Knowledge-Based Systems 89 (2015) 301–313

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Multi-kernel multi-criteria optimization classifier with fuzzification and penalty factors for predicting biological activity

Zhiwang Zhang [a,*], Guangxia Gao [b], Yingjie Tian [c]

[a] School of Information and Electrical Engineering, Ludong University, Yantai 264025, China
[b] Shandong Institute of Business and Technology, Yantai 264005, China
[c] Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 14 April 2015
Received in revised form 12 July 2015
Accepted 13 July 2015
Available online 17 July 2015

Keywords:
Multi-kernel learning
Multi-criteria optimization
Fuzzification
Class-imbalanced learning
Classification
Bioassay

ABSTRACT

Nowadays it is important to develop effective computational methods for accurately identifying and predicting biological activity in the virtual screening of bioassay data so as to speed up the process of drug development. Among these methods, multi-criteria optimization classifier (MCOC) is a classifier which can find a trade-off between the overlapping degree of different classes and the total distance from input points to the decision hyperplane. The former should be minimized while the latter should be maximized. Then a decision function is derived from training data and this function is subsequently used to predict the class label of an unseen instance. However, due to outliers, anomalies, highly imbalanced classes, high dimension, nonlinear separability and other uncertainties in data, MCOC and other methods often give the poor predictive performance. In this paper, we introduce a new fuzzy contribution to each input point based on class median, by defining the new row and column kernel functions the linear combination of different feature kernels to replace the single kernel function in the kernel-induced feature space and penalty factors to imbalanced classes, thus a novel multi-kernel multi-criteria optimization classifier with fuzzification and penalty factors (MK–MCOC–FP) is proposed and the effects of the aforementioned problems are significantly reduced. The experimental results of predicting active compounds in the virtual screening and comparison with linear and quadratic MCOCs, support vector machines (SVM), fuzzy SVM and neural network, the conclusions show that MK–MCOC–FP evidently increased the ability of resisting noise interference, the predictive accuracy of highly class-imbalanced bioassay data, the separation of active compounds and inactive compounds, the interpretability of importance or contributions of different features to classification, the efficiency of classification with feature selection or dimensionality reduction for high-dimensional data, and the generalization of predicting the biological activity of new compounds.

1. Introduction

We know that in the process of drug discovery and development a large amount of time and efforts are devoted to the primary-screening and the confirmatory-screening for extracting the relevant compounds from the bioassays. For high-throughput screening (HTS) a large number of compounds are screened against a biological target to test whether the compound is capable of binding to the target, if the compound binds then it is an active for that target otherwise it is an inactive one. In order to increase the efficiency and accuracy of HTS, virtual screening (VS) can be employed by using some effective computational methods [43]. These methods mainly include protein-based approaches, ligand-based approaches and some data mining approaches [36,44,22]. However, owing to the challenges of the growing complexity of bioassay data, for instance, the higher dimensionality, the highly imbalanced class and the massive data, these methods often give the poor predictive performance, such as the high false positive rates, the low classification accuracy, the poor generalization of predicting the bioactivity of new compounds, the weak interpretability and the overfitting majority of inactive compounds. Recently, more and more data mining and machine learning techniques are used to aid the prediction and selection of active compounds in VS, for example, naïve Bayes classifier, support vector machine (SVM), decision tree, random forest, and so on.

SVM based on optimization method and statistical learning theory is recently applied to prediction of bioactivity [43]. At the same time, the SVM classifier has been extensively utilized to a variety of applications because of its better generalization than

* Corresponding author. Tel.: +86 15053532980.
E-mail address: zzwmis@163.com (Z. Zhang).

http://dx.doi.org/10.1016/j.knosys.2015.07.011
0950-7051/

some traditional data mining methods [49,9,32,42,12]. The main idea of the SVM classifier is to partition instances into different classes by finding a separating hyperplane that maximizes the margin between two supporting hyperplanes of two classes and minimizes the misclassification simultaneously. For the linearly separable case, the separating hyperplane is built in the input space. For the nonlinearly separable case, kernel techniques are used to map input points from the input space into a feature space, and the separating hyperplane is positioned in the new feature space.

However, in many real-world applications SVM is very sensitive to noise, outliers and anomalies in data set so that the separating hyperplane severely deviates from the right position and direction [5,6,51]. Thus several methods have been proposed to solve the problem by introducing a proper fuzzy membership function to the SVM model [45,31,48,41,25,47,18,23,3,14]. Additionally, in real world applications, it is very common that one class is more important than others, and the class distribution is imbalanced, which results in a rapidly degenerating classification precision and accuracy for instances from the minority class. In order to deal effectively with the class imbalance problem, some penalty techniques based on cost-sensitive learning and other methods are used [46,52,53,27,38,10].

