Optimizing Support Vector Machine Parameters with Genetic Algorithm for Credit Risk Assessment

by Elviawaty13 M. Zamzami
Optimizing Support Vector Machine Parameters with Genetic Algorithm for Credit Risk Assessment

To cite this article: Jorson Manurung et al 2017 J. Phys.: Conf. Ser. 930 012026

View the article online for updates and enhancements.
Optimizing Support Vector Machine Parameters with Genetic Algorithm for Credit Risk Assessment

Jonson Manurung1, Herman Mawenkang2, Elvisawat Zamzami2

1,2,3 Department of Computer Science,
Faculty of Computer Science and Information Technology,
Universitas Sumatera Utara,
Medan, Indonesia, 20155
Email: jonson.gee@gmail.com, hmawenkang@yahoo.com, elvi_zamzami@usu.ac.id

Abstract. Support vector machine (SVM) is a popular classification method known to have strong generalization capabilities. SVM can solve the problem of classification and linear regression or nonlinear kernel which can be a learning algorithm for the ability of classification and regression. However, SVM also has a weakness that is difficult to determine the optimal parameter value. SVM calculates the best linear separator on the input feature space according to the training data. To classify data which are not linearly separable, SVM uses kernel tricks to transform the data into a linearly separable data on a higher dimension feature space. The kernel tricks using various kinds of kernel functions, such as: linear kernel, polynomial, radial base function (RBF) and sigmoid. Each function has parameters which affect the accuracy of SVM classification. To solve the problem genetic algorithms are proposed to be applied as the optimal parameter value search algorithm thus increasing the best classification accuracy on SVM. Data taken from UCI repository of machine learning database: Australian Credit Approval. The results show that the combination of SVM and genetic algorithms is effective in improving classification accuracy. Genetic algorithms have been shown to be effective in systematically finding optimal kernel parameters for SVM, instead of randomly selected kernel parameters. The best accuracy for data has been upgraded from kernel Linear: 85.12%, polynomial: 81.56%, RBF: 77.22% Sigmoid: 78.70%.

Keywords: support vector machine (SVM), genetic algorithm, parameter optimization, credit risk.

1. Introduction

Classification is the process of finding models or functions that explain and distinguish classes or concepts, with the aim that the obtained model can be used to identify classes or objects that have unknown class labels. The derived model is based on the analysis of the training data. Classification falls into the category of predictive data mining [1].

In the classification process is divided into two phases namely, learning and testing. In the learning phase, some data that has been known to the data class (training set) is used to form the model. Furthermore, in the testing phase, the established model is tested with some other data (test set) to determine the accuracy of the model. If the accuracy is sufficient then the model can be used for prediction of unknown data class [2].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Published under licence by IOP Publishing Ltd.
1.1. Support Vector Machine

Support Vector Machine (SVM) is a learning system for classifying data into two groups of data using hypothetical space in the form of linear functions in a feature space (feature space) high dimension. SVM has properties that are not owned by the learning machine in general that is in the process of finding the best dividing line (hyperplane) so as to obtain the maximum margin size between the non-linear input space with the feature space using the kernel rules [3]. Margin is twice the distance between hyperplane and support vector. The point closest to the hyperplane is called the support vector. Suppose the data is denoted as \( x_i \in \mathbb{R}^n \), for the class label of data \( x_i \) denoted \( y_i \in \{-1, 1\} \) with \( i = 1, 2, ..., \) \( l \) where \( l \) is a lot of data. Separation of data linearly on the SVM method can be seen in Figure 1.

![Figure 1. The data are linearly separated](image)

The example above shows that the two classes can be separated by a pair of parallel bounding planes. The first delimiter field limits the first class while the second delimiter field limits the second class, so that it is obtained:

\[
\begin{align*}
    x_i, w + b & \geq +1 \text{ for } y_i = +1 \\
    x_i, w + b & \leq -1 \text{ for } y_i = -1
\end{align*}
\]

(1)

(2)

\( w \) = Vector weights that are perpendicular to the hyperplane (normal plane)

\( b \) = Position of the field relative to the coordinate center

The margin value between the two classes is

\( m = \frac{2}{|w|} \). Margin Can be maximized using the optimization function

Lagrange:

\[
\min_{w,b} L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{l} \alpha_i \left( y_i ((x_i \cdot w) + b) - 1 \right)
\]

(3)

By minimizing \( L \) against \( w \) and \( b \), then it is obtained:

\[
\begin{align*}
    \frac{\partial L}{\partial w} & = w - \sum_{i=1}^{l} \alpha_i y_i x_i = 0 \\
    \frac{\partial L}{\partial b} & = \sum_{i=1}^{l} \alpha_i y_i = 0
\end{align*}
\]

