Survey on General Classification Techniques for Effective Bug Triage

Nitu Bhardwaj[#1], A.S Bhattacharya[*2]

[1]*ME Student, Computer Science and Engineering, G. H. Raisoni College of Engineering for Women, RTMNU, Nagpur India*

[2]*Lecturer, Computer Science and Engineering, G. H. Raisoni College of Engineering for Women, RTMNU, Nagpur 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. Data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data, data warehouse data and transactional data. This paper presents a survey on several classification techniques for effective bug triage which are generally used for data mining such as naïve bayes, decision tree, K- nearest neighbor, Rule based, neural network etc.*

Keywords —*Bug report, classification, naïve bayes, decisiontree, K-nearest neighbor, Rule based, neural network.*

I. INTRODUCTION

With the increasing growth of data in every application, data mining meets the imminent need for effective, scalable, and flexible data analysis in our society. Data mining is the process of discovering interesting patterns from huge amounts of data[2]. 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 labelled 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 [7].

This paper presents a study on Statistical and Software 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 neighbour, Naïve Bayes, Rule base learning, Decision tree, Neural Network learning and Support vector machines. Software Computing consists of different 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 neighbour, Rule based neural network, support vector machines, fuzzy logic, genetic algorithm and rough sets.

II. CLASSIFICATION METHODS

A. *Naive Bayes Classifier*

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifier. Naïve Bayesian classification is based on Bayes theorem of posterior probability. It assumes class conditional independence that the effect of an attribute value on a given class is independent of the values of the other attributes.

Jiangtao Ren & Sau Dan Lee [4] proposed a novel naive Bayes classification algorithm for uncertain data with a 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 naïve 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 learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. These tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.
Earlier studies of decision tree B VChowdary& Annapurna Gummadi [8] 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 [9] performs a comparison analysis of various decision tree classifiers such as ID3, C4.5 and C5.0. C5.0 classifier performs 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.

B. *K-NN (K-NEAREST NEIGHBOR*

In pattern recognition, the *k*-Nearest Neighbors algorithm is a non-parametric method used for classification.N. Suguna and Dr. K. Thanushkodi [5] presented 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 [5] explored 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 domains was news articles. He applied three experiments on these datasets to verify the validity of evidence learnt from the heterogeneous distance metric sources.

C. *Rule Based*

Rules are a good way of representing information or bits of knowledge. Rule based algorithm provide process that generates rule by concentrating on a specific class at onetime 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 form: "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] presented 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, and number of foreign-keys, classification accuracy and runtime

D. *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 byhuman brains.
B.Madasamy&Dr.J.JebamalarTamilselvi [13] 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 biomedical 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 method for the classification of both linear and nonlinear data. An SVM is an algorithm that work as follows: It uses a nonlinear mapping to transform the original training data into higher dimension. With this new dimension, it searches for the linear optimal separating hyper plane i.e. a "decision boundary" separating the tuples of one class from another.

Trevor Hastie and SaharonRosset [1] 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

Fuzzy logic uses truth values between 0.0 and 1.0 to represent the degree of membership that a certain values has in a given category. Each category then represents a fuzzy set. Fuzzy set theory is also called as possibility theory .It was proposed by It lets us work at a high abstraction level and offers a means for dealing with imprecise data measurement.S. Sendhil kumar and K. Selva kumar [16] proposed a fuzzy based user classification model to suit a personalized web search environment. The data is fuzzified and fuzzy rules are generated by applying decision trees. Sainani Arpitha and P.Raja Prakash Rao [10] proposed a fuzzy similarity based self-constructing algorithm for feature clustering. These words in the feature vector of a document set are grouped into clusters, based on the similarity test.

H. Genetic Algorithm

Genetic algorithmincorporates ideas of natural evolution. An initial population is created consisting of randomly generated rules. Each rule can be represented by a string of bits. The process of generating new populations based on prior population P, evolves where each rule in P satisfies a pre-specified fitness threshold. A genetic algorithm is easily parallelization and have been used for classification as well as other optimization problems. In data mining, they may be used to evaluate the fitness of other algorithms.

Revathi N and Anjana Pete [14] 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.

I. Rough Set

Rough set theory can be used for classification to discover structural relationships within imprecise or noisy data. It applies to discrete-valued attributes. Continuous-valued attributes must therefore be discredited before its use. A rough set definition for a given class, C, is approximated by two sets –a lower approximation of C and an upper approximation of C. Rough Set Theory (RST) is based on mathematical concept can handle vagueness in classification of data. However, priorto applying RST, the data is discredited 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.

NanditaSengupta and Jaya Sil [11] showed in their paper, network traffic data was classified using rough set theory where discretization of data is an important preprocessing step. 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.COMPARATIVE ANALYSIS

Different technique has got its own feature and limitations. Each classification technique have its advantages and disadvantages.C4.5 Algorithm produces the accurate result. It takes the less memory to large program execution and takes less model build time. Its searching time is also less. ID3 Algorithm produces the more accuracy result than the C4.5 algorithm. ID3 generally uses nominal attributes for classification with no missing values. It produces false alarm rate and omission rate decreased, increasing the detection rate and reducing the space Consumption and has long searching time. It takes the more memory than the C4.5 to large program execution. Naive Bayes Algorithm improves the classification performance by removing the irrelevant features and have good performance with short computational time. The naive Bayes classifier requires a very large number of records to obtain good results. It is instance-based or lazy in that they store all of the training samples. Support vector machine Algorithm Produces very accurate classifiers, less over fitting. Especially popular in text classification problems where very high-dimensional spaces are the norm. SVM is a binary classifier. To do a multi-class classification, pair-wise

classifications can be used (one class against all others, for all classes). Computationally expensive, thus runs slow. K-Nearest Neighbor Algorithm is an easy to understand and easy to implement classification technique. KNN robust to noisy training data. It is particularly well suited for multimodal classes and sensitive to the local structure of the data. Being a supervised learning lazy Algorithm i.e., runs slowly.

