A Novel LASSO-Based Feature weighting Selection method for Microarray Data Classification

XiaoLi1, 2， Beiji Zou1, 2, Lei Wang1, 2, Min Zeng1, 2, Kejuan Yue1, 2, Faran Wei1, 2*

1 School of Information Science and Engineering, Central South University, Changsha 410083, People's Republic of China
2 Mobile Health Ministry of Education-China Mobile Joint Laboratory, Central South University, Changsha 410083, People's Republic of China
{lxgrac@163.com, bjzou@csu.edu.cn, wanglei@csu.edu.cn, zengmin1990@163.com, yuekejuan@163.com, franwee@163.com,}

Keywords: Microarray data, Feature Weighting, Feature Selection, Data Balance

Abstract

The biological data, especially microarray data has become more and more important for medical diagnostics. Microarray data usually has high dimension with small sample size and the positive samples are scarce, which makes the data severely imbalanced. In this paper, we propose a new feature selection model for high dimensions and imbalance data. Firstly, by computing the feature-label correlation we remove the low-score features which we consider are irrelevant. Then, a combined feature selection is employed in which the first stage is to remove the redundant features by using a correlation-based feature selection and then we propose a LASSO-based feature weighting approach to increase the weight of the key data. Finally, considering the imbalance problem, we process the selected data to obtain the balanced one by Synthetic Minority Over-sampling Technique (SMOTE), which will be used before classifier training. We use the Ten-fold cross-validationon the datasets. The experimental results on different classifiers show that the proposed method can achieve a higher accuracy and Area Under Curve (AUC) than the traditional feature selection methods.

1 Introduction and related work

The DNA microarray is a new-type and landmark biology technology. Since its birth in the 1990s, the DNA microarray data has received more and more extensive attention in bio-informatics and bio-medical diagnosis, especially in cancer diagnosis [1]. The analyzing and mining of DNA microarray data is of great significance in the field of medical diagnosis, bio-medical research, medicine and pharmacological research [2]. The research on DNA microarray data makes it possible to study the cure mechanism from the molecular level, and conveniences the exploration of new drug targets [6]. However, it poses a difficult challenge for machine learning researchers due to its large number of features, small sample size along with many redundant and noise genes. So, before dealing with microarray data, we first need to do the feature selection to choose the representative data. Because of the imbalance, the standard classifier learning algorithms have a bias toward the classes with a greater number of instances, whereas specific rules that predict examples from the minority class are usually ignored. In these cases, standard classifier learning algorithms are always second to the classes with a greater number of instances, whereas specific rules that predict examples from the minority class are usually treated as noise.

Feature selection is a very important data processing technology in data mining. For the DNA microarray data it usually refers to selecting a subset of genes, that is, selecting the optimal feature subset from the existing features. Using the feature selection we can reduce the dimension of microarray dataset, remove redundant and irrelevant features. By doing this we can improve the accuracy of diagnostics. Therefore, in recent years, the study of feature selection for microarray data have become one of the most popular research direction in biological diseases, especially in the field of cancer classification.

At present, the feature selection methods can mainly be divided into three categories： Filter, Wrapper, and Embedded [11]. The Filter methods are independent of the subsequent learning algorithm, Filtering the high score features as a feature subset. The Filter methods have good generality and can quickly rule out a lot of noise data. However, it neglected the inheritance of features and classifiers. We can't be sure to select an optimal feature subset. The Wrapper method depends on the study and use of the subsequent learning algorithm as the feature selection evaluation criteria. It can narrow down the scope of feature subset, which is good for the identification of key features and can promote the accuracy. Compared with the Filter method, the wrapper method has high computing complexity and bad generalization ability, and maybe excessive fitting learning algorithm. The embedded method aims to deal with the instability observed in many techniques for feature selection when the features of the data change, but it has not been widely used.

Nowadays, how to use effective feature selection algorithm to select appropriate gene subset in microarray dataset has

gradually become a hot research topic in the bio-medical information processing. The literature about feature selection for microarray data is abundant. Lu et al. Lu et al. proposed a new method of gene features extraction in the identification of cancer. He extracted the cancerogenic factors them and built a relative space for the cancer and with these factors, and then extracted gene features for cancer [6]. Ferreira and Figueiredo [10] proposed to combine unsupervised feature discretization and feature selection techniques. By using this method we can improve previous related techniques over several microarray datasets. Yu et al. proposed a feature selection method for gene datasets that combines improved discrete particle swarm optimization with support vector machine [14]. Cho et al. tried to study the relationship between machine learning classifiers and feature selection methods. The author proposed an algorithm which used to select features by penalizing each feature's use in the double formulation of SVM during creation of the classifier [4]. Because of the different searching mechanism and evaluation strategies, the selected significant genes with different approaches are extremely different. There have not come out one feature selection methods can be proved to be the most optimal one for all microarray data [7].

