ISSN:0975-9646



# Survey on Types of Bug Reports and General Classification Techniques in Data Mining

Smita Mishra, Somesh Kumar

*Noida Institute of Engineering & Technology, Noida*
*Affiliated with UPTU, Lucknow, UP, India*

***Abstract*- Data mining is the process of extraction of hidden and useful information from huge data. It is also called knowledge discovery process from data. Bug tracking systems are made to manage bug reports, which are collected from various sources. These bug reports are needed to be labeled as security bug reports or non security bug reports. Data mining uses to apply mining algorithm to extract information which is stored in bug tracking systems. Classification is a task of data mining. A data mining system can be classified according to the kinds of databases mined. Database systems can be classified according to different criteria (such as data models, or the types of data or applications involved), each of which may require its own data mining technique. Data mining systems can therefore be classified accordingly. This paper present a survey on several classification techniques which are generally used for data mining such as naïve bayes, decision tree, K- nearest neighbor, Rule based, neural network etc.**

***Key words*- Bug report, classification, naïve bayes, decision tree, K-nearest neighbor, Rule based, neural network.**

## I. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. The key properties of data mining are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases

All the software bug related information is kept in bug tracking system. Bug tracking systems contains lots of useful information related to the bug which is called bug report, is collected from various sources like testing team, end users etc. Software organizations use these types of bug tracking system to make effective and proper development of the software. Bug reports are mainly two types: **Security bug reports (SBRs) and non security bug reports (NSBRs).**Bug report need to be labeled as security bug reports (SBRs) or non security bug reports (NSBRs).These SBRs have higher potential risk than NSBRs. A **security bug** is a software bug that can be exploited to gain unauthorized access or privileges on a computer system. Security bugs introduce by compromising one or more of:

- Authentication of users and other entities
- Authorization of access rights and privileges
- Data confidentiality
- Data integrity

Security bug report need to check by security team of the software development or Information security management system. Non security bug is related to hardware, site, personnel vulnerabilities etc. These are not much harmful like security bug.

This paper presents a study on **Statistical** and **Soft computing** approaches for classification which are commonly used. Statistical approach learns from examples, where classification is done with similar procedure and categories to be predicted are known as outcomes, which are considered to be possible values of the dependent variables. Statistical method includes: K-nearest neighbor, Naïve bayes, Rule base learning, Decision tree, Neural Network learning and Support vector machines. Soft computing is an emerging approach to computing which parallels remarkable ability of human mind to reason and learn in an environment of uncertainty and imprecision. Soft Computing consists of several computing paradigms like Neural Networks, Fuzzy Logic, Genetic algorithms and Rough sets.

The next section deals with a study on various classification techniques such as naïve bayes classifier, decision tree, K-nearest neighbor, Rule based neural network, support vector machines, fuzzy logic, genetic algorithm and rough sets.

## II. CLASSIFICATION METHODS

### *A. Naive Bayes Classifier*

A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting.

In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

Abstractly, the probability model for a classifier is a conditional model : $\boldsymbol{p}(C/F1,\ldots\ldots,Fn)$

Using Bayes' theorem, we write

$$p(C/F1,\ldots\ldots\ldots,Fn) = \frac{p(C)\,p(F1,\ldots\ldots,Fn/C)}{p(F1,\ldots\ldots\ldots,Fn)}$$

Posterior = $\frac{prior \times likelihood}{evidence}$

This means that under the above independence assumptions, the conditional distribution over the class variable can be expressed like this:

$$p(C/F1,\ldots\ldots\ldots,Fn) = 1/Z\ p(C) \prod_{i=1}^{n} p(Fi/C)$$

where Z (the evidence) is a scaling factor dependent only on i.e., a constant if the values of the feature variables are known.

Various research have been focused on naïve bayes domain, Paedro Domingos & Michael Pazzani [4] Presented the study on the optimality of the simple Bayesian classifier under zero-one loss and verified that Bayesian classifier performs good even when strong attribute dependences are present and even show the Bayesian classifier does not require attribute independence to the optimal under zero-one loss.

