Asian Journal of Informational Technology 5 (10): 1088-1094, 2006

A Hybrid Feature Selection Method Using Modular Perceptron Networks

Yen-Po Lee, Wei-Yu Han, Wu-Ja Lin and Kuang-Shyr Wu
Department of Computer Science and Information Engineering,
Ching Yun University, Jung-Li, Taiwan

Abstract: In this study, we propose an efficient hybrid feature subset selection method to overcome the curse of dimensionality and to obtain good learning performance on classification problems. The proposed method includes two steps: Scheming an evaluation function to create the rank of the feature significance, constructing a new Binary Search Feature Subset (BSFS) algorithm to generate the optimum feature subset. We have applied the proposed method on a Modular Perceptron Network (MPN) to learn the realworld datasets. It shows that from the experimental results the feature of the input data can be decreased largely (less 75%~88%), the data presentations are reduced (less 67%~91%) and a small size MPNs can be procured with learning and testing performance maintained as the good level as before.

Key words: Modular perception, hybrid feature, selection method

INTRODUCTION

Many factors affect the accuracy and efficiency of pattern recognition/classification tasks. The quality of the data is one such factor. Generally the features of data have variant significance, such as relevant, irrelevant, redundant or noisy. If the information is irrelevant, redundant, noisy or unreliable, then the knowledge discovery during the training is more difficult or has bad performances. Feature selection method is an important process of searching and removing as much of the irrelevant and redundant features as possible, from a larger set of candidate features, ideally necessary and sufficient to perform the Pattern Recognition[1], Classification[2], Clustering[3] and Data Mining[4] problems.

Feature selection is a process that selects a subset of original features. The optimality of a feature subset is measured by an evaluation criterion. A general feature selection process consists of four basic steps, namely, searching subset, evaluating subset, stopping criterion and testing/validating result. Subset searching is a computing procedure that produces candidate feature subsets for evaluation based on a certain computation and search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation criterion. If the new subset turns out to be better, it replaces the previous best subset. The process of subset search and evaluation is repeated until a given stopping criterion is satisfied. Then, the selected optimal subset usually needs to be validated using one of the learning machine.

One popular categorization has grouped different feature selection methods into two broad groups: the filter approach and the wrapper approach[5,6]. In the wrapper approach[7-9], the feature reduction/selection algorithm exists as a wrapper around the learning algorithm, such as neural network[1,7,10] Bayesian classifiers[11], support vector machine[12,13]. The Filter approach[14-17] attempts to assess the merits of feature from the data, ignoring the learning algorithm, such as Distance measure[18], Information theory[19,20], Dependency[16], Consistency[21]. Wrappers generally give better results of performance than filters because of the interaction between the best feature subset search and the learning scheme's inductive bias. But improved performance forks out the cost of computational expense. Recently, many researchers[22-25] attempt to take the advantages of the filters and wrappers approaches and then integrate a new hybrid approach.

A new novel hybrid method is proposed in this study that the features of the input space is reduced and the complexity of the created neural network is simplified applied to MPN+DCL+WE methodology[26]. The feature subset selection method includes two steps:

- scheming an evaluation function to create the rank of the feature significance,

Corresponding Author: Yen-Po Lee, Department of Computer Science and Information Engineering, Ching Yun University, Jung-Li, Taiwan

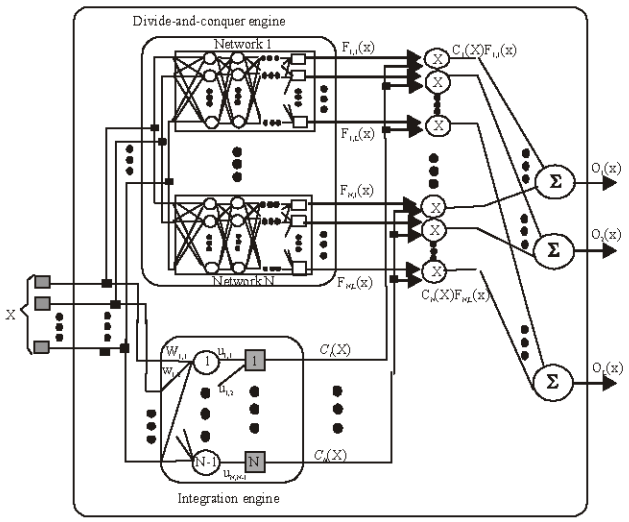

Fig. 1: The architecture of a Modular Perceptron Network: (a) divide-and-conquer learning engine, (b) intergration engine

