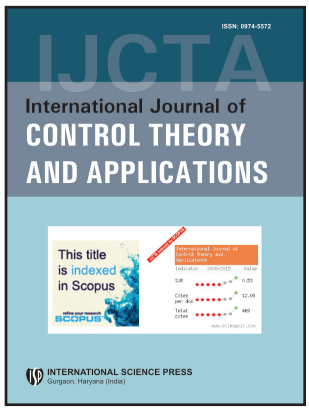


International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 12 • 2017

Optimal Knowledge Extraction System Based on GSA and AANN

S. Jagadeesh Soundappan[a] and R. Sugumar[b]

[a]*Department of CSE, St. Peter's University, Avadi, Chennai-600 054, India*
[b]*Department of CSE, Velammal Institute of Technology, Panchetti, Chennai-601204, India*

Abstract: Knowledge discovery in database (KDD) is an active area of research that resolves the non-trivial process of identifying valid, potentially useful, and ultimately understandable patterns in data. In this paper to extract knowledge from different datasets, we will propose a hybrid mining technique. The knowledge extraction can be done by decision tree algorithm such as ID3 with the combination of Gravitational search Algorithm (GSA). The main aim of this technique is to identify best rules from the decision tree algorithm. Based on that final classification will be done by employing Auto Associate Neural Network (AANN). The implementation will be done in MATLAB with various data bases.

Keyword: Knowledge Discovery Database, Gravitational Search Algorithm, Auto Associate Neural Network, Data Mining and Automatic Optical Inspection.

1. EXISTING SYSTEM

Knowledge discovery in databases (KDD) is very useful in economic and scientific domains. KDD techniques are used to reveal critical information hidden in the data sets [7]. A fast target maneuver detecting and highly accurate tracking technique using a incremental neural learning neural fuzzy network based on Kalman filter is proposed in this paper [2]. Moments have been used to distinguish between aircraft shapes, character recognition, and scene-matching applications. For instance, the area of an object is the (0, 0)th order moment of the object [5]. The Sound Event Classification is a research direction with wide potential for applications in music search, automatic broadcasting, meeting transcription, surveillance, and security [4]. Optimizing nonlinear and non-stationary processes with multiple objectives presents a challenge for traditional solution approaches [3]. Stereo vision has been an important problem in computer vision for decades, but it is still an unsolved problem due to its complexity. It is a process that transforms the information of two photographic images to a three dimensional description of the world [6]. However, the growing complexity of control systems, accompanied by high levels of inherent uncertainty in modelling and estimation and intrinsic nonlinear dynamics involving unknown functional make achieving the aforementioned objective impossible except under idealized conditions [1].

2. RELATED WORK

Haifeng Chen et. al., [9], have proposed, a new strategy, the experience transfer, to facilitate the management of large-scale computing systems. The dependencies between system configuration parameters are treated as transferable experiences in the configuration tuning for two reasons: (i) knowledge is helpful to the efficiency of the optimal configuration search, and (ii) The parameter dependencies were typically unchanged between two similar systems. We use the Bayesian network to model configuration.

To improve the performance of automatic optical inspection (AOI), a new inspection method for chip component of mounted components on printed circuit boards is developed. Xie Hongwei et. al., [10] proposed the inspection procedure is divided into training stage and test stage. In the training stage, first, the solder joint is divided into several sub-regions according to priori knowledge, second various features in every sub-region are extracted, then, for every sub-region the optimal features are selected with an improved AdaBoost.

For multi-agent reinforcement learning in Markov games, knowledge extraction and sharing are key research problems. Min Fang1 et. al., [11], have proposed, A state list extracting algorithm checks cyclic state lists of a current state in the state trajectory, condensing the optimal action set of the current state. By reinforcing the optimal action selected, the action policy of cyclic states is optimized gradually.

Hiroyuki Kasai et. al., [12], have presented a proposal of a music image generator for users to grasp music easily and instantaneously on consumer electronics devices such as music players and smartphones. It provides users faster and more efficient music browsing functionality using a technique that automatically generates a music image that matches lyric contents.

Problem Statement

In the Existing [12], The image materials for the image synthesizer might be insufficient. The scene knowledge might be too small for application to practical services. For future research, usability evaluations will be conducted to confirm the practical effectiveness of the proposed method.

