1- Instance_2)
   return (dist)
end
```

C. Weight estimation and selecting the relevance features

It is noted that the relevant features F_i's are selected on the basis of their weight value and it is calculated as

$$\begin{aligned} \Delta W(F_i) &= diff(F_i, I_i, NM) - diff(F_i, I_i, NH) \\ W(F_i) &= W(F_i) + \Delta W(F_i) \\ W(F_i) &= W(F_i) / m \end{aligned} \quad (2)$$

Where $diff(F_i,I_i,NH)$ and $diff(F_i,I_i,NM)$ is calculated as

$$diff(F_i, I_i, NH) = | I_i - NH / (\max(F_i) - \min(F_i)) | \quad (3)$$

$$diff(F_i, I_i, NM) = | I_i - NM | / (\max(F_i) - \min(F_i)) | \quad (4)$$

After finding weight of each feature, the top ranked 'k' features are selected as the relevance features. The pseudo code of the proposed methodology is shown in RELIEF_NCM_PS_1.

Pseudo code RELIEF_NCM_PS_1

1. Let D be the training data set with 'n' features F_1, F_2, F_3,...,F_n and a class label C
2. Let k be the size of F_{subset}, the subset of features and $k \leftarrow 0.75 \times n$
3. Initialize the weights of all features F_i as 0.0 ie., $W(F_i) \leftarrow 0.0$ for $1 \le i \le n$
4. $F_{subset} \leftarrow \emptyset$
5. for every feature F_i do
6. if F_i is nominal then
7. Encode the nominal values as numeric using target based encoding
8. end if
9. Sort the values of F_i in ascending order
10. Split the values of F_i into finite number of intervals 'm' with the values of the features have different classes
11. for each interval $I_j, 1 \le j \le m$ in F_i do
12. Select the current instance I randomly in the current interval
13. Find the nearest hit *NH* and miss *NM*
14. Calculate $diff(F_i,I,NH)$ and $diff(F_i,I,NM)$
15. Update the weight of the features F_i as $W(F_i) \leftarrow W(F_i) - diff(F_i,I,NH)/m + diff(F_i,I,NM)/m$
16. end for
17. end for
18. Sort all the features F_i in descending order of $W(F_i)$
19. repeat
20. $F_{subset} \leftarrow F_{subset} \cup F_i$
21. until $|F_{subset}| \le k$
22. return F_{subset}

The proposed method also paves way for multiclass problems by solving the multiclass problems as multiple binary class problems. i.e., the weight of each feature is the mean of multiple binary class weights. It provides support for nominal features by converting them into numeric using target-based encoding method. Further it considers all the values of the feature rather than the unique values as in RELIEF_DISC because the feature with same value may belong to different class.

D. RELIEF-NCM: An Example

In order to understand the relevance of the work, the weather dataset shown in Table 1 has been taken from the UCI machine learning repository [11]. It contains 5 fields viz., outlook, temperature, humidity, windy and play. Among them temperature and humidity are continuous attributes and outlook and windy fields are nominal type. The unique values of outlook viz., Overcast, Rainy and Sunny are encoded as 1, 0.6 and 0.4 respectively. Similarly the TRUE and FALSE values in windy attribute are encoded as 0.5 and 0.75 respectively using the target-based encoding method. After encoding the entire dataset contains now only numerical values and it is shown in Table 2.

The weight value for temperature is calculated based on RELIEF_NCM_PS_1. The temperature has 8 intervals after sorting and they are shown in Table 3. The current instance *I*, *NH, NM* and the weight value for each interval are shown in Table 4.

