Asian Journal of Information Technology 7 (11): 489-493, 2008
ISSN: 1682-3915

Optimal Design of Radial Basis Function for Intrusion Detection Data

M. Govindarajan and R.M. Chandrasekaran
Department of Computer Science and Engineering,
University of Annamalai, Annamalai Nagar-608002, Tamil Nadu, India

Abstract: Data Mining is the use of algorithms to extract the information and patterns derived by the knowledge discovery in databases process. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. In many data mining applications that address classification problems, feature and model selection are considered as key tasks. That is, appropriate input features of the classifier must be selected from a given set of possible features and structure parameters of the classifier must be adapted with respect to these features and a given data set. This study describes an Evolutionary Algorithm (EA) that performs feature selection and model selection simultaneously for Radial Basis Function (RBF) classifiers. In order to reduce the optimization effort, various techniques are integrated that accelerate and improve the EA significantly: hybrid training of RBF networks, comparative cross validation. The feasibility and the benefits of the proposed approach are demonstrated by means of data mining problem: intrusion detection in computer networks. It is shown that, compared to earlier RBF technique, the run time is reduced by up to 0.13 and 0.06% while, error rates are lowered by up to 0.01 and 0.01% for normal and abnormal behavior, respectively. The algorithm is independent of specific applications so that many ideas and solutions can be transferred to other classifier paradigms.

Key words: Data mining, classification, radial basis function neural network, run time, error rate, intrusion detection

INTRODUCTION

Information technology has become a key component to support critical infrastructure services in various sectors of our society. In effort to share information and streamline operations, organizations are creating complex networked systems and opening their networks to customers, suppliers and other business partners. While, most users of these networks are legitimate users, an open network exposes the network to illegitimate access and use. Increased network complexity, greater access and a growing emphasis on the internet have made network security a major concern for organizations. The number of computer security breaches has risen significantly in the last 3 years. While, traditional approaches to network security have focused on prevention, network intrusion detection has become increasingly important in recent years to enable firms to reduce undetected intrusion. Typically, network intrusion is detected by examining the data trail left by user and searching for abnormal user behavior.

Hybrid models have been suggested to overcome the defects of using a single supervised learning method, such as multilayer perceptron and radial basis function techniques. Hybrid models combine different methods to improve prediction accuracy. The term combined model is usually used to refer to a concept similar to a hybrid model. Combined models apply the same algorithm repeatedly through partitioning and weighting of a training data set. Combined models also have been called Ensembles. Ensemble improves prediction performance by the combined use of two effects: reduction of errors due to bias and variance (Haykin, 1999).

Hybrid models and combined models, terms often used synonymously, have been developed to improve prediction accuracy by using several supervised learning methods together. Some studies on hybrid or combined models utilize different supervised learning methods sequentially.

In addition to hybrid methods that have tried to combine two completely different methods, hybrid models that use one method in multiple ways have also been studied. Hansen and Salaman (1990) show that the generalization ability of a neural network system can be significantly improved through ensembling a number of neural networks. Indurkhya and Weiss (1998) show the

Corresponding Author: M. Govindarajan, Department of Computer Science and Engineering, University of Annamalai, Annamalai Nagar-608002, Tamil Nadu, India

improvement of predicted gain values of the final nodes in decision trees by multiple re-sampling of decision tree induction methods and combination of them using the voting method. Kuncheva *et al.* (1998) presented cases, in which prediction accuracy was improved using hybrid models. With combinations of RFM, neural networks and logistic regression models, Suh *et al.* (1999) showed that performance of hybrid techniques improves when the correlation between hybrid models is low. Zhang and Zhang (2004) explain that a single data mining technique has not been proved appropriate for every domain and data set. Instead, several techniques may need to be integrated into hybrid systems that can be used cooperatively during a particular data mining operation.

MATERIALS AND METHODS

Dataset used: The data used in this study is based on an immune system developed at the University of New Mexico. It is for one privileged program-send mail (Blake and Merz, 1998). The data includes both normal and abnormal traces. The normal trace is a trace of the send mail daemon and several invocations of the send mail programs. During the period of collecting these traces, there are no intrusions or any suspicious activities happening. The abnormal traces contain several traces including intrusions that exploit well-known problems in Unix systems. For example, *Sunsendmailcp* (SSCP) is a script that send mail uses to append an email message to a file, but when used on a file such as/.rhosts, a local user may obtain root access. Syslog attack uses the syslog interface to overflow a buffer in send mail. Forwarding loops occur in send mail when a set of files in $home/.forward form a logical circle. In our study, intrusion traces include 5 error conditions of forwarding loops, 3 *sunsendmailcp* (SSCP) attacks, 2 traces of the syslog-remote attacks, 2 traces of syslog-local attacks and 2 traces of decode attacks and 2 traces of unsuccessful intrusion attempts-sm565a. Detailed descriptions of these intrusions can be found in Hofmeyr *et al.* (1998). Each trace has 2 attributes: the first one is the process ID, indicating the process the system call belongs to and the second one is the system call value

