Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Selection of Candidate Support Vectors in incremental SVM for network intrusion detection☆

*Roshan Chitrakar**, *Chuanhe Huang*

School of Computer, Wuhan University, Wuhan, Hubei, China

ARTICLE INFO

Article history:
Received 24 September 2013
Received in revised form
25 April 2014
Accepted 10 June 2014
Available online 19 June 2014

Keywords:
Incremental support vector machine
Karush–Kuhn–Tucker condition
Candidate Support Vector
Half-partition strategy
Network intrusion detection

ABSTRACT

In an Incremental Support Vector Machine classification, the data objects labelled as non-support vectors by the previous classification are re-used as training data in the next classification along with new data samples verified by Karush–Kuhn–Tucker (KKT) condition. This paper proposes Half-partition strategy of selecting and retaining non-support vectors of the current increment of classification – named as Candidate Support Vectors (CSV) – which are likely to become support vectors in the next increment of classification. This research work also designs an algorithm named the Candidate Support Vector based Incremental SVM (CSV-ISVM) algorithm that implements the proposed strategy and materializes the whole process of incremental SVM classification. This work also suggests modifications to the previously proposed concentric-ring method and reserved set strategy. Performance of the proposed method is evaluated with experiments and also by comparing it with other ISVM techniques. Experimental results and performance analyses show that the proposed algorithm CSV-ISVM is better than general ISVM classifications for real-time network intrusion detection.

1. Introduction

Network intrusion detection is also considered as a pattern recognition problem of classifying the network traffic patterns into two classes – normal and abnormal; according to the similarity between them. Nowadays, in the field of intrusion detection, Support Vector Machine (SVM) is becoming a popular classification tool based on statistical machine learning (Mohammad et al., 2011). There are two issues in machine learning – training of large-scale data sets and availability of a complete data set (Le and Nguyen, 2011; Du et al., 2009a,b). Computer's memory will not be enough and training time will be too long if training data set is very large. Next, when we capture data packets from a stream of a network, we cannot obtain the complete network information in the very first time and hence a continuous online learning is required for high learning precision with increasing number of samples. The challenge of incremental learning is to decide what and how much information from the previous learning should be selected for training in the

☆ This work is supported by the National Science Foundation of China (No. 61373040, No. 61173137), The Ph.D. Programs Foundation of Ministry of Education of China (20120141110073), Key Project of Natural Science Foundation of Hubei Province (No. 2010CDA004).

* Corresponding author.
E-mail addresses: roshanchi@gmail.com, roshanchi@whu.edu.cn (R. Chitrakar), huangch@whu.edu.cn (C. Huang).

http://dx.doi.org/10.1016/j.cose.2014.06.006
0167-4048/

next learning phase and how to deal with new data sets being added in that phase. So, the key of incremental learning is to cope with increasing data samples while retaining the information of original data samples in the meantime.

Most of the intrusion detection methods use non-incremental learning algorithms. With accumulation of new data samples, their training time increases continuously, and at the same time, they have difficulties in adjusting themselves with changing network environment. On the contrary, incremental learning has ability of rapidly learning from new samples and modifying their original model (Yi et al., 2011). Incremental learning methods can better meet requirements of real-time intrusion detection, and can also improve computational accuracy of real-time applications.

A simple Incremental Support Vector Machine (ISVM) algorithm acquires the support vectors by training the initial sample set. Both new data sets and the previous support vectors are merged to form a new sample set, and train them to produce new support vectors. The process is repeated till the final data set (Makili et al., 2013). Chances of new data samples becoming support vectors are tested using Karush–Kuhn–Tucker theory (Wang et al., 2006). Samples that violate the KKT conditions may change the previous support vector set, and hence they are added to previous data set. Samples that meet KKT conditions are discarded because they don't change the previous support vector set (Karasuyama and Takeuchi, 2010).

This paper proposes an improved and efficient learning approach of Incremental Support Vector Machine based on the idea of retaining the original and current data samples throughout the whole learning process. In this proposed approach, data points, that are not the support vectors in the current increment of classification but have chances of becoming support vectors in the next increment of classification, are selected and retained as the Candidate Support Vectors (CSVs) so that they can be combined with the new training data set that will be added in the next training. This approach also introduces Half-partition strategy as a part of the CSV selection method and devises an algorithm named CSV-ISVM using the strategy.

The rest of the paper is organized as follows. Section 2 presents some of the works related to this paper. In Section 3, CSV based incremental SVM is introduced and background knowledge for the approach (viz. KKT-conditions and SVM hyperplane rotations) is also described. Section 4 proposes modification to the concentric circle method and Section 5 explains the main work of this paper i.e. the Half-partition method. All the experiments of this research are illustrated in Section 6 along with necessary analyses. At last, Section 7 draws conclusion of the paper and suggests further work too.

2. Related work

Many research works have been done regarding incremental learning methods using SVM classification and, of course, various modifications have also been proposed. Most of the works seem to have suggested improvements on detection performance, accuracy etc.

A basic incremental SVM based approach called Incremental Batch Learning with SVM was suggested by Liu et al. (2004), in which only the support vectors were preserved for the next increment, while all other data samples were discarded. In fact, the discarded samples carry some amount of information about classification. With the addition of new data samples in the following increments of the learning process, these data samples may become support vectors or vice-versa. Therefore, in simple incremental SVM like this, the classification accuracy is seriously affected in the later increments.

WenJian and Wang explained that data samples lying near the hyperplane had greater possibilities of becoming support vectors after adding new training set (Wang, 2008). So, they proposed the Redundant Incremental Learning Algorithm that retained the redundant samples lying near the hyperplane and added them to the new data sets in the next increment to check if they become the support vectors.

Preserving support vectors in each increment of SVM classification increases the overall classification time. So, execution time reduction becomes primarily important. A typical solution to accelerate SVM training is to decompose the quadratic programming (QP) into a number of sub-problems so that the overall SVM training complexity can be reduced from $O(N^3)$ to $O(N^2)$ (Joachims, 1998; Platt, 1998). However, when the number of data points N is very large, the time complexity is still unsatisfactory and the training needs further improvement (Zhang et al., 2009).

A number of research works have also made contributions in decreasing execution time of incremental SVM classification. An incremental learning approach with SVM proposed by Yao et al. (2012) in classification of network data stream, an incremental learning method proposed by Sun and Guo (2012) in multi-model structure design, a fast incremental learning algorithm for SVM named the Active Set Iteration method (Tao, 2006) etc. are a few among such works.

Again, an incremental learning algorithm of SVM based on clustering was proposed by Du et al. (2009a,b) considering the fact that the boundary support vectors may change into support vectors after adding new samples. This work also used KKT theory in the learning process. The cluster centres obtained from clustering process were checked for the KKT conditions and the cluster centres violating the conditions were added to the support vector set for further training. The necessary and sufficient conditions violating the KKT were given by Wang, Zheng, Wu, and Zhang (Wang et al., 2006). They presented and proved that if there were any new samples contrary to the KKT conditions, then the non-support vectors of the original SVM would have the chance to become support vectors.

An incremental learning algorithm based on the Karush–Kuhn–Tucker (KKT) conditions was also proposed in the literature (Wen-hua & Jian, 2001). In this method, the whole learning process was divided into the initial process and the incremental processes. The optimal solution of the QP allowed each sample to satisfy the KKT conditions.