Feature Selection for Classification Using Extreme Learning Machine

Mrs. Chapala Maharana, Ch. Sanjeev Kumar Dash, Bijan Bihari Mishra, Satchidananda Dehuri
Assistant Professor, Associate Professor, Professor, Professor
Computer Science & Engineeering,
Eastern Academy of Science & Technology, Bhubaneswar, Odisha.

Abstract--- *It* has been realized that feature subset selection is running with a lot of challenges even though many efforts have been made so far for small to large scale dimensional datasets. But for larger dimensional datasets a very rare feature subset selection techniques have been developed with their own set of constraints. In this work, we propose a novel two steps feature subset selection for high dimensional datasets. In the first step the features in the database are ranked and are arranged as per their rank from maximum to minimum. Then few maximum ranking features are presented to Extreme Learning Machine (ELM) for classification and the minimum subset that performs well is selected. The proposed approach is evaluated with a few benchmarking highly skewed dataset retrieved from University of California, Irvine (UCI) repository. In addition to classification accuracy, four other performance metrics such as sensitivity, specificity, Jaccard index and M-Estimate also used to validate the results. The experimental study is encouraging us to pursue further research in high dimensional skewed data.

Keywords - Extreme Learning Machine; signal to noise ration; classification; feature selection.

I. INTRODUCTION

It is being accepted by researcher the accuracy of the newly exposed model (i.e. neural networks NNs)[1] strongly depends on the quality of data being mined. Feature selection one of the preprocessing tasks to obtain quality data brings lots of attention of many researchers [17]. It is the process of selecting a subset of available features to use in experimental modeling. Feature selection can be broadly classified into two categories: i) filter approach (it depends on standard statistical measurement); and ii) wrapper approach (based on the accuracy of a specific classifier) [20]. Over the decade ELM have attracted many researchers in various domains; one of the reason is that usually a too small classifier network lacks the capability of learning satisfactorily. On the other hand, a network that is too large could also overfit the training data, thus producing the poor generalization performance. In addition, particularly in large network also brings about better prediction responses and unnecessary requirement for large memory as well as high cost for hardware implementation. A new fast learning algorithm referred to as extreme learning machine with additive hidden [9] nodes and radial basis function (RBF) [13] kernels has been developed. ELM has been applied to many real world applications [6] and has been publicized to generate good generalization performance at extremely high learning speed.

This paper is organized as follows. Section I is the introduction. In Section II, the ELM is discussed. In Section III, the signal to noise feature ranking is discussed. In Section IV, proposed method is presented. In Section V, a number of numerical experiments using this newly proposed model are conducted and some experimental results and remarks are illustrated. Section VI concludes this paper.

II. ELM NETWORK

ELM network is a popular artificial neural network architecture that has found wide applications in different fields of engineering. It is used in pattern recognition, function approximation, and time series prediction.

In case of feed forward-type neural networks the parameters are to be determined by using learning algorithms. The back propagation (BP) training algorithm based on the gradient descent method has been most widely used for training of feed forward type neural networks. In order to overcome the slow training performance, the extreme learning machine (ELM) has been proposed [3, 5]. Major advantage in ELM [7] over conventional BP is that ELM learns without iteratively tuning hidden weights. When the weights connecting the hidden layer, β and the output layer are solved by minimizing the following approximation error:

$$\min \left\| H\beta - T \right\|^2 \qquad (1)$$

Whereas H is the hidden layer output matrix and T is the training data target matrix.

The optimal solution of β can be given as:

$$\beta^{+} = H^{+}T \qquad (2)$$

Where H^{+} denotes a generalized inverse of the matrix H.