Comparative cross validation: Holdout, random subsampling, cross-validation and bootstrap are common techniques (Vapnik, 1998). For accessing accuracy based on randomly sampled partitions of the given data. The use of such techniques to estimate accuracy increase the overall computation time, yet is useful for model selection. Apart from these techniques in our case, we have proposed a technique, "comparative cross validation" which, involves accuracy estimation by either stratified k-fold cross-validation or equivalent repeated random subsampling.

As per cross validation initial dataset (S) is divided into parts-training [Str] and test [Stst]. Subsequently, k-fold cross validation should divide data [Str] into a secondary training set [(k-1) folds] and a validation set [1 fold]. After training with cross validation, the overall prediction accuracy for Str was always significantly higher than that of Stst.

By increasing the size of the Str dataset so that it is more representative of the dataset as a whole (S). That is increasing the number of training vectors, we seem to be getting much more similar training/test accuracy results.

Our goal is to calculate the expectation of the classification accuracy, as given by either Stratified k-fold cross-validation or repeated random subsampling (Jiawei and Micheline, 2003). The classification accuracy obtained using Stratified k-fold cross-validation or repeated random subsampling

$$|S|T| = N/K_S$$

where:

N = Size of S (|S|)
c(x) = The class label associated with x
C = Number of class labels in S
Ni = Number of elements in class i

$$Ni = |fx : c(x) = i\}|$$

k = Number of folds in k-fold Cross Validation (CV)

Let, $D = (d_1, d_2 \ldots d_{ks})$ be a partition of S for Stratified k-fold cross-validation.

Two accepted techniques for estimating the generalization accuracy are repeated random subsampling and Stratified k-fold cross-validation. In the former is repeated random subsampling, the validation of holdout method, in which the holdout method is repeated K times. In this hold out method, S is randomly partitioned in to 2 independent sets, a training set and test set. Typically ⅔ of data are allocated to training set and the remaining ⅓ is allocated to test set. The training set is used to derive the model, whose accuracy is estimated with test set.

In latter Stratified k-fold cross-validation, the folds are stratified so that the class distribution of the tuples in each fold is approximately, the same as that in the initial data.

Repeated random subsampling (T) be the classification accuracy (Jovanovic *et al*., 2002) computed by repeated random subsampling with training set T and Stratified k-fold cross-validation (D) be the classification accuracy computed by Stratified k-fold cross-validation with partition D.

Then by definition, we have

$$CV(D) = 1/K_S \sum_{i=1}^{KS} \text{Repeated random subsampling } (S/d_i)$$

The expectation is, by substitution and linearity:

$$E[CV] = 1/K_S \sum_{i=1}^{KS} E[\text{Repeated random subsampling } (S/d_i)]$$

$$= 1/K_S \sum_{i=1}^{KS} E[E[\text{Repeated random subsampling } (S/d_i) \mid di = d]]$$

By Proposition 6.1 in Ross (1988)
Now:

$$E[CV] = 1/K_S \sum_{i=1}^{KS} E[\text{Repeated random subsampling } (S/d)]$$

$$= E[\text{Repeated random subsampling } (S/d)]$$

Because E [Repeated random subsampling (S/d)] is independent of i and E[CV] = E [Repeated random subsampling (T)] by a simple correspondence of a test set d and the training set T = S/d.

Let, T be the set of permissible training sets. The expectation of the classification accuracy using repeated random subsampling is simply the proportion of possible classified (overall T). The number of possible classification is

Σ |S/T|, while the total number of T ε T correct classification is

$$A = \sum_{T \varepsilon T} \sum_{x \varepsilon T} \text{correct}(x, T)$$

Where the binary function, correct (x,T), returns 1 iff x is correctly labeled by a classifier trained on T.