TABLE 1. Weather Dataset

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	FALSE	No
Sunny	80	90	TRUE	No
Overcast	83	86	FALSE	Yes
Rainy	70	96	FALSE	Yes
Rainy	68	80	FALSE	Yes
Rainy	65	70	TRUE	No
Overcast	64	65	TRUE	Yes
Sunny	72	95	FALSE	No
Sunny	69	70	FALSE	Yes
Rainy	75	80	FALSE	Yes
Sunny	75	70	TRUE	Yes
Overcast	72	90	TRUE	Yes
Overcast	81	75	FALSE	Yes
Rainy	71	91	TRUE	No

TABLE 2. Weather Dataset after Target-based Encoding

Outlook	Temperature	Humidity	Windy	Play
0.4	85	85	0.75	No
0.4	80	90	0.50	No
1.0	83	86	0.75	Yes
0.6	70	96	0.75	Yes
0.6	68	80	0.75	Yes
0.6	65	70	0.50	No
1.0	64	65	0.50	Yes
0.4	72	95	0.75	No
0.4	69	70	0.75	Yes
0.6	75	80	0.75	Yes
0.4	75	70	0.50	Yes
1.0	72	90	0.50	Yes
1.0	81	75	0.75	Yes
0.6	71	91	0.50	No

TABLE 3. Interval Split-Up of Temperature

Interval	1	2	3	4	5	6	7	8
Temperature	64	65	68 69 70	71 72	72 75 75	80	81 83	85
Play*	Y	N	Y	N	Y	N	Y	N

***Y-Yes, N-No**

From table 4, the weight of temperature attribute in the first run is -0.07143, the process is simulated for 20-runs and the final weight of temperature computed as -0.0807. Similar process can also be performed for other features with 20-runs and the results are shown in Table 5. From table 5, the top ranked attributes viz., outlook, temperature and humidity are selected as the more relevant features because k=75% of total attributes in the dataset.

TABLE 4. Weight Computation of each Interval

Interval	Generated random number	I	NH	NM	Weight
1	--	64	68	65	-0.01786
2	--	65	71	68	-0.04762
3	29	68	69	71	-0.03571
4	05	72	71	72	-0.04167
5	74	75	75	72	-0.02381
6	--	80	85	81	-0.04762
7	04	81	83	80	-0.05357
8	--	85	80	83	-0.07143

TABLE 5. Feature Weights of Weather Dataset

Feature (F_i)	Weight ($W(F_i)$)	Rank
Outlook	0.0000	1
Temperature	-0.0807	2
Humidity	-0.1124	3
Windy	-0.2500	4

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In order to analyze RELIEF-NCM, 8 datasets have been taken from UCI Machine Learning Repository. Each dataset contains both continuous and nominal features. The datasets taken for the experiments contain two or three classes. The missing values for each attribute in the datasets are filled with their corresponding mean. The detailed specification of these datasets is shown in Table 6. The existing RELIEF-DISC and the proposed RELIEF-NCM have been implemented in PYTHON. The number of features selected and the selected features are shown in Table 7.

TABLE 6. Description of the Datasets

Dataset	No. of Attributes	No. of Instances	No. of Classes
Pima Indian Diabetes	9	768	2
Breast Cancer	11	699	2
Statlog Heart	14	270	2
Eeg	15	14979	2
Weather	5	14	2
Ann-train	22	3772	3
Lung Cancer	57	32	3
SPECTF_train	45	80	2

The selected attributes obtained from both methods and the original attributes are then fed into NBC for determining the predictive accuracy, precision and recall using WEKA tool

with 10-fold cross validation. The predictive accuracies, precision and recall of all datasets are shown in Tables 8, 9 and 10 respectively. The number of original features and selected features using RELIEF-NCM is shown in Fig. 1.

TABLE 7. Number of Selected Features Vs. the Selected Features

Data sets	No. of Features Selected	Selected Features excluding class label
Pima Indian Diabetes	7	3,8,1,2,7,6
Breast cancer	9	1,3,7,5,9,2,6,10
Statlog Heart	11	7,8,1,5,4,10,3,12,13,11
Eeg	12	12,10,6,7,13,14,4,1,9,8,2
Weather	4	2,3,1
Ann-train	17	1,2,18,20,19,21,17,3,4,6,7,8,13,15,16,5
Lung Cancer	42	47,21,32,33,4,2,8,30,31,37,39,40,41,48,49,52,53,54,13,7,16,29,38,42,43,5,34,3,44,45,46,50,15,26,27,35,36,1,6,11,12,14
SPECTF_train	33	8,14,31,5,26,3,43,28,34,17,1,36,19,13,7,18,27,25,6,35,21,4,39,20,12,29,22,24,23,9,2,10,44

TABLE 8. Accuracy of NBC with All Features and Selected Features using RELIEF-DISC and RELIEF-NCM

Data sets	Accuracy (%) of NBC with		
	All Features	Selected Features using	
		RELIEF-DISC	RELIEF-NCM
Pima Indian Diabetes	76.3021	67.1875	**76.8229**
Breast cancer	95.9943	**96.2804**	95.9943
Statlog Heart	**83.7037**	83.3333	82.9630
Eeg	48.0406	**48.5880**	48.1808
Weather	57.1429	NA*	**64.2857**
Ann-train	**95.6522**	NA*	95.5992
Lung Cancer	84.3750	78.1250	**85.1852**
SPECTF_train	76.2500	**77.5000**	**77.5000**
Average	77.1826	76.1690	**78.3164**

*NA-Not Applicable

From Tables 8, it is observed that the accuracy of NBC for Pima Indian Diabetes, weather and Lung Cancer are more with the selected features using RELIEF-NCM than the others. Similarly the accuracy of NBC for Breast Cancer and Eeg using RELIEF-DISC is more than that of the others. Similarly the accuracy of Statlog Heart dataset and Ann-train is slightly less with selected features using both methods than the original features. It is also noted that the accuracy of NBC for SPECTF_train remains the same with the selected features using RELIEF-DISC and RELIEF-NCM and it is more than the original features.