In the Existing [8], Another problem of the minimum variance method is its numerical instability due to the singularity of the covariance matrix. In addition, some of these statistical fusion methods are difficult to be implemented adaptively for real-time applications such as signal and image processing.

Proposed Methodology

The challenges for instance, the traditional methods usually discover homogeneous features from a single source of data while it is not effective to mine for patterns combining components from multiple data sources. In order to extract knowledge from different datasets, we will propose a hybrid mining technique. The knowledge extraction can be done by decision tree algorithm such as ID3 with the combination of Gravitational search Algorithm (GSA). The main aim of this technique is to identify best rules from the decision tree algorithm. The final classification will be done by employing Auto Associate Neural Network (AANN).

KDD Dataset

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modelling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. Data Mining (DM) is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and

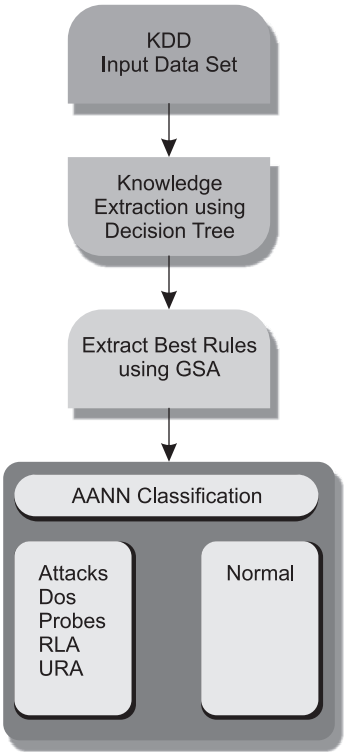


Figure 1: Block Diagram for Proposed AANN Classification

prediction. The accessibility and abundance of data today makes knowledge discovery and Data Mining a matter of considerable importance and necessity. Given the recent growth of the field, it is not surprising that a wide variety of methods is now available to the researchers and practitioners. No one method is superior to others for all cases. The handbook of Data Mining and Knowledge Discovery from Data aims to organize all significant methods developed in the field into a coherent and unified catalogue; presents performance evaluation approaches and techniques; and explains with cases and software tools the use of the different methods. In our work the KDD dataset contain 2000 for training and 1000 for testing. In that dataset we are extracted three attacks and one normal, the attacks are DOS, Probe, and RLA.

Probing Attack (PROBE): Probing is a collection of attacks where an attacker scrutinizes a network to gather information or to conclude prominent vulnerabilities. In this category the attacker attempt to gather information about network of computers for the apparent purpose of circumventing its security. Probe contains the attacks: 'port sweep', 'Satan', 'Nmap', and 'Ip sweep'.

Random Link Attack (RLA): In an RLA, the malicious user creates a set of false identities and uses them to communicate with a large, random set of innocent users. Attackers create some fake nodes and randomly connect to regular nodes. Fake nodes form some inner structure among themselves to evade detection. From the dataset we are taken 2000 for training and 1000 for testing then that data's are given in to the decision tree to extract knowledge, they are given below.

Decision Tree ID3 ID3 is a greedy algorithm which builds decision trees based on an up to down approach. The input and output data in ID3 are categorical. All categories of attributes can be applied in generating decision trees by ID3, thus creates wide and shallow trees. It builds trees in 3 phases.

1. Creating splits in a multi-way manner, for example for all of attributes a split is made and subdivisions of the proposed split are attributes categories.

2. Estimation of the greatest split for tree branching according to information gain metric.
3. Testing the stop criterion, and repeating the steps recursively for new subdivisions. These three steps are done iteratively for all of the nodes of the tree. The below formula represents the information gain measure.

$$\text{Entrophy (S)} = \sum_{i=1}^{k} -p_i \log_2 p_i \tag{1}$$

S denotes the dataset. K denotes the number of output variable classes, and Pi the possibility of the class *i*. In this algorithm the quality of the split is represented by information gain.

$$\text{Gain (S, A)} = \text{Entropy(S)} - \sum_{v \in \text{values (S)}} \frac{|s_v|}{s} \text{Entropy}(s_v) \tag{2}$$

Values (A) represent probable values of attribute A, S_v represents the subdivision of dataset S which contains value *v* in S. Entropy (S) calculates the entropy of an input attribute A which has *k* categories, Entropy (S_V) is the entropy of an attributes category with respect to the output attribute, and $|S_V|/|S|$ is the probability of the *j*-th category in the attribute. The difference between entropy of the node and an attribute is Information gain of an attribute. Information gain shows the information an attribute convey for disambiguation of the class.

