ISSN: 2319-345X

IJMRBS

International Journal of
Management Research and
Business Strategy

www.ijmrbs.org

E-mail
editor@ijmrbs.org
editor.ijmrbs@gmail.com

ISSN 2319-345X www.ijmrbs.com

Vol. 10, Issuse. 1, Feb 2022

A Hybrid Network Intrusion Detection using Genetic Algorithm

M.Govindarajan

Abstract— Extracting patterns and information from a database using algorithms is known as data mining. Organizing data into a set of distinct categories or subgroups is known as classification. Because the classifications are predetermined before looking at the data, this method is frequently referred to as supervised learning. The selection of features and models is a critical step in many data mining applications that attempt to solve classification challenges. Classifier input features must be selected from a given collection of features, and classifier structure parameters must be adjusted in relation to these features and a particular data set. This is what it means. Genetic Algorithm feature selection and model selection are discussed in this study at the same time (GA). Several strategies that expedite and improve the classifier greatly are incorporated to reduce the optimization effort: hybrid GA, comparative cross validation. Data mining problem: Intrusion Detection in Computer Networks is used to demonstrate the feasibility and benefits of the proposed approach. With normal traces, the suggested Genetic Algorithm reduces the run time by up to 0.24s, and with aberrant traces, it reduces the run time by 4.82s compared to the previous GA technique (GA). It's also worth noting that the proposed hybrid model has a little lower error rate than the original classifiers. In the intrusion detection dataset, the error rate is low by up to 0.9449 percent for normal traces and 0.5859 percent for aberrant traces. Because the approach isn't tied to any one application, it can be applied to a variety of different classifier paradigms.

Keywords-Intrusion Detection, Classification, Genetic Algorithm, and Data Mining.

INTRODUCTION

In a variety of spheres of modern society, information technology has emerged as a fundamental support for essential infrastructure functions. Organizations are constructing complex networked systems and extending their networks to consumers, suppliers, and other business partners in an effort to share information and streamline operations. There are many genuine users of these networks but an open network exposes the network to unauthorized access and misuse. Companies are increasingly concerned about network security due to an increase in network complexity and wider availability as well as a rising reliance on the internet (Denning DE, 1987). Three years ago, the number of computer security breaches was quite low. Network intrusion detection has grown increasingly essential in recent years as a means for companies to reduce undiscovered intrusions, whereas previous approaches to network security have concentrated on prevention. Exploiting the data trails left by users and looking for anomalous user activity are the most common methods of discovering network intrusions.

AssistantProfessor,DepartmentofComputerScienceandEngineeringAnnamalaiUniversity,AnnamalaiN
agar –608002,TamilNadu,India
govind_aucse@yahoo.com

Data mining (Margaret H.Dunham, 2003) has emerged as a highly effective method for reducing
information overload and enhancing decision-making through the extraction and refinement of relevant knowledge from vast amounts of data gathered by organizations. Using the extracted data, it is possible to forecast and categorize, model, and summarize the mined data.

Rule induction and neural networks (Michie, D., Spiegelhalter & Taylor C. 1994), genetic algorithms, fuzzy logic, and rough sets are utilized in various sectors for categorization and pattern recognition. They've been widely utilized to tell apart normal conduct from abnormal behavior in a wide range of situations. Recently, data mining techniques have been employed in the field of network intrusion prevention. When it comes to intrusion detection, an ensemble of genetic algorithms outperforms an individual approach by a wide margin.

A.IntrusiondetectionMethods

The signatures of some attacks are known, where as other attacks only reflect some deviation fromnormal patterns. Consequently, two main approaches have been devised to detect intruders (Dan Zhu and et al.,2001).

1) **AnomalyDetection**

The assumption behind anomaly detection is that any incursion will show some deviations from the usual. This can be separated into static and dynamic anomaly detection. It is assumed that a section of the system being monitored does not change when developing a static anomaly detector. In most cases, static detectors primarily focus on the software side of a system, and they assume that the hardware does not need to be inspected. The code and constant data that make up a system's static section are referred to as its "static portion." For example, the operating system software and data needed to get a computer up and running are always the same. Errors have happened or an intruder has altered the system's static part if it ever deviates from its initial shape. The integrity of a system is the primary concern of static anomaly detectors. Audit records or network traffic data are frequently used for dynamic detection. The audit records of operating systems do not record all events that are documented in the audit, and these events may occur in a sequential order. Detection of events in distributed systems can be achieved with only a partial ordering of events. The order isn't always displayed in this way; sometimes only aggregate data is kept, such as the total amount of processing resources consumed during a period of time. In this situation, thresholds are used to distinguish between normal and abnormal utilization of resources.

