A Feature Selection Method Based on Feature Grouping and Genetic Algorithm

Xiaohui Lin(✉), Xiaomei Wang, Niyi Xiao, Xin Huang, and Jue Wang

School of Computer Science & Technology,
Dalian University of Technology, 116024 Dalian, China
datas@dlut.edu.cn

Abstract. Feature selection technique has shown its power in analyzing the high dimensional data and building the efficient learning models. This study proposes a feature selection method based on feature grouping and genetic algorithm (FS-FGGA) to get a discriminative feature subset and reduce the irrelevant and redundancy data. Firstly, it eliminates the irrelevant features using the symmetrical uncertainty between features and class labels. Then, it groups the features by Approximate Markov blanket. Finally, genetic algorithm is applied to search the optimal feature subset from the different groups. Experiments on the eight public datasets demonstrate the effectiveness and superiority of FS-FGGA in comparison with SVM-RFE and ECBGS in most cases.

Keywords: Feature selection · Symmetrical uncertainty · Feature grouping · Genetic algorithm

1 Introduction

As the quick development of genomic, proteomics and metabolomics techniques, they have been widely applied in the study of pathology, diagnostics and prognosis. Since the bioinformatic data are often high dimensional and contain noise and redundant variables, finding the interested features to get an efficient classification model is becoming very important. Many feature selection methods, such as Support Vector Machine-Recursive Feature Elimination (SVM-RFE) [1], Random Forests (RF) [2], Genetic Algorithm (GA) [3], Relief-F [4], and Mutual Information (MI) [5, 6], have been applied to select the meaningful feature subset from the high dimensional data to induce a classification model with a high performance [7, 8].

SVM [9] is a supervised machine learning technique. It is suitable to analyze the high dimensional data [10]. Originally, SVM was proposed for binary problems. And it could solve the multi-class problems by means of "one-versus-all" and "one-versus-one" methods[11], etc. SVM-RFE [12] is a popular feature selection approach based on SVM. It calculates the weights of the features according to the SVM learning model and removes the features with the smallest weights iteratively. GA is a stochastic global search technique [13] and has got a promising performance. Many feature selection techniques have been proposed based on GA [14, 15].

X. He et al. (Eds.): IScIDE 2015, Part II, LNCS 9243, pp. 150–158, 2015.
DOI: 10.1007/978-3-319-23862-3_15

To filter out noise and redundant data simultaneously, several techniques have been proposed, such as min-redundancy and max-relevance (mRMR) [16], a method combining SVM-RFE and correlation coefficient [17], a method where SVM-RFE and mRMR work together [18], and a dynamic weighting-based feature selection algorithm [19].

To select the meaningful feature subset from the high dimensional data, this paper proposes a new feature selection method based on feature grouping and GA (FS-FGGA). It removes the irrelevant data which has small relevance with the class label, groups the features, and applies GA to search the optimal combination feature subset from different feature groups. The applications on eight public data verify the effectiveness of FS-FGGA.

2 Methods

To improve the performance of the learning model, FS-FGGA selects the meaningful non-redundant features from the original data. It eliminates the irrelevant features by symmetrical uncertainty [20, 21] and groups the features according to the relevance among the features. The features lying in the same group have the similar information related to the class label. Hence each group contributes one feature to the final feature subset. But selecting different features from each group may induce different learning models which may have different classification performance. GA is adopted to search the optimal combination feature subset. Fig. 1 shows the main framework of FS-FGGA.

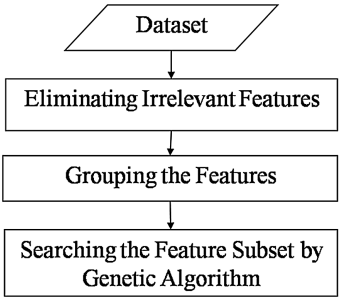

Fig. 1. Framework of FS-FGGA.