GSA

GSA's effectiveness has been testified against other well-known methods, but there is room for improvement. In this section, memory strategy and iterative chaotic perturbation operator are performed to avoid premature convergence and improve the search speed compared to GSA. The steps of an gravitational search algorithm (GSA). During GSA searching process, all agents gradually converge to a small local zone, which results in a low searching efficiency in the late period, so an effective mechanism should be established to help poor agents jump out of the local minimum. Influenced by the gravitational attraction, artificial satellites and space ships (including space station) fall at a speed of 100 m/d, which will hamper their normal operation. So during the flying, orbital change is always required. Based on the concept mentioned above, the paper performs an orbital change operation upon the poor agents (in the paper, the worst 10 agents are chosen according to the fitness) in the late search period of the algorithm in order to prevent them from falling into the local minimum and improve the algorithm performance.

$$\text{XM}_i = x_i + \text{rands}\, x_i \quad x = 1, 2, ..., \text{N} \tag{3}$$

Where rands is a random number between −1 and 1. Further search of optimal agent position. The GSA algorithm generally converges quickly in the early 70% iterations, and then the convergence speed becomes slow. In order to further intensify the optimal searching ability of the algorithm in the late period, the optimal agent is further optimized by coordinate descent method and it transforms the multi-variable optimization problem into some single-variable sub-problems. It helps optimize further the position of the optimal agent, establish an effective local search mechanism and thus improve the algorithm performance further. The detailed steps for coordinate descent method are as follows.

Step 1: The variable that needs further optimization is the optimal agent *s* position x_{best}. Define the initial unit orthogonal search direction, generally the coordinate axis direction, as the candidate, i.e., d_1, ..., d_{min}; the range of the variable d_{best} is $[\text{low, up}]^{\text{dim}}$, where dim is the dimension of x_{best}.

Step 2: Solving sub-problem

$$\text{for } (j = 1, j \le \text{dim}, j{+}{+}) \tag{4}$$

$$\min: f(x_{\text{best}}, j + \lambda_j d_j) \tag{5}$$

Where, λ_j is the coordinate parameter in the direction of the j-th coordinate and is required to meet the feasible condition.

$$\text{low} - x_{\text{best}}, j \leq \lambda_j \leq \text{up} - x_{\text{best},j} \tag{6}$$

Step 3: By precise linear search, we can obtain the optimal solution and update the position of optimal agent by

$$x_{\text{best}} = x_{\text{best}} + \lambda_j d_j \tag{7}$$

Update optimal agent using trial-and-error method. In GSA, all current agents change at each step; if the optimal agent s fitness becomes bad, the next search will begin from a worse position. The optimal position of those historical search steps, L_{best}, and its fitness F_{best}, only play a role for comparison, rather than participate into each step of iterative search. In order to utilize the information of L_{best}, the optimal agent is updated using the trial-and-error method, i.e., after each iteration, the search will continue to the next step if the fitness of the optimal agent turns better. Otherwise, the position of optimal agent s position and fitness will be replaced by L_{best} and F_{best}.

Auto-Associative Neural Network

Neural Networks are information processing models that simulate the manner in which biological nervous systems process information. There are four main constituents of any neural network and they include the processing units, activation functions, weighted interconnections and the activation rules. Auto-associative neural networks (AANN) are network models in which the network is trained to recall the inputs as the outputs thus guaranteeing the networks are able to predict the inputs as outputs whenever new inputs are presented. These networks have been used in a variety of applications. An auto-associative network encoder also referred to as an auto-encoder consists of an input and output layer with the number of inputs being equal to the number of outputs, hence the name auto-associative. In addition to these two layers, there also is a narrow hidden layer. It is necessary that the hidden layer uses a lower dimension, so as to enforce the encoding and decoding processess. The narrow hidden layer forces the network to reduce any redundancies that may occur in the data whilst allowing the network to detect non-redundant data. Figure 2 depicts the framework of an AANN. For the purpose of this work, one hidden layer is used as it has been proven that a single hidden layer network is capable of approximating any continuous multivariate function to any suitable degree of accuracy. The number of nodes in the hidden layer is determined by the ability of the network to approximate the error function.

