VOL. 10, NO. 16, SEPTEMBER 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A NOVEL HYBRID MEDICAL DIAGNOSIS SYSTEM BASED ON GENETIC DATA ADAPTATION DECISION TREE AND CLUSTERING

P. S. Jeetha Lakshmi[1] and S. Saravan Kumar[2] and A. Suresh[3]
[1]St.Peter's University, Chennai, India
[2]Panimalar Institute of Technology, Chennai, India
[3]S.A.Engineering College, Chennai, India
E-Mail: jeethaps@gmail.com

ABSTRACT

The medical expert system is a special type of recommender systems that plays a major role in decision making process by the medical doctors nowadays. This kind of expert systems often provides the medical diagnosis activity based on the handling various patients in various situations by the medical doctors and clinical symptoms of patients to give a list of possible diseases attended with the membership values. Many acquiring diseases from that list are then determined by medical doctors experience expressed through specific combinations of features in the clinical dataset. The major issue of the expert system is increasing the accuracy of the medical diagnosis attributes that involves the cooperation of decision making systems and recommender systems in the sense that predict the behaviors of disease symptoms and the doctors experience are represented by rules whilst the prediction of the possible diseases is identified by the prediction capability of medical expert systems. From the past observation, the accuracy of features similarity could be improved by the integration with the information of possibility of patients belonging to clusters specified by a weighted K-means clustering method. For improving the performance of medical expert system, a new hybrid intrusion detection framework is introduced to improve the classification accuracy. This hybrid system is combining the proposed genetic based data adaptation decision tree (GDADT) and the existing weighted K-Means clustering. Moreover, we have used the existing cluster and decision tree based classifier called Intelligent Agent based Enhanced Multiclass Support Vector Machine (IAEMSVM) for improving the prediction accuracy. The experimental results of the proposed system show that this system achieved high-detection rate with less time and low false alarm rate when tested with UCI Machine Learning data set.

Keywords: weighted K-Means clustering, medical diagnosis, genetic based data adaptation tree, support vector machine.

INTRODUCTION

Recently, the medical expert system or the clinical diagnosis system has emerged as an important tool in medical sciences to assist medical doctors in decision making especially identifying the diseases from the past symptoms and their experience in the particular area. Medical doctors, nurses and other medical professionals use the medical expert system to prepare a diagnosis and to review the diagnosis as a means of improving the final result. According to [R.Sethukkarasi*et al*., 2014] [S.Ganapathy*et al*, 2014], the medical expert system can be defined as software package that helps to the doctors for improved decision-making process by providing proof according to the patient data. This kind of medical application consists of three major components such as language system, a knowledge system and a processing system for solving the problem. This medical application helps to handle complex problems, applying domain-specific expertise to measure the significances of executing its recommendations.

Machine learning methods are used to examine patients' medical treatment records in aggregation with relevant medical researches, which are able to predict potential events ranging from drug interactions to disease symptoms. Utilizing the medical diagnosis process, characteristics of an individual patient are matched to a computerized medicalknowledge base and patient-specific assessment and recommendations are then presented to the medical diagnosis or the particular patient for a decision [R.Sethukkarasi*et al*., 2014] [S.Ganapathy*et al*, 2014]. The major issue of medical expert system is increasing the accuracy of the medical diagnosis. The existing researchers concentrated on improving the performance of machine learning methods in medical diagnosis process.

Clustering are used to perform investigative data analysis technique, it attempts to partition a given data set into dissimilar groups such that data patterns within a group are more similar to one another than those belonging to different groups. The clustering techniques are categorized into supervised and unsupervised methods.

The unsupervised clustering methods are used to detect the underlying structure in the data set for classification. Supervised clustering method is involved with the human interaction. The unsupervised clustering techniques are quite famous because of the reason that they do not need much knowledge about the data sets [Reda M Elbasiony*et al*, 2013]. The various clustering algorithms like K-Means s [Reda M Elbasiony*et al*, 2013] and Weighted K-Means [Reda M Elbasiony*et al*, 2013]. Analysis is very important for identifying the diseases in huge volume of medical data.