Existing radial basis function: The RBF (Margaret and Dunham, 2003) networks used here may be defined as follows:

- RBF networks have three layers of nodes: Input layer, hidden layer and output layer
- Feed-forward connections exist between input and hidden layers, between input and output layers (shortcut connections) and between hidden and output layers. Additionally, there are connections between a bias node and each output node. A scalar weight is associated with the connection between nodes
- The activation of each input node (fanout) is equal to its external input where is the th element of the external input vector (pattern) of the network (denotes the number of the pattern)
- Each hidden node (neuron) determines the Euclidean distance between "its own" weight vector and the activations of the input nodes, i.e., the external input vector. The distance is used as an input of a radial basis function in order to determine the activation of node. Here, Gaussian functions are employed the parameter of node is the radius of the basis function; the vector is its center
- Each output node (neuron) computes its activation as a weighted sum (Dietterich, 1998). The external output vector of the network, consists of the activations of output nodes, i.e. The activation of a hidden node is high if the current input vector of the network is "similar" (depending on the value of the radius) to the center of its basis function. The center of a basis function can, therefore, be regarded as a prototype of a hyper spherical cluster in the input space of the network. The radius of the cluster is given by the value of the radius parameter. In the literature, some variants of this network structure can be found, some of which do not contain shortcut connections or bias neurons. Parameters (centers, radii and weights) of the RBF (Oliver *et al*., 2005) networks must be determined by means of a set of training patterns with a target vector and (supervised training)

Proposed radial basis function: Evolutionary optimization of RBF architecture is in no way a new idea, but existing approaches (Mitchell, 1997) typically suffer from the problems of a high run time. This study describes an Evolutionary Algorithm (EA) that performs feature selection and model selection (Kohavi, 1995) simultaneously for Radial Basis Function (RBF) classifiers. In order to reduce the optimization effort, various techniques are integrated that accelerate and improve the EA significantly: hybrid training of RBF networks, comparative cross validation. Comparative Cross-validation involves estimation of classification rate by either stratified k-fold cross-validation (Jiawei and Micheline, 2003) or equivalent repeated random subsampling. From the view point of model and

feature selection, this approach can be characterized as hybrid training of RBF networks. Bagging (Naohiro *et al*., 2005) is performed with radial basis function Classifier to obtain a very good generalization performance. The main objective of the hybrid RBF training and comparative cross validation of individuals is a substantial reduction in runtime and error rate. Due to a significantly reduced rum time and a goal-oriented search more and fitter solutions can be evaluated within shorter time. Therefore, it can be expected that better solutions with higher classification rates can be obtained. We show that proposed ensemble of radial basis function classifier are superior to individual approach for intrusion detection problem in terms of classification rate.

RESULTS AND DISCUSSION

In this study, we demonstrated the properties and advantages of our approach by means of normal and abnormal intrusion data sets and also we present the performance of radial basis Function Neural Networks. Here, we constructed the base classifier of radial basis function Neural Network. Comparative cross validation technique is applied to the base classifiers and evaluated run time and error rate. Bagging is performed with radial basis function Classifier to obtain a very good generalization performance. We show that proposed ensemble of radial basis function classifier are superior to individual approach for intrusion detection problem in terms of classification rate (Table 1).

Figure 1 shows the run time for normal and abnormal datasets with existing radial basis function and proposed radial basis function. Figure 2 shows error rate for normal and abnormal datasets using existing radial basis function neural network and proposed radial basis function neural network.

According to Fig. 1, the proposed radial basis function neural network shows better improvement of run time than the existing radial basis function neural network. The run time is reduced by up to 0.13 and 0.06% with respect to proposed radial basis function classifier for normal and abnormal behavior, respectively. According to Fig. 2, the proposed hybrid model shows small reduction in error rate than the base classifiers. The error rate is relatively low by up to 0.01% with respect to proposed radial basis function classifier for both the normal and abnormal datasets. This means that the hybrid method is more accurate than the individual methods.

The experimental results shows that proposed radial basis function classifier is found to be effective compared with existing radial basis function classifier in the intrusion detection dataset in terms of both run time and classification rate.

Table 1: Properties of dataset

System call	Instances	Attributes
Normal	2000	2
Abnormal	373	2

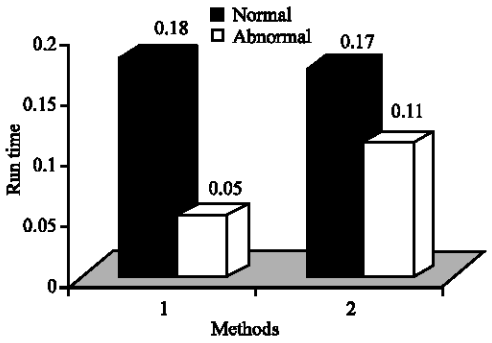

Fig. 1: Run time (Sec)

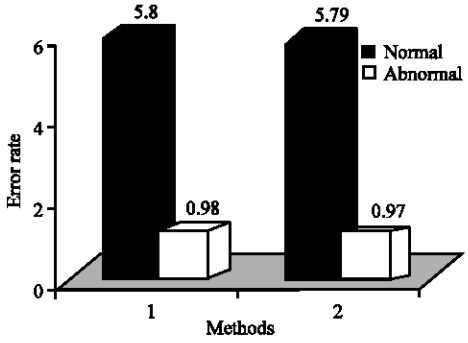

Fig. 2: Error rate (%)

An analysis of development of classification rate shows that the shorter runtimes are possible. The classification error for the intrusion detection problem is low, which indicates the good generalization ability. Finally the improvements to classification rate and run time of the new approach are outlined by means of a comparison to own, earlier approach. Thus, rum time reduction as well as improvements to the classification rate are achieved by combination of various techniques (hybrid training, comparative cross validation).