2) **MisuseDetection**

Systems with known vulnerabilities and attack patterns can be used to detect misuse. The goal of misuse detection is to identify attackers who are trying to get into a system by exploiting a known vulnerability. In an ideal world, a system security administrator should be aware of all known vulnerabilities and eliminate them. The term "incursion scenario" is used to describe a well-known type of intrusion: it is a sequence of circumstances that would result in an intrusion if no preventive action were taken from the outside. For the sake of preventing the exploitation of known vulnerabilities, an intrusion detection system compares recent activity to known intrusion scenarios. Each possible intrusion scenario must be specified or modeled in detail in order to carry this out. Misuse strategies differ in how they characterize or model what constitutes an invasion of privacy. Rule-based intrusion detection systems were used in the early days of the Internet to specify the events that a security administrator was looking for in the system. The interpretation of a large number of regulations can be complicated. To make rule changes, if-then rules are not grouped by incursion scenarios, therefore the affected rules might be spread out all throughout the rule set. New rule organization and state transition diagrams have been developed to alleviate these issues. It is possible to identify intrusions based on the rules used by misuse detection systems. System calls can be monitored in real time, or audit data can be used later. Cross-validation of base classifier and hybrid classifier error rates and run times is proposed in this research to increase the classification rate.

I. STATEOFART

Inthissection,thestateoftheartconcerningcom parativecrossvalidationofGeneticalgorithmand hybridGAisinvestigated.Theresultsofthissurvey willmotivateanewapproach.

ARelatedWork

To improve prediction accuracy, a variety of supervised learning approaches are combined to create hybrid models and combined models, names that are commonly used interchangeably. Some research on hybrid or blended models employs a sequence of supervised learning approaches. For example, Coenen, Swinnen, Vanhoof, and Wets (2000) offer a hybrid strategy to improve the response rate of direct mailing. A hybrid model comprising the optimal tree model discovered by association analysis with categorical data and the tree model directly applied on continuous variables was used to study the improvement of students' learning capabilities, according to Hsu, Lai, Chui, and Hsu (2003).

The above hybrid models employ a phased approach to include diverse models. As a result, one data mining approach is employed first, and the other way is used in a subsequent phase. There is yet another hybrid technique at work here. To put it another way, a technique or method is incorporated into a primary method in order to enhance the primary method's performance. Using a SOM embedded in fuzzy theory, Chen (2003) proposes a hybrid framework for textual categorization in text mining (self-organized map). An artificial neural network and a genetic algorithm are combined in a model proposed by Versace, Bhatt, Hinds, and Shifier (2003). When used as chromosomes in the genetic algorithm cycle, these neural networks can increase prediction performance on stock closing prices by being created, selected, and regenerated. Some hybrid models also combine the results of two or more approaches to get the final anticipated values. In a study by Lin and McClean (2001), they used a combination of multivariate statistical analysis and artificial intelligence to forecast the likelihood of a company's collapse. By merging the parameters derived from several statistical methods, the model proposed by Conversano, Roberta, and Francesco (2002) is said to improve performance (regression analysis, discriminant analysis, non-parametric statistical method, C&RT, and so on).

Other hybrid models have been studied that employ one method in numerous ways, rather than combining two wholly different methods. The generalization ability of a neural network system can be considerably improved by assembling many neural networks, as demonstrated by Hansen and Salaman (1990). By re-sampling multiple decision tree induction methods and combining them using the voting method, Indurkhya and Weiss (1998) show that predicted gain values of the final nodes in decision trees can be improved. Using hybrid models, Kuncheva, Bezdek, and Shutton (1998) show how prediction accuracy can be increased. RFM, neural networks, and logistic regression models can be used in conjunction. There is no single data mining technique that can be applied to every domain and dataset, according to Zhang and Zhang (2004, Chapter 8). Instead, hybrid systems that can be employed cooperatively during a certain data mining activity may need to be combined with a number of methodologies.

BDatasetUsed

At many levels, the system can be inspected. The decision is influenced by a number of variables, such as price, accuracy, and the system's capacity to discern typical behavior from deviant. Both human behavior and the privileges of privileged programs are typically the focus of intrusion detection systems (MIT Lincoln Laboratory). However, recent research have employed the latter strategy instead of the former, which was more widely used in the past. Privileged processes are applications that require access to system resources that are normally unavailable to the general public.

