Extensible Multi-criteria Optimization Classifier for Prediction of Chinese Semantic Word-formation Patterns

Zhiwang Zhang\textsuperscript{a,*}, Xinseng Li\textsuperscript{b}, Hailin Zou\textsuperscript{a}, Shiyong Kang\textsuperscript{c}, Guangxia Gao\textsuperscript{d}

\textsuperscript{a}School of Information and Electrical Engineering, Ludong University, Yantai 264025, China
\textsuperscript{b}Ningbo Institute of Technology, Zhejiang University, Ningbo 315100, China
\textsuperscript{c}Shandong Key Laboratory of Language Resource Development and Applications, Yantai 264025, China
\textsuperscript{d}Shandong Institute of Business and Technology, Yantai 264005, China

Abstract

Data mining has become crucial in modern science and industry. Data mining problems raise interesting challenges for different research domains. For the supervised learning methods, owing to the class-overlapping, inconsistent, non-compatible, and contradictory problems in real world applications, the predictive performance of multi-criteria optimization classifier (MCOC) and other traditional data mining approaches will rapidly degenerate. In this paper we put forward an novel extensible MCOC (EMCOC) based on the dependent function: firstly the matter-element models of input data and test data are built, then the weighted dependent degree of data are computed according to the definition of the dependent function, and EMCOC based on the matter-element models is built for predicting the patterns of Chinese semantic word-formation. Our experimental results and comparison with MCOC show that our proposed approach can increase the separation of different patterns, the predictive performance of semantic pattern of a new compound word. In conclusion, we know that some extension-based methods can be used to effectively deal with such challenging problems in data mining.

Keywords: Data Mining; Extension; Multi-criteria Optimization; Classification

1. Introduction

Data mining is very important for us to extract the useful knowledge for some critical decision-making. As for classification there are many different methods for solving this kind of problem, such as neural networks, decision tree, Bayesian networks, SVM, MCOC, extension classifier, and so on.

* Corresponding author. Tel.: +86-150-535-32980.
E-mail address: zzwmis@163.com.
Extension classifier (EC) is a new classification method which based on extensible transformation. The matter-element models of different classes and unlabeled data are built, and the dependent degree values between the unlabeled data matter-element and class matter-elements are calculated, respectively. According to the dependent degree values, the unlabeled data is classified as the class with the maximum dependent degree. Because of its advantages over other classification methods for solving inconsistent, non-compatible, and contradictory problems in practical applications, extension methods are also used for clustering, association, and other problems in data mining [3, 8].

Recently the interest in the optimization-based data mining is increasing, and MCOC can be used to solve the classification problem in data mining. The classifier tries to reach a tradeoff between the overlapping degree of different classes and the total distance from the input data to decision boundary. Then a linear MCOC [4, 5] based on comprise solution was proposed and used to analyze the behaviors of credit cardholders. Then a multiple phase fuzzy linear programming approach [6] was proposed and used to analyze the behaviors of credit cardholders. A rough set-based MCOC [9] was put forward and applied to the medical diagnosis and prognosis, and a linear MCOC with fuzzy parameters [10] was used to improve the generalization of MCOC. Besides, a kernel-based MCOC classifier [11] was provided just like the use of the kernel method in SVM. Then MCOC was used to analyze the behaviors of VIP E-mail users [12]. Besides, the above rough set-based MCOC was employed to predict the protein interaction hot spots [13]. In these applications, MCOC has showed that it can provide better performance than some traditional data mining methods. And MCOC is extensively used for solving classification, regression and other problems.

However, MCOC is not good at dealing with the correlation of different conditional attributes and their contribution to class attribute, especially when data set is composed of many attributes and a large number of data with the aforementioned problems. Fortunately, EC can effectively find the intrinsic relation among different attributes and efficiently obtain the dependent degree between an object and its class [1, 2]. Contrarily, the class-overlap and loss induced by misclassification in data set are not considered by EC, so generalization of classification badly degenerates. To some extent, MCOC can tradeoff the overlap of different classes and misclassification loss and obtain a good generalization. Consequently, the combination of the advantages of the former two approaches is promising to overcome the above problems and challenges.

The rest of this paper is organized as follows: First, Section 2 is the basic principle of MCOC. Then EMOC is detailed in Section 3. The experiment on the pattern analysis of Chinese semantic word-formation, and the results and comparisons are demonstrated in Section 4. Finally, conclusions will be given in Section 5.