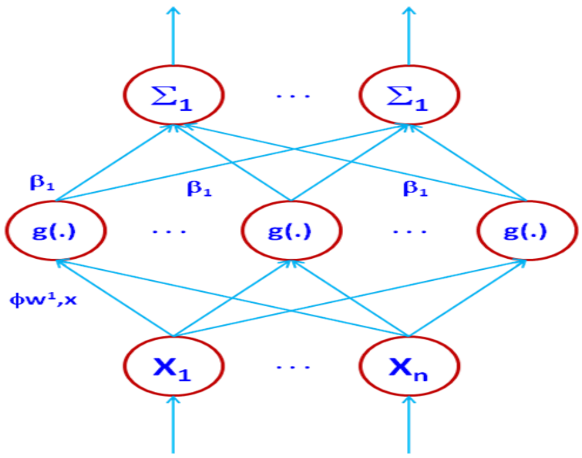

Fig. 1. Artitecture of Extreme Learning Machine

ELM has the advantage in training speed and generalization performance while having problems including robustness. Regularized ELM was proposed to alleviate the robustness problem in ELM by minimizing the regularized cost function of least squared estimated regularization with the following formulation

$$\min L_{REM} = \frac{1}{2}\|\beta\|^2 + \frac{c}{2}\|H\beta - T\|^2 \qquad (3)$$

Where C is a scale parameter.

By setting the gradient of L_{REM} with respect to β to zero, we can find the output weight matrix β when the number of training samples is larger than the number of hidden neurons as follows:

$$\beta = \left(\frac{1}{c} + H^T H\right)^{-1} H^T Y \qquad (4)$$

The output weight matrix β when the number of training samples is smaller than the number of hidden neurons is given as:

$$\beta = H^T \left(\frac{1}{c} + H^T H\right)^{-1} Y \qquad (5)$$

Rong, H. J., Ong, Y. S., Tan, A. H., & Zhu, Z.. in [2] have addressed the architectural design of the ELM classifier network, since too few/many hidden nodes employed would lead to underfitting/overfitting issues in pattern classification. In particular, they describe the proposed pruned ELM (P-ELM) algorithm as a systematic and automated approach for designing ELM classifier network. P-ELM uses statistical methods to measure the relevance of hidden nodes. Beginning from an initial large number of hidden nodes, irrelevant nodes are then pruned by considering their relevance to the class labels. Huang, G. B., Zhou, H., Ding, X., & Zhang, R. in [4] shows that both LS-SVM and PSVM can be simplified further and a unified learning framework of LS-SVM, PSVM, and other regularization algorithms referred to extreme learning machine (ELM) can be built. ELM works for the “generalized” single-hidden-layer feed-forward networks (SLFNs), but the hidden layer (or called feature mapping) in ELM need not be tuned. Huang, G. B., Zhu, Q. Y., & Siew, C. K. in [6] have proposed a new learning algorithm called extreme learning machine (ELM) for single-hidden layer feed-forward neural networks (SLFNs) which randomly chooses hidden nodes and analytically determines the output weights of SLFNs. Huang, G. B., Zhu, Q. Y., & Siew, C. K in [7] proposed a new learning algorithm called extreme learning machine (ELM) for single hidden layer feed-forward neural networks (SLFNs) which randomly chooses the input weights and analytically determines the output weights of SLFNs. A hybrid learning algorithm is proposed by Zhu, Q. Y., Qin, A. K., Suganthan, P. N., & Huang, G. B.. in [8] which uses the differential evolutionary algorithm to select the input weights and Moore–Penrose (MP) generalized inverse to analytically determine the output weights. Feng, G., Huang, G. B., Lin, Q., & Gay, R. [9] have given a approach that referred to as error minimized extreme learning machine (EM-ELM) can add random hidden nodes to SLFNs one by one or group by group (with varying group size). During the growth of the networks, the output weights are updated incrementally. The convergence of this approach is proved in this brief as well. Li, M. B., Huang, G. B., Saratchandran, P., & Sundararajan, N. in [10] have extended the ELM algorithm from the real domain to the complex domain, and then apply the fully complex extreme learning machine (C-ELM) for nonlinear channel equalization applications. Huang, G. B., Ding, X., & Zhou, H. in [11] have shown that (1) under the ELM learning framework, SVM's maximal margin property and the minimal norm of weights theory of feed forward neural networks are actually consistent; (2) from the standard optimization method point of view ELM for classification and SVM are equivalent but ELM has less optimization constraints due to its special separability feature; (3) as analyzed in theory and further verified by the simulation results, ELM for classification tends to achieve better generalization performance than traditional SVM. Huang, G. B., & Chen, L. [12] in their study found that some of the hidden nodes in such networks may play a very minor role in the network output and thus may eventually increase the network complexity. Huang, G. B., & Siew, C. K. [13] have proposed a new algorithm that was based on a single-hidden layer feed forward neural networks (SLFNs) with additive neurons to easily achieve good generalization performance at extremely fast learning speed. Deng, W., Zheng, Q., & Chen, L. in [14] have proposed a novel algorithm called Regularized Extreme Learning Machine based on structural risk minimization principle and weighted least square. The generalization performance of the proposed algorithm was improved significantly in most cases without increasing training time. Chacko, B. P., Krishnan, V. V., Raju, G., & Anto, P. B. in [18] have used wavelet energy feature (WEF) and extreme learning machine (ELM) for reorganization of handwritten Malayalam character .