Due to the narrow range of actions they execute, privileged processes are easier to identify than users with a large range of actions; the range of behavior is limited compared to that of users and is fairly constant over time. The super-user status is required to launch a privileged process on a Unix system. Super-user status gives the user access to a wide variety of permissions, allowing them to carry out operations that are normally prohibited. Normally, the processes are trusted to only access the resources they need, but this trust might be violated if the code is misconfigured or modified.

The Unix process uses'system calls' to access system resources, making it possible to witness the privileged process in action. Short sequences of system calls can be used to identify a variety of intrusions, according to Hofmeyr et al. (1998). When it comes to detecting intrusions, a detection system must be able to distinguish between acceptable and inappropriate behavior while using minimal amounts of computer resources. Information such as temporal parameters, incursion sequence, and interactions with other processes can all be recorded from system calls. The temporal ordering of system calls was employed by Hofmeyr et al. to detect intrusions. Data from system calls are compared to a database of typical behavior in order to discover aberrant behavior in intrusion detection and other data mining processes. As an example, the send mail software generates system calls, which are examined and a database is formed of all unique sequences that are at least one character long.

Based on an immunological system established at the University of New Mexico, the data used in this investigation. Send mail is the only software that can access this privilege. Typical as well as atypical traces are included in the data. The send mail daemon and multiple invocations of the send mail programs are traced in the regular trace. No intrusions or suspicious actions are taking place during this time period. Traces of unusual activity include intrusions that exploit well-known Unix vulnerabilities. Use of SSCP (Sunsendmailcp) on files like /.rhosts, for example, can grant root access to a local user because SSCP appends emails to files. Using syslog, an attacker can overflow a send mail buffer. When a series of files in $home/.forward form a logical circle, a forwarding loop occurs in send mail. Five forwarding loop errors, three sscp attacks, two syslog remote attacks, two syslog local attacks, two decode assaults, and two unsuccessful intrusion attempts—sm565a — were all included in our analysis of intrusion traces. In Hofmeyr et al., you may read more about these incursions (1998). It is possible to identify a system call by its process ID and its value, both of which can be found in a trace.

Thenextsectiondescribesanewcomparativecrossvalidationtechnique.Section4describeshybridmodelforintrusion detection system and Section 5 describes an extensive empirical analysis of classification methods.Section 6 focuses on the experimental results on Existing Genetic Algorithm (EGA) and Proposed GeneticAlgorithm(PGA).Finally,resultsaresummarizedandbriefdescriptionofproposedworkisgiveninsection7.

A. SelectingaTemplate(Heading2)

First, confirm that you have the correct template for your paper size. This template has been tailored foroutput on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file for“MSW_USltr_format”.

B. MaintainingtheIntegrityoftheSpecifications

Formatting and text styling can both be accomplished using a template. Please do not alter any of the prescribed margins, column widths, line spacings, or text typefaces. You'll notice a few oddities. A good example of this is the template's very large head margin. Using standards that anticipate your paper as an integral component of the proceedings rather than a stand-alone document, we took this measurement and others into consideration. All current designations should remain unchanged.

II. COMPARATIVECROSSVALIDATION

Holdout, random subsampling, cross-validation (Kohavi, R, 1995) and bootstrap are common techniques foraccessing accuracy based on randomly sampled partitions of the given data. The use of such techniques toestimate accuracy increase the overall computation time yet is useful for model selection. Apart from thesetechniques in our case, we have proposed a technique, “comparative cross validation” which involves accuracyestimationbyeitherstratifiedk-foldcross-validationorequivalentrepeatedrandomsubsampling.Aspercrossvalidation, initial dataset (S) is divided into parts - training [Str] and test [Stst]. Subsequently, k-fold crossvalidationshoulddividedata[Str]intoasecondarytrainingset[(k-1)folds]andavalidationset[1fold].Aftertraining with cross validation, the overall prediction accuracy for Str was always significantly higher than that ofStst.ByincreasingthesizeoftheStrdatasetsothatitismorerepresentativeofthedatasetasawhole(S).Thatis increasing the number of training vectors seem to be getting much more similar training / test accuracyresults.

