Extraction of Highly Interpretable Classification Rules from Nonlinear Support Vector Machines

Hiroshi TENMOTO[1] Ganma KATO[2]

Abstract — The authors are aiming to decrease vaguely fear and anxiousness on the people to AI based on machine learning technology. In this study, we utilize hyper-rectangle rule extraction algorithm and its improved version for the classification rule extraction from nonlinear support vector machines. We evaluated the methods by using classification accuracy rate, the number of classification rules (for interpretability) and fidelity on a concentric circles form artificial dataset. We could confirm that the proposed method can extract highly interpretable classification rules from nonlinear dataset through the experiments.

Keywords: Machine Learning, Pattern Recognition, Classification Rule Interpretability, Support Vector Machine

1 Introduction

Machine learning is a study field and technology, which gives automatic learning ability to computers without explicit programming [1]. Recently, machine learning technology is utilized in wide range of fields. Representative examples are pattern recognition [2] and data mining [3]. In pattern recognition, photo images or sound signals are distinguished by finding characteristic patterns. In data mining, valuable patterns such as typical combinations of items are extracted from big datasets.

Machine learning technologies can be classified into two groups as supervised learning [4] and unsupervised learning. In supervised learning, given input data and its corresponding output data, called training dataset, algorithms find relations in function forms between the input data and output data. On the other hands, in unsupervised learning, given input data without corresponding output data, algorithms find hidden structures in the input data, such as clusters.

In this study, SVM (support vector machine) [5] is focused among supervised learning algorithms in machine learning technology. The biggest advantageous point of SVM is to learn so as to gain generalization ability, that is, the power to predict correctly for unknown data, which is not included in the training dataset.

However, the result of learning of SVM is not interpretable for humans. In recent years, AI (artificial intelligence) technology including machine learning is hoped that it can be applied to problems concerning human life, such as automatic car driving and automatic diagnosis of sickness. So, uninterpretability of learning machine is very important and serious problem. Therefore, in this study, developing technologies that can relax people's incredibility and terrority to AI is aimed.

2 Supervised Learning, Linear SVM and Nonlinear SVM

In supervised learning, given input data $\boldsymbol{x}_i$ $(i = 1, 2, \cdots, n, \boldsymbol{x}_i \in \mathbb{R}^d)$ and corresponding output data y_i $(i = 1, 2, \cdots, n, y_i \in \mathbb{Z})$, learning algorithms find relations in function form $y = f(\boldsymbol{x})$ between the input data and output data. The inferred function is called learning model. After learning phase, the algorithm predicts output y for the unknown data $\boldsymbol{x}$ which is not included in the training dataset. Therefore, the learning model has to hold relations $y = f(\boldsymbol{x})$ between unknown inputs and outputs that are not included in the training dataset. One of the purposes of studies in the supervised learning field is to obtain prediction function which can predict correctly for unknown input.

[1]NITKC (National Institute of Technology, Kushiro College)
[2]TUT (Toyohashi University of Technology)
The original report of this work was written by Ganma KATO as his graduation thesis in NITKC under Hiroshi TENMOTO's supervision, then it is edited and translated to English by Hiroshi TENMOTO.

Linear SVM was proposed as a supervised learning algorithm that can solve linear two class classification problem. Given a training dataset consists of the feature vectors and class labels $\{\boldsymbol{x}_i, y_i\}$ $(i = 1, 2, \cdots, n,\ \boldsymbol{x}_i \in \mathbb{R}^d,\ y_i \in \{1, -1\})$, SVM's discriminant function is expressed as Eq. (1):

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b \tag{1}$$

Here, $\boldsymbol{w}$ is called weight vector and b is called bias. By adjusting the values of the parameters, the algorithm find a hyper-plane that can separate the training dataset correctly. This process is called learning. In SVM, the algorithm find the values of parameters of the separating hyper-plane so as to maximize the distance between classes, called margin.

In linear SVM, the algorithm finds $d-1$ dimensional hyper-plane as the separating hyper-plane for d dimensional input dataset. Therefore, the algorithm cannot find separating hyper-plane for linearly nonseparable dataset. For this problem, the algorithm can be extended to nonlinear SVM. In nonlinear SVM, the inputs $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ are mapped onto higher dimensional space by kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, then linear SVM is applied in the higher dimensional space (Fig. 1).