2. MCO Classifier

Different from the traditional methods in data mining, MCOC utilizes the flexibly multi-objective and multi-constraint equation to fit decision function for separating the points of different classes. For a binary classification problem, given a training set \( T = \{ (x_i, y_i), \cdots, (x_n, y_n) \} \), each input point \( x_i \in \mathbb{R}^d \) belongs to either of the two classes with a label \( y_i \in \{ -1, 1 \} \), \( i = 1, \cdots, m \) for \( y_i = -1 \); \( i = m + 1, \cdots, n \) for \( y_i = 1 \), where \( d \) is the dimensionality of the input space, and \( n \) is the sample size. For the linearly separable dataset \( T \), we can write the MCOC model as

\[
\begin{align*}
\min C \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \beta_i \\
\text{subject to} \quad w^T x_i - b = y_i (\beta_i - \alpha_i), \quad \alpha_i \geq 0, \quad \beta_i \geq 0, \quad \forall i.
\end{align*}
\] (1)

In the model (1) the parameter \( \alpha_i \) (\( \alpha_i \geq 0 \)) is the distance where the point \( x_i \) deviates from the separating hyperplane, and the sum of \( \alpha_i \) should be minimized. At the same time, the parameter \( \beta_i \) (\( \beta_i \geq 0 \)) is the distance where the point \( x_i \) departs from the decision hyperplane, then the sum of \( \beta_i \) should be maximized. Besides, the points \( x_i \) are given training data, \( w \) and \( b \) are unrestricted variables, \( C > 0, i = 1, 2, \cdots, n \).
For the nonlinearly separable case, we suppose that $\phi(x)$ is a basic function mapping the input data into a higher dimensional feature space. Given the dataset $T' = \{(\phi(x_1), y_1), \ldots, (\phi(x_n), y_n)\}$, the weight vector $w$ can be denoted as the linear combination of $\phi(x_j)$ and $y_j$ with respect to the positive coefficient $\lambda_j$ ($\lambda_j \geq 0$), that is

$$w = \sum_{j=1}^n \lambda_j y_j \phi(x_j).$$

Plugging the above weight vector $w$ into the model (1), we have

$$\min \ C \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \beta_i$$

subject to $\sum_{j=1}^n \lambda_j y_j \phi(x_j) - b = y_i (\beta_i - \alpha_i)$, $\alpha_i \geq 0$, $\beta_i \geq 0$, $0 \leq \lambda_j \leq C$, $\forall i$.

In the case the dot product of basic functions is replaced by the kernel function $K(x_i, x_j)$, we get the below MCOC model

$$\min \ C \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \beta_i$$

subject to $\sum_{j=1}^n \lambda_j y_j K(x_j, x_i) - b = y_i (\beta_i - \alpha_i)$, $\alpha_i \geq 0$, $\beta_i \geq 0$, $0 \leq \lambda_j \leq C$, $\forall i$.

By solving the model (4), we can obtain the coefficient $\lambda_j$ ($j = 1, 2, \ldots, n$). For all training points $x_i$ ($i = 1, 2, \ldots, n$) which satisfy $\alpha_i = 0$ or $\beta_i > 0$, according to the decision hyperplane $w^T \phi(x_i) = b$, we have $b = \sum_{j=1}^n \lambda_j y_j K(x_j, x_i)$, then an average of $b$ is taken. The decision function of MCOC is denoted as

$$f(x) = \text{sign}(w^T \phi(x) - b) = \text{sign}(\sum_{j=1}^n \lambda_j y_j K(x_j, x) - b).$$

Besides, the radial basis function (RBF) kernel as $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2)$ ($\sigma > 0$) is often chosen for a nonlinear mapping in MCOC.

3. Extensible MCO Classifier

3.1. The Disadvantages of MCOC and EC

Although EC has many advantages for solving inconsistent, non-compatible, and contradictory problems, it is short of the tradeoff between fault tolerance and generalization in the unlabeled data. However, MCOC is good at those aspects. That is to say MCOC can gain the better tradeoff between minimizing the overlapping degree and maximizing the distance departed from adjusted boundary. It does not attempt to get the optimal solution but to seek for the non inferior solution in order to gain the better generalization.