Thegoalistocalculatetheexpectationoftheclassificationaccuracy,asgivenbyeitherStratifiedk-foldcross-validation or repeated random subsampling (Jiawei Han, Micheline Kamber 2003). The classification accuracyobtainedusingStratifiedk-foldcross-validationorrepeatedrandomsubsamplingwhere|S|T|=N/KS

N-SizeofS(|S|)
c(x)- The class label associated with xC-NumberofclasslabelsinS
Ni-Numberofelementsinclassi.
Ni= |fx :c(x)= i}|
k-Numberoffoldsink-foldcrossvalidation(CV).
LetD=(d1,d2…dks)beapartitionofSforStratified k-foldcross-validation

Two accepted techniques for estimating the generalization accuracy are repeated random subsampling andstratifiedk-foldcross-validation.Theformerisrepeatedrandomsubsampling,thevalidationofholdoutmethodinwhich theholdoutmethodisrepeatedKtimes.Inthisholdoutmethod,Sisrandomlypartitionedintotwoindependent sets, a training set and test set. Typically two third of data are allocated to training set and theremaining one third is allocated to test set. The training set is used to derive the model, whose accuracy isestimated with test set.In latter Stratified k-fold cross-validation, the folds are stratified so that the classdistributionofthetuplesineachfoldisapproximatelythesameasthatintheinitialdata.

Repeated random subsampling (T) be the classification accuracy computed by repeated random subsamplingwith training set T and stratified k-fold cross-validation (D) be the classification accuracy computed by stratifiedk-foldcross-validationwithpartitionD.

Thenbydefinition,
Ks
CV(D) 1Ks Repeatedrandomsubsampling(Sdi)
i 1
Theexpectationis,bysubstitutionandlinearity:

Ks
 [CV] 1Ks E[Repeatedrandomsubsampling(Sdi)]
i1
ByProposition6.1in Ross,1988(p.285).

Ks
 1Ks E[E[Repeatedrandomsubsampling(Sdi)di d]]i 1

Now:
Ks

 [CV] 1Ks E[Repeatedrandomsubsampling(Sd)]i 1

 E[Repeatedrandomsubsampling(Sd)]

Because E [Repeated random subsampling (S/d)] is independent of i and E[CV] = E [Repeated randomsubsampling(T)]byasimplecorrespondenceofatestsetdandthetrainingsetT=S/d.

III. HYBRIDMODELFORINTRUSIONDETECTION

Voting is a simple and popular hybrid model for combining the results of several methods (Ali & Pazzani, 1996).In the case of classification, for a tiebreak, the prediction probabilities of each method are calculated andconsidered to make final predictions. Bagging (bootstrap aggregation) and Boosting are commonly usedtechniques for combined models. Bagging generates multiple training data sets by bootstrapping (resamplingrandomly with replacement), and combines the results of modeling with each separated set (Brieman, 1996).Brieman (1996) reports that prediction accuracy can be improved from 57% to 94% by applying Bagging to theC&RT algorithm. In summary, Bagging is one of the methods for improving prediction performance bydeducing not a single logic but multiple logics from a data set, combining them, and supplementing themisclassified portion. Bagging utilizes data set separation through arbitrary extraction and merges using a simplevotingmethod.Asmentionedearlier,thedifferencebetweenhybridmodelsandcombined modelsisgenerallythat hybrid models use different learning methods in a mixed manner, whereas combined models generatemultiple models using one learning method with repeatedly sampled data sets and combining the generatedmodels. Even though the approach of the two methods is a little different, the basic assumption of the twomethodsisthatsupervisedlearningalgorithmshaveselectivesuperiority,andtheirobjectivesarethesame:toimprove the accuracy of final data mining results. That is, a final data mining model that has higher accuracy canbegeneratedbydisclosingdiverselogics(rules)usingasinglemethod(philosophyofcombined model)orbycombiningcompletelydifferentmodels(philosophyofhybridmodel).

ADefinitionandNotations

Ensemble approaches, such as bagging and boosting (David Opitz and Richard Maclin, 1999), make use of several models. A composite model, M*, is created by combining a set of M1-MK trained models (classifiers or predictors) into a single, better model. Classification and forecasting can both benefit from bagging and boosting (Nikunj C. Oza, Kagan Tumer, 2008).BBasicidea

As a "bootstrap" ensemble method, bagging (Jiawei Han and Micheline Kamber, 2003) trains each classifier on a random redistribution of the training set, resulting in an individual for the ensemble. It is possible that some of the original instances will be repeated in the final training set, while others may be omitted. The training set for each classifier is formed by randomly picking N examples with replacements, where N is the number of examples in the original training set. Random sampling is used to generate each individual classifier in the ensemble.

CProceduresofhybridmodelingusingbaggingclassifiers

Givenaset,D,ofdtuples,baggingworksasfollows.ForiterationI(i=1,2,…k),atrainingset,Di,oftuplesissampled with replacement from the original set of tuples, D. Note that the term bagging stands for bootstrapaggregation.