TABLE I: FEATURES

SNO.	ALGORITHM	FEATURES	LIMITATIONS
1.	Naïve Bayes Algorithm	Simple to implement. Good Computation efficiency & classification rate. Predicts accurate results for most classification & prediction problems.	The prediction of algorithm decreases if amount of data is less. For obtaining good results it requires large number of records
2.	Support Vector Machine	High accuracy. Work even if data is not linearly separable in the base feature space.	Speed and Size requirement both in training and testing is more High complexity and memory requirements for classification in many cases
3.	Artificial Neural Network Algorithm	A neural network learns and reprogramming is not needed. Easy to implement Applicable to wide range of problems in real life.	Requires high processing time if network is large Learning can be slow Difficult to know how many neurons & layers are necessary
4.	K- Nearest Neighbor Algorithm	Classes need not be linearly separable. Zero cost of the learning process. Sometimes its robust with regard to noisy training data Well suited for multimodal classes.	Time to find nearest neighbor in a large training data set can be excessive Sensitive to noisy or irrelevant attributes. Performance depends on number of dimensions used

TABLE II: ADVANTAGES & LIMITATIONS

Technique	Advantages	Limitations
Naive Bayes Classifier	Super simple, you're just doing a bunch of counts. If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. Even if the NB assumption doesn't hold, a NB classifier still often performs surprisingly well in practice.	Very simple representation doesn't allow for rich hypotheses Assumption of independence of attributes is too constraining
Decision Tree	Easy to interpret and explain Non-parametric, so you don't have to worry about outliers or whether the data is linearly separable (e.g., decision trees easily take care of cases where you have class A at the low end of some feature x, class B in the mid-range of feature x, and A again at the high end).	May over fit data May get stuck in local minima so need ensembles to help reduce the variance
K-NN (K-nearest neighbour)	The main disadvantage of the KNN algorithm is that it is a *lazy learner*. To predict the label of a new instance the KNN algorithm will find the *K* closest neighbours to the new instance from the training data, the predicted class label will then be set as the most common label among the *K* closest neighbouring points. The algorithm must compute the distance and sort all the training data at each prediction, which can be slow if there are a large number of training examples. Further, changing *K* can change the resulting predicted class label.	Take up a lot of memory to run (storing all the instances) Work well for a small number of dimensions, but not a high number of dimensions
Neural Network	Neural Network is required to model the given problem. The advantage area: NN does not require any model Apriori. The model need not assume any model structure before starting the ANN model. It can be used for non-linear problems? It is a non-parametric	Picking the correct topology is difficult Training takes a long time/requires a lot of data Output/issues are

	method, thus eliminates the error in parameter estimation.	incomprehensible
Support Vector Machine	High accuracy, Nice theoretical guarantees regarding over fitting and with an appropriate kernel they can work well even if you're data isn't linearly separable in the base feature space. Especially popular in text classification problems where very high-dimensional spaces are the norm. Memory-intensive and kind of annoying to run and tune, though, so I think random forests are starting to steal the crown.	Picking/finding the right kernel can be a challenge Results/output are incomprehensible No standardized way for dealing with multi-class problems; fundamentally a binary classifier

IV. 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 form of data analysis that extracts models describing important data classes called as classifier, predict categorical class label. The classification algorithm described in an interesting combination of approaches. This study presents various data classification methods which are commonly used in data mining to improve the accuracy of bug reports. Each technique has got its own feature and limitations as given in the paper. Based on the Conditions, corresponding performance and feature each one as needed can be selectedfor effective bug triage.

REFERENCES

[1] Trevor Hastie and SaharonRosset," *The Entire Regularization Pathfor the Support Vector Machine*", Journal of Machine LearningResearch, pp 1391-1415, 2004

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

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

[4] JiangtaoRen and Sau Dan Lee, "Naïve *Bayes Classification ofUncertain Data",* IEEE, pp 944-949, 2009

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

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

[7] KanhaiyaLal, N.C.Mahanti, "*Role of soft computing as a tool indata mining",* IJCSIT, pp 526- 537, 2011.

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

[9] A. S. Galathiya& A. P. Ganatra, "*Improved Decision Tree InductionAlgorithm with Feature Selection, Cross Validation, ModelComplexity and Reduced Error Pruning ",* IJCSIT, pp 3427-3431,2012.

[10] SainaniArpitha and P.Raja Prakash Rao,"*Clustering Algorithm forText Classification Using Fuzzy Logic*", IJARCSSE, pp 258-262,2012.

[11] NanditaSengupta and Jaya Sil,*" Evaluation of Rough Set TheoryBased Network Traffic Data Classifier Using DifferentDiscretization Method*", IJIEE, pp 338-341, 2012.

[12] M. Thangaraj&C.R.Vijayalakshmi, "*Performance Study on RulebasedClassification Techniques across Multiple DatabaseRelations*", IJAIS, pp 1-7, 2013.

[13] B.Madasamy&Dr.J.JebamalarTamilselvi, "*ImprovingClassification Accuracy of Neural Network through ClusteringAlgorithms*", IJCTT, pp 3242-3246, 2013.

[14] Revathi N and Anjana Pete," *Web Text Classification Using GeneticAlgorithm and a Dynamic Neural Network Model*", IJARCET, pp436-442, 2013

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

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

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