(4)

(5)

The above equation can be modified as \( L \) maximization containing only \( \alpha_i \) as the equation below.

\[
\max_{\alpha} L = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j x_i \cdot x_j
\]

(6)
\[ s.t. \sum_{i=1}^{\alpha_i}, y_i = 0, \alpha_i \geq 0 \] (7)

The resulting \( \alpha \) is used to find \( w \). Data that have \( \alpha_i \) value \( \geq 0 \) is support vector while the rest have value \( \alpha_i = 0 \). After the \( \alpha_i \) value is found, then the class of the test data \( x \) can be determined based on the value of the decision function:

\[ f(x) = \sum_{i=1}^{N} \alpha_i y_i x_i, x_d + b \] (8)

dengan
\( x_i = \) support vector
\( n_i = \) amount support vector
\( x_d = \) Data to be classified.

If data cannot be perfectly separated by linear separation, SVM is modified by adding variables \( \xi \) (\( \xi_i \geq 0, y_i \xi_i = 0 \) if \( x \) classified correctly) so that the best divider form search formula becomes:

\[ \min \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i \] (9)
\[ s.t. y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \] (10)

The search for the best splitter field with the addition of \( \xi \) variable is called soft margin hyperplane. \( C \) is the parameter that determines the magnitude of the penalty due to errors in the classification. Thus, the dual problem generated in a non-linear problem equals the dual problem generated by linear problem only the \( \alpha_i \) range between \( 0 \leq \alpha_i \leq C \).

![Figure 2. Soft margin hyperplane](image)

Some commonly used Kernel functions are:
1. Linear kernel
   \[ K(x_i, x) = x_i^T \cdot x \] (11)
2. Polynomial kernel
   \[ K(x_i, x) = (y_i \cdot x_i^T \cdot x + r)^d, \gamma > 0 \] (12)
3. Radial Basic Function
   \[ K(x_i, x) = \exp(-\gamma||x_i - x||^2), \gamma > 0 \] (13)
4. Sigmoid kernel
   \[ K(x_i, x) = \tanh(y_i \cdot x_i^T \cdot x + r)^d, \gamma > 0 \] (14)
In this case γ, r, and d are kernel parameters, as well as parameter C as a penalty due to errors in the classification for each kernel.

1.2. Algorithm Genetika

Defines a genetic algorithm as a search algorithm based on the mechanism of natural selection and genetics [4]. Genetic algorithms are widely used on practical problems that focus on the search for optimum parameters. The emergence of this algorithm is inspired by theories in the biological sciences because initially genetic algorithms are used by biologists to simulate biological systems, so many terms and concepts of biology are used in genetic algorithms [5].

2. Computational Analysis

Data consists of 14 attributes with Categorical, Integer, and Real type. The amount of data is 690 lines. The data are classified into 2 classes, namely Positive (+) and Negative (-) classes. Class data + consists of 307 data or 44.5% of the total data, while class data - consists of 383 data or as much as 55.5% of the overall data. Australian Credit Approval data does not have the value of a missing value attribute or a missing value label. So the missing missing value technique in pre-processing data is not done.

2.1. Classification

In SVM training process for Linear Kernel function, Polynomial Kernel, Radial Basis Function (RBF) Kernel and Sigmoid Kernel, parameters C, γ, r, and d. Parameter C for Linear Kernel function, parameters C, γ, r, and d for Kernel Polynomial function, and parameters C and γ for RBF Kernel function. As for the function of Sigmoid kernel required parameter values C, γ, and r. The SVM training process uses the LIBSVM library developed by National Taiwan University Computer Science students, Chih-Yuan Yang and Chih-Huai Cheng. This study classifies Australian Credit Approval Data with all four Kernel functions by inputting the specified parameter values. From the classification of data with SVM yields an accuracy level using the 5-fold cross validation method.

2.2. Parameter SVM Optimisation

For this study we determine the probability value of crossover of 0.8, and the probability of mutation is also very small, about 1 divided by the number of genes. To search the parameter values C, γ, and r are limited in range 2-10 to 20, whereas parameter d is limited in range 2 to 5. The process of calculating the accuracy level will stop if on 10 successive generations have a small difference of accuracy value equal to 0.1. This means that an increase in the accuracy level of 0.1 in a generation will be considered the same as the preceding generation. Here is the best fit for SVM classification, obtained using the genetic algorithm along with the values of the kernel parameters.