Bootstrap samples are used in every training set. In Di, some of the original tuples of D may not be included, while others appear many times because of sampling with replacement. Each training set (Di) yields a classifier model (Mi). Each classifier, M*, collects the votes and assigns X to the class with the most votes in order to categorize an unknown tuple, X. In the case of continuous values, a method known as "bagging" can be used to take the average of each forecast for each test tuple.

In many cases, the bagged classifier outperforms a single classifier trained solely on data set D. It is less sensitive to the impacts of noisy data and will not be much poorer. As a result of a reduction in the separate classifiers' variance, the composite model achieves its improved accuracy. Theoretically, a bagged predictor will always outperform a single D-derived predictor in terms of accuracy. As a result, this research proposes a hybrid approach using bagging and a genetic algorithm..

Algorithm:Bagging.

The bagging algorithm creates an ensemble of models (classifiers or predictors) for a learning scheme whereeachmodelgivesanequallyweightedprediction.

```
Bagging({(x1,y1),(x2,y2),.           ,(xN,yN)},M)
Foreachm=1,2,….,M
Tm=Sample_With_Replacement({(x1,y1),(x2,y2),.,(xN,yN)},N)
hm=Lb(Tm)

Returnhfin(x) argmaxy Y



 I(hm(x) y)
m 1

Sample_With_Replacement(T,N)
S {}
Fori=1,2,….,N
R=random_integer(1,N)AddT[r]toS.
ReturnS.
```

Weights are assigned to each training tuple in boosting (An H.Witten and Eibe Frank, 2005). Iteratively, a set of k classifiers are learnt. Classifier Mi's weights are adjusted after learning so that Mi+1 can "pay more attention" to the training tuples that Mi misclassified. It is possible to merge the votes of all the different classifiers into a single boosted classifier, which is called M*.

Continuous values can be predicted using the boosting technique. A number of scholars have looked into the creation of an ensemble classifier by combining various classifiers together (Haykin, S., 1999). Improved overall generalization is a major benefit of combining redundant and complementary classifiers.

V CLASSIFICATIONMETHODS

A. ExistingGeneticAlgorithm(EGA)

A quick overview of genetic algorithms is provided in this section. As a machine learning model, the genetic algorithm is based on metaphors for some of nature's methods of evolution. By creating a population of individuals, which are actually character strings represented by chromosomes, within a machine, this is accomplished.

Candidates for the answer to the optimization issue are represented by the persons. M.Mitchell, 1996) uses -bit binary vectors to represent people in evolutionary algorithms. The resulting search space is analogous to a boolean space with dimensions –1..–2.. An assumption is made here: a fitness function can be used to assess the quality of each possible solution to the problem at hand.

Some form of fitness-dependent probabilistic selection of individuals from the current population is used by genetic algorithms (Man K.F., Tang S., and K.W., 1999) to produce individuals for the next generation. To create the next generation, the selected individuals are subjected to the actions of genetic operators. Operators like mutation and crossover are often utilized in genetic algorithms that represent individuals as binary strings. A single string can be mutated at random, whereas two parent strings can be crossed to produce two children. Using the correct genetic operators is necessary for other genetic representations.

Selection and application of genetic operators to produce consecutive generations is repeated until a good solution is discovered. On the real-world level, a number factors influence genetic algorithm performance, including but not limited to genetic representation and operators, fitness function details and fitness-dependent selection procedure details and various user-specified parameters like population size and the likelihood that different genetic operators will be used. The following is an explanation of how the genetic algorithm works:

Procedure:

```
begint<-0
```

```
initializeP(t)
while(notterminationcondition)t<-t+1
selectP(t)fromp(t-1)crossoverP(t)
mutate P(t)evaluateP(t)
endend
```

Sincegeneticalgorithms(J.YangandV.Honavar,1998)weredesignedtoefficientlysearchlargespaces,theyhave been used for a number of different application areas such as camera calibration (Q.Ji and Y.Zhang, 2001),signatureverification,medicaldiagnosis,facialmodelingandhandwrittenrecognition.

1) RepresentationandOperators

Inthissubsectionwepresentthechoiceofarepresentationforencodingcandidatesolutionstobemanipulatedbythegeneticalgorithm(Hsuandetal, 2003).

The feature subset selection problem can be solved by any one of the individuals in the population. There are an infinite number of features that can be used as a representation of a pattern. A binary vector of dimension m represents an individual (chromosome). Bits with a value of 1 indicate that the relevant feature is selected; if the value is 0, the feature is not selected. I think this is the most simple and straightforward representation method. Other genetic representations, as previously noted, necessitate the use of appropriate genetic operators.

