早稲田大学大学院情報生産システム研究科

博 士 論 文 概 要

論 文 題 目

Parameter Estimation for Binary Classification by Particle Swarm Optimizations and Its Applications

申 請 者
Zhenyuan XU

情報生産システム工学専攻
経営工学研究

2015 年 6 月

In a daily life we face various datasets in different research fields, for example, in business, medical, security and industry. It is important in decision making in many cases to classify or label the data into two groups. This is named binary classification problem. Binary classification is the procedure to classify the instances of a given dataset into two groups based on some classification rules. Binary classification is required when such situation is met for certain purposes: the systems seek to separate two sets of instances in the dataset space, each of them corresponding to a given class, once the dataset is separated under the rule of classification (the boundary), the systems can use it to predict which class a newly coming instance belongs to, by observing which side of the boundary it falls. Briefly, the binary classification is a task of classifying target class (important for certain purpose) from the other class in a dataset. The target problem of this thesis is to select a better classifier to solve difficulties of binary classification, so sufficiently effect and efficient results can be obtained to satisfy the accuracy requirements of decision making. In various real applications, binary classification methods play a core role in research studies. For example, in pattern recognition field, the data of importance should be classified and labeled in a huge dataset. Forecasting based on historical data is also important in a power system, where binary classification can discriminate each datum according to whether or not it follows the historical tendency. In human tracking field to analyze human behaviors, the positive samples should be classified from the negative samples to finish the tracking step.

The development of binary classification technology greatly affects various decision making in human society. Although binary classification plays a pivotal role in decision making in every field of industries and social life, there exist some difficulties. The datasets become incredibly huge and complicated (linear inseparability). Moreover, different production ways changed the datasets into multi-dimensional and quite noisy datasets (complex distribution). In many binary classification problems, the two groups are not symmetric: the relative proportion of different types of instances may be with great imbalance. As performances of binary classification greatly affect the decision making, quite high accurate rate of classification is required, for example, classification for encrypted semiconductor datasets from INTEL is required an accurate rate as 100%.

Such as decision tree, regression analysis and over-sampling technique, the existing methods of binary classifications ignore that both parameter selection and training space selection are important in classifiers to achieve high classification accurate. Instead, they emphasize too much on finding relationships between different instances. Therefore, those conventional methods are sensitive to noise, influenced by raising and reducing instances and overfitting. Most conventional methods did not show satisfied results for the high accuracy requirements.

In this thesis, the particle swarm optimization (PSO) based support vector machine (SVM) systems are proposed to overcome the difficulties introduced above. When the SVM uses hyperplane technique, it obtains good performance for linear inseparability problems. As hyperplane technique only relies on few instances that named support vectors, there is little influence from the complicated distribution of the dataset. But the conventional SVM cannot solve imbalance datasets problem and the accurate rate is not high enough because of its fixed parameters. The PSO, as a higher-level procedure of metaheuristic

framework that can provide a sufficiently good solution to a selection problem method, can search optimal parameters in the SVM. The SVM, as a set of related supervised learning methods used for classification, classifies predicted samples directly from a dataset by hyperplane construction, may efficiently avoid noise intrusion and reduce unnecessary dataset at the same time. It is rarely overfitting. With these advantages of PSO and SVM, an SVM based binary classification with PSO parameter selector may outperform conventional methods. The structure of this thesis can be summarized as follows:

Chapter 1 [Introduction] introduces the background, motivations, objectives and the thesis' structure.

Chapter 2 [preliminary concepts] provides some preliminary concepts for understanding the proposed methods.

Chapter 3 [Approach of the thesis] explains the target problem, difficulties of solving the problem and necessary description of the proposed method. Besides, originality of the improvement done for the proposed method is explained: we improve a double PSO-SVM to solve both parameter selection problems and training space selection problems to achieve the high accurate rate. Also, we originally apply a PSO-NN-SVM model to solve the prediction problem for a huge historical dataset for which the conventional SVM cannot work well. Additionally, we use a proposed PSO-particle filter to improve the search accuracy of PSO itself to satisfy the strict accuracy requirements of binary classifications.