- constructing a new Binary Search Feature Subset (BSFS) algorithm to generate the optimum feature subset. We have applied the proposed method on a Modular Perceptron Network (MPN) to learn the realworld datasets. It shows that from the experimental results the feature of the input data can be decreased largely (less 75%~88%), the data presentations are reduced (less 67%~ 91%) and a small size MPNs can be procured with learning and testing performance maintained as the good level as before.

*This study was supported by the National Science Council, Republic of China, under contract NSC 93-2213-E-231 -004

Modular perceptron networks: The architecture of the proposed Modular Perceptron Network (MPN) is shown in Fig. 1. As mentioned, the MPN consists of two modules: the Divide-and-Conquer Learning (DCL) Engine and the Integration Engine (IE). The DCL Engine, which consists of a set of self-growing MLP subnets, performs training data partitioning, subnet self-growing and weight learning. The Integration Engine is a self-growing two-layer feed-forward neural network with the Heaviside (hard-limit) activation function at each hidden and output neurons. Acting as a gating network of the MPN, the Integration Engine performs the function of a mediator among the subnets in the DCL Engine. The number of output neurons of the Integration Engine is equal to the number of subnets in the DCL Engine.

We have proposed two learning algorithms: one is the error correlation based Divide-and-Conquer Learning (DCL) scheme[21] and the other is the Weight Estimation method[26,28]. According to the error correlation scheme, the DCL divides a complex training data set into two subsets, one is an easy to learn region, and the other is a hard to learn region. And, then a new MLP is created to learn the hard region, in the meantime the original MLP continues to learn the easy region. The divide and conquer process continues until all the training data are learned successfully. It can be seen that the number of input data subregions and the number of MLP subnets were created by the system itself, instead of being predicted by neural network designers or users. Before the standard error backpropagation learning process is conducted in the subset, Weight Estimation for the subnet is applied first. The Weight Estimation method, which is motivated by Oriented principal component analysis, seeks to guide the initial weight vector toward the desired orientation so that faster weight learning and less subnet creating can be achieved during the construction of the MPNs. DCL with the WE scheme can effectively deal with the slow learning and the unpredictable network size (i.e., the number of hidden units) problems in the design of an MLP-based system.

The evaluation and reduction algorithm: We proposed a fast and efficient feature reduction method. The procedure is to combine the merits of filter approach and wrapper approach. First, we use evaluation function to determine the significance of features, and then the fast selecting subset algorithm provides an candidate feature set , finally, learning algorithm makes sure the performance of optimum subset of features.

The concept of Fisher Discriminate[29] is chosen as a evaluation principle to reduce the features of the dataset. The method is to preprocess the data so as to reduce its features before applying a classification algorithm.

We consider the generalization of the Fisher discriminate to several classes, and we will assume that the features of the input space is greater than the number of classes. The input vector x is projected onto a vector y given by

$$y = W^T x \tag{1}$$

where, W is a projection matrix
The mean vectors of the K'th class are given by

$$m_k = \frac{1}{N_k} \sum_{n \in C_k} x^n \tag{2}$$

The within-class scatter of the transformed data from class C_k is described the within-class covariance given by

$$S_k = \sum_{n \in C_k} (x^n - m_k)(x^n - m_k)^T \tag{3}$$

S_W is the total within-class covariance matrix, given by

$$S_W = \sum_{k=1}^{c} S_k \tag{4}$$

S_B is the between-class covariance matrix and is given by

$$S_B = \sum_{k=1}^{c} N_k (m_k - m)(m_k - m)^T \tag{5}$$

where m is the mean of the total data set

$$m = \frac{1}{N} \sum_{n=1}^{N} x^n = \frac{1}{N} \sum_{k=1}^{c} N_k m_k \tag{6}$$