A binary string is used to represent a chromosome, therefore the operator's mutation and crossover works as follows: On a single string, mutation is able to modify things up a bit at random. As a result, a random mutation could result in the string 11010 becoming 11110. In order to produce two children, you must cross over two parent strings When the fourth crossing position is randomly selected, the two strings 01101 and 11000 give rise to the children 01100 and 11001.

2) ParameterSettings

Ourexperimentsusedthefollowingparametersettings:

Population size: 100Number of generation: 20Probabilityofcrossover:0.9
Probabilityofmutation:0.07
Theparametersettings(A.E.Eibenandetal,1999) werebasedonresultsofseveralpreliminaryruns.

3) FeatureEnsembleSelection

Themainideaofensemblemethodologyistocombineasetofmodels,eachofwhichsolvesthesameoriginaltask,inordertoobtainabettercompositeglobalmodel,withmoreaccurateandreliableestimatesordecisionsthatcanbemadefromusingasinglemodel.Someofthedrawbacksofthefilters(RokachL.,ChiziB.,MaimonO,2007)andwrapperscanbesolvedbyusingensemble.

4) ObjectiveFunctionandFitnessEvaluation

Thefitnessevaluationisamechanismusedtodeterminetheconfidenceleveloftheoptimizedsolutionstotheproblem.Usually,thereisafitnessvalueassociatedwitheachchromosome,e.g.,inaminimizationproblem,alower fitnessvalue means that the chromosome or solution is more optimized to the problem while a highervalue of fitness indicates a less optimized chromosome. Our problem consists of optimizing two objectives:Minimizationoftheerrorrateandtherebymaximizingtheclassificationrateoftheclassifier.

BProposed Genetic Algorithm(PGA)

EvolutionaryoptimizationofGAarchitectureisinnowayanewidea,butexistingapproachestypicallysufferfromtheproblemsofahighruntime.

This paper describes an evolutionary algorithm (EA) that performs feature selection and model selectionsimultaneouslyforGeneticAlgorithm(GA).Inordertoreducetheoptimizationeffort,varioustechniquesareintegrated that accelerate and improve the EA significantly: hybrid training of GA, Comparative Crossvalidation. Comparative Cross-validation involves estimation of classification rate by either stratified k-foldcross-validationorequivalentrepeatedrandomsubsampling.Theerrorrateandruntimeofthebaseclas sifierareestimatedusingcomparativecrossvalidation.BaggingisperformedwithGeneticAlgorithm(GA)toobtainaverygoodgeneralizationperformance.ThemainobjectiveofthehybridGAandcomparativecrossvalidationofindividualsisasubstantialreductioninerrorrateandruntime.Duetoasignificantlyreducedrumtimeandagoal-oriented search more and fitter solutions can be evaluated within shorter time. Therefore, it can beexpected that better solutions with higher classification rates can be obtained. It is shown that proposedensembleofGeneticalgorithmissuperiortoindividualapproachforintrusiondetectionintermsofclassificationrate.

VI RESULTSANDDISCUSSION

Thissectiondemonstratesthepropertiesandadvantagesofproposedapproachbymeansofintrusiondetectiondata set with two categories: normal traces and Abnormal Traces. The normal traces contain 373 instances and 2attributes.Similarlytheabnormaltracescontain2000instancesand2attributes.ItalsopresentstheperformanceofGeneticAlgorithm(GA)intermsofruntimeanderrorrate.Here,thebaseclassifierofGeneticAlgorithm(GA)isconstructed.Comparativecrossvalidationtechniqueisappliedtothebaseclassifiersandevaluatedruntime and error rate. Bagging is performed with Genetic Algorithm (GA) to obtain a very good generalizationperformance.WeshowthatproposedensembleGeneticAlgorithm(GA)issuperiortoindividualapproachforintrusiondetectionintermsofclassificationrate.

TABLE1:RUNTIMEANDERRORRATEFORGENETICALGORITHM

Intrusiondetection	Existing GeneticAlgorithm		Proposed GeneticAlgorithm	
	RunTime(Seconds)	ErrorRate (%)	RunTime(Seconds)	ErrorRate(%)
NormalTraces	1.08	7.1803	0.84	6.2354
AbnormalTraces	8.06	1.6120	3.24	1.0261

Table1showstheruntimeanderrorrateforintrusiondetectiondatasetwithexistingGeneticAlgorithm(GA) and proposed Genetic Algorithm (GA). According to table 1, the proposed Genetic Algorithm (GA) showsbetter improvement of run time than the existing Genetic Algorithm (GA). The run time is reduced by up to0.24swithrespecttonormaltracesand4.82swithabnormaltracesfortheproposedGeneticAlgorithm(GA).Similarly,theproposedhybridmodelshowssmallreductioninerrorratethanthebaseclassifiers.Theerrorrateis relatively low by up to 0.9449 % with respect to normal traces and 0.5859 % with abnormal traces. This meansthatthehybridmodel(M.Govindarajanandetal,2011)ismoreaccuratethantheindividualclassifier.