It is further identified that on an average the accuracy, precision and recall of RELIEF-NCM is increased by 1.1338%, 0.013% and 0.01% respectively when they are compared with all features. Similarly RELIEF_NCM increases the accuracy, precision and recall by 2.1474%, 0.029% and 0.03% respectively than RELIEF-DISC.

TABLE 9. Precision of NBC with All Features and Selected Features using RELIEF-DISC and RELIEF-NCM

Data sets	Precision of NBC with		
	All Features	Selected Features using	
		RELIEF-DISC	RELIEF-NCM
Pima Indian Diabetes	0.759	0.653	**0.764**
Breast cancer	0.962	**0.964**	0.962
Statlog Heart	**0.837**	0.833	0.829
Eeg	0.529	**0.536**	0.534
Weather	0.528	NA*	**0.629**
Ann-train	**0.950**	NA*	**0.950**
Lung Cancer	0.84	0.768	**0.841**
SPECTF_train	0.776	**0.786**	0.781
Average	0.773	0.757	**0.786**

*NA-Not Applicable

TABLE 10. Recall of NBC with All Features and Selected Features using RELIEF-DISC and RELIEF-NCM

Data sets	Recall of NBC with		
	All Features	Selected Features using	
		RELIEF-DISC	RELIEF-NCM
Pima Indian Diabetes	0.763	0.672	**0.768**
Breast cancer	0.960	**0.963**	0.960
Statlog Heart	**0.837**	0.833	0.830
Eeg	0.480	**0.486**	0.482
Weather	0.571	NA	**0.643**
Ann-train	**0.957**	NA	0.956
Lung Cancer	**0.844**	0.781	0.841
SPECTF_train	0.763	**0.775**	**0.775**
Average	0.772	0.752	**0.782**

*NA-Not Applicable

The reason for the increasing the predictive accuracy, precision and recall is that the RELIEF-NCM selects the nearest hit and miss using Manhattan distance by considering all possibilities than the one specified in RELIEF-DISC. Also

it picks the more relevant features and provides better performance of NBC. Since the proposed methodology is based on RELIEF-DISC, the quality of features is also maintained. As in RELIEF-DISC, the RELIEF-NCM does not require the domain expert to specify the number of samples to be selected and it will always depend on the number of intervals for each feature.

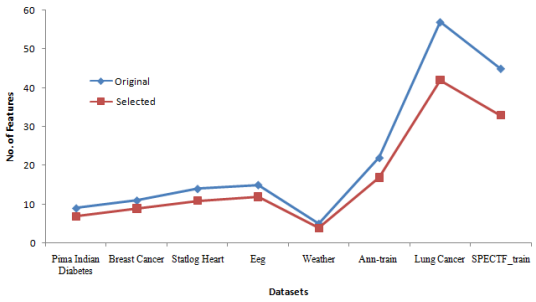

Fig. 1. Number of original features vs. selected features

V. CONCLUSION

This paper provides a modified approach called RELIEF-NCM, which eradicates the issues in RELIEF-DISC. The proposed approach enhances the accuracy of NBC by considering the hit and miss values in both directions but the RELIEF-DISC consider only the values towards forward direction for continuous features. It converts the nominal values as numeric using target-based encoding method and uses the same procedure of finding the feature weight as continuous case. Further it solves the multiclass problems. The experimental results clearly show that the proposed method RELIEF-NCM enhances the accuracy of NBC with the selected features by 1.1684% approximately than the others.