So in this paper, after fully considering the characteristics of microarray data, we propose a new feature selection model for microarray data. By using it we can improve the accuracy of results for microarray data. Firstly, we use the feature selection method that we proposed on the microarray dataset to deal with the dataset. Then we select an optimal feature subset. Because the datasets vary greatly in the number of positive and negative samples, so we are going to use imbalance data processing on them. We use different classifiers to determine the accuracy of the feature selection model we proposed. The experimental results from three different datasets on different classifiers show that the proposed method can achieve higher accuracy and Area Under Curve (AUC) than the traditional feature selection methods.

The rest of the paper is organized as follows: section 2 introduces our feature weighting Selection method, section 3 introduces the Experiments and results, and finally section 4 provides the conclusions.

2 Methodology

There are three main steps in our paper.

Step1: Pre-processing the data, filling all the missing value of the dataset (if the dataset does not have missing value, then go to step2 directly), then we can get the set D1.

Step2: Using the feature selection method LWSS which is shown in **Algorithm 1** to select the optimal feature subset W from the dataset D1.

Step3: Choosing several kinds of classifies to test the datasets W_1 using Ten-fold cross-validation experiment. Fig. 1 shows the flow chart of the proposed algorithms.

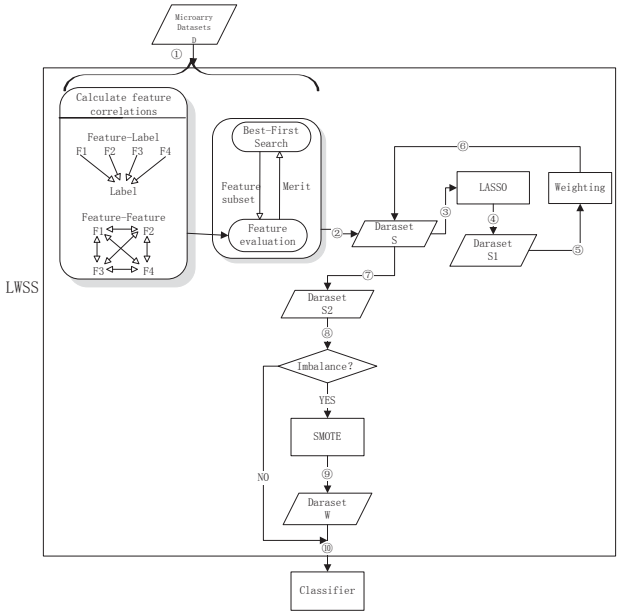

Fig 1: Flow chart of proposed algorithms

Step2 is the main part of this paper. During this step we use the LASSO-Based Feature weighting Selection method with SMOTE, named the method LWSS, to deal with the datasets. The steps of LWSS select feature are introduced as follow:
In our method we first rank the feature subsets according to a correlation-based heuristic evaluation function. The bias of the evaluation function is toward subsets which contain features that are uncorrelated with each other and highly correlated with the label.

For origin dataset D, we use Pearson's linear correlation to calculate the feature-feature inter-correlation and feature-label correlation of the data, which is defined as follows.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad (1)$$

Suppose n is the size of sample, X_i is feature and Y_i is label when compute feature-label correlation, X_i and Y_i are features when compute feature-feature correlation, $\bar{X}$ and $\bar{Y}$ are the mean values. We search the space of feature subsets through the best-first strategy, and then we evaluate the merit of a feature subset S consisting of k features [12]:

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (2)$$

Here, k is the number of features, S_k is a feature subset consisting of k features, $\overline{r_{cf}}$ is the average value of all feature-

label correlations, and $\overline{r_{ff}}$ is the average value of all feature-feature correlations. A feature subset with greater feature-label correlations and less feature-feature correlations will achieve the best "Merit" value.

