Journal of Biomedical Informatics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Classifier ensemble selection based on affinity propagation clustering

Jun Meng [a], Han Hao [a], Yushi Luan [b,*]

[a] School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116023, China
[b] School of Life Science and Biotechnology, Dalian University of Technology, Dalian, Liaoning 116023, China

ARTICLE INFO

Article history:
Received 5 August 2015
Revised 11 February 2016
Accepted 16 February 2016
Available online xxxx

Keywords:
Classification
Ranking aggregation
Affinity propagation clustering
Kappa correlation
Ensemble feature selection

ABSTRACT

A small number of features are significantly correlated with classification in high-dimensional data. An ensemble feature selection method based on cluster grouping is proposed in this paper. Classification-related features are chosen using a ranking aggregation technique. These features are divided into unrelated groups by an affinity propagation clustering algorithm with a *bicor* correlation coefficient. Some diversity and distinguishing feature subsets are constructed by randomly selecting a feature from each group and are used to train base classifiers. Finally, some base classifiers that have better classification performance are selected using a kappa coefficient and integrated using a majority voting strategy. The experimental results based on five gene expression datasets show that the proposed method has low classification error rates, stable classification performance and strong scalability in terms of sensitivity, specificity, accuracy and G-Mean criteria.

1. Introduction

Ensemble feature selection has the goal of finding a set of feature subsets that will promote disagreement among the component members of the ensemble. Bagging [1] and AdaBoost [2] are two powerful ensemble techniques. Numerous empirical studies have shown that they almost always produce better classifiers than do their base predictors. Ho [3] proposed an ensemble selection method based on the random subspace method (RSM) derived from random partitioning and stochastic discrimination (SD) theory. Ensemble feature selection based on genetic algorithm (GA) is focused on the ensemble perspective rather than that of traditional feature selection – finding one appropriate set for learning [4]. A state-of-the-art random forest (RF)-based feature selection method [5] has been found to perform well in high-dimensional data. In recent years, the RSM and Bagging have been used to construct the feature subspace to generate a plurality of different base classifiers based on the generalized additive model (GAM), such as the GAMbag, GAMrsm and GAMens ensemble learning methods [6]. A novel feature selection method based on the normalization of the well-known mutual information measurement is presented to ensure that the features inside the feature subset with large difference also have little relevance [7]. This normalization ensures that redundancy is eliminated.

The ensemble feature selection method improves the performance of the learning algorithm by producing diverse classifier sets. It forms an effective ensemble learning method that is suitable for classification problems with high-dimensional data. Moon et al. [8] proposed an ensemble classification method that used a random partition method to divide the feature space to train base classifiers for high-dimensional data. A simple, high-performance and easy-to-implement group ensemble gene selection method (EGSG) based on fast correlation-based filter (FCBF) was proposed [9]. This method uses the approximate Markov blanket for gene grouping so that genes in the same group are mutually correlated. A gene subset is chosen randomly from the top t genes, which are closely associated with class labels, to ensure the superiority of the gene subset. The ensemble classifier that was trained by the feature subspace obtained a higher classification accuracy in the cancer dataset.

In our previous work [10], we proposed a clustering method combined with GO-term semantic similarity. An affinity propagation clustering algorithm was chosen to analyze the impact of the biological similarity on the results. Based on the clustering results, a neighborhood rough set was applied to select representative genes for each cluster [11].

Experiments showed that this method ensures the feature diversity of subsets and enhances the distinguishability, which improves the classification ability of the ensemble learning algorithm. Many proposed ensemble feature selection methods mainly search the whole feature space to construct a feature subset, but the sample classification was only associated with a few features

* Corresponding author.
E-mail address: luanyush@dlut.edu.cn (Y. Luan).

http://dx.doi.org/10.1016/j.jbi.2016.02.010
1532-0464/

in high-dimensional data. We propose the affinity-propagation-based classifier ensemble selection (APCES) method to address this problem. This method uses ranking aggregation technology [12] to filter the features and selects those with a strong influence on sample classification. The features are grouped by affinity propagation clustering (AP) [13] with a *bicor* correlation coefficient [14], and the feature subsets are generated by randomly selecting a feature from each group to ensure a considerable difference between any two feature subsets, as well as to ensure that the features within the subset are uncorrelated. Finally, the kappa correlation coefficient is used to select the base classifiers, and the support vector machine (SVM), which performs well on a diverse set of datasets [15], is used to train base classifiers. The proposed method has the advantage of classifying and effectively improving the performance of the ensemble learning. The related datasets and codes for our APCES method are available on the supporting website (https://github.com/Garyapple/apces.git).

2. Methodology

2.1. Ranking aggregation technology

Many feature filtering methods are regarded as useful in sorting problems. Feature ranking is described as follows: dataset $D = (X, Y)$, $X = \{x_{ij} \mid i = 1,2,\ldots,N;\ j = 1,2,\ldots,M\}$ is the sample observation value, where N is the number of samples, M is the number of features, and $Y = (y_1, y_2, \ldots, y_N)$ is the set of class labels. A scoring function $S(x)$ is defined to measure the differences of feature space $F = (f_1, f_2, \ldots, f_N)$ in different sample groups. Then, the statistical significance (SS) is calculated by the estimated value and is sorted to obtain feature ranking $R = (r_1, r_2, \ldots, r_M)$, where r_i $(1 \leqslant i \leqslant M)$ is the position serial number of feature f_i in the ranking. An ordered feature set $L = (l_1, l_2, \ldots, l_M)$ is obtained by sorting R, l_i $(1 \leqslant i \leqslant M)$ represents the feature serial number of position i, $l_p = q \Leftrightarrow r_q = p$ $(p, q \in [1, M])$, and the top K features are selected as a feature subset. This method is usually simple, fast and easy to implement. Therefore, it is widely used to analyze all kinds of high-dimensional data.