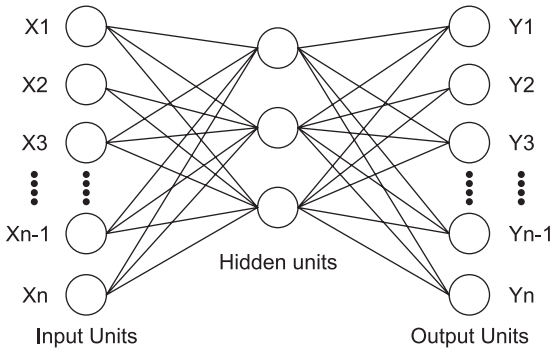


Figure 2: AANN Diagram

3. RESULT AND DISCUSSION

The experimental result of AANN based classifier is discussed below. The proposed system is implemented using MATLAB 2014 and it is performed with i5 processor of 3GB RAM.

Dataset Description: The proposed AANN based classifier is experimented with the dataset namely KDD dataset. These dataset are given as input to identify the Attacks.

KDD Dataset: This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and "good'' normal connections.

Evaluation metrics: An evaluation metric is used to evaluate the effectiveness of the proposed system. It consists of a set of measures that follow a common underlying evaluation methodology some of the metrics that we have choose for our evaluation purpose are True Positive, True Negative, False Positive and False Negative, Specificity, Sensitivity, Accuracy, F measure.

Sensitivity: The measure of the sensitivity is the proportion of actual positives which are accurately recognized. It relates to the capacity of test to recognize positive results. TP stands for True Positive and FN stands for False Negative.

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{8}$$

Specificity: The measure of the specificity is the extent of negatives which are properly recognized. It relates to the capacity of test to recognize negative results. Where TN stands for True Negative and FP stands for False Positive.

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{9}$$

Accuracy: Accuracy of the proposed method is the ratio of the total number of TP and TN to the total number of data.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{(\text{TN} + \text{TP} + \text{FN} + \text{FP})} \tag{10}$$

Experimental Outcome	*Condition as determined by the Standard of Truth*	
	Positive	*Negative*
Positive	TP	FP
Negative	FN	TN

Performance Analysis: The performance of the proposed knowledge extraction in attack and normal prediction methods evaluated by the three metrics Sensitivity, Specificity and Accuracy. The results of proposed work help to analyze the efficiency of the prediction process. The subsequent Table 2 tabulates the results. Here, only the results of dataset given in Table 1.

Table 1
Results of the proposed Optimal Knowledge Extraction System

	TP	*TN*	*FP*	*FN*	*Accuracy*	*Sensitivity*	*Specificity*
DOS	24909	24884	116	91	0.99586	0.99636	0.99536
Probes	24770	3584	523	230	0.97413	0.9908	0.872656
RLA	24988	51	26	12	0.998485	0.99952	0.662338
URA	25000	0	42	0	0.998323	1	0

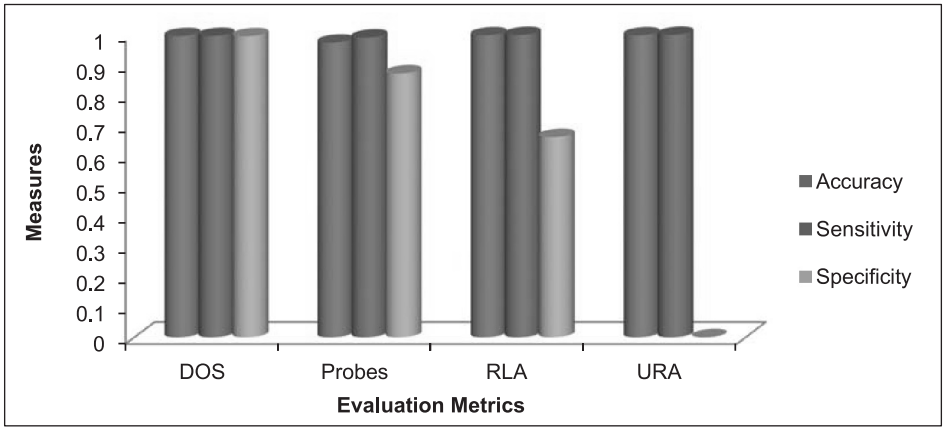


Figure 3: Graph for results with the Performance measures Specificity and Sensitivity, Accuracy