Irina Rish, IBM Research Division [5] Understand the data characteristics which affect the performance of naïve bayes. In their approach they used Monte Carlo simulations which allow systematic study of classification accuracy for several classes of randomly generated problems and analyze the impact of the distribution entropy on the classification error and verified that naïve bayes works best in two cases: completely independent features and functionally dependent features.

Jiangtao Ren & Sau Dan Lee [6] Propose a novel naive Bayes classification algorithm for uncertain data with a pdf. Uncertainty arising from measurement error and repeated measurement etc. With uncertainty, the value of each data item is represented by a probability distribution function (pdf). They extended the class conditional probability estimation in the bayes model to handle pdf's. They did experiments on UCI dataset and show that the accuracy of naïve bayes model can be improved by taking uncertain information. Uncertain naive Bayes model considering the full pdf information of uncertain data can produce classifiers with higher accuracy than the traditional model using the mean as the representative value of uncertain data.

### B. *Decision Tree*

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Earlier studies of decision tree B V Chowdary & Annapurna Gummadi [7] focused to develop a new method for decision tree for classification of data using a data structure called Peano Count Tree (P-tree) which enhances the efficiency and scalability of the classification. They apply data smoothing and attribute relevance techniques along with classifier and result shows that P-tree method is faster than existing classification method.

A. S. Galathiya & A. P. Ganatra [8] performs a comparison analysis of various decision tree classifiers such as ID3, C4.5 and C5.0. C5.0 classifier perform feature selection , cross validation , reduce error pruning and model complexity to reduce the optimization of error ratio .Feature selection is used to remove irrelevant data attributes cross-validation is used to get more reliable estimation of prediction , by increasing the model complexity, accuracy of the classification increases and apply pruning technique over fitting problem of decision tree can be solved, accuracy gained about 1-3%, classification error rate is reduced.

### C. *K-NN (K-nearest neighbor)*

Instance-based classifiers such as the KNN classifier operate on the premises that classification of unknown instances can be done by relating the unknown to the known according to some distance/ similarity function. Classification (*generalization*) using an instance-based classifier can be a simple matter of locating the nearest neighbor in instance space and labeling the unknown instance with the same class label as that of the located (known) neighbor.

Li Xiong & Subramanyam Chitti [9] present a frame-work including a general model as well as multi-round algorithms for mining horizontally partitioned databases using a privacy preserving k-Nearest Neighbor (kNN) classifier. A salient feature of this approach is that it offers a trade-off between accuracy, efficiency and privacy through multi-round protocols. They conducted experimental evaluation in terms of correctness, efficiency and privacy characteristics. To conduct this experiment they used 3 datasets. The first data set, GLASS, which contain various characteristics of different types of glass. The second dataset was PIMA, which is a medical data set, and third dataset was ABALONE which is used to predict the age of abalone from its physical characteristics. In their experiments they compare the accuracy of private kNN against a distributed kNN classifier.

N. Suguna and Dr. K. Thanushkodi [10] present a model in which genetic algorithm is combined with K-NN to improve its classification performance. GA is employed to take k-neighbors straightaway and then calculate the distance to privacy preserving K-Nearest Neighbor classifier. Classify the test samples and compare the performance with the traditional KNN, CART and SVM classifier

Ming Yao [11] explore the widely used distance metrics (such as Euclidean) in Text Classification problems, and find that these metrics may not be appropriate for highly skewed dataset like text categorization. Therefore, a novel method of learning evidence from multiple distances metric is proposed. Based on DS theory, the evidences learnt from these distance metric are combined for improving the effectiveness of kNN based text classifier. The ensemble of distance metric is tested on three standard

benchmark data sets. First dataset was Reuters and domain was new articles, second data set WebKB domain WebPages, and third dataset was 20 News group domain was news articles. He applied three experiments on these datasets to verify the validity of evidence learnt from the heterogeneous distance metric sources.

### *D. Rule Based*

In Rule based algorithm each classification method uses an algorithm to generate rules from the sample data or training data. These rules are then applied to new data or test data. Rule based algorithm provide process that generates rule by concentrating on a specific class at one time and by maximizing the probability of the desired classification.