Reversely, MCOC is not good at dealing with the correlation of different conditional attributes and finding the corresponding contribution to class attribute, especially when data set is composed of many attributes and a large number of data with the inconsistent, non-compatible, and contradictory problems. Nevertheless, EC can effectively find the intrinsic relation among different attributes and efficiently obtain the dependent degree between an object and its class by means of matter-element relation analysis.

Consequently, combining the advantages of MCOC and EC can help to gain the promising classifier for solving the aforementioned problem.

3.2. EMCO Classifier

Based on given dataset $T' = \{(\phi(x_1), y_1), \ldots, (\phi(x_n), y_n)\}$, the matter-element model of class $y_i$ ($y_i \in \{-1, 1\}$, $i = 1, \ldots, n$) is defined as
\[ R_{yi} = (X_{yi}, x_i, V_{yi}) = \begin{pmatrix} X_{yi} & x_{i1} & V_{yi1} \\ \vdots & \vdots & \vdots \\ x_{id} & V_{yid} \\ \end{pmatrix} \]

where \( X_{yi} \) is the name of class \( y_i \), \( x_{ij} (j = 1, \ldots, d) \) is the \( j \)th attribute of \( X_{yi} \), \( V_{yi} \) is the domain of the attribute \( x_{yi} \) and \( < l_{yij}, u_{yij} > \) is the its interval with the lower and upper boundaries.

Without regard to class \( y_i \), the matter-element models of the whole set \( X(x_i \in X) \) and any input data \( x \) is respectively denoted as

\[ R_x = (X, x_i, V_x) = \begin{pmatrix} X & x_{i1} & V_{x1} \\ \vdots & \vdots & \vdots \\ x_{id} & V_{xd} \\ \end{pmatrix} \]

where \( v_i \) is the value of the attribute \( x_{yi} \).

In addition, we suppose that the \( c_j \) is the weight coefficient of the attribute \( x_{ij} \), and it is obvious that we have \( \sum_{j=1}^{d} c_j = 1 \). Thus, for any input data \( x \), its dependent degree value of all the attributes with respect to different class \( y_i \) is defined as

\[ K_{yi}(x_i) = \begin{pmatrix} \rho(x_{yi}, X_{yi}) / D(x_{yi}, X_{yi}, X), & \text{for } \rho(x_{yi}, X_{yi}) = \rho(x_{yi}, X) \text{ and } x_i \notin V_{yi} \\ \rho(x_{yi}, X_{yi}) / D(x_{yi}, X_{yi}, X), & \text{otherwise} \end{pmatrix} \]

where the functions \( \rho(x, \cdot) \) are denoted as

\[ \rho(x_{yi}, X_{yi}) = \begin{pmatrix} x_i - (l_{yij} + u_{yij}) / 2 & (u_{yij} - l_{yij}) / 2 & \rho(x_{yi}, X) = \begin{pmatrix} x_i - (l_{yij} + u_{yij}) / 2 & (u_{yij} - l_{yij}) / 2 \end{pmatrix} \end{pmatrix} \]

And the function \( D(x_{yi}, X_{yi}, X) \) is expressed as

\[ D(x_{yi}, X_{yi}, X) = \begin{pmatrix} \rho(x_{yi}, X) - \rho(x_{yi}, X_{yi}), & \text{for } \rho(x_{yi}, X_{yi}) \neq \rho(x_{yi}, X) \text{ and } x_i \notin V_{yi} \\ \rho(x_{yi}, X) - \rho(x_{yi}, X_{yi}) + l_{yij} - u_{yij}, & \text{for } \rho(x_{yi}, X_{yi}) \neq \rho(x_{yi}, X) \text{ and } x_i \in V_{yi} \end{pmatrix} \]

Thus, the dependent degree value of the input data \( x \) with respect to the \( R_{yi} \) of class \( y_i \) is obtained by

\[ K_{yi}(x) = \sum_{j=1}^{d} c_j K_{yi}(x_{yi}) \]

According to the function (11), for given dataset \( T^* \), we may compute the dependent degree factor of input data \( x \) with respect to class \( y_i \), that is, the dependent degree factor is defined as

\[ s_i = 1 - (K_{yi}(x) - \min(K_{yi}(x))) / (\max(K_{yi}(x)) - \min(K_{yi}(x))) \]

