Feature selection algorithm for high dimensional biomedical data classification based on redundant removal

Bingtao Zhang[1,2], Peng Cao[3], Yi Zhang[1], Chaochao Zhang[1], Zhe Li[1], Zhidiao Qu[1], Xiaopeng Wang[2], Tao lei[4], Hanshu Cai[1,*], Bin Hu[1,*]
[1] School of Information Science and Engineering, Lanzhou University, Lanzhou, China
[2] School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, China
[3]School of Architecture and Urban Planning, Lanzhou Jiaotong University, Lanzhou, China
[4]College of Electronical and Information Engineering, Shaanxi University of Science and Technology, 710021, Xi'an, China
Address: School of Information Science and Engineering, Lanzhou University, Tianshui South Road, Lanzhou, Gansu Province, China
E-mail: zhangbt14@lzu.edu.cn, 174601298@qq.com, zhangyi2014@lzu.edu.cn, zhangchch2017@lzu.edu.cn, lizh2014@lzu.edu.cn, quzhd17@lzu.edu.cn, wangxiaopeng@mail.lzjtu.cn leitao@sust.edu.cn，caihsh13@lzu.edu.cn, bh@lzu.edu.cn
** Corresponding author*

High dimensional biomedical data contain thousands of features, and accurate identification of the main features in these data can be used to classification related data. However, it is usually a large number of irrelevant or redundant features seriously influence classification accuracy. To solve this problem, a new feature selection algorithm based on redundant removal is proposed in this study. Firstly, two redundant criteria are determined by vertical relevance and horizontal relevance. Secondly, an approximate redundancy feature framework based on mutual information (MI) is defined to remove redundant and irrelevant features. Finally, to evaluate the effectiveness of our proposed method, contrast experiments based on the classic feature selection algorithm are conducted using (K-nearest neighbour) KNN classifiers, and the results show that our algorithm can effectively improve the classification accuracy.

Feature selection, high dimensional data, KNN, classification accuracy

1. INTRODUCTION

High dimensional data analysis (Tamaresis et al., 2014) is a very hot research area, especially in cancer data (Lee et al., 2013), or mental illness data (Jiang et al., 2017; Li et al., 2017). Usually high dimension data contain many weak relevance or irrelevance features. Hypothesis all the features are treated equally, computational complexity and accuracy of the prediction can be seriously affected. Therefore, feature selection is considered to be an essential procedure in high dimension data processing.

Feature selection (Saeys et al., 2007) refers to selecting relevant features while to remove irrelevant and redundant features. As one of the important part of knowledge discovery technology, feature selection can effectively improve the computing speed of subsequent prediction algorithm, enhance the compactness of the prediction model, increase the generalization ability of the corresponding model. Additionally, the major purpose of high dimensional data feature selection is to overcome the curse of dimensionality (Li et al., 2016; Zhang et al., 2018).

In general, the process of feature selection mainly consists (Mafarja et al., 2018): search strategy and evaluation criterion. Evaluation criterion can be categorized into the wrapper method and the filter method. The wrapper method (Chrysostomou et al., 2017) to evaluate superiority and inferiority of the optimal feature subset under the premise of keeping the classification algorithm unchanged. And the corresponding classification accuracy is adopted as an index to select optimal feature subset. It is necessary to execute the feature selection process again when the classification algorithm is changed. Hence, the complexity is too high, especially for high dimension data. The filter method (Hancer et al., 2018; Lei et al., 2018), the search of feature space depends on the intrinsic correlation of the data itself rather than the

classification algorithm. The filter method is increasingly attractive because of its simplicity and fast speed. So filter method is more general applied than the wrapper method.

According to the above discussion, in this paper, a filter feature selection method is proposed. First of all, four kinds of boundary extremes are analyzed, and then two redundant criterions are proposed. Meanwhile, in order to quantify the redundancy criterion, the core module based on mutual information (MI) (Estevez et al., 2009) is proposed: the definition of approximate redundancy feature. Finally, the experiment is given.

The remainder of this article is organized as follows: Section 2 provides basic concepts related to this research. A feature selection algorithm based on redundancy removal is proposed in Section 3. In Section 4, we describe our experimental design, experimental results. Finally, in section 5 concludes the work of this study.