Rule based algorithms are based on If-Then rules and these rules generate from decision tree. We can express the rule in the from: "IF condition THEN conclusion"

The IF part of the rule is called rule antecedent or precondition. The THEN part of the rule is called rule consequent.

In reviewed papers M. Thangaraj & C.R.Vijayalakshmi [12] present the performance comparison of different rule based classification techniques namely Decision tree, PART, RIPPER and RIDOR based on tuple-id propagation techniques. They compare the performance of four rule based classifier across multiple database relations. Tuple-id propagation technique is based on five criteria: number of tuples, number of relations, number of foreign-keys, classification accuracy and runtime

Ritika & Aman Paul [13] examine different classification algorithms and to find out a classification technique with best accuracy rate & least error for the prediction of blood donors. Different classification algorithms like Naïve Bayes, Bayes net, J48, Prism, PART, Ridor, ZeroR and CBA are discussed and compared. These algorithms are applied on blood donor's dataset to find out their accuracy and error rate. The accuracy is computed as the sum of true positive and true negative over the sum of true positive, true negative, false positive and false negative or total number of correctly classified instances over total number of instances.

### *E. Neural Network*

Neural network is commonly known as artificial neural network, which is non-liner statistical data modeling tool. This is used to model relationships between input and output. A neural network structure consisting **processing elements** which are connected through unidirectional signal channels called **connections.** This structure is inspired by human brains. The word network in the term 'artificial neural network' refers to the inter–connections between the neurons in the different layers of each system. An example system has three layers. The first layer has input neurons which send data via synapses to the second
layer of neurons, and then via more synapses to the third layer of output neurons. More complex systems will have more layers of neurons with some having increased layers of input neurons and output neurons. The synapses store parameters called "weights" that manipulate the data in the calculations.

Hongium Lu [14] presents a method to discover symbolic classification rules using neural networks. With the proposed approach the activation values of the hidden units in the network are analyzed and classification rules are generated using the result of the analysis. With this approach, concise symbolic rules with high accuracy can be extracted from a neural network. The network is first trained to achieve the required accuracy rate. Redundant connections of the network are then removed by a network pruning algorithm. The activation values of the hidden units in the network are analyzed, and classification rules are generated using the result of this analysis. The effectiveness of the proposed approach is clearly demonstrated by the experimental results on a set of standard data mining test problems.

B.Madasamy & Dr.J.Jebamalar Tamilselvi [15] proposed a method to combine neural network and data mining techniques to automate biomedical classification process to support decision. For improving the classification ability and behavior of neural network is used by pre-processing and pre-clustered data with the help of Rule based induction, Multilayer perception model, nearest neighbor, radial basics function and back propagation learning algorithm is employed to classify such complex tasks. The proposed clustering algorithm applied to the bio-medical dataset to reduce the amount of samples to be presented to the neural networks to automate biomedical classifier. This proposed method improves accuracy, computation time and performance of the classifier when applied to the publicly available bench-mark biomedical dataset.

### *F. Support Vector Machine*

Support Vector Machine (SVM) is a classification technique based on statistical learning theory.A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Issam El-Naqa and Yongyi Yang investigated [16] a method based on support vector machines (SVMs) for detection of micro calcification (MC) clusters in digital mammograms, and propose a successive enhancement learning scheme for improved performance. SVM is a machine-learning method, based on the principle of structural risk minimization, which performs well when applied to data outside the training set. They formulate MC detection as a supervised-learning problem and apply SVM to develop the detection algorithm. They use the SVM to detect at each location in the image whether an MC is present or not. They tested the proposed method using a database of 76 clinical mammograms containing 1120 MCs, use free-response receiver operating characteristic curves to evaluate detection performance, and compare the proposed

algorithm with several existing methods. In their experiments, the proposed SVM framework outperformed all the other methods tested. In particular, a sensitivity as high as 94% was achieved by the SVM method at an error rate of one false-positive cluster per image. The ability of SVM to outperform several well-known methods developed for the widely studied problem of MC detection suggests that SVM is a promising technique for object detection in a medical imaging application.

Trevor Hastie and Saharon Rosset [20] derive an algorithm that can fit the entire path of SVM solutions for every value of the cost parameter, with essentially the same computational cost as fitting one SVM model. They illustrate their algorithm on some examples, and use their representation to give further insight into the range of SVM solutions.