Classification [R.Sethukkarasi*et al*., 2014] [S.Ganapathy*et al*, 2014] [Ganapathy S et al, 2012] is used to train a model called classifier from a set of labeled data instances called training and then to classify a test instance into one of the classes using the learned model known as testing. Classification-based medical diagnosis systems operate in a similar two phase fashion. The training phase learns a classifier using the available labeled training data. The testing phase classifies a test instance as normal or disease affected, using the classifier. Classification-based diagnosis systems operate under either one-class classifier or multi-class classifier. One-class-classification-based disease identification methods assume that all training instances have only one class label. Such techniques learn a discriminative boundary around the normal instances using a one-class classification algorithm. Any test instance that does not fall within the learned boundary is declared as disease affected data. Multi-class classification-based medical expert systems assume that the training data contains labeled instances belonging to multiple normal classes [Ganapathy S et al, 2012]. Such anomaly detection techniques teach a classifier to distinguish between each normal class and the rest of the classes. Many classification techniques such as decision trees [Dewan Md. Farid*et al*., 2014], SVM [S Mukkamala*et al*., 2003], MSVM [SA Mulay*et al*., 2010], EMSVM [S Ganapathy*et al*., 2011], neural networks (NN) [S.Ganapathy*et al*, 2014], Naive Bayes [Dewan Md. Farid*et al.*, 2014] and incremental Particle swarm Optimization (PSO) [Chun-Wei Tsai *et al.*, 2013] have been proposed by various researchers for detecting the disease affected records in clinical data.

In this paper, a new framework has been designed for detecting intruders effectively in computer networks. This framework is the combination of the proposed Genetic based Data Adaptation Decision Tree (GDADT) and the existing Weighted K-Means Clustering [11]. In addition to that we have used the existing cluster based decision tree called IAEMSVM algorithm for better classification accuracy. Rest of this paper is organizes follows: Chapter 2 discusses about various past works done in this direction. Chapter 3 explains the overall system architecture. Chapter 4 described the proposed method. Chapter 5 contains the results and discussion. Chapter 6 gives the conclusion and future works.

LITERATURE SURVEY

There are many works have been proposed for medical diagnosis using feature selection, classification and clustering techniques by various researchers in the past. Among them, [Tomoharu Nakashima *et al*., 2005] introduced a novel approach to fuzzy logic based classifier with weighted training patterns for medical diagnosis. Their classifier provides better classification accuracy on medical dataset. [M. Benkaci*et al*., 2010] proposed a fuzzy-ARTMAP classification for medical diagnosis.[J. Novakovic*et al*., 2011] investigated the impact of kernel function and parameters of C-Support Vector Classification (C-SVC) to solve biomedical problems in a variety of clinical domains. Their experimental results demonstrated the effectiveness of optimizing parameters for C-SVC with different basic kernel. [Agnieszka*et al*., 2014] proposed the methodology of searching for features which are less informative while considering independently, but still meaningful in the process of diagnosis. Their approach is mainly useful when new attributes derived from new diagnostics techniques are introduced.

An intelligent knowledge representation model has been proposed [Sethukkarasi*et al* 2014] medical diagnosis and decision support system. They designed four layer temporal neuro fuzzy inference networks with linguistic fuzzifier and temporal fuzzy mutual subset hood activation spread, for automatic identification and quantification of causalities. They identified the cause effect relationships automatically which reduces the number of iterations and obtained accurate result. Their model tested with diabetic dataset and it reduces the excessive dependence on expert knowledge and provides better performance for prediction than the existing approach. A pattern classification and rule extraction system called TFMM-PSO has been proposed by [Ganapathy*et al* 2014]. Their proposed system comprises a Temporal FMM network and a PSO rule extractor. The performances of the proposed system evaluated using various problems and a real medical diagnosis task. Their system achieved better detection accuracy and extracted comprehensible rules.