III. Feature Selection Using Signal To Noise Ratio

Feature selection (FS) [19] is essentially a task to remove irrelevant and/or redundant features. In simple words, feature selection techniques study how to select a subset of attributes or variables in a dataset. The selection of features can be achieved in two ways:
(a)Filter Method: It precedes the actual classification process. The filter approach is independent of the learning algorithm, computationally simple fast and scalable.

(b) *Wrapper Method*: These methods generally result in better performance than filter methods because the feature selection process is optimized for the classification algorithm to be used. However, wrapper methods are too expensive for large dimensional database in terms of computational complexity and time since each feature set considered must be evaluated with the classifier algorithm used.

The signal to noise ratio (SNR)[15,16] test identifies the expression patterns with a maximal difference in mean expression between two groups and minimal variation of expression within each group

$$SNR = (\mu^1 / \mu^2)(\sigma_1 + \sigma_2), \qquad (6)$$

where μ^1 and μ^2 denote the mean expression values for the sample class 1 and class 2 respectively. σ_1 and σ_2 are the standard deviations for the samples in each class. Mishra et al.[15] have provided a model for feature selection using signal-to-noise ratio (SNR) ranking. Basically we have proposed two approaches of feature selection. In first approach, the genes of microarray data is clustered by k-means clustering and then SNR ranking is implemented to get top ranked features from each cluster and given to two classifiers for validation such as SVM and k-NN. In the second approach the features (genes) of microarray data set is ranked by implementing only SNR ranking and top scored feature are given to the classifier and validated. We have tested Leukemia data set for the proposed approach and 10fold cross validation method to validate the classifiers. Lakshmi et al. [16] have produced excellent accuracy with reduced feature set by a simple method. When the profile built using a feature selection method called MSNR (maximized signal to noise ratio) combined with modified fractional similarity method, it performs in a competitive manner. MSNR identifies the highly contributing features and increases the distance between the profiles.

IV. Proposed Method

In the first step of this approach, the attributes of the dataset is ranked using signal to noise ratio technique as discussed in Section III. Then the database is rearranged as per the ranks i.e., starting with maximum rank to minimum rank. It is considered that the higher ranked attributes are more suitable for classification than lower ranking attributes. We have added one by one attribute to feature set until the classifier accuracy is better. The pseudocodes for our proposed model is as follows.

```
1. Rank the database using signal to noise ratio.
2. Rearrange the attributes with maximum rank to minimum rank.
3. Let mxAtr=100.
4. If number of  attributes in the database is less than 100
       then mxAtr= number of  attributes in the database.
5. for A=1: mxAtr
       select subset of attributes with index1toindexA for both the sets of database. Present set1 with subset of attributes to ELM for training and use set2 with the same subset of attributes for testing and record the performance as confusion matrix for subsetA of attributes. Then present set2 for training and set1 for testing with same subset of attributes and the result is recorded for subsetA of attributes.
6. Evaluate performance metrics such as accuracy, sensitivity, specificity, Jaccard index from the confusion matrices for each subset of attributes and for each test set.
7. Compare the different performance metrics and find suitable subset of attributes which yields overall better performance of all the performance metrics.
```

V. Experimental Study

The datasets used in this work were obtained from the UCI machine learning repository. Four two class databases are used to test the effectiveness of the proposed method. Twofold cross validation technique is applied here. The detail of data distribution in each set of the data is presented in Table 1.