From Table I, the evaluation metrics are analyzed for the dataset, by which we can observe the efficiency of proposed detection system. The results of the measures Sensitivity, Specificity, Accuracy are graphically represented in Figure 3. The sensitivity of four attacks is explained DOS, Probes, RLA, URA, 0.9963, 0.9908, 0.99952, and 1. With these metrics, the specificity and accuracy are the main measures for evaluating the detection accuracy of our proposed system The values of specificity for four attacks are DOS, Probes, RLA, URA, 0.99536, 0.87265644, 0.662337662, 0, and the values of accuracy is 0.99586, 0.974129934, 0.998484667, 0.998322818. The results get high accuracy results on behalf of the reduced error rates in the proposed system. From the Figure 3 also, we find out the minimal value of error rates for the three dataset.

Comparative Analysis

The literature review works are compared in this section with the proposed work to show that our proposed work is better than the state-of-art works. We can establish that our proposed work helps to attain very good accuracy for the attack prediction of database using AANN classifier. And also we can establish this prediction accuracy outcome by comparing other classifiers. We have utilize Neural network and cuckoo search for our Comparison in our work. The Comparison outcomes are presented in the following Table 2.

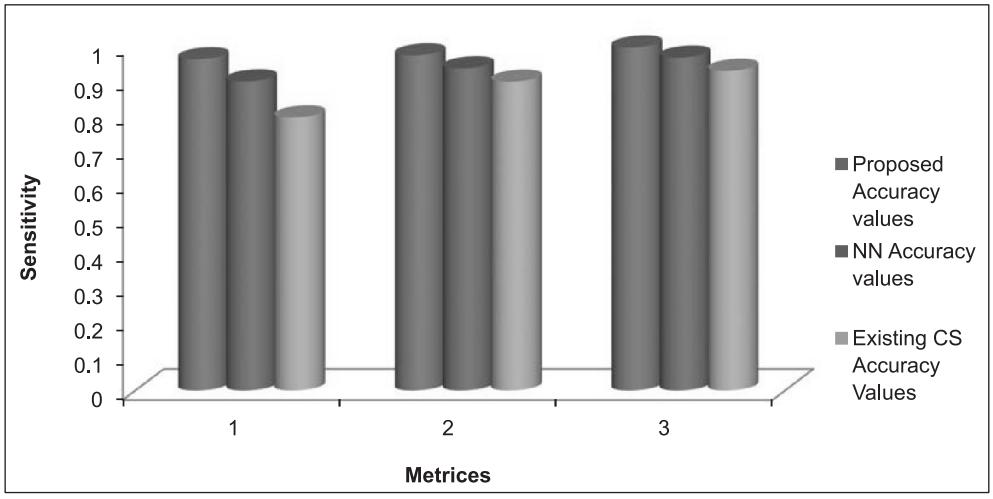


Figure 4: Comparison for proposed Method vs Existing Method for Accuracy

Table 2
Comparison of Accuracy values in proposed vs existing method

Trial	*Proposed Accuracy values*	*NN Accuracy values*	*Existing CS Accuracy Values*
1	0.962827	0.89866	0.793554
2	0.97413	0.936504	0.897851
3	0.99686	0.967699	0.929091

The accuracy for the Neural Network is 0.89866, 0.936504, and 0.967699 which is low in compared with our classifier, AANN for our dataset are 0.962827, 0.97413, and 0.99686. We have also compared with our classifier in Cuckoo search it will also shows a lower result which is 0.793554, 0.897851, and 0.929091.

Table 3
Comparison of Sensitivity values in proposed vs existing method

Trial	*Proposed Sensitivity values*	*NN Sensitivity values*	*Existing CS Sensitivity Values*
1	0.99136	0.89944	0.858797
2	0.98948	0.944	0.972031
3	0.99744	0.8998	0.978178