### *G. Fuzzy Logic*

The simplest fuzzy rule-based classifier is a fuzzy if-then system, similar to that used in fuzzy control. Accordingly, fuzzy classification is the process of grouping individuals having the same characteristics into a fuzzy set. A fuzzy classification corresponds to a membership function μ that indicates whether an individual is a member of a class, given its fuzzy classification predicate ~Π.

$$\mu: \sim PF \times U \longrightarrow \sim T$$

Here, ~T is the set of fuzzy truth values (the interval between zero an one). The fuzzy classification predicate ~Π corresponds to a fuzzy restriction "i is R"of U, where R is a fuzzy set defined by a truth function. The degree of membership of an individual i in the fuzzy class ~C is defined by the truth value of the corresponding fuzzy predicate.

$$\mu \sim C(i) := \tau(\sim \Pi(i))$$

Sainani Arpitha and P.Raja Prakash Rao [21] proposed a fuzzy similarity based self constructing algorithm for feature clustering. The words in the feature vector of a document set are grouped into clusters, based on similarity test. Words that are similar to each other are grouped into the same cluster. Each cluster is characterized by a membership function with statistical mean and deviation. When all the words have been fed in, a desired number of clusters are formed automatically. They have one extracted feature for each cluster. The extracted feature, corresponding to a cluster, is a weighted combination of the words contained in the cluster. By this algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data. Besides, the user need not specify the number of extracted features in advance, and trial and error for determining the appropriate number of extracted features can then be avoided. Results show that method can run faster and obtain better extracted features than other methods.

S. Sendhilkumar and K. Selvakumar [24] proposed a fuzzy based user classification model to suit a personalized web search environment. The user browsing data is collected using an established customized designed to suit personalization. The data is fuzzified and fuzzy rules are generated by applying decision trees. Using fuzzy rules, the search pages are labelled to aid grouping of user search interests. Evaluation of the proposed approach proves to be better when compared with Bayesian classifier.

### *H. Genetic Algorithm*

A genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic (also sometimes called a metaheuristic) is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

Mohd Najwadi Yusoff and Aman Jantan [25] proposed the usage of Genetic Algorithm (GA) as an approach to optimize Decision Tree (DT) in malware classification. New classifier is developed by combining GA with DT that called as Anti-Malware System (AMS) Classifier. Experimental results obtained from AMS Classifier and DT is compared. AMS Classifier shows an accuracy increase from 4.5% to 6.5% from DT Classifier. Outcome from this paper is a new Anti-Malware Classification System (AMCS) consists of AMS Classifier and new malware classes that named as Class Target Operation (CTO). Malware is classified by using CTO which are mainly based on malware target and its operation behavior.

Revathi N and Anjana Pete [26] presented a genetic algorithm and dynamic neural network based approach for web text classification. The introduction of GA in this system has helped in retrieving the most optimal features. The dynamic neural network is scalable and can be used for document text classification without requiring experimentation with parameter settings or network architectural configurations. Overall, results show that DAN2 performs better than both kNN and SVM, for Reuters-21578 dataset. The excellent uniform performance of dynamic neural network indicates the robustness of the approach and its generality. The transformation and dimensionality reduction property of dynamic neural network makes its application to text classification especially important.

### *I. Rough Sets*

A rough set is a formal approximation of a crisp set (i.e., conventional set) in terms of a pair of sets which give the lower and the upper approximation of the original set. In the standard version of rough set theory the lower- and upper-approximation sets are crisp sets, but in other variations, the approximating sets may be fuzzy sets. Rough Set Theory (RST) is based on mathematical concept can handle vagueness in classification of data. However, prior to applying RST, the data is discretized and selection of discretization procedure has great impact on classification accuracy. The theory of RS can be used to find dependence relationship among data, discover the patterns of data, learn common decision-making rules, reduce all redundant objects and attributes and seek the minimum subset of attributes so as to attain satisfying classification

Aboul Ella Hassanien and Jafar M.H. Ali [27] presented a rough set method for generating classification

rules from a set of observed 360 samples of the breast cancer data. The attributes are selected, normalized and then the rough set dependency rules are generated directly from the real value attribute vector. Then the rough set reduction technique is applied to find all reduces of the data which contains the minimal subset of attributes that are associated with a class label for classification. Experimental results from applying the rough set analysis to the set of data samples are given and evaluated.. The study showed that the theory of rough sets seems to be a useful tool for inductive learning and a valuable aid for building expert systems.

