

A New Approach to Feature Selection for Data Mining

Dr. B.M. Vidyavathi

*Department of Computer Science,
Sir M. Visvesvaraya Institute of Technology,
Bangalore, Karnataka State, India
E-mail: vidyabml@yahoo.co.in*

Abstract

In recent years many applications of data mining deal with a high-dimensional data (very large number of features) impose a high computational cost as well as the risk of “over fitting”. In these cases, it is common practice to adopt feature selection method to improve the generalization accuracy. Feature selection method has become the focus of research in the area of data mining where there exists a high-dimensional data. In this paper we propose a novel two-phase feature selection approach using both filter and wrapper. It begins by running artificial neural network weight analysis (ANNWA) as a filter approach to remove irrelevant features, then it runs genetic algorithm as a wrapper approach to remove redundant or useless features. We demonstrate the usefulness of the proposed approach on real world datasets from machine learning repository. Our algorithm reduces the size of feature space significantly without compromising the classification or the prediction performance. The experimental results confirm that the method leads to promising improvement on feature selection and classification accuracy.

Keywords: Data mining, two-phase feature selection, Pattern recognition, Genetic algorithm, Neural networks.

Introduction

Developing an accurate classifier for high dimensional datasets is a challenging task due to availability of small sample size. Therefore, it is important to determine a set of relevant features that classify the data well. From the classification point of view it is well known that when the number of samples is much smaller than the number of features, classification methods may lead to over fitting. Moreover high dimensional

data requires inevitably large processing time. So for analyzing high dimensional data, it is necessary to reduce the data dimensionality by selecting a subset of features that are relevant for classification.

Feature Selection is often used as preprocessing technique in machine learning and data mining. It is often effective in reducing dimensionality, improving mining accuracy and enhancing accuracy of the classifier. There are two major approaches to feature selection: filter and wrapper approach [1], [2]. Most filter methods have adopted statistical feature selection, which needs less computation than the others do. It independently measures the importance of features to select good features. Since, the filter approach does not take into account the learning bias introduced by the final learning algorithm, it may not be able to select the most suitable set of features for the learning algorithm. The disadvantage of filter approach is that the features could be correlated among themselves [3], [4]. On the other hand, wrapper methods tend to find features better suited to the predetermined learning algorithm resulting in better performance. But, it also tends to be more computationally expensive since the classifier must be trained for each candidate subset. In literature, several strategies were considered to explore the space of possible subsets. Some of them are evolutionary algorithms used with a k-nearest neighbor classifier [5], parallel genetic algorithms using adaptive operators [6], SVM Wrapper with standard GA [7] and filtered and supported sequential forward search (FS_SFS) in the context of support vector machines (SVM) [10]. The conventional wrapper methods using genetic algorithm have been applied to feature selection of small or middle scale feature datasets [1], [8]. But, it is hard to apply them directly to high dimensional datasets due to much processing time [9]. Reducing the search space for genetic algorithm will decrease the computation time. This can be achieved by selecting a reduced set of important features from high dimensional data without losing any informative feature.

This paper explores a two-phase feature selection method that can take the advantages of both filter and wrapper. In the first phase of the filter approach, the important features are ranked by artificial neural network weights analysis (ANNWA), and most of the irrelevant features will be removed. In the second phase of the wrapper approach, feature importance estimation gotten by ANNWA are adopted as the heuristic information, and genetic algorithm (GA) is used for the wrapper selector, which makes use of induction algorithm as part of its evaluation.

The remainder of this paper is organized as follows. In the next section, the ANNWA approach is described. Section 3 introduces and describes genetic algorithm for feature selection in detail. The results of the experiments are given in sections 4. The conclusions are presented in the final section.

ANNWA as a Filter Approach

The ANNWA [11] ranks the importance of features by relevance based on the weights analysis of a trained multilayer feed-forward network. The reasoning behind this approach is that NN weights can be viewed as representing the gain of the input signal to the output node. Input signals that are noisy or irrelevant to the output will have a high error rate if they have high associated weights. Therefore, training

algorithms must reduce their weights such that they do not contribute to the output. In a similar manner, the weights of relevant and noise-free signals will be increased.