TABLE 1. DESCRIPTION OF DATASETS

Database	Samples In				Attributes
	Set1		*Set2*		
	Class1	*Class2*	*Class1*	*Class2*	
Mushroom	1066	1732	1065	1733	22
HorseColic	91	59	90	60	27
LSVT voice rehabilitation	21	42	21	42	310
Arcene	22	28	22	28	10000

V.I Results and Analysis

TABLE 2.Accuracy of Different Dataset with Feature Number

Dataset	Number of attributes selected	Percentage of total attributes selected
Mushroom	3	13.64
Horse Colic	1- 4	16.67
LSVT Voice Rehabilitation	6	01.94
Arcene	14	0.14

From the above table it can be seen that in arcane database there are 10,000 attributes out of which only 14 attributes are selected with acceptable level of performance. In other databases also it selects few attributes with good performance. When number of attributes is very large, selection of appropriate attributes is also difficult. This technique performs well for database with few attributes to very large number of attributes.

Classification accuracy obtained for each subset of attributes is presented in Figure 2-5 for different datasets. The results obtained for set1, set 2, and their average is presented in these figures.

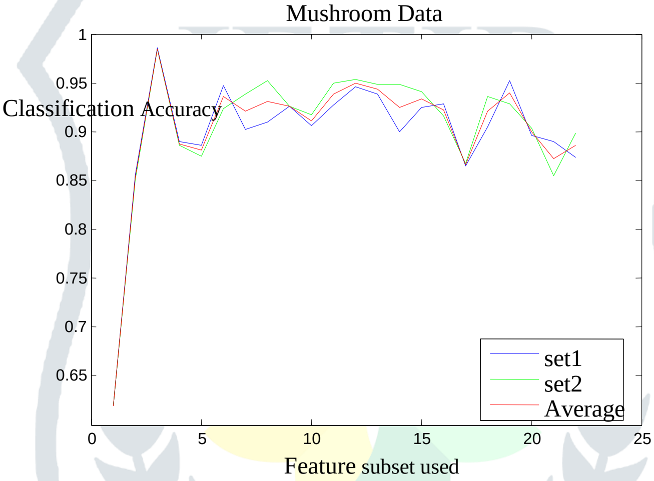

Fig. 2. Comparison of Classification Accuracy of Mushroom Dataset with Different Set of Features

From Figure 2, it can be revealed that in case of Mushroom database, the maximum average accuracy level is reached only with 3 attributes. Further increase in set of attributes does not improve the performance.

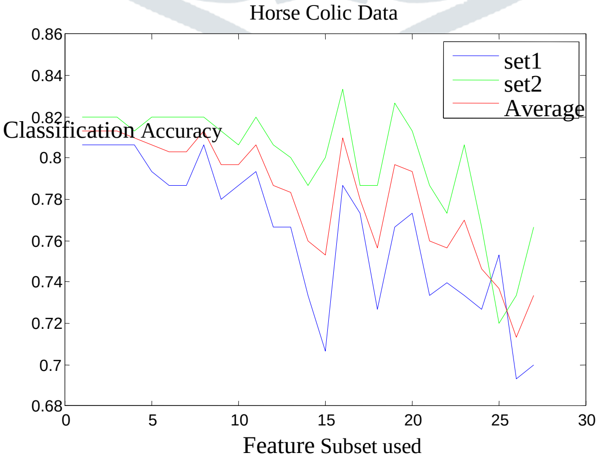

Fig. 3. Comparison of Glassification Accuracy of Horse Colic Dataset with Different Set of Features

It is observed from the Horse Colic database in Figure 3 that the maximum level of average accuracy is reached only with one attribute and the same level of average accuracy is maintained up to four attributes. Further increase in set of attributes does not improve the average accuracy.