We can make the dependence on W explicit by using above equations to construct the Fisher discriminate criterion as follows:

$$J(W) = Tr\{(WS_W W^T)^{-1}(WS_B W^T)\} \tag{7}$$

To explain the criterion simply, a two-class problem which has N1 data of class C_1 and N_2 data of class C_2 is considered. From (2) m_1, m_2 are obtained. It might be thought of defining the separation of the classes, when projected onto w, as being the separation of the projected class means. This suggests that w is chosen to be the maximum.

$$m_2 - m_1 = w^T(m_2 - m_1) \tag{8}$$

where S_B is the between-class covariance matrix and is given by

$$S_B = (m_2 - m_1)(m_2 - m_1)^T \tag{9}$$

The resolution proposed by Fisher is to maximize a function which represents the difference between the projected class means, normalized by a measure of the within-class scatter along the direction of w. The within-class scatter of the transformed data from class C_k is described the within-class covariance, given by

$$s_k^2 = \sum_{n \in C_k} \left(y^n - m_k\right)^2 \tag{10}$$

and is the total within-class covariance matrix, given by

$$S_W = \sum_{n \in C_1} (x^n - m_1)(x^n - m_1)^T + \sum_{n \in C_2} (x^n - m_2)(x^n - m_2)^T \tag{11}$$

and the total within-class covariance for the whole data set is defined to be simply $s^2_1 + s^2_2$. It therefore arrives at the Fisher criterion given by

$$J(w) = (m_2 - m_1)^2 / (s_1^2 + s_2^2) \tag{12}$$

The evaluation function is used to determine the significance of each feature. The features are ranked depending on the significance value. According to the order of decreasing progressively, we obtain a new ranking feature set. This is a filter approach, the first part, in we proposed the hybrid procedures (Fig. 2.). To construct a candidate ranking features. From the feature set, we understand the importance of each feature, but we do not know that the boundary of the feature set is optimum. For the reason, we need to construct a fast search algorithm for the good feature selection and for higher performance, the wrapper approach is adopted in my proposed hybrid procedure.

The Modular Perceptron Network (MPN) is used and a Binary Search Feature Subset algorithm (BSFS) is developed.

Given a trained MPN with the set of all features $F = \{f_1, f_2, \ldots\ldots f_N\}$ as its input, the learning and testing performance are computed and treated as a reference. In iterative training process, if the learning/testing performance of the input dataset can keep fitting the system requirements, then the updated feature set has to be evaluated and the weight vector of the MPN is also

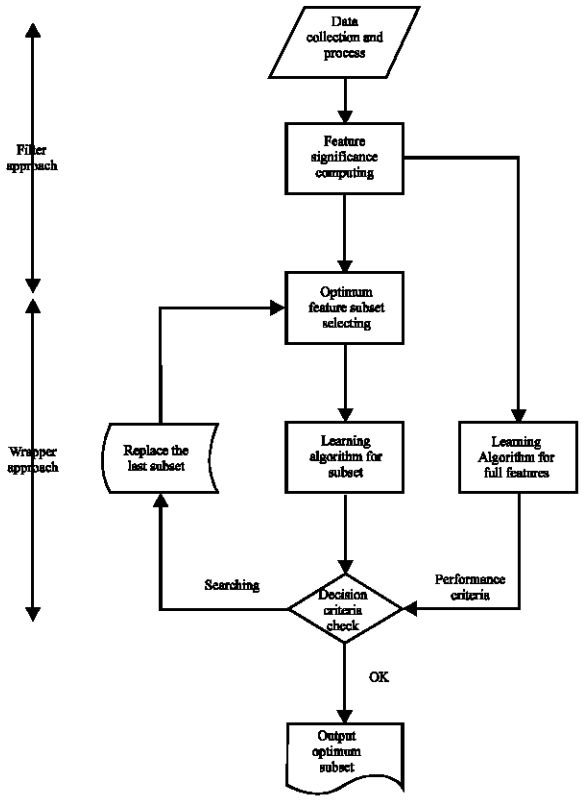

Fig. 2: The efficient hybrid feature subset selection method procedure flowchart

estimated again. For reducing the computational time, we use the binary search concept to construct a feature reduction algorithm named Binary Search Feature Set algorithm (BSFS). The algorithm can decrease the searching optimum feature set time in wrapper approach. The procedures of the algorithm are described as follow:

- Let $F = \{f_1, f_2, \ldots\ldots f_N\}$ be the set of all input features. We separate the dataset into two sets based on normal distribution: the training set and the testing set. Let the? be the allowable maximum drop in performance on the testing set.
- Compute the Significance Index (SI) of the features in training dataset and arrange for full features to a decreasing order sequence array depending on SI.
- Initially, the MPN is trained based on previous section techniques using full features dataset. The $P_{training}$ performance and $P_{testing}$ performance are obtained as a reference of the evaluation procedure.
- all flag(i) are zero (i=1…N) and Left=1; Right=N;
- Adopt binary search concept to find the index in a adequate range.

$$m = \left\lfloor \frac{Left + Right}{2} \right\rfloor$$

- Create a candidate subset: $F'_m = \{f_1, f_2, \ldots\ldots f_m\}$
- The MPN is trained with F'_m. The training performance $P'_{training}$ and $P'_{testing}$ performance are found.
- If the difference of the testing performance $\left|P_{testing} - P'_{testing}\right|$ is larger than Δ

Then flag (m) = -1 and calculating
flag (m)* flag (m+1) = α
If α = -1
Then F'_{m+1} is the optimum feature subset and the performance $P_{testing}$ is accepted Finish the process
Else Left = m + 1
Feedback step 5
End
Else
flag (m) = +1 and
flag (m)* flag (m-1) = α
If α = -1
Then F'_m is the optimum feature subset.
and the performance $P_{testing}$ is accepted
Finish the process
Else
Right = m-1
Feedback step 5
End
End

In the above procedures, we can attention three points:

- We use the concept of Fisher Discriminate to decide the importance in each feature fast.
- We adopt a learning machine (MPN) to generate the higher accuracy performance.
- We construct a Binary Search Feature Set algorithm (BSFS) to decrease the longer computation time of the wrapper approach. The efficient hybrid feature subset selection method can overcome the curse of dimensionality and to obtain a better learning performance on classification problems.

RESULTS AND DISCUSSION

The experiments explore the classification ability of the MPN on the realworld datasets[30]. We select three datasets to proceed the experiments to identify the proposed algorithm. The selected datasets include the Pima Indians. Diabetes dataset(768 instances of eight realvalued features), the Credit Approval dataset

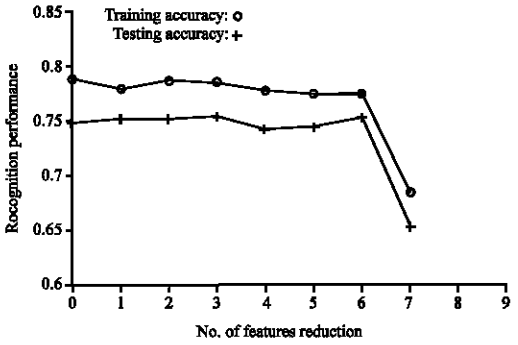

Fig. 3: Recognition performance of the Pima Indians Diabetes dataset vs. number of feature reduction for a simple MLP based MPN

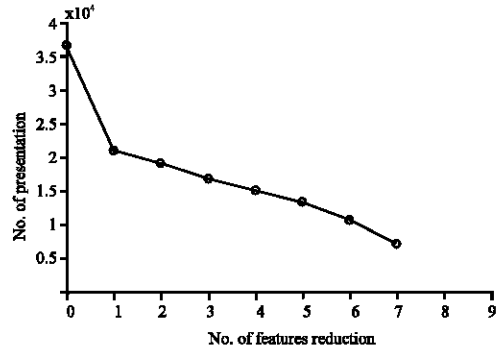

Fig. 4: Number of subnetworks of the Pima Indians Diabetes dataset vs. number of feature reduction for a simple MLP based MPN

Table 1: Performance of feature reduction of different neural networks on the pima indians diabetes dataset[30]. Numbers in ( ) indicates the standard deviation

Types of NN	No. of Features	Training accuracy(%)	Testing accuracy(%)
MPN	8.00(full)	78.75(1.37)	74.78(1.89)
MPN	5.00	78.48(1.55)	75.32(1.78)
MPN	3.00	77.44(2.31)	74.38(2.53)
MPN	2.00	77.41(2.11)	75.27(1.96)
Setiono[7]	2.03(0.18)	74.02(6.00)	74.29(3.25)
NBBFS[10]	4.4		76.03(1.60)
Paetz[31]	3.0		73.11(1.74)