Although the feature ranking method can obtain satisfactory results in most cases, the feature selection results may be unsatisfied in the case of a slight perturbation of the dataset. Discrepancies in the results among different methods on the same dataset might arise. To a certain extent, ranking aggregation technology [12,16,18] solves the problem using ensemble learning methods. It performs a ranking of multiple features and integrates the ranking results to select feature subsets. This method can effectively improve the stability of feature selection.

According to the different feature ranking methods, aggregation technology can be divided into two categories [16]: ranking criteria and data perturbation. The ranking criteria method uses several different ranking methods to rank the features in the same dataset. The ranking results are then aggregated to obtain optimal results. For five ranking methods – eBayes ($R^{(e)}$), Fold-Change ($R^{(f)}$), SAM ($R^{(S)}$), maxT ($R^{(m)}$) and Welch T-test ($R^{(w)}$) – each method has a feature ranking in dataset D. We generate feature ranking aggregation observed values $\overline{R} = (\overline{r_1}, \overline{r_2}, \ldots, \overline{r_M})$ by the mean aggregation (MA) method, where $\overline{r_j} = \left(r_j^{(e)} + r_j^{(f)} + r_j^{(s)} + r_j^{(m)} + r_j^{(w)}\right)/5$. Then, an optimized ordered list $L = (l_1, l_2, \ldots, l_M)$ is obtained by sorting $\overline{R}$, and the top K features are chosen as the final feature subset. The ranking results of this method are associated with multiple ranking methods and have different results with different combination methods.

The data perturbation method repeatedly uses the bootstrap or sub-sampling disturbance in the original dataset to obtain multiple disturbance datasets, and then ranks the features based on a specific ranking method. An optimized ranking is formed based on the results of the aggregation. The five ordered tables, $R^{(1)}$, $R^{(2)}$, $R^{(3)}$, $R^{(4)}$ and $R^{(5)}$, represent the results of five disturbances on the dataset by the Welch T-test method. A feature ranking aggregation observation, $\overline{R} = (\overline{r_1}, \overline{r_2}, \ldots, \overline{r_M})$, where $\overline{r_j} = \left(r_j^{(1)} + r_j^{(2)} + r_j^{(3)} + r_j^{(4)} + r_j^{(5)}\right)/5$, is obtained by the simple average aggregation method. Finally, $\overline{R}$ is sorted to optimize the table of ordered features $L = (l_1, l_2, \ldots, l_M)$, and the top K features are selected as the feature subset. The ranking result of this method is often associated with the selected method and perturbation frequency, and a stable ranking is obtained when the dataset has a small change in disturbance and fewer disturbances.

The feature ranking aggregation method mainly includes Mean, Median, Quantile, Markov Chain Model and Robust Rank, although other features are occasionally used. Wald et al. [12] and Slawski [16] used the aggregation experiments on 11 datasets with five classification algorithms. Their results show that the average aggregation method is simple and effective and has relatively low computational cost, which is suitable for high-dimensional data.

2.2. Affinity propagation cluster algorithm

Affinity propagation (AP) is a clustering algorithm proposed by Frey and Dueck [13]. It takes all of the data points as potential class-represented points (Exemplar), selects a representative point set by transferring and updating information, and finally moves each data point to the nearest representative data point. Compared with the traditional K-means and K-center [17] methods, the AP algorithm has three advantages [19]: (1) the number of classes is automatically decided by the algorithm, (2) it produces a more stable and accurate clustering result, and (3) it needs less time to generate the same clustering accuracy.

The AP algorithm is based on a similarity matrix. The distance between the data points, such as the negative Euclidean distance, is used to construct the similarity matrix $S_{N\times N}$. The matrix can be symmetrical or asymmetrical. The bias parameter P is the diagonal value ($S(k,k)$) of the matrix and determines whether the corresponding data point k is a representative point or not. The greater the value of P, the higher the probability that point k is a representative point. Usually, the P values of all of the data points are set to the same value, meaning that all of the data points have the same chance of being the representative point. The P value determines the number of the clusters produced by the algorithm, and the number of clusters is much greater when P is large.

The AP algorithm transfers two important values: responsibility (R) and availability (A). Responsibility $r(i,k)$ represents the degree of point x_k as the representative point of point x_i, and availability $a(i,k)$ represents the appropriateness of point x_i; point x_k is selected as the representative point. The iterative formula is as follows:

$$r(i,k) = s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\} \tag{1}$$

$$a(i,k) = \begin{cases} \min\left\{0, r(k,k) + \sum\limits_{i's.t\ \ i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\}, & i \neq k \\ \sum\limits_{i's.t\ \ i' \notin \{k\}} \max\{0, r(i',k)\}, & i = k \end{cases} \tag{2}$$

In the iterative process of the algorithm, oscillation and non-convergence occur when two or more points are suitable as representative points in the same class cluster. In this case, the damping