2. BASIC CONCEPTS

In order to facilitate follow-up research, some basic concepts (John et al., 1994) used in this study are listed as follows.

(i) Strong relevance: Fi is strongly relevant feature iff there exists $P(F_i, A_i) > 0$ such that

$$P(C \mid F_i, A_i) \neq P(C \mid A_i) \quad (1)$$

(ii) Weak relevance: Fi is weakly relevant feature iff it is not strongly relevant (i.e. $P(C \mid F_i, A_i) = P(C \mid A_i)$), there exists $A_i' \subset A_i$ and $P(F_i, A_i') > 0$ such that

$$P(C \mid F_i, A_i') \neq P(C \mid A_i') \quad (2)$$

(iii) Irrelevance: Fi is irrelevant feature iff it are not strongly relevant and weakly relevant, there all $A_i' \subset A_i$ and $P(F_i, A_i') > 0$ such that

$$P(C \mid F_i, A_i') = P(C \mid A_i') \quad (3)$$

where P is a probability measure. F is feature set, $=\{F_1,F_2,\ldots,F_i,\ldots,F_n\}$. F_i: $F_i=\{f_{i,1}, f_{i,2},\ldots, f_{i,n}\}$. A_i: $A_i = F-\{F_i\}$. C is class attribute, $C=\{C_1,C_2,\ldots,C_i,\ldots,C_m\}$.

Strong relevance shows that the feature is very important for classification accuracy, so it can't be arbitrarily removed. Weak relevance indicates that this feature can sometimes contribute to improve prediction accuracy. Irrelevance indicates that this feature is useless on the improvement of classification accuracy, so it can be directly deleted.

3 METHOD

3.1 Redundancy criterion

A redundancy criterion based on the correlation is proposed to lay the foundation for further feature selection. Based on three basic concepts in the section 2, the redundancy of feature F_i is analyzed under four extreme values of $R_{i,c}$ (the relevance between any feature F_i and class attribute C) and $R_{i,j}$ (the relevance between any pair of feature F_i and F_j, $i \neq j$). And four extreme values are shown in table 1.

Table 1: *Four cases of extreme value*

	$R_{i,c}$	$R_{i,j}$
1	*large*	*large*
2	*large*	*small*
3	*small*	*large*
4	*small*	*small*

From table 4, it is easy to draw the following conclusions:

Conclusion 1：$R_{i,c}$ is large, which means that F_i contains more information about C. $R_{i,j}$. is large, which means that the correlation between F_i and F_j is strong. If $R_{i,j}$.=1, then F_i and F_j is complete correlation, hence F_i is redundant. If $R_{i,j}$.≠1, it is difficult to determine the feature F_i whether or not is redundant.

Conclusion 2：$R_{i,j}$ is small, which means that the correlation between F_i and F_j is weak. Hence F_j can't replace F_i. In other words, no matter the size of the $R_{i,c}$, the feature F_i is not redundant.

Conclusion 3：$R_{i,c}$ is small, which means that F_i contains less information about C. $R_{i,j}$ is large, which means that the correlation between F_i and F_j is strong. In this case, the feature F_i is redundant with higher probability. With the increase of $R_{i,j}$, this probability is also increasing.

Conclusion 4：$R_{i,j}$ is small, which means that the correlation between F_i and F_j is weak. This conclusion is consistent with the conclusions 2, no matter the size of the $R_{i,c}$, the feature F_i is not redundant.

Based on the above four conclusions, two redundant criteria can be obtained:

Criteria 1: when $R_{i,j}$ is large, whether F_i is redundant is uncertain.

Criteria 2: when $R_{i,j}$ is small, no matter the size of the $R_{i,c}$, the feature F_i is not redundant.