[Nguyen and Le *et al*., 2015] proposed a novel hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical

diagnosis so-called HIFCF (Hybrid Intuitionistic Fuzzy Collaborative Filtering). Their experimental results reveal that HIFCF obtains better accuracy than the existing IFCF and other classifiers. The same authors [Le Hoang Son *et al*., 2015] concentrated on the problem of enhancing the accuracy of medical diagnosis and presented a novel intuitionistic fuzzy recommender system (IFRS) consisting of the new definitions of single-criterion IFRS (SC-IFRS), multi-criteria IFRS (MC-IFRS), intuitionistic fuzzy matrix (IFM), intuitionistic fuzzy composition matrix (IFCM), intuitionistic fuzzy similarity matrix (IFSM) and the intuitionistic fuzzy similarity degree (IFSD). Some interesting theorems and properties of the proposed components were also investigated. The proposed IFCF algorithm was used mainly for the medical diagnosis problem.

[Dewan Md. Farid*et al.* 2003] introduce two independent hybrid mining algorithms to improve the classification accuracy rates of decision tree (DT) and naive Bayes (NB) classifiers for the classification of multi-class problems. [LeventKoc*et al* 2012] explained the necessary to apply data mining methods to classify network attacks. They summarized their results of existing methods on the performance improvement of the Naïve Bayes model in data mining. They proved that the HNB model is a best method to achieve better detection accuracy to the intrusion detection. A hybrid intrusion detection framework was proposed [Reda*et al*., 2013] that depends on classification and clustering techniques. They used random forests classification algorithm for detecting misuse detection and used weighted k-means for detecting anomalies in computer networks. [Gisung Kim *et al*., 2014] proposed a new hybrid intrusion detection method that hierarchically integrates misuse detection and an anomaly detection model. First, the C4.5 decision tree (DT) was used to create the misuse detection model for decomposing the training data into several subsets. In each decomposed region, the one-class support vector machine was used to create anomaly detection. A Novel hybrid KPCA SVM with GAs model is proposed for intrusion detection [FangjunKuang*et al*., 2015]. In their model, KPCA is adopted to extract the necessary features of intrusion detection data, and multiclass SVM is employed to estimate whether the action is an attack or not.

In this paper, a new hybrid medical expert system has been proposed for effective intrusion detection which is the combination of the proposed Genetic based Data Adaptation Decision Tree (GDADT) and the existing weighted K-means clustering [Reda M Elbasiony*et al*., 2013]. The proposed decision tree has been built by using the Genetic algorithm and Enhanced Multiclass Support Vector Machine (EMSVM) [S Ganapathy*et al*., 2011]. For making hybrid system, we have applied an enhanced multiclass decision tree called IAEMSVM [Ganapathy S *et al*., 2011] for achieving better detection accuracy.

SYSTEM ARCHITECTURE

The architecture of the proposed system in this paper consists of four major components namely, user interface module, hybrid intrusion detection system module and result module as shown in Figure-1. Hybrid intrusion detection module consists of three sub modules namely, Genetic Algorithm, EMSVM and Clustering.

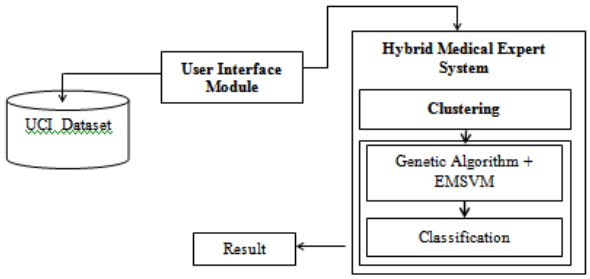

Figure-1. System architecture.

