Dual Membership Fuzzy Support Vector Machine Algorithm Based on SVDD

Yingcheng Xu¹, Wei Feng¹*, Fei Pei², and Haiyan Wang³

¹Quality Research Branch, National Institute of Standardization, Beijing 100191, China
²National Institute of Standardization, Beijing 100191, China
³School of Management and E-Business, Zhejiang Gongshang University, Hangzhou 310018, China

*Corresponding author email: fengw@cnis.ac.cn

Abstract. In the case of excessive overlap between positive and negative samples in data set, the deviation in the category of reconstructed sample points will lead to unsatisfactory discrimination of SVM, no matter what methods are used to reconstruct the sample set. A dual membership fuzzy support vector machine algorithm based on support vector data domain description was thus proposed, followed by a simulation analysis of common data set. Experimental results show that the proposed algorithm can work well in classification when the sample set is overlapped.

1. Introduction
In traditional support vector machine (SVM) models, the training sample set is absolutely divided into two mutually exclusive classes, resulting in the “sample overlap” problem in reality. In the event of excessive overlap between positive and negative samples in data set, traditional SVM models will always fail to perform well in classification, regardless of the methods used to reconstruct the sample set. Support vector data domain description (SVDD) is a method proposed by D. Tax and R. Duin to describe the data by calculating the boundary of the smallest hypersphere in a high dimensional space that contains training samples. The present study intends to build a dual membership fuzzy support vector machine (D-FSVM), so that each training sample belongs to the two classes according to its dual membership. In this way, the problem caused by excessive overlap between positive and negative samples can be effectively solved to provide better classification.

2. Build D-FSVM Model
2.1. Basic Model
The basic model of D-FSVM is as follows:

\[
\begin{align*}
\min_{\alpha, b} & \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{L} (\mu_i^A \xi_i + \mu_i^B \eta_i) \\
\text{s.t.} & \quad w^T \phi(x_i) + b \geq 1 - \xi_i \\
& \quad w^T \phi(x_i) + b \leq -1 + \eta_i \\
& \quad \mu_i^A + \mu_i^B = 1
\end{align*}
\] (1)
In the model above, \( \xi_i \) and \( \eta_i \) are slack variables that reflect the width of error band of each sample point. The Lagrange function of this problem is:

\[
L(w, b, \xi, \eta, \alpha, \beta, v_i, u_i) = \frac{1}{2} w^T w + C \sum_{i=1}^l \mu_i \xi_i + C \sum_{i=1}^l (1 - \mu_i \eta_i)
-
\sum_{i=1}^l \alpha_i (w^T \phi(x_i) + b - 1 + \xi_i)
+
\sum_{i=1}^l \beta_i (w^T \phi(x_i) + b + 1 - \eta_i)
-
\sum_{i=1}^l v_i \xi_i - \sum_{i=1}^l u_i \eta_i
\]  

(2)

In the function above, \( \alpha \), \( \beta \), \( v \), \( u \) are non-negative Lagrange multipliers. The objective function of the optimal solution is as follows:

\[
\max_{\alpha, \beta} \sum_{i=1}^l \alpha_i + \sum_{i=1}^l \beta_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i - \beta_i)(\alpha_j - \beta_j)\phi(x_i)^T \phi(x_j)
\]

s.t.

\[
\sum_{i=1}^l \alpha_i = \sum_{i=1}^l \beta_i
\]

\[
0 \leq \alpha_i \leq C \mu_i
\]

\[
0 \leq \beta_i \leq C (1 - \mu_i)
\]

\[
i = 1, 2, ..., l
\]

The above equation is also a function to solve the optimal solution of its dual problem. This function may cause "curse of dimensionality" in the operation process, it is necessary to use Equation 4 to replace the operation of high-dimensional space:

\[
K(x_i, x_j) = \phi(x_i)^T \phi(x_j)
\]  

(4)

The obtained operator for classification is:

\[
f(x) = \text{sign}(\sum_{i=1}^l (\alpha_i - \beta_i)K(x_i, x_j) + b)
\]

(5)

The distance-based double membership computing method is then used to describe the degree of membership of training sample points relative to the two classes of samples:

\[
\mu_i^A = \frac{d_i^B}{d_i^A + d_i^B}, \mu_i^B = \frac{d_i^A}{d_i^A + d_i^B}
\]  

(6)

In the equation above, \( d_i^A = \frac{\|\Phi^+(x_i) - a^+\|}{R^+}, d_i^B = \frac{\|\Phi^-(x_i) - a^-\|}{R^-} \) represent the ratio of the distance between the sample located in overlapping region and the center of two types of the smallest hyperspheres to their radius, respectively.

2.2. Steps of Algorithm
The steps of the SVDD-based D-FSVM algorithm are as follows:
SVDD is used for one-class learning of the two classes of samples in training set, respectively. The discriminant decision functions for the two classes, $f^+(x)$ and $f^-(x)$, are obtained to identify the noise points, positive samples, negative samples and samples in overlapping region;

(2) Based on $f^+(x)$, $f^-(x)$ and the smallest hypersphere of two classes of samples, the degree of dual membership of samples in overlapping region is calculated;

(3) The D-FSVM model is used to train the samples in overlapping region. The classification decision function for the samples in overlapping region, $f(x)$, is obtained;

(4) The samples in test set $x_i$ are first identified as noise point, positive samples, negative samples or samples in overlapping region by using $f^+(x)$ and $f^-(x)$;

(5) The degree of dual membership of overlapped samples in test set is calculated. The decision function of D-FSVM model, $f(x)$, is then used for discrimination.