After the above steps, we can remove the redundant data and noise data, reduce the dimension of the original data, and obtain the feature subset S. We then propose a LASSO-based feature weighting approach for S. Our approach firstly employs LASSO to select the best feature subset from dataset S. LASSO is a selection method based on a linear model and L1-norm constraint, the criterion of LASSO can be defined as [13]:

$$\arg\min_{\beta}\left\{\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2\right\} \quad (3)$$

$$\text{Subject to } \sum_{j=1}^{p}|\beta_j|\le s \quad (4)$$

Here $\beta=\beta_0,\beta_1,\ldots,\beta_p$ is the coefficient vector, n is the size of samples and p is the number of features, x_{ij} is the value for the jth feature of the ith sample, y_i is the observation for the ith sample, s is a constant that is used to control the amount of shrinkage. With the decrease of s value, some coefficients will be set to zero, which means the corresponding features are pruned from the model. These pruned features are irrelevant (or less relevant) to the observations, or just redundant ones, so this method is also used for feature selection. In this paper we use Least Angle Regression (LARS) algorithm to compute the LASSO solutions for all values of s, the best value for s is specified when the cross-validation error reaches the minimum.

By using the LASSO algorithm, we obtain the selected feature subset $S_1$and then assign larger weights to the features in $S_1$and smaller weights to others features. We set the weights of the selected features to 2 and 1 to others. In our experiments, we also observed the performance of our proposed approach using different weights of the selected attributes and found that the weight of 2 is almost the best. Due to the limited space, we have not presented the detailed experimental results here.

As mentioned above, the dataset is imbalanced. The dataset is considered unbalanced when the classification categories are not approximately equally represented. So, before making the classification, we use SMOTE to balance the dataset, which is an over-sampling method and create artificial database on the feature space of the existing minority examples [8]. For a sample x_i belonging to the minority class, we randomly select one of the K-nearest neighbours (K is adjustable), then multiply the corresponding feature vector difference with a random number between [0, 1], and finally add this vector to x_i [9].

$$x_{new}=x_i+(\hat{x}_i-x_i)\times\delta \quad (5)$$

Here, x_i is a existing minority sample, $\hat{x}_i$ is one of the K-nearest neighbors for x_i, $\delta\in[0,1]$ is the random number, x_{new} is the new artificial sample along the line segment between x_i and $\hat{x}_i$. With SMOTE, the original imbalance dataset will include more synthetic samples that represent characters of positive class, which will improve the prediction accuracy of both positive samples and negative samples.

Now, let us give the detailed description of our algorithm as follows.

Algorithm 1: The pseudo-code of the novel LASSO-Based Feature weighting Selection(LBWW)

Input: the original microarray dataset D
Output: the best subset W
For: the original microarray dataset
1. Calculate feature-label and feature-feature correlations
2. Use best-first strategy to search the space of feature subsets, iteratively expand the subset according to the rule of best "Merit" value
3 Subset S=select for the dataset D which have the best Merit
4: Subset S_1 = Select the subset S_1 from S by using LASSO algorithm
5: for each feature w_i from S
    If w_i is the elements of S_1
        Then, set the weight of w_i to 2
    otherwise ,set the weight of w_i to 1
  Obtain the weighted subset S_2
6: if the dataset is unbalance
    Then use the SMOTE algorithm to handle the dataset S_2, get subsetS_3
    Return ,subsets W= S_3
  else
    Return, subset W= S_2

3 Experiments and results

In this section we will present the microarray datasets chosen for testing the method described before. All the classifiers and Filters are executed using the WEKA platform. The classification results of the Microarray dataset selected by LWSS were compared with the results of the feature set selected by Filter，Consistency-based Filter(Cons), LASSO, Logistics. The classifiers of Support Vector Machine (SVM), KNN, Bayesian network(BN), Decision tree (J48) were used to recognize the samples in the experiment. Ten-fold cross validation experiment was performed in these three datasets. Each experiment was run ten times, and the mean of these ten times were calculated. For limited pages, we will not introduce all kinds of classical feature selection method and Filter in detail.

3.1 Dataset

The performance of the novel LASSO-Based Feature weighting Selection method will be tested over DNA microarray data. However, this type of dataset is always huge, multi-dimensional and with small sample size. How to classify the information properly and efficiently is a difficult problem. Three well-known binary microarray datasets (Colon Cancer dataset, CNS dataset, and Leukemia) are used in this paper, listed in Table 1.

Dataset	Attribute	samples	distribution
CNS	7129	60	39:21
Colon Cancer	2000	62	40:22
Leukemia	7129	72	47:25

Table1: The microarray datasets used in this paper

3.2 Result and analysis

Now we present and analyze the experiential results over the three microarray datasets.
(a) The number of selected features;
(b) The classification accuracy of the three datasets use different feature selects methods in different classifiers;
(c) The value of AUC acquired in the experiment.