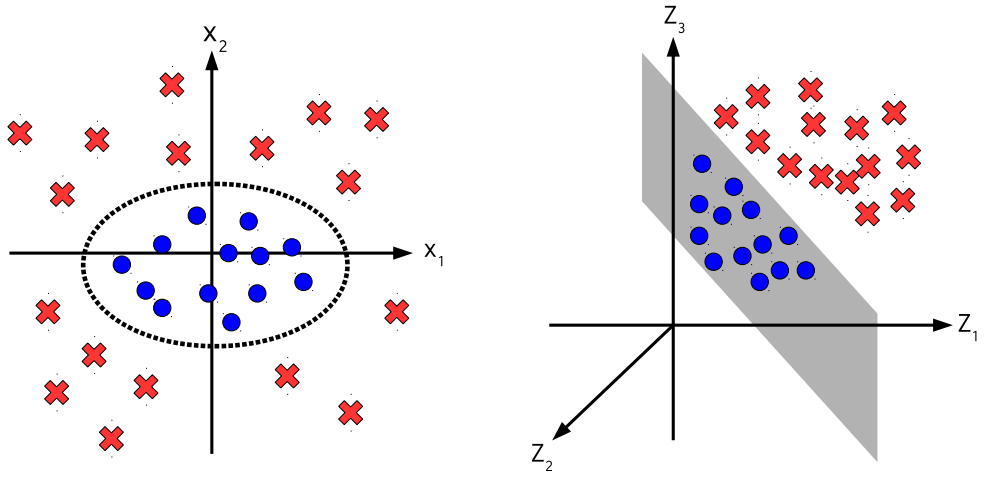

Figure 1: Mapping the data points in original space onto higher dimensional space in nonlinear SVM.

The kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ has wide variations such as linear, polynomial, RBF (radial basis function), histogram intersection, and so on. So the users of the algorithm have to select appropriate kernel function according to the application. In this study, RBF kernel is adopted for nonlinear SVM.

3 Unsupervised Learning and DBSCAN

In unsupervised learning, given input data without corresponding output data, algorithms find hidden structures in the input data. The difference from supervised learning is that true output, called ground truth, is not given for the algorithm. The typical application of unsupervised learning is clustering that forms some groups on the given dataset.

In this study, DBSCAN (Density-based spatial clustering of applications with noise) [6] is employed for clustering phase. DBSCAN clustering algorithm was proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996. This algorithm has advantages on the following two points: there is no need to give the number of clusters as hyper-parameters explicitly, and it can find nonlinearly separable clusters.

4 Extraction of Highly Interpretable Classification Rules

This study aims at extracting highly interpretable classification rules by analyzing the learning result of nonlinear SVM. In this study, we utilize hyper-rectangle rule extraction (HRE) algorithm [7] and its improved version for analysis of the learning result of nonlinear SVM.

4.1 Hyper-rectangle Rule Extraction

HRE was proposed by Ying Zhang, HongYe Su, Tao Jia and Jian Chu in 2005 [7]. The algorithm of HRE is shown below and Fig. 2.

1. Find prototype vectors on clusters without regarding the decision boundary by nonlinear SVM.
2. Define small hyper-rectangles around the prototype vectors.
3. Expand the hyper-rectangles based on the decision boundary of nonlinear SVM.

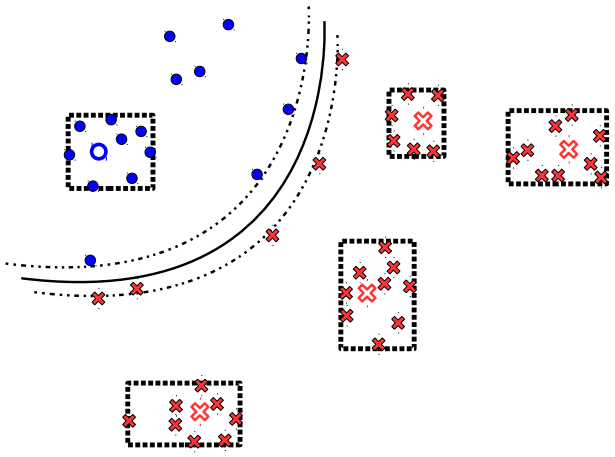

Figure 2: Concept of expansion of hyper-rectangles around the prototype vectors in HRE algorithm.

HRE algorithm minimizes the number of classification rules by approximating the clusters by hyper-rectangles and controls the quality of classification rules by a stopping condition for the expansion of hyper-rectangles. However, HRE algorithm holds hyper-parameter adjusting problem on clustering. In addition, HRE algorithm cannot approximate complex form clusters well.

4.2 Proposed Method

The proposed method in this study is shown below.

1. Select prototype vectors from training dataset randomly.
2. Expand the hyper-rectangles around the prototype vectors according to the decision boundary of nonlinear SVM.
3. Find optimal solution on the resultant hyper-rectangles as combinatorial optimization problem.