The equation for a three-layer NN with the second layer having a logistic activation function $S(x)=(1/(1+\exp(-x)))$ and the third layer having a linear output is

$$O_k = L_k \times \sum_j S(\sum_i A_i \times W_{ij}) \times W_{jk} \quad (1)$$

where i,j,k are the input, hidden, and output layers node indexes respectively. L is the third layer linear multiplier value; A is the input node (feature); O is the output node; and W is the weight between the layers. The output O_k as a function of a single input A_i can be expressed as

$$O_k = L_k \times \sum_j S(A_i \times W_{ij} + C_{ij}) \times W_{jk} \quad (2)$$

where C_{ij} represents the constant value of all the other inputs, including biases. C_{ij} here acts as a “setpoint” on the logistic function curve. This equation, with all of the inputs processed through the logistic function, is too complex to analyze directly. An approximation can be made of the total relative gain G of a particular input node i to a particular output node j . The approximation substitutes a linear factor for the logistic activation function. The approximation’s error is reduced when the inputs are all in the same range. If we substitute a linear factor F for the activation function, we have

$$O_k \cong L_k \times \sum_j F \times (A_i \times W_{ij} + C_{ij}) \times W_{jk} \quad (3)$$

The local gain LG is defined to be

$$LG_{ik} = \left| \frac{\Delta O_k}{\Delta A_i} \right| \quad (4)$$

Since L_k and F are common factors to ANNWA’s numerator and denominator, they can be dropped in the calculation of LG , i.e.

$$LG_{ik} = \sum_j |W_{ij} \times W_{jk}| \quad (5)$$

The importance ω_i for input feature i is defined to be the function of LG_{ik} as

$$\omega_i = \frac{\sum_k LG_{ik}}{\sum_k LG_{ik}} \quad (6)$$

The importance $\{ \omega_i \}$ for input features will be ranked decreasingly, and then M features will be selected approximately from all N features when

$$\sum_i^M \omega_i \geq \alpha,$$

where α is a threshold and $M \leq N$

Genetic feature selection as a wrapper approach

Although the first phase dramatically reduces the feature number, there are maybe still some redundant features. In this wrapper phase, GA is used for feature selection based on NNs as the induction algorithm. The performance of the trained NN is used to provide a measure of fitness used to guide the GA [12].

Encoding scheme

Each individual in the population represents a candidate solution. In this step, a binary vector of dimension M represents the individual in the population. A value of 0 indicates that the corresponding feature is not selected, and a value of 1 means that the feature is selected.

Initial population

The population of GA is initialized based on the feature importance rank list according to the results of ANNWA. The procedure is as follows:

1. Get the importance rank list of the selected features after the first phase.
2. Generate the selection probabilities of each feature: set the probability to be p_1 ($p_1 > 0.5$) for the feature ranking first and p_2 ($p_2 < 0.5$) for the feature ranking last, and then generate probabilities for the other features according to specified rules. In this work, p_1 and p_2 are set to be 0.8 and 0.4 respectively.
3. Individuals are initialized according to the selection probabilities of each feature obtained in step 2.

Fitness evaluation

The goal of feature selection is to use fewer features to achieve the same or better performance. Therefore, the fitness evaluation contains two terms: (1) the accuracy and (2) the number of features used. Between the accuracy and the size of the feature subset, the accuracy is a more important factor. Combining these two terms, the fitness function is given as:

$$fitness = 10^4 Accuracy + 0.5 \times Zeros \quad (7)$$

where *Accuracy* is the accuracy that an individual achieves, and *Zeros* is the number of zeros in the chromosome. Overall, the higher the accuracy is, the higher the fitness is, and the fewer the number of features used, the higher the fitness is. Randomly divide the dataset into training set D_{train} and testing set D_{test} . For a classification task, the NN is taken as a pattern classifier, and *Accuracy* (CA) is calculated as a percentage of the successfully classified testing cases:

$$CA = \frac{|D_{test_suc}|}{|D_{test}|} \quad (8)$$

where $|D_{suc_test}|$ denotes the number of the successfully classified test cases and $|D_{test}|$ denotes the number of all the test cases. For a regression task, the NN is taken as a regression predictor, and CA can be measured by the mean squared error (MSE) for

the testing set:

$$CA = 1 - \text{MSE} = 1 - \frac{1}{|D_{test}|} \sum_{i \in D_{test}} (y_i - o_i)^2 \quad (9)$$

where y_i and o_i is the predictive and expected output of sample i in D_{test} respectively.

The genetic feature selection algorithm

The genetic feature selection procedure is summarized as follows:

1. Let S_f be the feature subset after the first phase. Let M be the number of features in S_f .
2. Create the initial population (see 3.2) and evaluate the *fitness* of the initial population given by (7).
3. Repeat step 4-5 until stopping criteria ()
4. Apply genetic operators such as selection, crossover and mutation to generate new population of feature subsets.
5. For every chromosome, evaluate the *fitness*, and then rank the population by fitness function.
6. Compute the best chromosome of the population and obtain the selected feature subset S .

Experimental results

In this section we report our experimental results to select features for the above datasets.

Experimental procedure

Experiments are carried out on well-known data sets from UCI Machine Learning Repository. In the experiments the original partition of the datasets into training and test sets is used whenever information about the data split is available.