Thus we obtain the training set \( T^* = \{(x_1, y_1, s_1), \ldots, (x_n, y_n, s_n)\} \), the dependent degree factors can be added MCOC. Besides, for the class-imbalanced case, let \( C_1 (C_1 > 0) \) be the misclassification cost or penalty factor if \( y_i = 1 \). Similarly, let \( C_2 (C_2 > 0) \) be the misclassification cost if \( y_i = 1 \). That is, EMCOC is written as

\[ \min \{ C_1 \sum_{y_i=-1} s_i \alpha_i + C_2 \sum_{y_i=-1} s_i \beta_i - \sum_{i=1}^{n} \lambda_i \} \]

subject to \( \sum_{j=1}^{n} \lambda_j y_j K(x_j, x_i) - b = y_i (\beta_i - \alpha_i) \), \( \alpha_i \geq 0, \beta_i \geq 0, \forall i \),

\[ 0 \leq \lambda_j \leq C_1, \text{ for } y_j = -1, 0 \leq \lambda_j \leq C_2, \text{ for } y_j = 1. \]

From the above model we know that the classifier has the advantages of both MCOC and EC owing to taking the dependent degree factors into account.

Finally, we have implemented the corresponding algorithms of MCOC and EMCOC and the experiments in the next section are carried out on Matlab 7.0 platform. The linear programming problems of MCOC and EMCOC are respectively solved by utilizing Matlab optimal tools.
4. Experiments of Chinese semantic word-formation

4.1. Datasets

In our experiment, database is sourced from the Chinese semantic word-formation corpus. The datasets are obtained by using computer to automatically label the real and large-scale Chinese compound word corpus and then checking and correcting it manually. The database contains 50562 disyllabic compound words. Each word-formation represents the three types of word senses: word sense, sense of morpheme 1, and sense of morpheme 2. Then each sense type consists of three semantic layers: the big class, medium class, and small class. Besides, each word-formation is classified as one of the predefined 8 patterns, which are listed in Table 1.

Table 1. The class distribution in Chinese semantic word-formation database (Only 4 classes are provided)

<table>
<thead>
<tr>
<th>Pattern (Class) labels</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types of patterns</td>
<td>A+B=A=B</td>
<td>A+B=A</td>
<td>A+B=B</td>
<td>A+B=C</td>
</tr>
<tr>
<td>Number of words (%)</td>
<td>3675 (7.27)</td>
<td>1073 (2.12)</td>
<td>319 (0.63)</td>
<td>4382 (8.67)</td>
</tr>
</tbody>
</table>

Owing to the fuzziness of language and some manual errors, the database may contain potential noise, the inconsistent, non-compatible, and contradictory values of attributes.

4.2. Experiments and Results analysis

For Chinese semantic word-formation analysis, the experiment is actually a multi-class classification problem, so it can be transformed into multiple paired two-class classification problems. Then we randomly and respectively select 250 positives and the same number of negatives from any two patterns as the training set and the remaining instances are used for the independent test set.

Based on the training sets, MCOC and EMCOC are trained using 10-fold cross-validation method, and the corresponding classifiers are tested on test sets respectively. Our parameter sets are set to: the penalty factors $C_1$ and $C_2$ from the set $\{2, 16, 64, 128, 256, 512, 1024, 12288\}$ the parameter $\sigma$ from the set $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ for the RBF kernel function. The predictive accuracy of different patterns is computed by Accuracy $= TP / (TP + FN)$ where true positive (TP) is the number of positives that are correctly predicted and false negative (FN) is the number of positives that are predicted as negatives.

Fig. 1. Evaluation of classifiers for predicting Chinese semantic word-formation patterns on the independent test set
As the experimental results shown in Fig. 1, for the predictive performance of Chinese semantic word-formation patterns, we find that EMCOC is slightly better than MCOC in classification accuracies. Obviously adding the dependent degree factor to MCOC effectively enhanced the prediction and separation of different patterns.

5. Conclusion

In this paper, we proposed a classification model based on extension classification method for Chinese word-formation analysis. We see that the EMCOC extends the capacities of MCOC and effectively deals with the class-overlapping, inconsistent, non-compatible, and contradictory problems in real world applications. Experimental results show that it is an effective classifier for predicting Chinese semantic word-formation patterns.

Acknowledgements

This research has been partially supported by the Scientific Research Foundation of Ludong University (LY2010013), the Natural Science Foundation of Shandong (ZR2012FL13) and the National Natural Science Foundation of China (#70871111, #61170161, #71271191, #61272215). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References