Nandita Sengupta and Jaya Sil [28] showed in their paper, network traffic data is classified using rough set theory where discretization of data is a necessary preprocessing step. Different discretization methods are available and selection of one has great impact on classification accuracy, time complexity and system adaptability. Three discretization methods are applied on continuous KDD network data namely, rough set exploration system (RSES), supervised and unsupervised discretization methods to evaluate the classifier accuracy. It has been observed that supervised discretization yields best accuracy for rough set classification and provides system adaptability.

## III. CONCLUSION

This paper shows a study of types of bug reports and various classification techniques widely used in data mining. Data mining is a process of knowledge discovery in data set. Data mining is widely used in business (insurance, banking, retail), and science research (astronomy, medicine). Classification is a data mining (machine learning) technique used to predict group membership for data instances. Classification is the task of generalizing known structure to apply to new data. The classification algorithm described in an interesting combination of approaches. This study presents various data classification methods which are commonly used in dataming.

### REFERENCES


[1] I.H. Witten, E. Frank and M.A. Hall, *Data mining practical machine learning tools and techniques,* Morgan Kaufmann publisher, Burlington 2011.

[2] J. Han and M. Kamber, Data *mining concepts and techniques*, Morgan Kaufmann, San Francisco 2006.

[3] A. Darwiche, *Modeling and Reasoning with Bayesian Networks*, Cambridge University Press, 2009.

[4] Paedro Domingos & Michael Pazzani, "On *the Optimality of the Simple Bayesian Classifier under Zero-One Loss*", Springer, pp 103-130, 1997

[5] Irina Rish," *An empirical study of the naïve bayes classifier*" , IBM Research Report, pp. 1-7, 2001.

[6] Jiangtao Ren and Sau Dan Lee, "Naïve *Bayes Classification of Uncertain Data",* IEEE, pp 944-949, 2009.

[7] B V Chowdary & Annapurna Gummadi, "*Decision Tree Induction Approach for Data Classification Using Peano Count Tree*", IJARCSSE, pp 475-479, 2012.

[8] A. S. Galathiya & A. P. Ganatra, "*Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning ",* IJCSIT, pp 3427-3431, 2012.

[9] Li Xiong & Subramanyam Chitti, "Mining *Multiple Private Database Using a k-NN Classifier",* SAC, pp 435-440, 2007.

[10] N. Suguna and Dr. K. Thanushkodi, "An *Improved k-Nearest Neighbor Classification Using Genetic Algorithm*", IJCSI, pp 18-21, 2010.

[11] Ming Yao," *Research on Learning Evidence Improvement for kNN Based Classification Algorithm*", IJDTA, pp 103-110, 2014.

[12] M. Thangaraj & C.R.Vijayalakshmi, "*Performance Study on Rule-based Classification Techniques across Multiple Database Relations*", IJAIS, pp 1-7, 2013.

[13] Ritika & Aman Paul," *Prediction of Blood Donors Population using Data Mining Classification Technique*", IJARCSSE, pp 634-638, 2014.

[14] Hongium Lu," *Effective Data Mining Using Neural Networks*", IEEE, pp 957-961, 1996.

[15] B.Madasamy & Dr.J.Jebamalar Tamilselvi, "*Improving Classification Accuracy of Neural Network through Clustering Algorithms*", IJCTT, pp 3242-3246, 2013.

[16] Issam El-Naqa and Yongyi Yang," *A Support Vector Machine Approach for Detection of Microcalcifications"*, IEEE, pp 1552-1563, 2002.

[17] B. Scholkopf, S. Kah-Kay, C. J. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "*Comparing support vector machines with Gaussian kernels to radial basis function classifiers*,"IEEE Trans. Signal Processing, vol. 45, pp. 2758–2765, Nov. 1997.