The user interface module collects the data from UCI Repository Machine Learning Data set. This data is sent to the hybrid intrusion detection framework for analysing and classifying. First, the data can be analysed by the clustering algorithm which is used in this framework. This module selects only the valuable attributes from the data set using projection. Data will be sent to the classification module for classifying the data. In this module, a new genetic based data adaptation decision tree (GDADT) algorithm has been proposed for classification. It provides the classification result to the results module. It can be accessed by the user interface module.

PROPOSED WORK

In this work, a new hybrid medical expert system has been proposed for effective medical diagnosis. The proposed hybrid system consists of two effective methods namely genetic based decision tree and weighted K-Means clustering [J. Novakovic*et al.*, 2011]. And also, we have used an effective classification algorithm called EMSVM [S Ganapathy*et al*., 2011] for achieving better detection accuracy by the proposed medical expert system. In addition, we have used the enhanced version of the EMSVM called IAEMSVM [Ganapathy S *et al*., 2011] for achieving better detection accuracy.

Weighted K-Means Clustering

The weighted k-means algorithm [Reda M Elbasiony*et al* 2013] is used as a data-mining clustering algorithm into a proposed hybrid method to partition the clinical records into a specified number of clusters, and then detect the disease affected clusters depending on their features [Leung K *et al*., 2005]. The ''KMlocal'' implementation of the k-means clustering algorithm [Mount D. KMlocal*et al*., 2005] is used to implement in the proposed hybrid method. The proposed hybrid method is evaluated over the UCI Repository datasets after solving the problems of categorical and different scales features.

Genetic based Data Adaptation Decision Tree

The proposed Genetic based Data Adaptation Decision Tree (GDADT) algorithm uses genetic algorithm with new fitness function to identify the best attributes using many iterations to obtain the optimal solution. The best solution is obtained by calculating the fitness value for finding the suitable necessary features from the given training data set. In addition, an enhanced Multiclass Support Vector Machine (EMSVM) [S Ganapathy*et al*., 2011] is also used for making effective decision tree in multiclass problems.

Genetic based Data Adopted Decision Tree (GDADT) Algorithm

Input : Data set, D = { a_1, a_2, …..a_n}

Output: Resulted Tree (Selected Instances)

Step-1: Selecting the best attributes of the given data set using genetic algorithm with fitness function,

$$\text{Fitness} = (W_1 \times \text{sensitivity}) + (W_2 \times \text{specificity})$$

Step-2: Initialize the set of unique values (UV) and the regular intervals (RI)

Step-3: Splitting data into many classes C_1,C_2….C_n, with different labels

Step-4: Set the Decision node (D) and Best Attributes (B_{best})

Step-5: If the unique value is belongs to the particular class then Split the regular intervals.

Else the unique value is belongs to the another particular class then

Assign the highest probability to the particular class members (instances).

Step-6: Splitting the instances into many groups and updates the intervals.

Step-7: Repeat step 5 until read all unique values

Step-8: Update all the records into the corresponding classes of dataset with best attributes

Step-9: Building the decision tree for updated records using EMSVM algorithm for making final decision.

The proposed algorithm reads the necessary data from the data set by the help of user interface as input and the selected instances are output of the system in the tree format. First, selects the best attributes from the given input dataset by using the genetic algorithm which is developed by the introduction of new fitness function in this algorithm. Second, split the datasets based on the regular intervals and unique values. The best attributes were selects from the data set as decision nodes and split into several classes based on the unique values and time intervals. This process is repeated until read all unique values. Finally, build the decision tree using an effective multiclass classifier called EMSVM for making final decision.

Intelligent Agent based Enhanced Multiclass Support Vector Machine (IAEMSVM)

We have used the enhanced version of EMSVM [S Ganapathy*et al*., 2011] called IAEMSVM [Ganapathy S *et al*., 2012] for improving the performance of the hybrid medical diagnosis system. This algorithm has been proposed for detecting intruders [Ganapathy S *et al*., 2012]. In this paper, we have used this IAEMSVM for medical diagnosis by the uses of medical data set and changing the decision tree rule for medical diagnosis process.