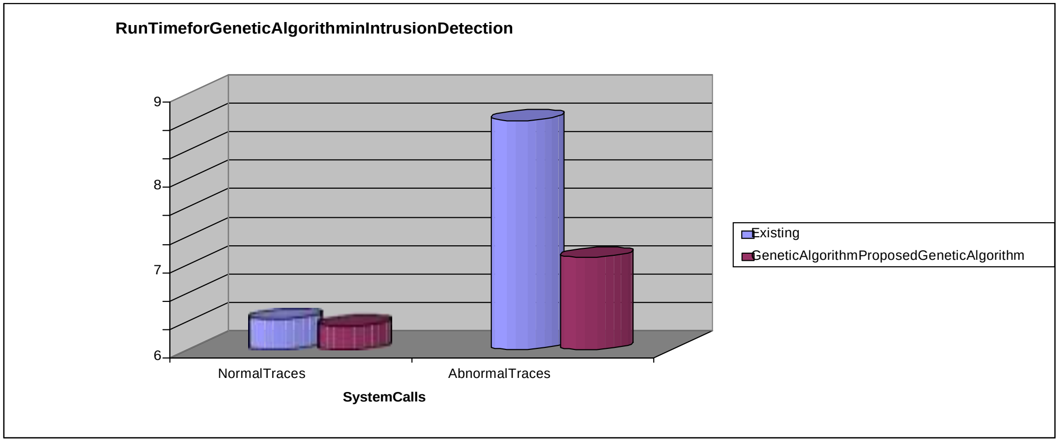

Figure2:RunTime(Seconds)

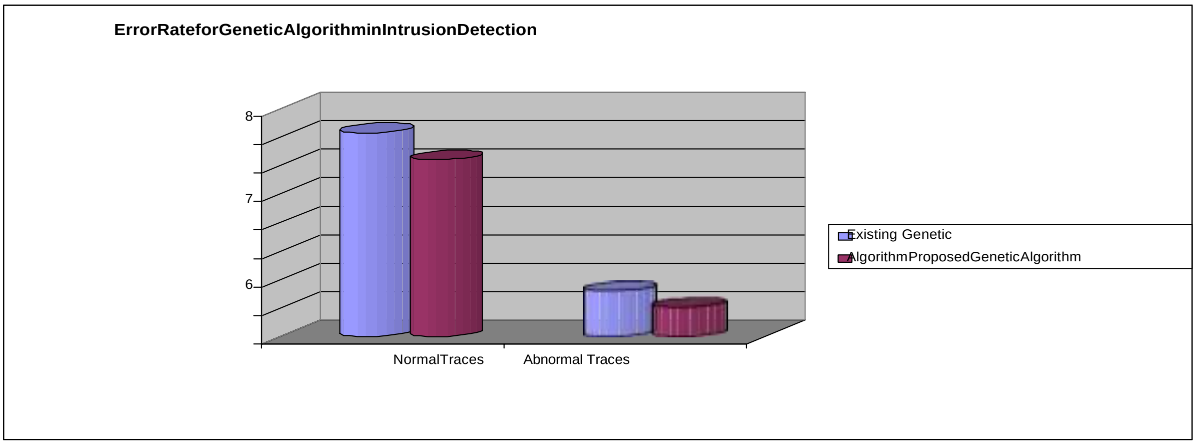

Figure3:ErrorRate(%)

In this case, the chi-squared statistic χ2 is determined and the critical value is found to be 0.5882 which liesbetween 0.455 – 2.706. The corresponding probability lies between 0.5 < p < 0.1. This is smaller than theconventionally accepted significance level of 0.05 or 5 %. Thus examining a chi-squared significance table, it isfound that this value is significant with a degree of freedom of 1. Thus the results of chi-squared statistic analysisisshowsthattheproposedGeneticAlgorithm(GA)issignificantatp<0.05thantheexistingGeneticAlgorithm(GA).