2.1 Symmetrical Uncertainty

Symmetrical Uncertainty (SU) [20, 21] is an effective technique to measure the correlation of two random variables. Let X and Y be two variables, their correlation $SU(X, Y)$ is defined as follows:

$$SU(X,Y) = 2 \cdot \frac{IG(X|Y)}{H(X) + H(Y)}. \quad (1)$$

$H(X)$ is the entropy of X, $IG(X|Y)$ is the information gain which reflects additional information about X provided by Y.

Let $F = \{f_1, f_2, \ldots, f_n\}$ denote the feature set, C denote the class label set. In order to filter out the irrelevant features, FS-FGGA adopts symmetrical uncertainty, $SU(f_i, C)$ ($1 \le i \le n$), to measure the relation between feature $f_i \in F$ and the class label C. If $SU(f_i, C)$ is low enough, i.e. it is lower than a threshold σ, feature f_i has little relevance with the class label, and is removed from the data [20, 21].

2.2 Grouping Features

Fast Correlation-Based Filter (FCBF) [21] is an efficient feature selection technique. It analyzes the relevance by symmetrical uncertainty, and removes the redundant data by means of Approximate Markov blanket (AMB). For two different features $f_i \in F$ and $f_j \in F$ ($1 \le i \ne j \le n$), f_i is an Approximate Markov blanket [21] of f_j, if and only if

$$SU(f_i, C) \ge SU(f_j, C) \ \ and \ \ SU(f_i, f_j) \ge SU(f_j, C). \quad (2)$$

FS-FGGA groups the features according to AMB. The features which are relevant to each other by FCBF [21] are put into the same group.

2.3 Searching the Optimal Feature Subset by GA

FCBF produces a feature subset which is formed by picking the center feature of the group [21]. But the center may be different as the training samples change [22]. Ensemble correlation-based gene selection (ECBGS) [23] method uses the different starting points and selects the best feature subset according to the corresponding classification performance.

Let $FG = \{FG_1, FG_2, \ldots, FG_k\}$ denote the feature group set. Since the features in the same group contain the similar information, only one feature is picked up from each group to constitute the selected feature subset. Further the combination of different features from different groups may have different classification performance. Hence FS-FGGA applies GA to search the optimal one. Initially, FS-FGGA randomly selects a feature from each group to form a feature subset as an individual and repeats this operation to get the initial population of GA. The flow chart of searching the optional feature subset is shown in Fig. 2.

The fitness of an individual in a population is assessed by the classification accuracy rate of SVM. Roulette wheel selection is adopted to select the parents from the population. A single-point crossover operation and a single-point mutation are also applied for the offspring individuals.

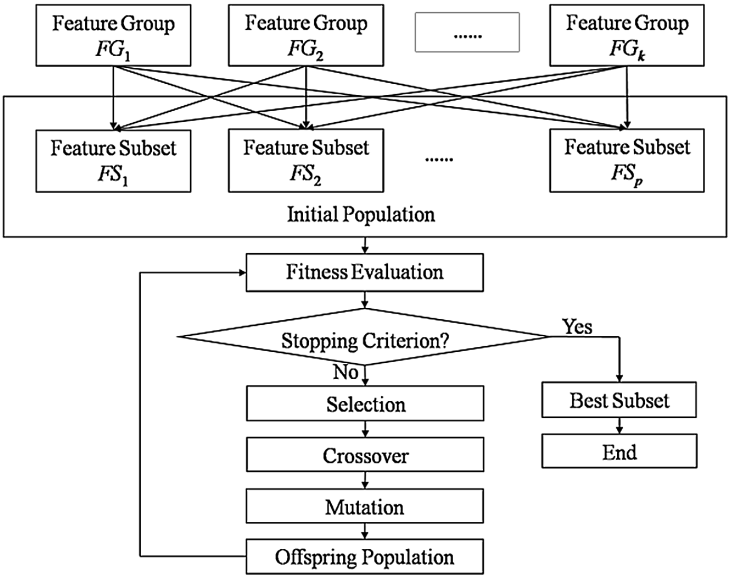

Fig. 2. Flow chart of searching the best feature subset.