The experiments are carried out as follows. For each dataset, the first step is to normalize the range of every input features into the interval $[0, 1]$. In the first phase, in order to avoid results of the relevance biased by a concrete neural network, which obviously depends on the initialization weights, we have trained 10 different neural networks (with different initialization weights) for every problem. Furthermore, the number of hidden units of the neural network for each problem was carefully obtained before training the 10 neural networks by a trial and error procedure.

Following the above methodology for each problem, we can obtain a final value of average importance ω_{aveage_i} for input feature by averaging the 10 values of ω_i . On the basis of ordination of feature importance, many redundant features can be removed. After the feature subset is selected in the first phase, the experiments were conducted by the classification or predictive accuracy for testing set to calculate the fitness for reproduction of genetic feature selection. Like the first phase, 5 different neural networks with different initialization weights are used to calculate the average fitness.

The neural networks used in both problems is a standard feed-forward three-layer NN trained with the back-propagation algorithm, starting from initial weights uniformly distributed in $[-1.0, 1.0]$. The GA parameters we used in both problems are as follows: population size: 200, number of generations: 400, crossover rate: 0.66 and mutation rate: 0.04.

The results and discussion

The proposed approach has been implemented in Matlab on Pentium 4 2.66 GHz PC's with 256 MB memory.

To compare the performance for each problem before and after feature selection, 10 networks with different initial weights were trained. Average accuracies of the 10 networks were calculated for the training set and the testing set with and without feature selection. Results are shown in Table 1. From these results, it can be seen that classification rate for testing set of datasets is improved significantly when the feature selection algorithm was used.

Table 1: Results for the datasets with and without the proposed feature selection method

Data set	With all the Features	With selected features
Iris	4	2
Average classification rate	89.31	92.67
Ionosphere	32	5
Average classification rate	85.43	90.56
Arrhythmia	195	8
Average classification rate	91.62	94.88
Multiple Features	649	10
Average classification rate	88.76	93.21

Conclusion

Many datasets include irrelevant and redundant information, which disturbs data mining process. By choosing only relevant features, better classification or prediction accuracy can be achieved. Because of the time complexity, wrapper methods are infeasible. On the other hand, filter approach is not always enough to get good accuracy. In this paper we have proposed a novel approach for feature selection by combining filter and wrapper approaches that use NNs and GA. We have tested the proposed method on real-world datasets. Our experimental results suggest that the approach is effective in eliminating unimportant and redundant features while improving classification or prediction accuracy noticeably. In the proposed approach, most of the irrelevant features are deleted after the first phase of filter approach, and it avoids the exponential computation problem of wrapper approach in the second phase.

References

- [1] Kohavi R., John G., Wrapper for feature subset selection, *Artificial Intelligence*, 97(1-2), pp.273-324, 1997.
- [2] Langley P., Selection of relevant features in machine learning, In *AAAI Fall Symposium on Relevance*, 1994.
- [3] Ding C., Peng HC., Minimum redundancy feature selection from microarray gene expression data, In *IEEE Computer Society Bioinformatics Conf*, pp. 523-528, 2003.
- [4] Jaeger J., Sengupta R., Ruzzo WL., Improved gene selection for classification of microarray, In *PSB*, pp. 53-64. 2003.
- [5] Li L., Weinberg CR., Darden TA., Pedersen LG. Gene Selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method, *Bioinformatics*, 17(12), pp.131-142, 2001.
- [6] Jourdan L., *Meatheuristics for knowledge discovery: Application to genetic data*, PhD thesis, University of Lille, 2003.
- [7] Peng S., Xu Q., Ling XB., Peng X., Du W., Chen L., Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines, *FEBS Letter*, 555(2), pp.358-362, 2003.
- [8] Deb K., Goldberg DE., An investigation of niche and species formation in genetic function optimization, In Schaffer J. D. (Ed) *Proc. 3rd Internat. Conf. Genetic Algorithm*, Morgan Kaufmann, San Mateo, pp. 42-50, 1989.
- [9] Bins J., Draper B., Feature selection from huge feature sets, In *Proc.Internat. Conf. Computer Vision*, 2, pp.159-165, 2001.
- [10] Yi Liu, Yuan F. Zheng ,FS_SFS: A novel feature selection method for support vector machines. *Journal of Pattern Recognition*,39(7) pp 1333-1345,2006.
- [11] Chun-Nan Hsu, et al, "The ANNIGMA-wrapper Approach to Fast Feature Selection for Neural Nets", *IEEE Transactions on Systems, Man and Cybernetics*, April 2002, Part B, vol. 32, Issue 2, pp. 207-212
- [12] Lanzi, P.L, "Fast Feature Selection with Genetic Algorithms: a Filter Approach", *IEEE International Conference on Evolutionary Computation*, April 1997, pp. 537 -540.