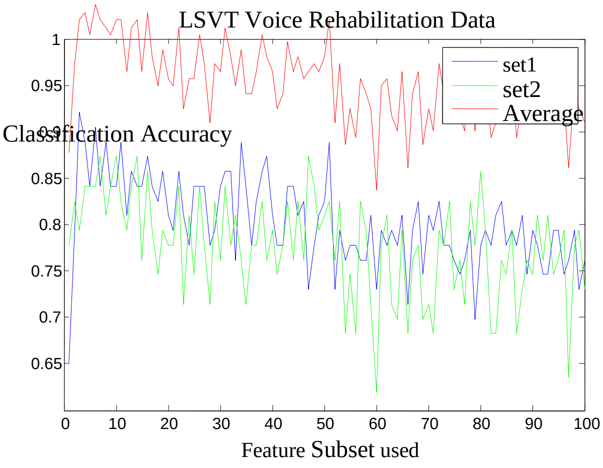

Fig. 4. Comparison of Classification Accuracy of LSVT_Voice_Rehabilitation Dataset with Different Set of Features

For LSVT Voice Rehabilitation Database in Figure 4, it can be seen that the maximum accuracy level is reached only with 6 attributes. Then it shows a decreasing trend for rest 100 subset of attributes considered.

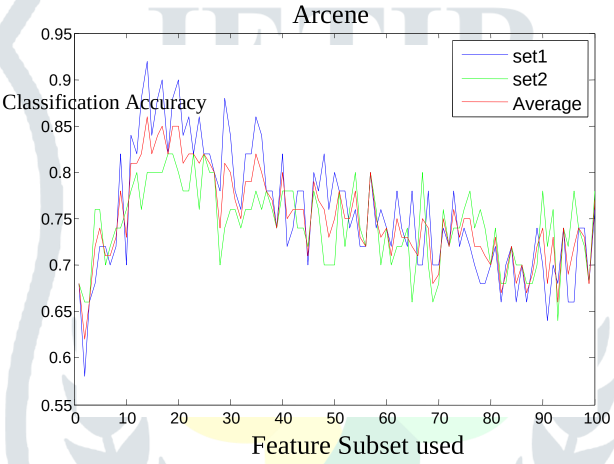

Fig. 5. Comparison of Classification Accuracy of Arcene Database with Different Set of Features

Similarly for Arcene Database in Figure 5, it can be seen that the maximum accuracy level is reached only with 14 attributes. Then it shows a decreasing trend for rest 100 subset of attributes considered.

Further to establish that the results obtained for accuracy is consistent with other performance metrics, four other performance metrics are evaluated such as sensitivity, specificity, Jaccard index and M-Estimate.

Table 3. Sensitivity, Specificity, Jaccard index and M-Estimate Values with Same Set of Attributes

Dataset	**Sensitivity**	**Specificity**	**Jaccard**	**M_Estimate**
Mushroom	0.96	1	0.96	.98
Horse Colic	0.87	0.71	0.73	0.76
LSVT Voice Rehabilitation	0.73	0.94	0.65	0.5
Arcene	0.88	0.83	0.74	0.61

It is observed that with the same set of attributes, best or close to best result for sensitivity, specificity, Jaccard index and M-Estimate is also obtained. Table 3 shows the results for sensitivity, specificity, Jaccard index and M-Estimate for the set of attributes selected as in Table 2.

Table 4. : klosgen , f_measure, kappaIndex with Same Set of Attributes .

Dataset	**klosgen**	**f_measure**	**kappaIndex**
Mushroom	0.401808	0.967926	0.967926
HorseColic	0.181254	0.8470808	0.596575

LSVT_voice_ rehabilitation Arcene0.278839	0.308049	0.7777778	0.693793
Arcene	0.278839	0.852381	0.720872

VI. CONCLUSIONS

In this paper, Signal to Noise ratio is used for feature ranking. Then different subsets of best ranked features are presented to ELM in different simulations to find the subset of features that contributes to better performance. It is observed that this technique works well for databases with few attributes to very large number of attributes. Especially, when the number of attributes is very large, selection of appropriate combination of attributes which yields substantially good result is a difficult task.

This technique is applied on four databases containing 22 to 10000 attributes. The result shows that the techniques works equally well for large number of attributes.

For this experiment, the number of attributes is selected based on the classification accuracy. Four other performance metrics are also evaluated such as sensitivity, specificity, Jaccard index and M-Estimate and it is observed that best or close to best results are also obtained for these performance metrics with the same set of attributes.

REFERENCES

[1] Theodoridis, S. , Koutroumbas, K., “Pattern Recognition,” Academic Press, San Diego, CA 92101--4495, USA (2008)

[2] Rong, H. J., Ong, Y. S., Tan, A. H. , & Zhu, Z., “A Fast Pruned-Extreme Learning Machine for Classification Problem,” Neurocomputing, vol. 72, no.1, pp. 359-366 (2008).