REFERENCES

[1] Kira, Kenji, and Larry A. Rendell. (1992). "The feature selection problem: Traditional methods and a new algorithm." ***In AAAI,*** vol. 2, pp. 129-134.

[2] Sun, Y., Lou, X., & Bao, B. (2011). A novel relief feature selection algorithm based on mean-variance model. ***Journal of Information & Computational Science***, vol. 8, issue. 16, pp. 3921-3929.

[3] S.S.Baskar, L Arockiam, (2013), "E LAS-Relief-A Novel Feature Selection Algorithm in Data Mining", *COMPUSOFT, **An international journal of advanced computer technology***, vol. 2 issue. 12, pp.391.

[4] Wenbing, F., Quanquan, W., & Hui, Z. (2012). "Feature Selection Method Based on Adaptive Relief Algorithm". ***In 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010) IPCSIT***, vol. 53, No. 2.

[5] Arauzo-Azofra, A., Benitez, J. M., & Castro, J. L. (2004). "A feature set measure based on relief". ***In Proceedings of the fifth international conference on Recent Advances in Soft Computing***, pp. 104-109.

[6] Draper, B., Kaito, C., & Bins, J. (2003), "Iterative relief". ***In Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03.*** vol. 6, pp. 62-62, IEEE.

[7] Sarrafzadeh, A., Atabay, H. A., Pedram, M. M., & Shanbehzadeh, J. (2012). "ReliefF Based Feature Selection in Content-Based Image Retrieval". ***In Proceedings of the International Multi Conference of Engineers and Computer Scientists***", vol. 1.

[8] Baskar, S. S., & Arockiam, L. (2013). "A Novel LAS-Relief Feature Selection Algorithms for Enhancing Classification Accuracy in Data Mining", ***International Journal of Computers and Technology***, vol. 11, no. 8, pp. 2922-2927.

[9] Wang, Y., & Makedon, F. (2004). "Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data". ***In Computational Systems Bioinformatics Conference***, ***IEEE,*** pp. 497-498.

[10] Blessie, E. C., & Karthikeyan, E. (2011). "RELIEF-DISC: An Extended RELIEF algorithm using Discretization approach for continuous features". ***In Second International Conference on Emerging Applications of Information Technology (EAIT)***, pp. 161-164, IEEE.

[11] UCI Machine Learning Repository - Center for Machine Learning and Intelligent System, available at: http://archive.ics.uci.edu.

[12] Jiawei Han and Micheline Kambar, "Data Mining: Concepts and Techniques", 3rd edition, Morgan Kaufmann Publisher, 2012

[13] Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. ***Machine learning,*** 53(1-2), 23-69.

[14] Baskar, S. S., & Arockiam, L. (2013). "C-LAS Relief-An Improved Feature Selection Technique in Data Mining". ***International Journal of Computer Applications***, vol. 83, no. 13.

P. Kalpana received B.Sc and M.Sc degrees in Computer Science from Seethalakshmi Ramaswami College, affiliated to Bharathidasan University, Tiruchirappalli, India in 1999 and 2001 respectively. She received M.Phil degree in Computer Science in 2004 from Bharathidasan University. She also received MBA degree in Human Resource Management from Bharathidasan University in 2007. She is presently working as an Assistant Professor in the Department of Computer Science, Nehru Memorial College, Puthanampatti, Tiruchirappalli. She is pursuing PhD degree in Computer Science at Bharathidasan University. She has published and presented around 6 research papers at international journals and conferences. Her research interests include Algorithms, Data Pre-processing and Data Mining techniques.

K. Mani received his MCA and M.Tech from the Bharathidasan University, Trichy, India in Computer Applications and Advanced Information Technology respectively. He completed his Graduation in Operations Research from Operational Research Society of India, Kolkata. Since 1989, he has been with the Department of Computer Science at the Nehru Memorial College, affiliated to Bharathidasan University where he is currently working as an Associate Professor. He completed his PhD in Cryptography with primary emphasis on evolution of framework for enhancing the security and optimizing the run time in cryptographic algorithms. He published and presented around 25 research papers at international journals and conferences. His research area includes Cryptography, Data Mining, Coding Theory, Computer Simulation and Optimization of Algorithms.