In the proposed method, prototype vectors are selected randomly without clustering and many hyper-rectangles are generated in advance. By generating such many hyper-rectangles, approximation of complex regions can be achieved. In addition to the generation of hyper-rectangles, combinatorial optimization reduces the number of classification rules on the generated hyper-rectangles. In this study, genetic algorithm (GA) [8] is utilized for the combinatorial optimization. Ranking selection is used for individual selection phase, two-point crossover for crossover phase and Eq. (2) is used for adaptation evaluation function. In Eq. (2), a_i is the classification accuracy for the training dataset, n_i is the number of selected hyper-rectangles and n_r is the total number of hyper-rectangles.

$$\text{maximize} \quad F(i) = a_i + \left(1 - \frac{n_i}{n_r}\right) \tag{2}$$

5 Evaluation Experiments

Evaluation experiments are carried out for the extraction of classification rules by the two methods on the result of nonlinear SVM. The classification rules are evaluated by the accuracy for validation dataset (not training dataset), the number of classification rules as interpretability index and consistency rate with the result of nonlinear SVM. Here, in the classification phase, data points outside of every hyper-rectangle are treated as unclassifiable. The experiments are carried out on Python 3.6, scikit-learn 0.19.1 and Ubuntu 16.04.2 LTS. scikit-learn is utilized for nonlinear SVM and DBSCAN, and hyper-parameters for nonlinear SVM are optimized by grid search method.

5.1 Extraction of Classification Rules from Concentric Circles Form Dataset

This concentric circles form dataset is generated artificially by using scikit-learn. In this datset, the number of classes (categories) c is 2, the number of data points n is 200, and the number of features (axes in feature space) d is 2.

The results of classification rule extraction with the concentric circles form dataset and nonlinear SVM are shown in Fig 3 and 4.

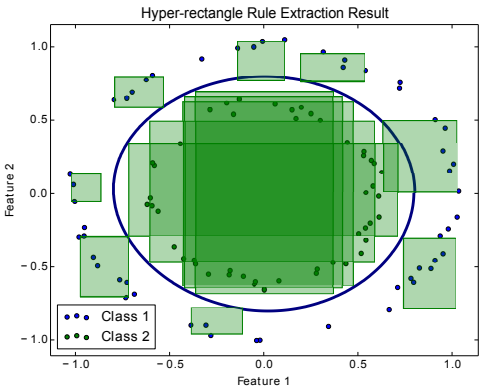

Figure 3: Classification rule extraction result for concentric circles form dataset by HRE.

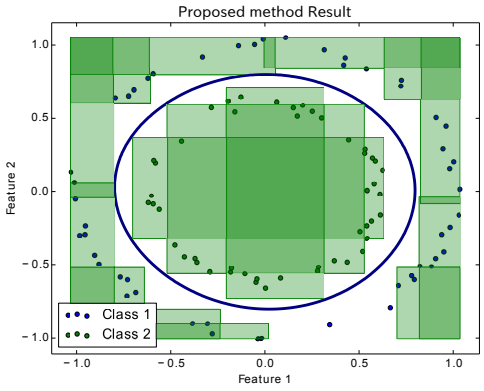

Figure 4: Classification rule extraction result for concentric circles form dataset by the proposed method.

Many hyper-rectangles are generated inside of the circle, while a few rectangles are generated outside of the circle in HRE algorithm. Therefore, leaked data points can be observed, while the number of rectangles is large. On the other hand, it can be observed that the proposed method can cover almost all of the dataset with many classification rules.

Next, classification accuracy rates for validation dataset, the numbers of classification rules and consistency rates with the result of nonlinear SVM are shown in Table 1. The classification accuracy

rates and consistency rates are decreased in HRE because of leakage. While, these problems cannot be seen in the proposed method because of combinatorial optimization.

Table 1: Evaluation results for both algorithms on concentric circles form dataset.

Indices	SVM	HRE	Proposed Method
Accuracy rates for validation dataset (%)	100	58	83
Numbers of classification rules	—	13	15
Consistency rates with result of nonlinear SVM (%)	—	74	96

The change of adaptation rate, classification accuracy rate and the number of classification rules in GA phase is shown in Fig. 5. According to the increase of generations in GA, it can be seen that the adaptation rate (red line) gets higher. In addition, it can be seen that keeping the consistency rate with the result of nonlinear SVM (green line) higher and also keeping the number of classification rules (blue line) lower.