Table 2: Performance of feature reduction of different neural networks on the credit approval dataset[30]. Numbers in ( ) indicates the standard deviation

Types of NN	No. of Features	Training accuracy(%)	Testing accuracy(%)
MPN	15.00(full)	85.81(1.42)	85.37(1.33)
MPN	8.00	85.64(1.28)	85.45(1.40)
MPN	5.00	85.59(1.32)	85.42(1.33)
MPN	3.00	85.60(1.32)	85.43(1.33)
MPN	2.00	85.59(1.49)	85.42(1.55)
ID3[11]	5		80.3
C4.5[11]	5		84.8

(690 instances each has 15 kinds of characteristics) and the Voting Records dataset (435 instances of 16 boolean valued features). All the experiments are done

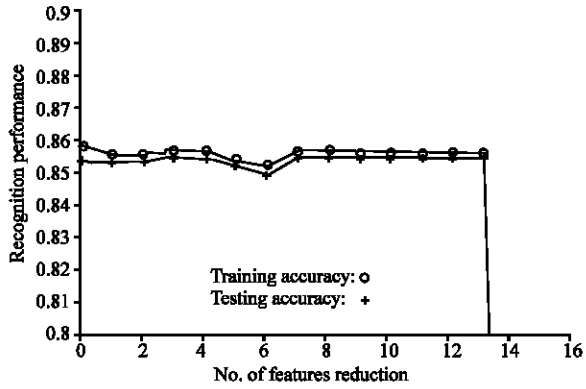

Fig. 5: Recognition performance of the Credit Approval dataset vs. number of feature reduction for a simple MLP based MPN

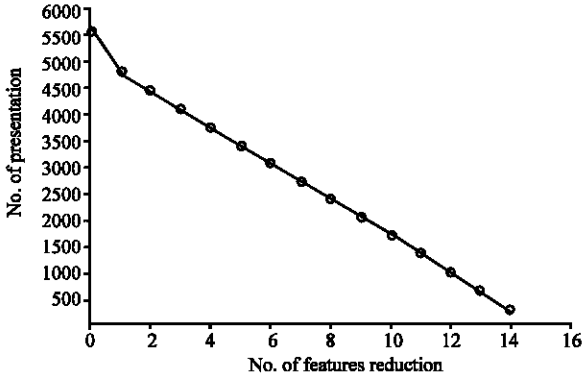

Fig. 6: Number of subnetworks of the credit approval dataset vs. number of feature reduction for a simple MLP based MPN

with a randomly selected partition of the data into half training and half testing data. The results are mean values of 30 repetitions, each with different random testing data.

The pima indians diabetes dataset: The results of this experiment are shown in Fig. 3. The solution is reached that two dimensions are enough on the dataset for classification problems. The performance of two dimensions dataset are compared with that of full dimensions dataset only little difference that can be ignored.

Figure 4, the dimensions are reduced and the number of the subnetworks are increased in the problem. It is concluded that the dataset is a more complex problem. For keeping a good recognition performance, the MPN must increase the number of the subnetworks to make up for the loose of the feature reduction. The number of the presentation also is

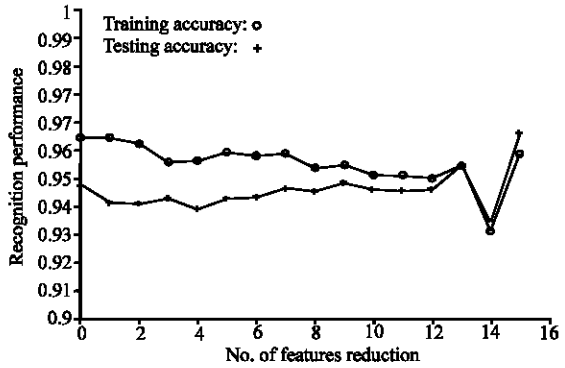

Fig. 7: Recognition performance ofthe Voting Records dataset vs. number of feature reduction for a simple MLP based MPN