From program execution using Australian Credit Approval data, Linear yield accuracy equal to 85.12% with parameter C worth 0.1503. Program execution with Polynomial kernel yields the highest accuracy value reaching 81.76% with parameter values C, γ, r and d respectively 0.0302, 0.0146, 0.5024, and 2. With RBF kernel yield accuracy of 77.22% and parameters C and γ obtained 0.7473 and 0.9883, while the execution result with optimum Sigmoid at 78.70% accuracy with parameter values C, γ, r and d respectively 0.8781, 0.8215, and 0.0556. The increase in fitness values can be seen in Figure 2 and the best fitness for each kernel can be seen in 1.
Table 1. Best fitness value of Australian Credit Approval data in Linear, Polynomial, RBF and Sigmoid

| Kernel   | Best Fitness |
|----------|--------------|
| Linear   | 85.12%       |
| Polynomial | 81.76%     |
| RBF      | 77.22%       |
| Sigmoid  | 78.70%       |

Based on the search results of SVM parameter values using genetic algorithm, there is an increase of accuracy value in the data used. Table 2 shows the magnitude of the increase in accuracy value.

Table 2. The magnitude of the increase in fitness value on SVM parameter optimization using genetic algorithm

| Kernel   | Kenaikan |
|----------|----------|
| Linear   | 5.29%    |
| Polynomial | 8.68%    |
| RBF      | 5.89%    |
| Sigmoid  | 10.56%   |

When compared to the usual SVM classification by inputting parameter values arbitrarily, the results of the SVM parameter value search that optimizes the fitness value by using the genetic algorithm show an increase in accuracy. This means, increased accuracy in data classification, shows the comparison of accuracy values that perform parameter optimization with the accuracy of the inputs of predefined parameter values. The result of comparison of accurate value obtained shows that the search for SVM parameter values using genetic algorithm yields greater accuracy value. This, proving that the optimization of SVM parameters using genetic algorithms succeeds in achieving optimal accuracy.

3. Conclusion
The result of the research shows that accurate value by searching the SVM parameter value systematically using genetic algorithm yields a better accuracy value compared to the accuracy value of SVM execution result without optimization with the highest accuracy increase that is 10.56%. Thus, the genetic algorithm successfully searches for SVM parameter values that optimize the accuracy value in the SVM classification method. This proves that genetic algorithm is effective in increasing the accuracy value of a classification model in the SVM classification method. Genetic algorithms are very effective and efficient for small data. However, the timing of genetic algorithm execution will increase with increasing population size, generation, parameter value range as well as data size. This causes the length of the execution process, making it less efficient when compared to the increase in accuracy value.

References
[1] Ross, D. T. 2006. Data Mining Methods and Models, John Wiley & Sons, Inc Hoboken.
[2] Han J, Kamber M. 2006. Data Mining: Concepts and Techniques, USA: Morgan Kaufman Publishers.
[3] Cortes, C. & Vapnik, V. 1995. Support-Vector Networks, Machine Learning, Vol. 2, pp 273-297.
[4] Goldberg, D. E. 1989. Genetic Algorithms in Search, Optimization, and Machine Learning England: Addison-Wesley Publishing Company.
[5] Michalewicz Z. 1996. Genetic Algorithms + Data Structures = Evolution Programs, USA: Department of Computer Science, University of North Carolina.
Optimizing Support Vector Machine Parameters with Genetic Algorithm for Credit Risk Assessment

**ORIGINALITY REPORT**

| %14 | %8 | %13 |
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS |
| % | | % |

**PRIMARY SOURCES**

1. discovery.ucl.ac.uk
   Internet Source

2. P.O. Abas Sunarya, Rina Refianti, Achmad Benny, Wiranti Octaviani. "Comparison of Accuracy between Convolutional Neural Networks and Naïve Bayes Classifiers in Sentiment Analysis on Twitter", International Journal of Advanced Computer Science and Applications, 2019

3. Yoga Pristyanto, Sumarni Adi, Andi Sunyoto. "The Effect of Feature Selection on Classification Algorithms in Credit Approval", 2019 International Conference on Information and Communications Technology (ICOIACT), 2019

4. A. Ramanan, S. Suppharangsan, M. Niranjan. "Unbalanced Decision Trees for multi-class classification", 2007 International Conference on
John Prakash, A K Mishra. "Fabrication, optimization and application of a dip-probe fluorescence spectrometer based on white-light excitation fluorescence", Measurement Science and Technology, 2013

Chusnul Khotimah, Santi Wulan Purnami, Dedy Dwi Prastyo. "Additive survival least square support vector machines and feature selection on health data in Indonesia", 2018 International Conference on Information and Communications Technology (ICOI Act), 2018