[18] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: Application to face detection," in Proc. Computer Vision and Pattern Recognition, Puerto Rico, pp. 130–136, 1997.

[19] C. J. Burges, "A tutorial on support vector machines for pattern recognition, "Knowledge Discovery and Data Mining, vol. 2, pp. 121–167, 1998.

[20] Trevor Hastie and Saharon Rosset," *The Entire Regularization Path for the Support Vector Machine*", Journal of Machine Learning Research, pp 1391-1415, 2004.

[21] Sainani Arpitha and P.Raja Prakash Rao,"*Clustering Algorithm for Text Classification Using Fuzzy Logic*", IJARCSSE, pp 258-262, 2012.

[22] H. Kim, P. Howland, and H. Park, "*Dimension Reduction in Text Classification with Support Vector Machines*," J.Machine Learning Research, vol. 6, pp 37-53, 2005.

[23] F. Sebastian, "*Machine Learning in Automated Text Categorization*,"ACM Computing Surveys, vol. 34,no. 1, pp. 1-47,2002.

[24] S. Sendhilkumar and K. Selvakumar," *Application of Fuzzy Logic for User Classification in Personalized Web Search",* IJCI, pp 23-49, 2014.

[25] Mohd Najwadi Yusoff and Aman Jantan,"*Optimizing Decision Tree in Malware Classification System by using Genetic Algorithm ",* IJNCCA, pp 694-713, 2011.

[26] Revathi N and Anjana Pete," *Web Text Classification Using Genetic Algorithm and a Dynamic Neural Network Model*", IJARCET, pp 436-442, 2013.

[27] Aboul Ella Hassanien and Jafar M.H. Ali, "Rough Set Approach for Generation of Classification Rules of Breast Cancer Data", Informatica, pp 23-38, 2004.

[28] Nandita Sengupta and Jaya Sil,*" Evaluation of Rough Set Theory Based Network Traffic Data Classifier Using Different Discretization Method*", IJIEE, pp 338-341, 2012.

[29] M. Garofalakis, D. Hyun, R. Rastogi and K. Shim, *"Building Decision Trees with Constraints"*, Data Mining and Knowledge Discovery, vol. 7, no. 2, 2003, pp. 187 – 214.

[30] Y. Singh Y, A.S. Chauhan, "*Neural Networks in Data Mining*", Journal of Theoretical and Applied Information Technology, 2005, pp. 37-42 International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012 ISSN: 2278-0181.

[31] "*Survey of Nearest Neighbor Techniques" Nitin Bhatia* (Corres. Author) Department of Computer Science DAV College Jalandhar, Vandana SSCS Deputy Commissioner's Office Jalandhar.

[32] "*Decision Trees*" Andrew W. Moore Professor School of Computer Science Carnegie Mellon University.

[33] "*Top 10 algorithms in data mining",* XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg © Springer-Verlag London Limited 2007.

[34] "*Artificial Neural Networks* "Ajith Abraham Oklahoma State University, Stillwater, OK, USA.

[35] Abraham, A. (2004) "*Meta-Learning Evolutionary Artificial Neural Networks, Neurocomputing* "Journal, Vol. 56c, Elsevier Science, Netherlands, (1–38).

[36] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein," *Introduction to Algorithms"*. MIT Press, 2001.

[37] Michael Gegick, Pete Rotella and Tao Xie, "*Identifying Security Bug Report Via Text Mining":* An Industrial case Study IEEE, pp.11-20, 2010.

[38] Drazen.S. "*Estimation of difficult to Measure process variables using neural networks*" Proceedings of IEEE MELECON , May 12-15, 2004.

[39] Imdad Ali Rizvi, B.Krishna Mohan "*Improving the Accuracy of Object Based Supervised Image Classification using Cloud Basis Function Neural Network for High Resolution Satellite Images*" International Journal of Image Processing (IJIP), Volume (4) Issue (4).

[40] Zhou Yong Li Youwen and Xia Shixiong "*An Improved KNN Text Classification Algorithm Based on Clustering*" Journal of computers, vol. 4, no. 3, march 2009.