3.2 Approximate redundancy feature

Assuming that the $R_{i,c}$ of the feature F_i is very close to R_{max} (the maximum value of $R_{i,c}$), it indicates that F_i contains a lot of information about class attribute C. In this condition, only if the value of $R_{i,j}$ is large

enough, F_i can be considered as an approximate redundancy feature. Otherwise, it can't be considered as redundancy feature. The reason is that F_i plays an important role in improving the accuracy of classification, and can't be easily removed as redundancy. By contrast, Assuming that the $R_{i,c}$ of feature F_i is not very close to R_{max}, it indicates that F_i contains relatively less information about class attribute C. In this condition, As long as the value of $R_{i,j}$ is relatively large, F_i is considered as an approximate redundancy feature. The reason is that F_i is not plays a main role in improving the accuracy of classification. In addition to the above conditions, F_i is removed as an approximate irrelevance feature when the difference between $R_{i,c}$ and R_{max} is quite large. Based on the above analysis and discussion, the approximate redundant feature is formally described in definition 1.

Definition 1 (approximate redundancy feature): There is any pair of correlation feature F_i and F_j, and $R_{j,c} \geqslant R_{i,c}$.

(i) F_i is an approximate redundancy feature iff there exists $R\max - R_{j,c} \leq \delta, 0 \leq \delta \leq 0.2$, such that

$$R_{i,j} \geq R\max \tag{4}$$

(ii) F_i is an approximate redundancy feature iff there exists $R\max - R_{j,c} > \delta \,\&\&\, R\max - R_{j,c} \leq \alpha$, $0 \leq \delta \leq 0.2, 0.5 \leq \alpha \leq 0.7$ such that

$$R_{i,j} > (\bar{R} + R_{j,c})/2 \tag{5}$$

Where $\bar{R}$ is a mean value of $R_{i,c}$, that is $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_{i,c}$.

In addition, definition 1 shows that F_j can be approximated as an alternative for F_i.

3.3 Correlation calculation

A nonlinear method based on MI is applied in this paper, and the reason is that the high dimensional data usually exist in the form of nonlinear in the real world. The correlation between any pair of variables (*X, Y*) can be calculated in the following formulas (6) or (7).

$$IG(X;Y) = H(X) + H(Y) - H(X,Y) \tag{6}$$

where *H(X)*, and *H(X,Y)* can be calculated on the basis of formulas (7) and (8).

$$H(X) = -\sum_i P(x_i)\log_2 P(x_i) \tag{7}$$

$$H(X,Y) = -\sum_j P(y_j)\sum_i P(x_i,y_j)\log_2 P(x_i,y_j) \tag{8}$$

To prevent the scale of data is not unified and to reduce the effect of extreme value, each *IG(X;Y)* is normalized to the range [0, 1] using formula (9).

$$R = \frac{2*IG(X;Y)}{H(X)+H(Y)} \tag{9}$$

3.4 Performance evaluation

In this paper, classification accuracy and the number of selected features are two indicators used to design the performance evaluation function (Hu et al., 2016; Chuang et al., 2008), which is shown in formula (10).

$$performance = w_1 * Acc + w_2 * (1 - \frac{n}{N}) \tag{10}$$

w_1 and w_2 are predefined weight coefficients, which are used to adjust the importance of two indicators in the performance evaluation function. In this study, the values of w_1 and w_2 are set to 0.999 and 0.001 respectively. *Acc* is classification accuracy as defined in formula (11). *n* is the number of selected features and *N* is the total number of features.

$$Acc = \frac{C_{num}}{C_{num}+I_{num}} * 100\% \tag{11}$$

C_{num} and I_{num} are the number of correct and incorrect classification features respectively.

4 EXPERIMENTS

4.1 Data description

Five well-known biomedical datasets (Table 2) were used to evaluate the performance of our proposed algorithm. These dataset includes three aspects of disease (cancer) diagnosis, such as gene expression, sera mass-spectrometric etc. The data dimension range was from 2000 to 10000. The first two datasets were taken from the Kent Ridge Biomedical dataset (Li & Liu, 2004), and the last one datasets were taken from the UCI dataset (Asuncion and Newman, 2007).

Table 2**:** *High dimension datasets*

Dataset	Attributes	Instances	Classes
ColonTumor	*2000*	*62*	*2*
DLBCL-Stanford	*4026*	*47*	*2*
Arcene	*10000*	*200*	*2*

4.2 Experimental procedure

We designed and conducted the following experiments: three kinds of high dimensional biomedical data were compared and analyzed by our proposed algorithm, Relief (a filter method based on the nearest neighbor distance) (Kononenko, 2004), maximum relevance and minimum redundancy (mRmR, a filter method based on MI) (Peng et al., 2005), under the same conditions, respectively. In this experiment, the same conditions refers to: Random forest (RF, numTrees=10) (Zhang et al., 2018) were adopted as classifier to evaluate classification accuracy respectively.