3. Simulation Analysis

3.1. Experimental Data

The databases such as Pima-indians, Breast-w and Inosphere in the UCI machine learning repository were selected. The details of each database are shown in the table below.

Table 1. Basic information of UCI data sets.

| Date sets    | Dimensions | Number of positive samples | Number of negative samples | Total number of samples | Non-equilibrium ratio |
|--------------|------------|----------------------------|---------------------------|------------------------|-----------------------|
| Pima-indians | 8          | 268                        | 500                       | 768                    | 1:2                   |
| Breast-w     | 9          | 241                        | 458                       | 699                    | 1:2                   |
| Inosphere    | 34         | 126                        | 225                       | 351                    | 1:2                   |

The UCI data sets were randomly divided in the way that 70% of data sets are used as the training set and 30% as the test set. The non-equilibrium ratio remained constant during the division process. To analyze the performance of the proposed D-FSVM algorithm based on SVDD, the data set reconstruction models including SVM and SVDD-based SVM were selected for a comparative analysis. The SVDD-based SVM algorithm was similar to the SVDD-based D-FSVM algorithm. Thus, the ordinary SVM model was used to distinguish the samples in overlapping region in Step 2 and Step 4. It was also unnecessary to assign the samples in overlapping region with dual membership.

3.2. Results and Discussion

The evaluation indexes of classification algorithm include sensitivity (SE), specificity (SP) and grand mean of classification accuracy (Gm). The experimental results are as follows:

Table 2. Experimental results.

| Date sets    | Evaluation indexes | SVM     | SVDD+SVM | SVDD+(D-FSVM) |
|--------------|---------------------|---------|----------|---------------|
| Pima-indians | SE(%)               | 53.04   | 68.49    | 78.37         |
|              | SP(%)               | 89.60   | 90.71    | 92.52         |
|              | Gm(%)               | 68.94   | 78.82    | 89.34         |
|              | SE(%)               | 59.03   | 70.14    | 82.11         |
| Breast-w     | SE(%)               | 65.30   | 73.55    | 81.63         |
| Inosphere    | SP(%)               | 73.65   | 85.19    | 92.07         |
|              | Gm(%)               | 69.35   | 79.16    | 86.69         |
The results indicate that among the three data sets, the SVDD+SVM algorithm and the SVDD+(D-FSVM) algorithm are both significantly better than the ordinary SVM model. Therefore, to provide better classification, it is necessary to identify the noise points, positive samples, negative samples and samples in overlapping region by using the SVDD algorithm, and then learn the overlapped samples by using the SVM model or the D-FSVM model.

Moreover, the proposed SVDD+(D-FSVM) algorithm has the highest SE, SP and Gm indexes among the three data set. Regarding the samples in overlapping region, the degree of dual membership can
better describe the relative extent to which the sample points belong to positive and negative classes, while the D-FSVM model can better classify the overlapped samples.

4. Conclusion
A SVDD-based D-FSVM algorithm is proposed to improve the poor performance of traditional SVM models in classification caused by excessive overlap between positive and negative samples in data set. Through the test using common data sets, the results show that the proposed D-FSVM algorithm can improve the classification of overlapped sample sets.

Acknowledgments
We would like to acknowledge that this Study is supported and funded by the Market Supervision Technology Assurance Project under Grant No. 2020YJ043, and the Market Supervision Policy Research Project "Research on Risk Multidimensional Analysis and Decision Early Warning System for Consumer Quality and Safety".

References
[1] H. Haibao, E. Garcia. Learning From Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 2009,21(9):1263-1284.
[2] M. Sabzekar, H.S. Yazdi, M. Naghibzadeh, et al. Emphatic constraints support vector machine. International Journal of Computer and Electrical Engineering. 2010, 2(2):296-310.
[3] Mehri Mangeli, Alireza Shahraki, Faranak Hosseinzadeh Saljooghi. Improvement of risk assessment in the FMEA using nonlinear model, revised fuzzy TOPSIS, and support vector machine. International Journal of Industrial Ergonomics, 2019,69:209-216.
[4] Rukshan B, Vasile P. FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning. IEEE Transaction on Fuzzy Systems, 2010,18(3):558-571.
[5] Snehamoy Chatterjee, Ansuman Dash, Sukumar Bandopadhyay. Ensemble Support Vector Machine Algorithm for Reliability Estimation of a Mining Machine[J]. Quality and Reliability Engineering International, 2015,31(8):137-154.
[6] Suad Mohammed Othman;, Fadl Mutaher Ba-Alwi. Intrusion detection model using machine learning algorithm on Big Data environment. Journal of Big Data, 2018,5(1): 1 -12.
[7] Xie Shangxin, Li Zhong, Hu Hailong. Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization. Gene, 2018,642:74-83.