[41] Ankit R. Deshmukh, P Sunil R. Gupta, "*Data Mining Based Softcomputing Methods for Web Intelligent",* IJAIEM, pp 376-382, 2014.

[42] Kanhaiya Lal, N.C.Mahanti, "*Role of soft computing as a tool in data mining",* IJCSIT, pp 526- 537, 2011.

[43] Ivo D., and G. Gunther "*Statistical evaluation of rough set dependency analysis*" International Journal of Human-Computer Studies, pp 589–604, 1997.

[44] Kent, R.E." *Rough concept analysis, rough sets, fuzzy sets knowledge discovery*" In W.P. Ziarko (Ed.), Proceedings of the International Workshop on Rough, Sets, Knowledge, Discovery Banff, Alta., Canada.Springer–Verlag. pp. 248–255, 1994.

[45] Kryszkiewicz, M., and H. Rybinski "*Finding reducts in composed information systems, rough sets, fuzzy sets knowledge discovery*" In W.P. Ziarko (Ed.), Proceedings of the International Workshop on Rough Sets,Knowledge, Discovery, Banff, Alta., Canada. Springer, pp. 261–273, 1994.

[46] P. Blajdo, J. W. Grzymala-Busse,Z. S. Hippe, M. Knap, T. Mroczek,and L. Piatek, "*A comparison of six approaches to discretization rough set perspective",* Rough Sets and Knowledge Technology, Lecture Notes in Computer Science, vol. 5009, pp. 31-38, 2008.

[47] L. Gaojun and Z. Yan, "*Credit assessment of contractors: a rough set method*," Tsinghua Science and Technology, vol. 11, no. 3, 2006.

[48] S. Kumar, S. Atri, and H. L. Mandoria, "A Combined Classifier to Detect Landmines Using Rough Set Theory and Hebb Net Learning and Fuzzy Filter as Neural Networks," in Proc. ICSPS, 2009.

[49] J. Zhang, J. Wang, D. Li, H. He, and J. Sun, "A new heuristic reduct algorithm base on rough sets theory," Advances in Web-Age Information Management, Lecture Notes in Computer Science, pp , 247-253, 2003

[50] Ning, S., H. Xiaohua, W. Ziarko and N. Cercone " *A generalized rough sets model*", In Proceedings of the 3rd Pacific Rim International Conference on Artificial Intelligence, pp. 437–443,1994.

[51] De Jong, K. A., Spears, W. M., and Gordon, D. F. "Using *genetic algorithms for concept learnin*". Machine Learning, 13, 161-188, 1993.

[52] Greene, D., P. and Smith, S. F. "*Competition-based induction of decision models from examples*" Machine Learning, 13, 229-257, 1993.

[53] Jacobs, R.A." *Increased Rates of Convergence Through Learning Rate Adaptation. Neural Networks*", pp 295-307, 1988

[54] Mingers, J." *Expert Systems – Rule Induction with Statistical Data*". Journal of the Operational Research Society, pp 39-47, 1987.

[55] C. J. C. Burges and B. Scholkopf."*Improving the accuracy and speed of support vector learning machines*", Advances in Neural Information Processing Systems 9, pp 375–381, Cambridge, MA, 1997. MIT Press.

[56] B. E. Boser, I. M. Guyon, and V .Vapnik "*A training algorithm for optimal margin classifier*", In Fifth Annual Workshop on Computational Learning Theory, pp 144–152, Pittsburgh, 1992. ACM.

[57] Luc Devroye, Laszlo Gyorfi, and G abor Lugosi"*A Probabilistic Theory of Pattern Recognition*." SpringerVerlag, Applications of Mathematics Vol. 31, 1996.

[58] M. Anthony and N. Biggs. Pac "*learning and neural networks*" InThe Handbook of Brain Theory and Neural Networks, pages 694–697, 1995.

[59] I. Guyon, V. Vapnik, B. Boser, L. Bottou, and S.A. Solla "*Structural risk minimization for character recognition*"Advances in Neural Information Processing Systems, pp 471–479, 1992.

[60] Jorge J. More and Gerardo Toraldo."*On the solution of large quadratic programming problems with bound constraints*" SIAM J. Optimization, pp 93–113, 1991.