Multi-criteria optimization classifier (MCOC) is another optimization-based method which can be used to solve the classification problems in data mining. The main idea of MCOC is to find a trade-off between the overlapping degree of input points belonging to different classes and the total distance from input points to their decision hyperplane. The first criterion should be minimized while the second criterion should be maximized. Then a linear MCOC model based on the comprise solution was proposed and applied to the credit card portfolio management and risk analysis [39,40]. A multiple phase fuzzy linear MCOC approach was provided and used for the behavior analysis of credit cardholders [20]. Subsequently, a linearly penalized MCOC was presented for improving the catch rate of the bad credit cardholders [29]. A quadratic MCOC in data mining was given and used to carry out credit analysis [28]. In order to increase accuracy and efficiency of classification, an appropriate fuzzy membership function is introduced to MCOC, that is to say, the objective functions and the constraints are transformed into the corresponding fuzzy decision sets, then the new MCOC with fuzzy parameters is constructed so as to improve the generalization in the unseen data [55]. For the nonlinearly separable case, a kernel-based MCOC was given just like the use of the kernel trick in the SVM classifier [54].

Besides, in many practical applications, owing to noise, outliers and anomalies, the highly imbalanced class distribution, and the nonlinear separability within a data set, MCOC will degenerate into an inefficient, instable, and inaccurate classifier when it is trained and tested with these data. To improve the predictive performance of MCOC, a novel MCOC using fuzzification, kernel and penalty factors was proposed and used for predicting protein–protein interaction hot spots, where a class mean-based fuzzy membership function was introduced to MCOC and associated with each input point in the input space, a kernel function was used to map input points into a high-dimensional feature space, and the unequal penalty factors are added to the input points of imbalanced classes [56]. At the same time, multi-criteria optimization classifier with kernel, fuzzification and penalty factors was put forward for credit risk evaluation. Different from the kernel trick in the former classifier, in this classifier a kernel function was firstly introduced, and then the fuzzy membership degree of each input point was calculated in kernel-induced feature space [57]. In a word, the results of the real world applications have shown that the two types of classifiers remarkably increased the predictive performance of classification. Therefore, the MCOC models and algorithms are gradually developing as a new alternative method for solving classification, regression and other problems in data mining and machine learning.

Although the improved MCOC has ability to discover noise in data, deal effectively with the nonlinearly separable case and overcome the case of overfitting majority-class instances for class imbalance, the interpretability of MCOC is weakened due to introduction of the single kernel function. Recently, in practice we also found that MCOC is still subjected to the effects of outliers and anomalies, and the classifier lacks of ability to efficiently process the highly dimensional data and obtain the better interpretability by means of feature selection or dimensionality reduction.

To this end, we reformulate the MCOC model and redesign the corresponding algorithm by introducing a new fuzzy membership function to each input point where class mean is not used but class median, substituting a linear combination of different feature kernels for the single kernel function, and integrating unequal penalty factors into highly imbalanced classes. Thus, the proposed MK–MCOC–FP approach can improved the overall performance of the original MCOC in stability, efficiency, separability and interpretability, that is to say, the aforementioned effects of outliers, anomalies, high dimension, class imbalance and nonlinearly separable problems can be reduced significantly.

Therefore, the main contribution of this paper can be summarized as follows: firstly, by using new fuzzification method based on class median the effects of noise and anomalies are removed to the greatest extent such that the stability and flexibility of MCOC are improved considerably. On the other hand, by defining new row and column kernels of different features in high-dimensional data and employing the new sparsification method based on multi-kernel learning the important features are selected and used for classification such that the interpretability and efficiency of MCOC are enhanced remarkably. In addition, by applying the cost-sensitive penalty factors to the highly class-imbalanced data the proposed MK–MCOC–FP method is used to identify active compounds in bioassay and achieves the better predictive performance than that of other classifiers as shown in the experimental results and comparative analysis.

The rest of this paper is organized as follows: Section 2 describes basic principles of MCOC. Then the new MK–MCOC–FP approach and the corresponding algorithm are illustrated in Section 3. The experimental results of predicting biological activity and comparison analysis are demonstrated in Section 4. Finally, discussions and conclusion will be given in Sections 5 and 6 respectively.

2. MCOCs

In this section the MCOCs are presented in detail. For a two-class classification problem, suppose a training set $T_1 = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ is given, each input point $\boldsymbol{x}_i (\boldsymbol{x}_i \in R^d)$ belongs to either of two classes with a label $y_i \in \{-1, 1\}$, d is the dimensionality of the input space, and n is the sample size. In order to separate different classes some researchers have chosen the two measures: the overlapping degree deviating from the separating hyperplane and the total distance departing from the decision hyperplane [15]. In the first case an input point is in the wrong side of the hyperplane and misclassified, while in the second case it is in the right side and correctly classified. Subsequently, Glover considered the above two factors in the classification models simultaneously [16].

Let $\alpha_i (\alpha_i \geqslant 0)$ be the distance where an input point $\boldsymbol{x}_i$ deviates from the separating hyperplane, and the sum (called as the overlapping degree) of the distance α_i should be minimized. We have

$$\begin{aligned} &\text{minimize } \sum_{i=1}^{n} \alpha_i \\ &\text{subject to } y_i(\boldsymbol{w}^T \boldsymbol{x}_i - b) \geqslant -\alpha_i, \quad \alpha_i \geqslant 0, \ i = 1, \ldots, n. \end{aligned} \tag{1}$$