3 Results and Discussion

3.1 Performance Metrics

Features selection technique aims at selecting a feature subset having the high classification ability. Meanwhile, the stability of the method is also very important. This study applied the classification accuracy and stability to evaluate the performance of the methods. The percentage of overlapping features related (POFR) [24] is used to measure the method stability. It is defined as follows [24]:

$$POFR_{F_1F_2} = \frac{|F_1 \cap F_2| + |R_{F_1F_2}|}{|F_1|}. \tag{3}$$

$$POFR_{F_2F_1} = \frac{|F_1 \cap F_2| + |R_{F_2F_1}|}{|F_2|}. \tag{4}$$

where F_1 and F_2 are two different feature subsets selected by the different running of a algorithm, $|F_1|$ (or $|F_2|$) is the number of the features in F_1 (or F_2), $R_{F_1F_2}$(or $R_{F_2F_1}$) is the set of the features in F_1 (or F_2) which are not in F_2 (or F_1) but have a strong correlation with at least one feature in F_2 (or F_1). The greater its value is, the more stable the feature selection algorithm is.

3.2 Experiment

To demonstrate the effectiveness of FS-FGGA, it is compared with SVM-RFE and ECBGS on eight public microarray datasets, which are gene expression data from various human cancers. Table 1 shows the basic information of the eight public datasets. Among them, Adenocarcinoma, Leukemia 2, Lymphoma 1 and Srbct datasets are from http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html, and the other four datasets come from http://linus.nci.nih.gov/~brb/DataArchive_New.html.

Table 1. The basic information of the eight public datasets

Datasets	Feature number	Sample number	Class number	Sample number of every class
Adenocarcinoma	9,868	76	2	64:12
Leukemia 1	7,129	72	2	47:25
Leukemia 2	3,051	38	2	27:11
Leukemia MLL	12,582	72	3	24:28:20
Lymphoma 1	4,026	62	3	42:9:11
Lymphoma 2	4,026	96	2	62:34
Prostate	12,600	102	2	50:52
Srbct	2,308	63	4	23:20:12:8

Auto scaling is used to reduce the differences of the magnitude of different features. To calculate SU, equal width discretization (EWD) [25, 26] is adopted, where the real data is divided into h (h is set to 3 in the experiments) intervals with equal width between the minimum value and the maximum value.

Parameter σ for FS-FGGA and ECBGS is set as follows:

$$\sigma = 0.5 * (SU_{\max} - SU_{\min}). \tag{5}$$

where $SU_{\max}$ and $SU_{\min}$ are the maximal and the minimal relevant values of the features with the class label, respectively.

For FS-FGGA, the maximal number of iterations and the size of population are set to 50 and 100, respectively. The crossover probability and mutation rate are set to 0.8 and 0.01, respectively. When the generation is up to the maximal number of iterations or the best fitness comes to 0.95, the GA search procedure stops.

Ten-fold cross-validation was run ten times. SVM is adopted as the classification method, and the RBF kernel function and the LINEAR kernel function are used respectively. The source code of SVM is from http://www.csie.ntu.edu.Tw/~cjlin/libsvm and the other algorithms were written in C++.

3.3 Results and Discussion

Tables 2 and 3 show the comparison of the three methods on the average classification accuracy rates. The bold face means the largest accuracy rate among the three methods

in a data set. The last row (W/T/L) of the two tables count the number of wins/ties/losses compared to the FS-FGGA over all data sets. It can be seen that FS-FGGA is superior to the other two feature selection methods in most cases.

In comparison with SVM-RFE, FS-FGGA ties with SVM-RFE on RBF kernel function (Table 2), but it shows a clear superiority over SVM-RFE on the LINEAR kernel function (Table 3), where FS-FGGA wins seven times to SVM-RFE. With LINEAR kernel function, the average classification accuracy rate of SVM-RFE is equal to that of FS-FGGA only on the Adenocarcinoma data, but the standard deviation of SVM-RFE is 1.55% higher than that of FS-FGGA.