RESULT AND DISCUSSIONS

This section discusses about the dataset used in this work for medical diagnosis, experimental scenario and also about obtained result and discussion of the proposed system and reason for achievements on decision making process.

UCI machine learning dataset

The UCI machine learning repository is a standard medical dataset which is used as input data for evaluation of the proposed system in this paper. This proposed approach uses diabetic data set, Heart dataset and Wisconsin Diagnostic Breast Cancer data (WDBC) for experimental results. These datasets includes data in the form of text files and also stores patient medical treatment details. Here, we build a knowledge base by the help of medical doctor advice for monitoring, controlling and

predicting the diseases based on their blood glucose level by proper insulin dosage.

Experimental setup

We have used the Pentium IV personal computer with Intel Core i3 Processor 2.20 GHz for evaluating the proposed system. Initially, feature selection based on weight and gain ratio method processed on standard dataset instances of heart disease, diabetic and WDBC data from UCI repository using WEKA tool. WEKA is a collection of machine learning algorithms for data mining tasks. The proposed algorithms applied to the dataset from Java code and it contains tools for data analysis and predictive modelling. The input dataset of the WEKA are used in the form of CSV file.

RESULTS AND DISCUSSIONS

The various experiments have been conducted for evaluating the proposed medical diagnosis system. This section discusses the various results obtained by the proposed system and other classifiers.

Figure-2 shows the performance of the proposed Genetic based Data Adaptation Decision Tree (GDADT) classifier. We have conducted five experiments for evaluating the algorithm. We have used three datasets namely Diabetic, Heart and WDBC for these experiments.

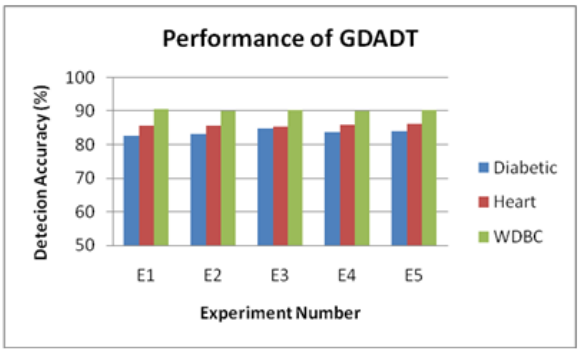

Figure-2. Performance of GDADT.

From Figure-2, it can be observed that the proposed classifier better performance on WDBC dataset when it is compared with other datasets.

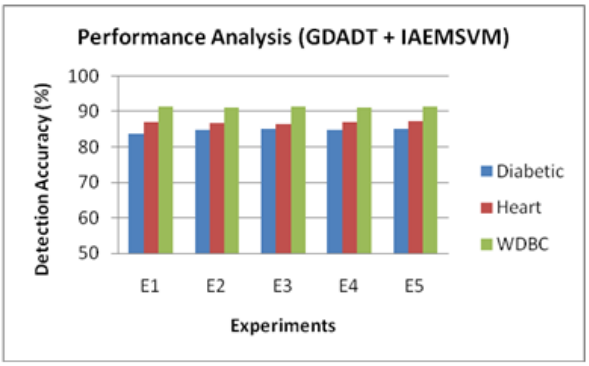

Figure-3. Performance analysis of GDADT + IAEMSVM.

Figure-3 shows the performance of the proposed hybrid medical diagnosis system which is the combination of Genetic based Data Adaptation Decision Tree (GDADT) classifier and IAEMSVM algorithm. We have conducted five experiments for evaluating this method. We have used three datasets namely Diabetic, Heart and WDBC for these experiments and obtained better results in all experiments.