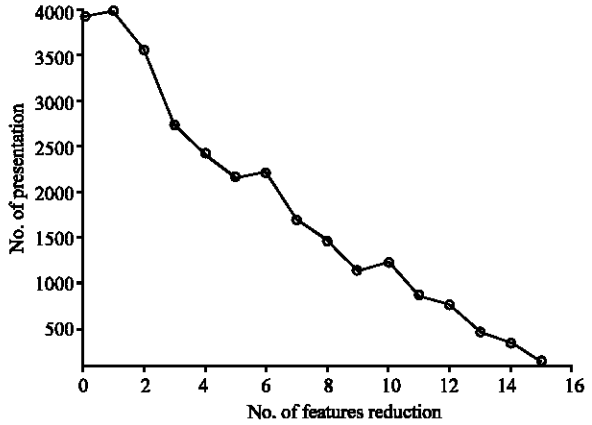

Fig. 8: Number of subnetworks of the Voting Records dataset vs. number of feature reduction for a simple MLP based MPN

decreased (from 3.61 × 104 to 1.23× 104), because the number of the feature of the weight vector is decreased. The learning results are compared with the results of other researches[7,10,31] (Table 1).

The Credit Approval dataset: The results of this experiment are presented in Fig. 5. The performance of two feature dataset is compared with that of full feature dataset only little difference. The feature are reduced but the number of the subnetworks only one in the problem (Fig. 6). For keeping a well recognition performance, the MPN may not increase the number of the subnetworks to make up for the loose of the feature reduction. The number of the presentation also is decreased (from 5,520 to 1,035). The learning results and the results of the other researches[11] are shown in Table 2.

Table 3: Performance of feature reduction of different neural networks on thevoting records dataset[30]. Numbers in ( ) indicates the standard deviation

Types of NN	No. of Features	Training accuracy(%)	Testing accuracy(%)
MPN	16.00(full)	95.85(1.38)	94.33(1.40)
MPN	9.00	95.32(1.41)	94.19(2.07)
MPN	5.00	94.62(1.63)	94.13(2.11)
MPN	3.00	94.92(1.18)	94.97(1.27)
MPN	2.00	92.83(3.32)	93.07(2.73)
MPN	1.00	95.31(1.05)	95.98(1.07)
Setiono[7]	2.03(0.18)	95.63(0.43)	94.79(1.60)
ID3[11]	8		94.7
C4.5[11]	8		94.5

The voting records dataset: The results of the experiments are presented in Fig. 7. The performance of training and testing can keep well until the fifteenth feature is removed. The solution is obtained that one feature is enough on the dataset for classification problem. The dimensions are reduced but the number of the subnetworks only one in the problem (Fig. 8). For keeping a well recognition performance, the MPN may not increase the number of the subnetworks to make up for the loose of the feature reduction. The number of the presentation also is decreased (from 3,900 to 460). The learning results and the results of the other researches[11] are arranged in Table 3

CONCLUSION

In this study, we proposed an efficient hybrid feature subset selection method to overcome the curse of dimensionality and to obtain good learning performance in classification problem. The method keeps the advantages of both the filter approach and wrapper approach, one is the fast computing of filter approach and the other is the high performance of wrapper approach. The ranking of significance of features and Binary Search Feature Subset algorithm are applied to MPN+DCL+WE architecture. The procedure can identify and remove a great deal of the minor, irrelevant and redundant features. The experimental results obtained by learning three realworld datasets show that the performance of training and testing can be maintained as the good level as before. The features of the input space are reduced significantly and the complexity of the divided subnet is simplified.

REFERENCES

1. Shen, L.J., Y.P. Lee, H.C. Fu, 1998. Feature Reduction for Face Recognition by PDBNN, International Computer Symposium Proceedings, pp: 63-66.