10 fold cross validation was adopted to evaluate the classification accuracy. Each data set was stratified into 10 folds, of which 9 folds were used as a training sample and the remaining 1 fold constitute a testing sample. The above experiments were implemented in Matlab 2017a.

4.3 Results

For the three datasets in Section 4.1, we have conducted the experiments described in Section 4.2. Three main statistical indicators were compared and analyzed in Table 3 which are: (1) Mean (%): the mean of performance, (2) Std: the standard deviation, and (3) MeanFN: the mean number of selected feature.

Table 3: *Comparison of experimental results based on different data sets*

Dataset	Algorithm	Mean (%)	Std	MeanFN
ColonTumor	*Full Set*	*78.21*	*3.51*	*2000*
	Relief	*85.49*	*3.16*	*38*
	mRmR	*86.61*	*3.85*	*42*
	Our proposed algorithm	***92.01***	***1.97***	***34***
DLBCL-Stanford	*Full Set*	*76.60*	*2.87*	*4026*
	Relief	*75.48*	*3.34*	*45*
	mRmR	*76.58*	*6.42*	*54*
	Our proposed algorithm	***82.99***	***2.33***	***29***
Arcene	*Full Set*	*73.14*	*4.87*	*10000*
	Relief	*81.01*	***2.06***	*73*
	mRmR	*78.80*	*4.82*	*71*
	Our proposed algorithm	***85.67***	*2.11*	***51***
Boldface *shows the best experimental result.*				

From Table 3, we can observe the following aspects: (1) our algorithm obtained the best Mean among all the three feature selection algorithms. The best Mean(s) are 92.01%, 82.99%, and 85.67%, respectively. In addition, we notice that the maximum Mean improvement of our proposed algorithm was 13.80% compared with the full set. (2) The Std among all the three feature selection algorithms for two out of three experimental results obtained by our proposed algorithm is smaller than other two algorithms. (3) The three feature selection algorithms can effectively reduce the feature dimension, and the dimensionality reduction of our proposed algorithm was the most obvious. In addition, Figure 1 was obtained by statistical analysis of table 3. Figure 1 shows that one average attribute values (avg(Mean)). From comparison results, we can observe that our proposed algorithm were superior to the other two algorithms.

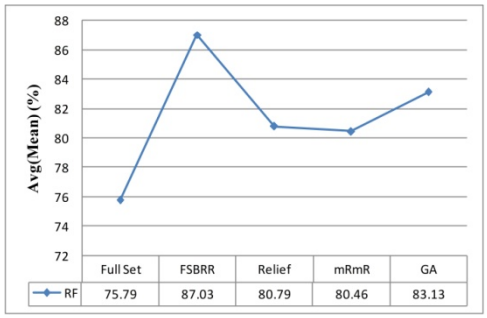

Figure 1: *The average attributes value* (*avg*(*Mean*))

5 CONCLUSIONS

In this study, the relationship between two kinds of correlation (the correlation between features and classes, and the correlation between features and features) is established to eliminate redundant features. Due to the determination of completely redundant features is difficult to realize, therefore we analyze four kinds of boundary conditions between $R_{i,c}$ and $R_{i,j}$, and then a redundancy feature criteria is proposed. On this basis, the approximate redundancy features are defined in this study. Finally, we have proposed a new feature selection algorithm based on redundancy removal for high dimensional data classification

ACKNOWLEDGMENT

This work was supported by the National Basic Research Program of China (973 Program) [2014CB744600]; the National Natural Science Foundation of China [61632014, 61210010, 61461025, 61871259, 61811530325,]; Program of Beijing Municipal Science & Technolog y Commission [Z171100000117005]; the Program of International S&T Cooperation of M OST [2013DFA11140]; the Yong Scholar Fund of Lanzhou Jiaotong University [2016004] and the Teaching and Reform Project of Lanzhou Jiaotong University [JGY201841].

REFERENCES

Asuncion, A., and Newman, D. J. (2007) UCI machine learning repository. Online available: http://archive.ics.uci.edu/ml/.