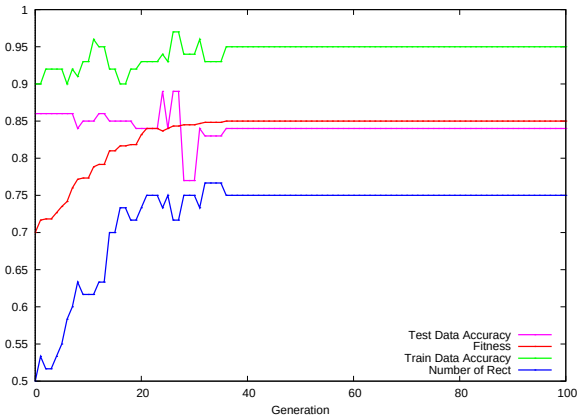

Figure 5: Change of adaptation rate, classification accuracy rate and the number of classification rules in GA for concentric circles form dataset.

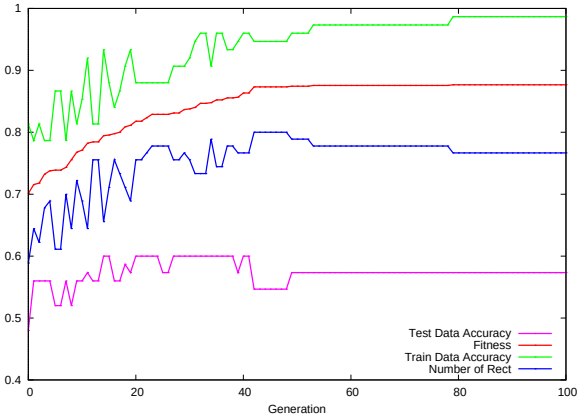

Figure 6: Change of adaptation rate, classification accuracy rate and the number of classification rules in GA for Fisher's Iris dataset.

5.2 Extraction of Classification Rules from Fisher's Iris Dataset

Iris dataset is a famous multivariate dataset which was presented by Ronald Fisher in 1936. In this dataset, the number of classes (categories) c is 3, the number of data points n is 150, and the number of features (axes in feature space) d is 4.

Classification accuracy rates for validation dataset, the numbers of classification rules and consistency rates with the result of nonlinear SVM are shown in Table 2. For this dataset, the classification accuracy rates and consistency rates are also better than HRE. On the other hand, classification accuracy rates for validation dataset was not so high in spite of the high number of classification rules. This means the proposed method adapted to the training dataset too much. This can be thought as occurring of overfitting, because of the dimensionality 4 is higher than 2 in the concentric circles form dataset.

Table 2: Evaluation results for both algorithms on Fisher's Iris dataset.

Indices	SVM	HRE	Proposed Method
Accuracy rates for validation dataset (%)	96	44	57
Numbers of classification rules	—	6	21
Consistency rates with result of nonlinear SVM (%)	—	74.67	98.67

Next, the change of adaptation rate, classification accuracy rate and the number of classification rules in GA phase is shown in Fig. 6. According to the increase of generations in GA, it can be seen that the adaptation rate (red line) gets higher, while classification accuracy rate for validation dataset was almost unchanged. Therefore, it can be thought that overfitting to the training dataset occurred in the hyper-rectangles generation phase in advance to optimization phase by GA.

6 Conclusion

In HRE algorithm, adjusting the values of hyper-parameters such as the number of clusters is required and complex form clusters cannot be approximated well. It was confirmed that the proposed method has advantages on these points.

However, in the proposed method, processing time grows according to the increase of the dimensionality and/or the number of training data points in order to generate and evaluate many hyper-rectangles. In addition, there is also a problem that the classification accuracy rate for validation dataset is not so high. In order to avoid overfitting to the training dataset, it can be thought that improving the way to expand hyper-rectangles and stopping condition is needed. Performing cross validation test in classification rule extraction phase is also desired.

Applying the proposed method to the learning result of DNN(deep neural network) to extract high interpretable classification rule will be challenged.

References

[1] Arthur L. Samuel, Some Studies in Machine Learning Using the Game of Checkers, *IBM Journal of Research and Development*, 3, **3**, 1959, 210–229.

[2] Christopher M. Bishop, Pattern Recognition and Machine Learning, *Springer*, 2006.

[3] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, *The MIT Press*, 2001.

[4] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar Foundations of Machine Learning, *The MIT Press*, 2012.

[5] V. Vapnik, A. Lerner, Pattern recognition using generalized portrait method, *Automation and Remote Control*, 24, 1963, 774-780.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, AAAI Press, 1996, 226–231.

[7] Y. Zhang, H. Su, T. jia, J. Chu, Rule Extraction from Trained Support Vector Machines, *Proceedings of the Advanced in Knowledge Discovery and Data Minig: Ninth Pacific-Asia Conference PAKDD2005*, 2005, 61–70.

[8] David E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley Professional, 1989.