(a). Numbers of selected features

Table2 shows the feature numbers selected by using LWSS and other feature selected methods on these three datasets. Through table 2 we know that the feature quantity of feature subset is less than the origin dataset, no matter what kind of feature select method used. So we know that any of the feature selection methods can reduce the dimension of data dramatically.

	CNS	Colon Cancer	Leukemia
Full set	7129	2000	7129
LWSS	39	26	81
LASSO	15	12	28
Cons	5	5	3
Filter subset	36	26	53

Table2: The feature number by using LWSS and other feature selected method

(b). Classification accuracy results

In this paper, the classification results on the feature set which are selected by LWSS were compared with the results of LASSO, cons, and Filter in different classifiers. The results of experiment are shown in table 3 to table 5. In the tables, the first row is the classifier we use and the first rank is the feature selection methods we use. The optimal result will be marked in the table.

From table 3 we can see, good result can be obtained by using different classifier in this paper on Leukemia dataset. Especially, on the Leukemia dataset, our LWSS method can reach 100% classification accuracy on the BN and Logistics classifiers.

Table 4 shows that on Colon Cancer dataset, LWSS can get very good result except that SVM is a little worse than Filter. For table 5, on CNS dataset, when using LWSS on J48, BN, Logistics, Filter, we can obtain optimal result, and a little worse on KNN than LASSO method, but far better than cons and Filter method. Filter method gets the highest accuracy (93. 33%) on sym classifier, and LWSS gets the second place by the accuracy of 90. 12%. From the result, not all the classifiers and methods are suitable for all the data. But through table 3 to 5, it can be seen that the presented method, for most of the classifier on the microarray data set can achieve good results, which has advantages for common.

	KNN	SVM	J48	BN	Logistics
LWSS	**97. 91**	**98. 96**	**91. 69**	**100**	**100**
LASSO	88. 33	91. 67	60.00	63. 33	90.00
Cons	93. 06	81. 94	90. 06	94. 44	88. 89
Filter	97. 22	97. 22	83. 33	100	95. 83

Table 3: The classification accuracy of ten-fold cross validation in Leukemia dataset(%)

	KNN	SVM	J48	BN	Logistics
LWSS	**91. 67**	**93. 48**	85. 54	**95. 24**	**94. 04**
LASSO	91.00	92. 25	80. 65	90. 32	85. 48
Cons	88. 71	82. 11	85. 48	85. 48	82. 26
Filter	83. 87	85. 48	**87. 09**	90. 32	75. 81

Table 4: The classification accuracy of ten-fold cross validation in Colon Cancer dataset(%)

	KNN	SVM	J48	BN	Logistics
LWSS	82. 72	90. 12	**85. 19**	**97. 53**	**90.00**
LASSO	**88. 33**	91. 67	60.00	63. 33	90.00
Cons	68.00	65.00	78.00	75.00	75.00
Filter	78. 33	**93. 33**	65.00	86. 67	75.00

Table 5: The classification accuracy of ten-fold cross validation in CNS dataset(%)

(c). The value of AUC

Fig 2, Fig 3 and Fig 4 show the AUC of three microarray datasets under the classifier KNN ,SVM ,J48 ,BN ,Logistics using LWSS, LASSO, CONS, Filter. Through these figures we can see that LWSS can get good results over three datasets, in the case of Leukemia dataset, Using the Bayesian network and Logistics, we can get the AUC value as 1.