CONCLUSION

In this research we have investigated new technique for intrusion detection model and evaluated their performance on the normal and abnormal intrusion datasets. We estimated run time and error rate using comparative cross validation method for base classifiers. Following this, we explored the general radial basis function as an intrusion detection model. We have also demonstrated performance comparisons using intrusion detection datasets. The proposed ensemble of radial basis

function combines the complementary features of the base classifiers. Finally, we proposed hybrid architecture involving ensemble and base classifiers for intrusion detection model. From the empirical results, it is shown that, compared to earlier RBF technique, the run time is reduced by up to 0.13 and 0.06% while, error rates are lowered by up to 0.01 and 0.01% for normal and abnormal behavior, respectively. This means that the hybrid method is more accurate than the individual methods. Thus, a suitable compromise between fast search (low run time) and exhaustive search (low classification error) may be effected. The algorithm is independent of specific applications so that many ideas and solutions can be transferred to other classifier paradigms. Our future research will be directed towards developing more accurate base classifiers particularly for the intrusion detection model.

ACKNOWLEDGEMENT

Authors gratefully acknowledge the authorities of Annamalai University for the facilities offered and encouragement to carry out this research. This part of study is supported in part by the first author got Career Award for Young Teachers (CAYT) grant from All India Council for Technical Education, New Delhi. They would also like to thank the reviewer's for their valuable remarks.

REFERENCES

Blake, C. and C. Merz, 1998. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Dietterich, T., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10: 1895-1923.

Hansen, L.K. and P. Salaman, 1990. Neural networks ensembles. Transactions on Pattern Analysis and Machine Intelligence, 12 (10): 993-1001.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. 2nd Edn. New Jersey: Prentice Hall.

Hofmeyr, S.A., S. Forrest and A. Somayaji, 1998. Intrusion detection using sequences of system calls. J. Comput. Security, 6: 151-180.

Indurkhya, N. and S.M. Weiss, 1998. Estimating performance gains for voted decision trees. Intelligent Data Anal., 2 (4): 303-310.

Jiawei, H., 2003. Micheline Kamber Data Mining-Concepts and Techniques. 2nd Edn. Elsevier, pp: 359-367. ISBN: 978-1-55860-901-3.

Jovanovic, N., V. Milutinovic and Z. Obradovic, 2002. Member, IEEE Foundations of Predictive Data Mining.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc. Int. Joint Conf. Artificial Intelligence, pp: 1137-1143.

Kuncheva, L.I., C. Bezdek and M.A. Shutton, 1998. On combining multiple classiers by fuzzy templates. International conference on articial neural networks. IEEE, pp: 193-197.

Margaret and H. Dunham, 2003. Data Mining-Introductory and Advanced Topics. 1st Edn. Pearson Education. Singapore, pp: 112. ISBN: 81-7808-996-3.

Mitchell, T., 1997. Machine learning. New York: McGraw-Hill.

Naohiro, L., S. Eisuke, Yongguangao and Y. Nobuhiko, 2005. Combining Classification Improvements by Ensemble Processing. Proceedings of the 2005 3rd ACIS Int. Conference on Software Engineering Research, Management and Applications (SERA), IEEE Computer Society, 0-7695-2297-1/05$20.00.

Oliver, B., M. Klimek and B. Sick, 2005. Member, IEEE Evolutionary Optimization of Radial Basis Function Classifier for Data Mining Applications. IEEE. Trans. Syst. Man and Cybernets, 35 (5): 928-947.

Ross, S., 1988. A first course in probability. New York: Macmillan.

Suh, E.H., K.C. Noh and C.K. Suh, 1999. Customer list segmentation using the combined response model. Expert Syst. Appl., 17 (2): 89-97.

Vapnik, V., 1998. Statistical learning theory. New York: Wiley.

Zhang, Z. and C. Zhang, 2004. Agent-based hybrid intelligent systems. Berlin, Heidelberg: Springer-Verlag, pp: 127-142.