TheexperimentalresultsshowsthatproposedGeneticAlgorithm(GA)isfoundtobeeffectivecomparedwithexistingGeneticAlgorithm(GA)intheintrusiondetectiondatasetintermsofbothruntimeandclassificationrate. An analysis of development of classification rate shows that the shorter runtimes are possible. Theclassification error for intrusion detection is low which indicates the good generalization ability. Finally theimprovementstoclassificationrateandruntimeofthenewapproachareoutlinedbymeansofacomparisontoown,earlierapproach.Thusrumtimereductionsaswellasimprovementstotheclassificationrateareachievedbycombinationofvarioustechniques(hybridtraining,comparativecrossvalidation).

VII **CONCLUSION**

In this study, novel techniques for intrusion detection datasets were studied and their performance was evaluated. Comparative cross validation is used to determine the execution time and error rate for base classifiers. An ensemble of Genetic Algorithms (GA) has been presented that incorporates the best aspects of the existing classifiers.

It is proposed to use ensemble and base classifier architectures in combination for intrusion detection datasets. The experimental results reveal that the suggested Genetic Algorithm (GA) has a run time reduction of up to 0.24s for normal traces and 4.82s for abnormal traces when compared to the previous GA technique (GA). Normal trace errors are less than 0.9449 percent; aberrant trace errors are less than 0.5859 percent, making for a low error rate across the board. Even while the difference in error rate is typically tiny, it may be statistically significant at p0.05, the suggested technique's run time is found to be significantly lower than the present technique's. As a result, the hybrid approach is more precise than each of the component approaches alone. An appropriate balance between speed and thoroughness can be achieved, resulting in a low categorization error rate. Because the approach isn't tied to any one application, it can be applied to a variety of different classifier paradigms. The focus of future work will be on improving the accuracy of the base classifiers used in intrusion detection models.

ACKNOWLEDGMENT(HEADING5)

AuthorsgratefullyacknowledgetheauthoritiesofAnnamalaiUniversityforthefacilitiesofferedandencouragementtocarryoutthiswork.

REFERENCES

[1]Ali,K.,&Pazzani,M.(1996).Errorreductionthroughlearningmultipledescriptions.MachineLearning,24(1):105–112.

[2] Brieman,L.(1996). Baggingpredictors. MachineLearning,24(2):123-140.

[3] Chen,Y.P.(2003).AhybridframeworkusingSOMandfuzzytheoryfortextualclassificationindatamining.Modelingwith Words, LNAI2873,153-167.

[4] Coenen,F.,Swinnen,G.,Vanhoof,K.,&Wets,G.(2000).Theimprovementofresponsemodeling:combiningrule-inductionandcase-basedreasoning,ExpertSystemswithApplications,18(4):307-313.

[5]C.Conversano,S.Roberta,M.Francesco,Generalizedadditivemultimixturemodelfordatamining,ComputationalStatistics andDataAnalysis38 (4) (2002) 487–500.

[6DanZhuandG.Premkimar,XiaoningZhang,Chao-HsienChu,(2001).DataMiningforNetworkIntrusionDetection:AComparisonofAlternativeMethods”,DecisionSciences,32(4),(2001)635–660.

[7]DenningDE.Anintrusion-detectionmodel.(1987).IEEETransactiononSoftwareEngineering;SE-13(2):222-32.

[8] David Opitz, Richard Maclin,(1999).PopularEnsembleMethods:An EmpiricalStudy,JournalofArtificialIntelligenceResearch 11, 169-198.

[9]A.E.Eiben,R.Hinterding,andMichalewicz.(1999).Parametercontrolinevolutionaryalgorithms.IEEETrans.onEvolutionaryComputation, 3(2):124–141.

[10]M.Govindarajan,RM.Chandrasekaran.(2011),Intrusiondetectionusingneuralbasedhybridclassificationmethods, ComputerNetworks,55(8):1662-1671

[11]Hansen,L.K.,&Salaman,P.(1990).Neuralnetworksensembles.TransactionsonPatternAnalysisandMachineIntelligence,12(10):993–1001.

[12]Haykin,S.(1999).Neuralnetworks:acomprehensivefoundation(secondedition).NewJersey: PrenticeHall.

[13]Hofmeyr,S.A.,Forrest,S.,&Somayaji,A(1998).Intrusiondetectionusingsequencesofsystemcalls.Journalofcomputer Security,6, 151-180.

[14]Hsu,P.L.,Lai,R.,Chui,C.C.,&Hsu,C.I.(2003).Thehybridofassociationrulealgorithmsandgeneticalgorithmfortreeinduction:anexampleofpredictingthestudentcourseperformance.ExpertSystemswithApplication,25(1):51-62.