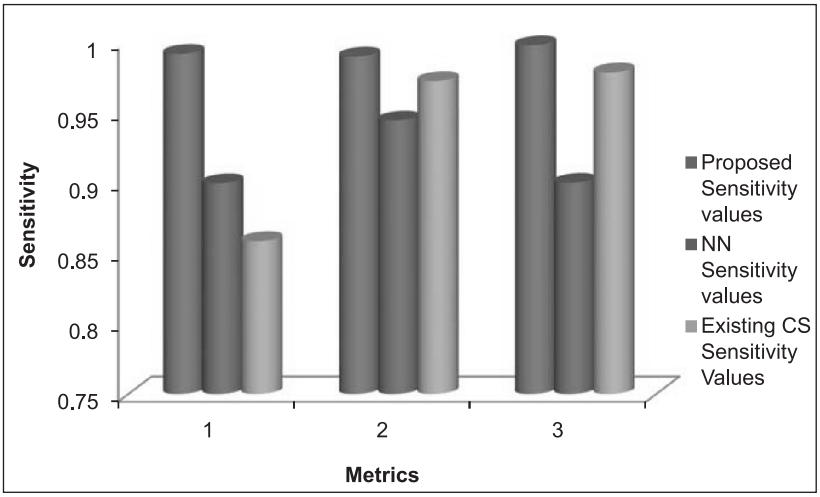


Figure 5: Comparison for proposed Method vs Existing Method for Sensitivity

The sensitivity for the Neural Network is 0.89944, 0.944, and 0.8998 which is low in compared with our classifier, AANN for our dataset are 0.99136, 0.98948, and 0.99744. We have also compared with our classifier in Cuckoo search it will also shows a lower result which is 0.858797, 0.972031, and 0.978178.

Table 4
Comparison of Specificity values in proposed vs existing method

Trial	*Proposed Specificity values*	*NN Specificity values*	*Existing CS Specificity Values*
1	0.99136	0.89944	0.858797
2	0.98948	0.944	0.972031
3	0.99744	0.8998	0.978178

The specificity for the Neural Network is 0.89944, 0.944, and 0.8998 which is low in compared with our classifier, AANN for our dataset are 0.99136, 0.98948, and 0.99744. We have also compared with our classifier in Cuckoo search it will also shows a lower result which is 0.858797, 0.972031, and 0.978178.

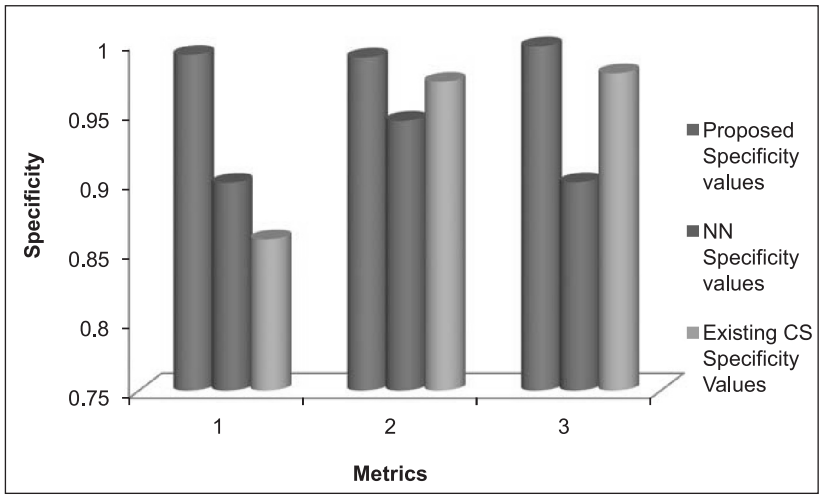
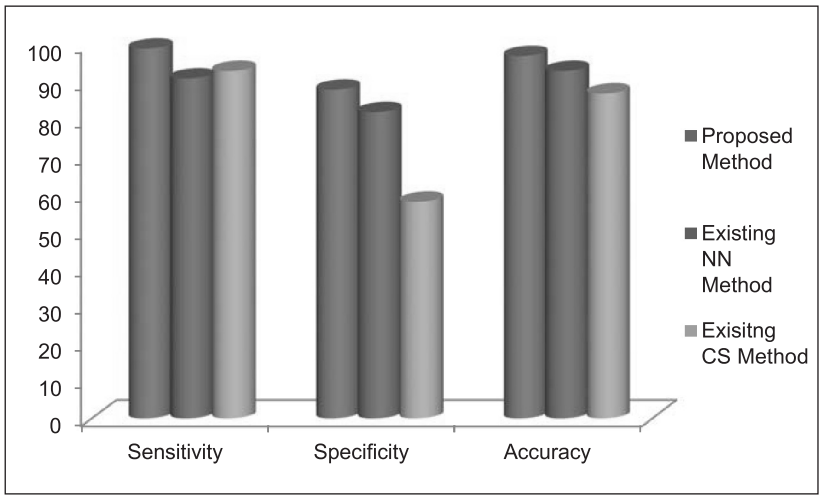