2. Dash, M. and H. Liu, 1997. Feature selection for classification, Intelligent Data Analysis.
3. Chris Ding, Xiaofeng He, Hongyuan Zha and D. Horst, 2002. Simon Adaptive feature reduction for clustering high feature data, 2002 IEEE International Conference on Data Mining (ICDM02),Maebashi City, Japan, pp: 147-154.
4. Mark A. Hall and Geoffrey Holmes, 2003. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, IEEE Trans. On Knowledge and Data Engineering.
5. Langley, P., 1994. Selection of relevant features in machine learning, Proceedings of the AAAI Fall Symposium on Relevance, pp: 1-5.
6. Huan Liu and Lei Yu, 2005. Toward Integrating Feature Selection Algorithms for Classification and Clustering IEEE Transactions on Knowledge and Data Engineering, pp: 491-502.
7. Rudy Setiono and Huan Liu, 1997. Neural-Network Feature Selector, IEEE Trans. on Neural Networks.
8. John, G.H., R. Kohavi and K. Pfleger, 1992. Irrelevant features and the subset selection problem, Proceedings of Ninth National Conference on Artificial Intelligence, pp: 129-134.
9. Kohavi, R. and G.H. John, 1997. Wrappers for Feature Subset Selection, Artificial Intelligence, pp: 273-324.
10. Paetz, J., 2002. Feature selection for RBF networks, Proceedings of 9th National Conference on Neural Information Processing, pp: 986-989.
11. Michael J. Pazzani, 1995. Searching for dependencies in Bayesian classifiers, Proceeding of the 5th Int. workshop on Artificial Intelligence and Statistics, pp: 129-134.
12. Burges, C.J.C., 1998. A Tutorial on Support Vector Machines for Pattern Recognition, Knowledge Discovery and Data Mining, pp: 1-43.
13. Vapnik, V., 1995. The Nature of Statistical Learning Theory. New York: Springer.
14. Almuallim, H. and T.G. Dietterich, 1994. Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence J., pp: 279-306.
15. Dash, M., K. Choi, P. Scheuermann and H. Liu, 2002. Feature Selection for Clustering-a Filter Solution, Proc. Second Intl Conf. Data Mining, pp: 115-122.
16. Hall, M.A., 2000. Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, Proc. 17th Intl. Conf. Machine Learning, pp: 359-366.
17. Yu, L. and H. Liu, 2003. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, Proc. 20th Intl Conf. Machine Learning, pp: 856-863.
18. Almuallim, H. and T.G. Dietterich, 1994. Learning Boolean Concepts in the Presence of Many Irrelevant Features, Artificial Intelligence, pp: 279-305.
19. Hanchuan Peng, Fuhui Long and Chris Ding Feature, 2005. Selection Based on Mutual Information; Criteria of Max-Dependency, Max-Relevance and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp: 1226-1238.
20. Ben-Bassat, M., 1982. Pattern Recognition and Reduction of Dimensionality, Handbook of Statistics-II, P.R. Krishnaiah and L.N. Kanal, Eds., North Holland, pp: 773-791.
21. Liu, H. and H. Motoda, 1998. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic.
22. Lee, Y.P. and Hsin-chia Fu, 2004. Dimensionality Reduction Using Modular Perceptron Networks 2004 IEEE International Workshop on Machine Learning for Signal Processing. São Luís Brazil, pp: 223-231.
23. Das S. Filters, 2001. Wrappers and a Boosting-Based Hybrid for Feature Selection, Proc. 18th Intl Conf. Machine Learning, pp: 74-81.
24. Dash, M. and H. Liu, 2000. Feature Selection for Clustering, Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, (PAKDD-2000), pp: 110-121.
25. Xing, E., M. Jordan and R. Karp, 2001. Feature Selection for High-Dimensional Genomic Microarray Data, Proc. 15th Intl Conf. Machine Learning, pp: 601-608.
26. Fu, H.C., Y.P. Lee, C.C. Chiang and H.T. Pao, 2001. Divide-and-Conquer Learning and Modular Perceptron Networks, IEEE Transactions on Neural Networks, pp: 250-263.
27. Chiang, C.C., H.C. Fu, 1994. A divide-and-conquer methodology for modular supervised neural network design, International Conference on Neural Networks, pp: 119-124.
28. Lee, Y.P. and C.H. Fu, 1999. Weight Estimation for the Learning of Modular Perceptron Networks, Proceedings of the 1999 IEEE Workshop on Neural Networks for Signal Processing IX. pp: 103-111.
29. Christopher M. Bishop, 1995. Neural Networks for Pattern Recognition. Published in the United States by Oxford University Press Inc., New York pp: 107-113.
30. Murphy, P.M. and D.W. Aha, UCI repository of machine learning databases, http://www.ics.uci.edu/ mlearn/MLRepository.html.
31. Kohavi, R. and G. H. John, 1997. Irrelevant features and the subset selection problem, Artifical Intelligence journal, special issue,97, pp: 273-324.