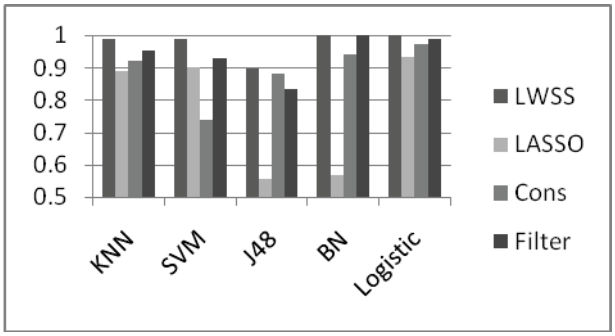

Fig2: The value of AUC by ten-fold cross validation in Leukemia dataset

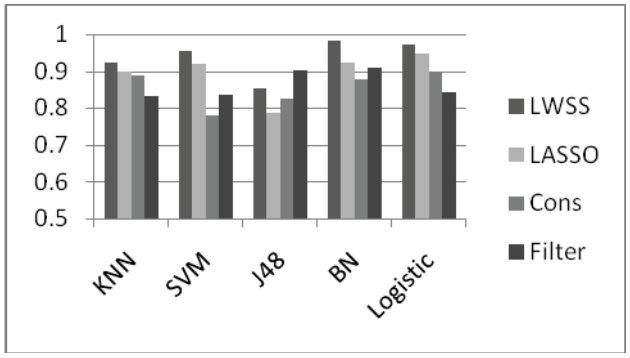

Fig3: The value of AUC by ten-fold cross validation in Colon Cancer dataset

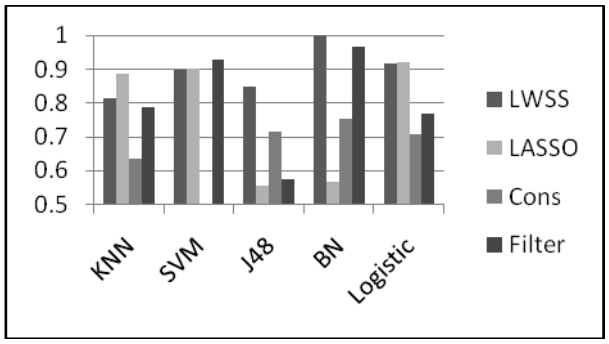

Fig4: The value of AUC by ten-fold cross validation in CNS dataset

4 Conclusions

In this paper, according to the characteristics of microarray data in biomedical data, we proposed LASSO-Based Feature weighting Selection method with SMOTE for Microarray Data. The feature selection algorithm can eliminate the noise and redundant data in the microarray data effectively, and can also well process the imbalance of the data. The method we propose in this paper can gain better AUC value and classification accuracy in comparison with other corresponding approaches.

Acknowledgements

This work is financially supported by the following foundation: Hunan Provincial Natural Science Foundation of China (No.09JJ6102). The Research Foundation of Education Bureau of Hunan Province, China (No.13C143).

References

[1] V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos. “Distributed feature selection: An application to microarray data classification”, *Applied Soft Computing*, **30**, pp. 136-150, (2015).

[2] A. Brazma, A. Robinson, G. Cameron, M. Ashburner. “One-stop shop for microarray data”, *Nature,* **403**, pp. 699-700, (2000).

[3] V. Bolon-Canedo, N. Sanchez-Marono, A.Alonso-Betanzos, J. M. Benitez. “A review of microarray datasets and applied feature selection methods”, *Information Sciences: an International Journal*, **282**, pp. 111-135, (2014).

[4] S. B. Cho, H. H. Won. “Machine Learning in DNA microarray analysis for cancer classification”, *Proc of Bioinformatics 2003 First Asia-Pacific Bioinfomatics Conference(APBC)*, **19,** pp. 189-198, (2003).

[5] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein. “Cluster analysis and display of genome-wide expression patterns”. *Proc.Natl.Acad.Sci USA*, **95**, pp. 14863-14868, (1998).

[6] A.J. Ferreira, M.A.T. Figueiredo. “An unsupervised approach to feature discreti-zation and selection”, *Pattern Recognit*, **45**, pp. 3048–3060, (2012).

[7] F. C. P. Hostege, E. G. Jennings, J. J. Wyrick, et al. “Dissecting the regulatory circuitry of a eukary genome”, *Cell*, **95**, pp. 717-728, (1998)

[8] H. Haibo, A.Edwardo. “Learning from Imbalanced Data”, *IEEE Transactions On Knowledge and Data Engineering*, **21**，pp. 1263-1284, (2009).

[9] E. M. Karabulut, T. Ibrikci. "Effective Automated Prediction of Vertebral Column Pathologies Based on Logistic Model Tree with SMOTE Preprocessing", *Journal of Medical Systems*, **38**, pp. 50-50, (2014).

[10] X.G. Lu, X. G. Peng, D. Li. “Novel method of gene features extraction in cancer recognition”, *Computer Engineering and Application*, **46**, pp. 237-240, (2010).

[11] Y. Saeys, I. Inza, P. Larranaga. “A review of feature selection techniques in bioinformatics”, *Bioinformatics*, **23**, pp. 2507-2517, (2007).

[12] Senliol, Baris. “Fast Correlation Based Filter (FCBF) with a different search strategy”, *International Symposium on Computer and Information Sciences (ISCIS 2008)*, pp. 27-29, (2008).

[13] R. Tibshirani. “Regression shrinkage and selection via the lasso”, *Journal of the Royal Statistical Society*, **58**, pp. 267-288, (1996).