In comparison with ECBGS, the average classification accuracy rates of FS-FGGA are higher than those of ECBGS on all the eight datasets with RBF kernel function (Table 2). While in Table 3, using the LINEAR kernel function, FS-FGGA wins ECBGS seven times. Only on the Leukemia 1 data, the average classification accuracy rate of ECBGS is higher than that of FS-FGGA a little.

Tables 4 and 5 show the average *POFR* of the three feature selection algorithms. From the two tables, FS-FGGA algorithm is more stable than the other two algorithms in the majority cases.

Table 2. The comparison on SVM with RBF kernel function

Datasets	SVM-RFE(%)	ECBGS(%)	FS-FGGA(%)
Adenocarcinoma	**82.37±3.11**	79.47±2.86	81.05±2.92
Leukemia 1	**94.44±2.45**	93.75±1.99	94.03±1.47
Leukemia 2	97.37±2.77	95.26±3.23	**98.16±2.50**
Leukemia MLL	93.61±2.55	92.50±2.19	**94.58±2.01**
Lymphoma 1	96.77±1.32	98.06±1.67	**99.84±0.51**
Lymphoma 2	**91.56±1.51**	90.52±1.04	91.35±1.56
Prostate	90.98±1.84	90.49±1.47	**91.18±1.46**
Srbct	**97.94±0.77**	94.60±3.76	95.40±1.58
W/T/L	4/0/4	0/0/8	-

Table 3. The comparison on SVM with LINEAR kernel function

Datasets	SVM-RFE(%)	ECBGS(%)	FS-FGGA(%)
Adenocarcinoma	**79.47±4.12**	76.84±2.99	**79.47±2.57**
Leukemia 1	91.94±2.91	**92.64±2.78**	92.36±1.76
Leukemia 2	95.79±3.96	96.58±2.50	**97.37±2.15**
Leukemia MLL	87.64±2.31	93.61±2.47	**95.28±2.09**
Lymphoma 1	93.87±2.82	97.90±1.87	**98.87±0.78**
Lymphoma 2	88.75±2.24	86.77±2.60	**89.38±2.90**
Prostate	88.73±2.70	89.61±1.74	**90.78±1.97**
Srbct	92.70±3.01	95.56±2.97	**96.83±1.06**
W/T/L	0/1/7	1/0/7	-

Table 4. The average *POFR* using RBF kernel function

Datasets	SVM-RFE	ECBGS	FS-FGGA
Adenocarcinoma	**0.6311**	0.3274	0.3590
Leukemia 1	**0.8320**	0.7409	0.8243
Leukemia 2	0.7316	0.6967	**0.9454**
Leukemia MLL	0.8419	0.8457	**0.8971**
Lymphoma 1	0.7360	0.7945	**0.9797**
Lymphoma 2	**0.7117**	0.5905	0.5653
Prostate	0.7792	0.7864	**0.8112**
Srbct	**0.9497**	0.6260	0.7466

Table 5. The average *POFR* using LINEAR kernel function

Datasets	SVM-RFE	ECBGS	FS-FGGA
Adenocarcinoma	**0.4904**	0.3131	0.3384
Leukemia 1	0.5501	0.7090	**0.8139**
Leukemia 2	0.6748	0.7101	**0.9457**
Leukemia MLL	0.6312	0.8296	**0.8931**
Lymphoma 1	0.5012	0.8274	**0.9797**
Lymphoma 2	0.4176	0.5667	**0.5757**
Prostate	0.4272	0.7514	**0.8051**
Srbct	**0.7769**	0.6034	0.7501

4 Conclusions

This paper proposes a new feature selection method based on feature group and genetic algorithm (FS-FGGA). The method can effectively eliminate the irrelevant features and reduce the redundant features. Applications on eight public microarray data show the effectiveness of FS-FGGA. It can select more discriminative feature subsets to build more efficient classification models than SVM-RFE and ECBGS in most cases.

Acknowledgments. The study has been supported by the State Key Science & Technology Project for Infectious Diseases (2012ZX10002011), the Sino-German Center for Research Promotion (GZ 753), National Natural Science Foundation of China (21375011).