Chrysostomou, K., Chen, S. Y., and Liu, X. (2017) Combining multiple classifiers for wrapper feature selection. International Journal of Data Mining Modelling & Management, 1, 91-102.

Chuang, Y. L., Chang, H. W., Tu, C. J. and Yang, C. H. (2008) Improved binary PSO for feature selection using gene expression data. Computational Biology and Chemistry, 32, 29-38.

Estevez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. (2009) Normalized Mutual Information Feature Selection. IEEE Transactions on Neural Networks, 20, 189-201.

Hancer, E., Xue, B., and Zhang, M. (2018) Differential evolution for filter feature selection based on information theory and feature ranking. Knowledge-Based Systems, 140, 103-119.

Hu, B., Dai, Y., Su, Y., Zhang, X. W., Mao, C. S., Chen, J., and Xu, L. X. (2016) Feature Selection for Optimized High-dimensional Biomedical Data using the Improved Shuffed Frog Leaping Algorithm. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 99, 1-10.

Jiang, H. H., Hu, B., Liu, Z. Y., Yan. L. H., Wang, T. Y., Liu, F., Kang, H. Y., and Li, X. Y. (2017) Investigation of different speech types and emotions for detecting depression using different classifiers. Speech Communication, 90, 39-46.

John, G. H., Kohavi, R., and Pfleger, K. (1994) Irrelevant Features and the Subset Selection Problem. Machine Learning Proceedings, 1994, 121-129.

Kononenko, I. (1994) Estimating attributes: analysis and extensions of RELIEF. European Conference on Machine Learning on Machine Learning. 23, 171-182.

Lee. K., Man, Z., Wang, D., and Cao, Z. (2013) Classification of bioinformatics dataset using finite impulse response extreme learning machine for cancer diagnosis. Neural Computing and Applications, 22, 457-468.

Lei, T., Jia, X., Zhang, Y., He, L. F., Meng, H. Y., and Nandi A. K., (2018) Significantly Fast and Robust Fuzzy C-Means Clustering Algorithm Based on Morphological Reconstruction and Membership Filtering. IEEE Transactions on Fuzzy Systems, 27, 1-15.

Li, J., and Liu, H. (2004) Kent Ridge Biomedical Data Set Repository. School of Computer Engineering, Nanyang Technological University, Singapore, Online available: http://datam.i2r.astar.edu.sg/datasets/krbd/index.html.

Li, X. W., Zhuang, J., Hu, B., Zhu, J., Zhong, N., Li, M., Ding, Z. J., Yang, J., Zhang, L., Feng, L., and Majoe, D. (2017) A Resting-State Brain Functional Network Study in MDD Based on Minimum Spanning Tree Analysis and the Hierarchical Clustering. Complexity, 22, 1-11.

Li, X., Hu, B., Sun, S., and Cai, H.S. (2016) EEG-based mild depressive detection using feature selection methods and classifiers. Computer Methods and Programs in Biomedicine, 36, 151-161.

Mafarja, M. M., and Mirjalili, S. (2018) Whale Optimization Approaches for Wrapper Feature Selection. Applied Soft Computing, 62, 441-453.

Peng, H., Long, F., and Ding, C. (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell, 27, 1226-1238.

Saeys, Y., Inza, I., and Larrañaga, P. (2007) A review of feature selection techniques in bioinformatics. Bioinformatics, 23, 2507-2517.

Tamaresis, J. S., Irwin, J. C., Goldfien, G. A. Rabban, J. T., Burney, R. O., Nezhat, C., DePaolo, L. V., and Giudice, L. C. (2014) Molecular classification of endometriosis and disease stage using high-dimensional genomic data. Endocrinology, 155, 4986-4999.

Zhang, B. T., Lei, T., Liu, H., and Cai, H.S. (2018) EEG-based automatic sleep staging using ontology and weighting feature analysis. Computational and Mathematical Methods in Medicine, 2018, 1-28.

Zhang, B. T., Wang, X. P., Wang, L. C., Zhang, Z. L., Li, Y. L., and Liu, H. (2018) Intrusion Detection Method for MANET Based on Graph Theory. Journal of Electronics and Information Technology, 40, 1446-1452.