[3] Dhanalakshmi, P., Palanivel, S., Ramalingam, V., “Classification of Audio Signals Using SVM and RBFNN.” Expert Systems with Applications, vol.36, no.3, pp. 6069-6075 (2009).

[4] Huang, G. B. , Zhou, H. , Ding, X., & Zhang, R., (2012), “Extreme Learning Machine for Regression and Multiclass Classification,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol.42,no.2,pp. 513-529 (2012).

[5] Del Campo, I. , Martínez, V., Orosa, F. , Echanobe, J., Asua, E., & Basterretxea, K. , (2017, May), “Piecewise Multi-linear Fuzzy Extreme Learning Machine for the Implementation of Intelligent Agents,” 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 3363-3370 (2017).

[6] Huang, G. B. , Zhu, Q. Y. , & Siew, C. K. , Extreme Learning Machine,” Theory and Applications. Neurocomputing, vol.70, no.1, pp. 489-501 (2006).

[7] Huang, G. B., Zhu, Q. Y. , & Siew, C. K.., Extreme Learning Machine, “A New Learning Scheme of Feedforward Neural Networks,”, in Proceedings of International Joint Conference on in Neural Networks, IEEE ,vol. 2, pp. 985-990 (2004).

[8] Zhu, Q. Y., Qin, A. K. , Suganthan, P. N. , & Huang, G. B., “Evolutionary Extreme Learning Machine,” Pattern Recognition, vol.38,no.10, pp. 1759-1763 (2005).

[9] Feng, G. , Huang, G. B. ,Lin, Q., & Gay, R., “Error Minimized Extreme Learning Machine With growth of Hidden Nodes and Incremental Learning,” IEEE Transactions on Neural Networks, vol. 20,no.8,pp. 1352-13579 (2009).

[10] Li, , M. B. , Huang, G. B., Saratchandran, P., & Sundararajan, N., ”Fully Complex Extreme Learning Machine,” Neurocomputing,vol.68, pp. 306-314 (2015).

[11] Huang, G. B., Ding, X., & Zhou, H., “Optimization Method Based extreme Learning Machine for Classification,” Neurocomputing, vol.74, no.1, pp. 155-163 (2010).

[12] Huang, G. B. , & Chen, L. , “Enhanced Random Search Based Incremental Extreme Learning Machine,” Neurocomputing, 71(16), pp. 3460-3468 (2008).

[13] Huang, G. B., & Siew, C. K., “Extreme Learning Machine with Randomly Assigned RBF Kernels,” International Journal of Information Technology, vol.11 no.1,pp.16-24 (2005).

[14] Deng, W., Zheng, Q., & Chen, L., ”Regularized Extreme Learning Machine,” in IEEE Symposium on Computational Intelligence and Data Mining, CIDM'09 ,pp. 389-395 (2009).

[15] Mishra, D., & Sahu, B., “Feature Selection for Cancer Classification: a Signal-to-Noise Ratio approach,” International Journal of Scientific & Engineering Research, vol.2,no.4, pp. 1-7 (2011).

[16] K. Lakshmi, & S. Mukherjee, (2006, April), “An Improved Feature selection Using Maximized Signal to Noise Ratio Technique for TC,” in Information Technology, Third International Conference on New Generations,IEEE, pp. 541-546(2006).

[17] S. Haykin, “Neural Networks,” A Comprehensive Foundation, Upper Saddle River, NJ: Prentice Hall, (1994).

[18] B. P. Chacko, V. V. Krishnan, G. Raju, & P. B., Anto, ”Handwritten Character Recognition Using Wavelet Energy and Extreme Learning Machine,” International Journal of Machine Learning and Cybernetics, vol.3,no.2,pp. 149-161 (2012).

[19] P. Langley, “Selection of Relevant Features in Machine Learning,” in Proceedings of the AAAI Fall Symposium on Relevance ,vol. 184, pp. 245-271 (1994).

[20] M. A. Hall, & L. A. Smith, ”Feature Subset Selection,” A Correlation based Filter Approach, (1997).