References

1. Tang, Y.C., Zhang, Y.Q., Huang, Z.: Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. **4**, 365–381 (2007)
2. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
3. Holland, J.H.: Adaptation in Natural and Artificial Systems. MIT Press, Cambridge, MA (1992)

4. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. Mach. Learn. **784**, 171–182 (1994)
5. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
6. Xu, J.C., Xu, T.H., Sun, L.: An efficient gene selection technique based on fuzzy C-means and neighborhood rough set. Appl. Math. Inf. Sci. **8**, 3101–3110 (2014)
7. Yassi, M., Moattar, M.H.: Robust and stable feature selection by integrating ranking methods and wrapper technique in genetic data classification. Biochem. Biophys. Res. Commun. **446**, 850–856 (2014)
8. Liu, X.M., Tang, J.S.: Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method. IEEE Syst. J. **8**, 910–920 (2014)
9. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
10. Shen, L., Tan, E.C.: Dimension reduction based penalized logistic regression for cancer classification using micro-array data. IEEE/ACM Trans. Comput. Biol. Bioinform. **2**, 166–175 (2005)
11. Zhou, X., Tuck, D.P.: MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. Bioinformatics **23**, 1106–1114 (2007)
12. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**, 389–422 (2002)
13. Arunachalam, J., Kanagasabai, V., Gautham, N.: Protein structure prediction using mutually orthogonal Latin squares and a genetic algorithm. Biochem. Biophys. Res. Commun. **342**, 424–433 (2006)
14. Ram, R., Chetty, M.: A Markov-Blanked-Based model for gene regulatory network inference. IEEE-ACM Trans. Comput. Biol. Bioinform. **8**, 353–367 (2011)
15. Abbasnia, R., Shayanfar, M., Khodam, A.: Reliability-based design optimization of structural systems using a hybrid genetic algorithm. Struct. Eng. Mech. **52**, 1099–1120 (2014)
16. Maji, P., Garai, P.: On fuzzy-rough attribute selection: criteria of max-dependency, max-relevance, min-redundancy, and max-significance. Applied Soft Computing. **13**, 3968–3980 (2013)
17. Xie, Z.X., Hu, Q.H., Yu, D.R.: Improved feature selection algorithm based on SVM and correlation. Adv. Neyral Netw. **3971**, 1373–1380 (2006)
18. Mundra, P.A., Rajapakse, M.J.: SVM-RFE with mRMR filter for gene selection. IEEE transactions on nano bioscience. **9**(1), 31–37 (2010)
19. Sun, X., Liu, Y.H., Xu, M.T., Chen, H.L., Han, J.W., Wang, K.H.: Feature selection using dynamic weights for classification. Knowl.-Based Syst. **37**, 541–549 (2013)
20. Shen, L.L., Zhu, Z.X., Jia, S.: Discriminative Gabor feature selection for hyper spectral image classification. IEEE Geosci. Remote Sens. Lett. **10**, 29–33 (2013)
21. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. **5**, 1205–1224 (2004)
22. Liu, H.W., Liu, L., Zhang, H.J.: Ensemble gene selection by grouping for microarray data classification. J. Biomed. Inform. **43**, 81–87 (2010)
23. Piao, Y.J., Piao, M.H., Park, K.J., Ryu, K.H.: An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. Bioinformatics **28**, 3306–3315 (2012)
24. Zhang, M., Zhang, L., Zou, J.F., Yan, C., Xiao, H., Liu, Q.: Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. Bioinformatics **25**, 1662–1668 (2009)

25. Bennasar, M., Setchi, R., Hicks, Y.: Unsupervised discretization method based on adjustable intervals. In: 16th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, vol. 243, pp. 79–87, San Sebastian (2012)
26. Orhan, U., Hekim, M., Ozer, M.: Epileptic seizure detection using artificial neural network and a new feature extraction approach based on equal width discretization. J. Fac. Eng. Archit. Gazi Univ. **26**, 575–580 (2011)