Figure-4 shows that the performance comparative analysis between GDADT and the proposed hybrid medical diagnosis system which is the combination of GDADT and IAEMSVM. From this figure, it can be observed that the performance of the proposed hybrid medical diagnosis system provides better performance than individual performance of GDADT.

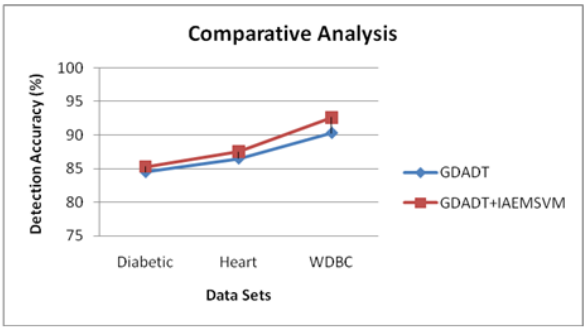

Figure-4. Comparative analysis.

The reason for this performance difference is the uses of weighted feature selection algorithm. By the uses of feature selection gives optimal features to the classifier, so classifier can perform well with optimal information and also effectively use the less number of rules.

Table-1 shows the results comparisons between the proposed method and the existing classifiers and

models. We have considered the same three datasets for this result comparison.

Table-1. Results comparisons.

Data Set	Detection Accuracy (%)			
	C4.5	DTC	GDADT	Hybrid System
Diabetic	74.1	77.7	84.5	85.3
Heart	72.1	79.5	86.5	87.5
WDBC	78.6	89.4	90.3	92.6

From Table-1, it can be observed that the performance of the proposed method provides better performance than other classifiers and models. This is due to the fact that the uses of effective weighted k-means cluster based feature selection and the uses of genetic algorithm in pre-processing process. Moreover, intelligent agent based EMSVM also used here for obtaining the medical doctor advice from the past experience and takes the necessary actions for effective feature selection and classification in the proposed algorithm. In classifier, the intelligent agent used for forming effective rules for decision making process and also fire the rules in proper time and place for making effective decision quickly.

The overall time taken also reduced when it is compared to the existing systems due to the uses of less number of features for calculating the information gain ration and uses those features for making decision by the proposed classifier. The classification accuracy is automatically increases when we used the less number of features.

CONCLUSION AND FUTURE ENHANCEMENTS

A new hybrid medical diagnosis system has been proposed and implemented for improving the disease prediction accuracy. The proposed hybrid system is combining the proposed Genetic based Data Adapted Decision Tree (GDADT), the existing Weighted K-Means clustering and Intelligent Agent based Enhanced Multiclass Support Vector Machine (IAEMSVM). From the experiments conducted in this work, it has been observed that the classification accuracy for Diabetics, Heart data and WDBC diseases are 85.3%, 87.5%, and 92.6%, respectively, when intelligent agents are added to the classifier and uses the knowledge base and rule base for making effective decisions. The main advantage of this method is that it reduces the time taken and false positive rates. Future works in this direction could be the use of fuzzy temporal logic enhancing the decision making capability of the proposed classifier.

REFERENCES

[1] Tomoharu Nakashima, Gerald Schaefer, Yasuyuki Yokota, Shao Ying Zhu and HisaoIshibuchi. 2005.Weighted fuzzy classification with integrated learning method for medical diagnosis. Proceedings of the IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China. pp. 5623-5626.

[2] M. Benkaci, B. Jammes, A. Doncescu. 2010. Feature selection for medical diagnosis using fuzzy artmap classification and intersection conflict.IEEE 24th International Conference on Advanced Information Networking and Applications Workshops. pp. 790-795.

[3] J. Novakovic and A. Veljovic. 2011. C-Support Vector Classification: Selection of Kernel and Parameters in Medical Diagnosis. IEEE 9th International Symposium on Intelligent Systems and Informatics, September 8-10, Subotica, Serbia, pp. 465-470.