Table 5
Comparison of Proposed Method vs Existing Method

Metrics	*Proposed Method*	*Existing NN Method*	*Existing CS Method*
Sensitivity	99	91	93
Specificity	88	82	58
Accuracy	97	93	87



The improved good accuracy outcomes of attack classification are presented by our proposed work. In comparison with the NN and CS gives very less accuracy values for the evaluation measures. The sensitivity values of NN gives 91% and CS is 93% but our proposed AANN gives 99%. The specificity values of existing NN gives 82% and CS is 858% but our proposed AANN method gives 88%. The accuracy values of existing NN gives 93% and CS gives 87% but our proposed AANN gives 97%. From these outcomes, it is known that by means of AANN classifier in our work provides very good for the classify the attack and normal as it gives improved accuracy outcomes.

4. CONCLUSION

AANN based classification with three phases – Knowledge Extraction, Best rules Extraction and classification was proposed in this paper. Knowledge was extracted with the aid of Decision tree algorithm. Then the optimal

rules are extracted with a aid os GSA algorithm. Finally the attacks are successfully classified by using AANN algorithm. The performance measures of sensitivity, specificity, accuracy, were evaluated for our proposed method. The efficiency of the classification is very high by presenting very good accuracy outcomes and also the classification of attacks is gives very accurate outcomes. From the outcomes, we have showed that the AANN classifier utilized in our proposed work outperforms the other classifiers NN and CS by facilitated very good accuracy. Thus, we can observe that our proposed work is better than other existing works for the attack classification.

REFERENCE

[1] Randa Herzallah and David Lowe, "A Bayesian Perspective on Stochastic Neurocontrol", In Proceedings of IEEE Transaction on Neural Network Vol. 19, No. 5, pp. 914-924, May 2008.

[2] Liu Mei, Quan Tai fan, "Tracking maneuvering target based on neural fuzzy network with incremental neural leaning", Journal of Systems Engineering and Electronics, Vol. 17, No. 2, pp. 343-349, 2006.

[3] Zhe Song, And Andrew Kusiak, "Multiobjective Optimization Of Temporal Processes", In Proceedings of IEEE Transactions On Systems, Man, and Cybernetics Part B: Cybernetics, Vol. 40, No. 3, pp. 845-856, June 2010.

[4] Huy Dat Tran, Haizhou Li, "Sound Event Classification Based On Feature Integration, Recursive Feature Elimination And Structured Classification", pp. 177-180, April 2009.

[5] Chin-Hsiung Wu And Shi-Jinn Horng, "Run-Length Chain Coding And Scalable Computation of A Shape's Moments Using Reconfigurable Optical Buses", In Proceedings of IEEE Transaction on Systems, Man, and Cybernetics Part B: Cybernetics, Vol. 34, No. 2, pp. 845-855, April 2004.

[6] Jun Yan, Ning Liu, Shuicheng Yan, "Trace-Oriented Feature Analysis for Large-Scale Text Data Dimension Reduction", In Proceedings of IEEE Transaction on Knowledge and Engineering, Vol. 23, No. 7, pp. 1103-1117, 2011.

[7] R. Yang, R.D. van der Mei, "On the Optimization of Resource Utilization in Distributed Multimedia Applications", In Proceedings of IEEE International Conference on Cluster Computing and the Grid, pp. 358-365, 2008.

[8] Analysis of E-commerce challenges in INDIA by using Weka Tool, Published in IJCTA. [Scopus Indexed].

[9] Perspectives on Educational Data Mining – A Study Published in Man in india. [Scopus Indexed]

[10] Xie Hongwei, Zhang Xianmin, "Solder Joint Inspection Method For Chip Component Using Improved Adaboost and Decision Tree", In Proceedings of IEEE Transaction on Components, Packaging and Manufacturing Technology, Vol. 1, No. 12, pp. 2018-2027, December 2011.

[11] Min Fang1, and Frans C.A. Groen, "Collaborative multi-agent reinforcement learning based on experience propagation", Journal of Systems Engineering and Electronics, Vol. 24, No. 4, pp. 683-689, August 2013.

[12] Hiroyuki Kasai, "Lyric-based Automatic Music Image Generator for Music Browser Using Scene Knowledge", In Proceedings of IEEE Transactions on Consumer Electronics, Vol. 59, No. 3, pp. 578-586, August 2013.