Chapter 4 [solving the imbalanced data classification problem with the particle swarm optimization based support vector machine]: Various machine learning methods have been proposed to solve binary classification problems according to various types of data distributions. Most conventional algorithms attempt to minimize the classification error rate in imbalanced dataset classification and show poor performance, since the difference between the majority and minority classes and the selection of training space are ignored. PSO was proposed as a metaheuristic framework for large and imbalanced dataset classification. Similarly, SVM has a high-level performance in handling balanced binary classifications. Therefore, based on the proposed method, we improve a system named double PSO-SVM to solve both parameter selection problem and training space selection problem. There are two parts of the double PSO-SVM: PSO-1 and PSO-2. In the proposed double PSO-SVM model, PSO-1 identifies an optimal partition of the original training space as a heuristic framework for training space selection, and SVM is used to perform classification within a specific partition. PSO-2 is used to determine the optimal parameters for SVM to obtain sufficient result with high accuracy.

Throughout the experiments, it is observed that the proposed model was able to treat the imbalanced classification problem with high accuracy. True Positive Rate (TPR) and True Negative Rate (TNR) were always higher than 80%. Accuracy of neither TPR nor TNR was ignored. Double PSO-S
is evaluated by comparing obtained results with ANN, SVM-RBF (Radial Basis Function) and LSSVM (Least Square SVM) methods. The double PSO-SVM showed better results with a G-mean 88.4% that outperformed ANN, SVM-RBF and LSSVM were 63.4%, 79.0% and 82.4%.

Chapter 5 [solving short term load forecasting problem by using particle swarm optimization based hybrid neural network and support vector machine]: Short term load forecasting (STLF) has become one of the core research topics for a secure and economic operation of the power system. Several methods have

been applied to improve STLF, as STLF's huge, time-variant and nonlinear characteristics, the result is not sufficient. SVM can almost accomplish all linear or nonlinear programming problems of small datasets, but for its low speed for training, it may be difficult for the proposed method to classify the huge time sequential datasets in STLF. The neural network (NN), based on simple learning role, can finish training with high efficiency, but various conventional methods which satisfied classification of huge datasets like NN can not work well when the information of instances are not abundant. Therefore an updated binary classification system named PSO-NN-SVM based on the proposed system will be introduced to solve those huge dataset by avoiding disadvantages of both NN and SVM in a complemental way. This chapter presents a hybridization of the radial basis function neural network (RBFNN) and SVM with the PSO named PSO-NN-SVM model to ameliorate the limitations of the existing method, at the same time, raise the accuracy and reliability of the forecasting. In the process of both RBFNN and SVM classification, PSO supports a selection framework of optimal parameters for the final prediction. Compared with some conventional algorithms, the PSO-NN-SVM is capable of better performance for the STLF problems.

Our proposed method is applied to forecast hourly load change in a short period with a high accuracy and reliability. This method not only takes advantage from the NN structure for tendency prediction, but also uses a successful binary classification characteristic from the SVM, which efficiently arranges the time series. In the results, a detailed diagram for daily and weekly load changes and a reliable prediction with high accuracy consumption values are obtained with the new method for both daily and weekly. Furthermore, the proposed method obtained a mean absolute percentage error (MAPE) value of 3.20%and a root mean square error (RMSE) value of 3.65% that are the lowest values outperformed other conventional methods such as RBFNN's 6.39%, 7.82%, Kalman filter's 6.63%, 7.20% and PSO-SVM's 14.26%, 17.27%.

Chapter 6 [PSO-particle filter-based biometric measurement for human tracking]: There are a number of works focusing on high requirement of human tracking. However, these works did not provide satisfied results, and some research works even search and use the value of every pixel to enhance the accuracy. It is time consuming that will limit the system. In order to face more strict accuracy requirements, we use a proposed PSO-particle filter to improve the search accuracy of PSO itself in this chapter. A human tracking model which requires quite high accuracy named biometric measurement system is introduced in chapter 6 to check the accuracy of the proposed PSO-particle. The objective of this chapter is to build a mathematical model to measure biometrics in human tracking to mark humans' and objects' size in each frame. To obtain more accurate results for biometric length surveying, the proposed particle swarm optimization (PSO)-particle filter provides a powerful framework for estimating the parameters of support vector machine (SVM) which is used for training and classifying to show a result with higher accuracy.

The experiment results show that in the same situation, the PSO-particle filter ran well, and the detection rate was 96.5% while the particle filter and was 91.7%, respectively. Also, the PSO-particle filter cost the least computation time for the whole processing, which spent only 36.5% of the most time consuming method in these experiments.

Chapter 7 [conclusion and future work]: concludes this thesis and shows the future works.