[4] AgnieszkaWosiak, DanutaZakrzewska. 2014. Feature Selection for Classification Incorporating Less Meaningful Attributes in Medical Diagnostics. Proceedings of the Federated Conference on Computer Science and Information Systems. 2: 235-240.

[5] R.Sethukkarasi, S.Ganapathy, P.Yogesh, A.Kannan. 2014.An Intelligent Neuro Fuzzy Temporal Knowledge Representation Model for Mining Temporal Pattern. Journal of Intelligent and Fuzzy Systems- IOS Press. 26: 1167-1178.

[6] S.Ganapathy, R.Sethukkarasi, P.Vijayakumar, P.Yogesh, A.Kannan. 2014. An Intelligent Temporal Pattern Classification System Using Fuzzy Temporal Rules and Particle Swarm Optimization.Sadhana, Springer.Vol. 39, Part-2, pp. 283-302.

[7] Nguyen Tho Thong, Le Hoang Son. 2015. HIFCF: An effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis. Expert Systems with Applications. 42: 3682-3701.

[8] Le Hoang Son, Nguyen Tho Thong. 2015. Intuitionistic fuzzy recommender systems: An effective tool for medical Diagnosis. Knowledge-Based Systems. 74: 133-150.

[9] LeventKoc, Thomas A Mazzuchi, ShahramSarkani. 2012. A network intrusion detection system based on a hidden naïve bayes multiclass classifier. Expert Systems with Applications. 39:13492-13500.

[10] Reda M Elbasiony, Elsayed A Sallam, Tarek E Eltobely, Mahmoud M Fahmy. 2013.A hybrid network intrusion detection framework based on random forests and weighted k-means. Ain Shams Engineering Journal: 753-762.

[11] Ganapathy S, Yogesh P, Kannan A. 2012. Intelligent agent-based intrusion detection system using enhanced multiclass svm. Computational Intelligence and Neuroscience.pp. 1-10.

[12] Gisung Kim, Seungmin Lee, Sehun Kim. 2014.A novel hybrid intrusion detection method integrating anomaly detection with misuse detection.Expert Systems with Applications. 41: 1690-1700.

[13] FangjunKuang, WeihongXu, Siyang Zhang. 2014.A novel hybrid kpca and svm with ga model for intrusion detection. Applied Soft Computing. 18:178-184.

[14] DewanMdFarid, Li Zhang, ChowdhuryMofizurRahman, M A Hossain, Rebecca Strachan. 2014.Hybrid decision tree and naive Bayes classifiers for multi-class classification tasks. Expert Systems with Applications.41:1937-1946.

[15] Chun-Wei Tsai. 2013. Incremental particle swarm optimisation for intrusion detection. IET Networks. 2:3:124-130.

[16] S Mukkamala, AH Sung. 2003. Detecting denial of service attacks using support vector machines. In Proceedings of the IEEE International conference on Fuzzy Systems.pp. 1231-1236.

[17] SA Mulay, PR Devale, GV Garje. 2010. Intrusion detection system using support vector machine and decision tree. International Journal of Computer Applications.3:3:0975-8887.

[18] S Ganapathy, P Yogesh, AKannan. 2011. An intelligent intrusion detection system for mobile ad-hoc networks using classification techniques," Communications in Computer and Information Science-Springer. 148:117-122.

[19] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/diabetes.

[20] G.V. Nadiammai, M. Hemalatha. 2014. Effective approach toward Intrusion Detection System using data mining techniques.Egyptian Informatics Journal. 15: 37-50.

[21] Mount D. 2005.KMlocal: a testbed for k-means clustering algorithms. <http://www.cs.umd.edu/~mount/Projects/-KMeans/kmlocal-doc.pdf>.

[22] Leung K, Leckie C. 2005. Unsupervised anomaly detection in network intrusion detection using clusters. In: Proc 28th Australasian CSconf.Vol. 38, Newcastle, Australia.pp. 333-42.