Classification of Bug Report Using Naïve Bayes Classifier with Gain Ratio

Smita Mishra and Somesh Kumar
Noida Institute of Engineering & Technology, Greater Noida
smita.mishra31dec@gmail.com
someshkumarrajput@rediffmail.com

Abstract- **Bug report is a report which contains the information about the defects in the system or in the software. Generally bug report contains the issues written by the wide variety of reporters, with different levels of training and knowledge about the system being discussed. 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, since security bug reports (SBRs) contain more risk than non-security bug reports (NSBRs). In this paper we are using Naïve Bayes classifier to classify the bug reports. With naïve bayes classifier, feature subset selection method such as Gain Ratio is applied to rank the attributes of the dataset. Gain Ratio is utilized as an iterative process where we select smaller sets of features in incremental manner. Result prove that the classification accuracy is high for attributes having high gain ratio and low for attributes having low gain ratio.**

Keywords- **Bug report, Classification, Naïve bayes, Feature selection, Gain ratio.**

I. INTRODUCTION

As new software systems are getting larger and more complex every day, software bugs are inevitable phenomenon. Bugs occur for a variety of reasons, ranging from ill-defined specifications, to carelessness, to a programmers misunderstanding of the problem, technical issues, non-functional qualities, corner cases, etc. There are several bug reports submitted by many users and tester for particular software or system.

Bug reports are mainly of two types: Security bug reports (SBRs) and non security bug reports (NSBRs). These report need to be labeled as security bug reports (SBRs) or non security bug reports (NSBRs).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 bug report is needed to be checked by security team of the software development or Information security management system. Non security bug is related to hardware, site, personnel vulnerabilities etc. They have lower potential to harm a system or software unlike security bug.

J. Han and M. Kamber [1] introduces Data mining and classification techniques of datamining. Data mining is the process of extraction of hidden and useful information from huge data. Classification is a task of data mining. Data mining involves various classification techniques like Naïve Bayes, neural network, fuzzy logic, GA, SVM, rough sets, decision tree, K-nearest neighbor and Rule based. In this paper we are using Naïve bayes classifier to classify our data. Naive Bayes is simple and flexible than other classification methods.

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 implemented very efficiently in a supervised learning setting.

Gain Ratio enhances Information Gain as it offers a normalized score of a feature's contribution to an optimal information gain based classification decision. Gain Ratio is utilized as an iterative process where we select smaller sets of features in incremental fashion. Gain ratio is used as one of disparity measures and the high gain ratio for selected feature implies that the feature will be useful for classification.

In this paper we are presenting a comparative analysis of classification accuracy based on those attributes which have high gain ratio and those having low gain ratio. Result show that classification accuracy is high with high gain ratio attributes and comparatively low with low gain ratio attributes.

TABLE I
Related Work List

No	Paper References	Data Set	Study Objective	Study Analysis
1	Marc Boulle (2007)[2]	Waveform dataset	Bayesian regularization technique to select the most probable subset of variables compliant with the Naive Bayes assumption.	The limits of Bayesian model averaging in the case of the Naive Bayes assumption and introduction a new weighting scheme based on the ability of the models to conditionally compress the class labels
2	Michael Gegick (2010) [3]	Cisco Software system	Automatic identification of security bugs based on natural language descriptions	Effectively automates the identification of SBRs
3	R.Praveena and S. Sivakumari (2011) [4]	UCI Machine Learning Repository	The accuracy of the privacy preserved reduced datasets and the original datasets are compared	Analysis slow classification and clustering accuracy are comparatively the same for reduced k-anonym zed and the original datasets.
4	Sangeeta Lal (2012) [5]	Google Chromium Browser	Identify several metrics and characteristics serving as dimensions on which various types of bug reports can be compared	Calculated metrics shows the similarities and difference on various dimensions for seven different types of bug reports
5	Astha Chharia (2013) [6]	Spam Assassin, PUI and lingspam	Enhance the performance of naïve bayes classifier in classifying spam mails by proposing a modification to Absolute discount smoothing method against the laplace method of traditional naïve bayes classifier.	Method not only achieves greater accuracy as compared to Laplace but also reduces false positives
6	R.S. Anu Gowsalya (2014) [7]	Real world traffic dataset	Improve the classification performance effectively by incorporating correlated information into the classification process.	Analyze the new classification approach and its performance benefit from both theoretical and empirical perspectives.

II. RELATED WORK AND RESEARCH CONTRIBUTIONS

In this section we presented related work and research contribution to our study in this paper. Table I shows a list of related papers with publication years. Table I characterizes six papers based on paper reference, experimental data set, study objective and study analysis.

Marc Boulle [2] introduces a Bayesian regularization technique to select the most probable subset of variables compliant with the naïve bayes assumption. He studied the limits of bayesian model averaging in the case of the naïve bayes assumption and introduced a new weighting scheme based on the ability of the models to conditionally compress the class labels. Experimental and theoretical results indicate that the posterior distribution of the models is exponentially peaked. Compression based model averaging scheme clearly out-performs the bayesian model averaging scheme.

M Gegick, P. Rotella, and T.Xie [3] developed a new approach that applies text mining on natural-language descriptions of bug reports to train a statistical model on already manually mislabeled as non-security bug reports (NSBRs). They evaluated the model's predictions on a large Cisco software system with over 10 million source lines of code. In their approach they identified a high percentage of SBRs mislabeled as NSBRs by bug reporters for a large Cisco software system. Their approach effectively automates the identification of SBRs based on natural language information present in bug reports.

R. Praveena Priyadarsini and S. Sivakumari [4] compared K- anonym zed original and reduced data sets for comparing the accuracy on both data mining task classification and clustering. Results show the accuracy level remained the same for K-anonym zed original data sets and reduced data sets for the both data mining functionalities.

Sangeeta Lal and Ashish Sureka [5] performed a case-study on Google Chromium Browser open-source project and conducted a series of experiments to calculate various metrics. They identified several metrics and characteristics serving as dimensions on which various types of bug reports can be compared. They presented a comparison study on different types of bug reports on metrics such as: statistics on close-time, number of stars, number of comments, discriminatory, entropy across reporters, entropy across component, opening and closing trend, continuity and debugging efficiency performance characteristics.

Astha Chharia and R. K. Gupta [6] they studied the performance of naïve bayes classifier and found that it largely depends on the smoothing method, which aims to adjust the probability of an unseen event from the seen event, that arises due to data sparseness. Therefore in that paper, they aim to enhancing the performance of naïve bayes classifier in classifying spam mails by proposing a modification to Absolute Discount smoothing method against the Laplace method of traditional naïve bayes classifier. In addition, they have introduced a cost metric to compare their approach with the traditional scheme. Their experimental results have shown that their method not only achieves greater accuracy as compared to Laplace but also reduces false positives, which is more serious problem in spam classification.

R.S. Anu Gowsalya and S. Miruna Joe Amali [7] presented a novel traffic classification scheme which is used to improve classification when few training data is available. In the proposed scheme, traffic flows are described using the discretized statistical features and flow correlation information is modeled by bag-of-flow (BoF). A novel parametric approach for traffic classification, which can improve the classification performance effectively by incorporating correlated information into the classification process. Then analyze the new classification approach and its performance benefit from both theoretical and empirical perspectives. Finally, a large number of experiments are carried out on large-scale real-world traffic datasets to evaluate the proposed scheme. The experimental results show that the proposed scheme can achieve much better classification performance than existing state of the art traffic classification methods. classification performance than existing state of the art traffic classification methods.

III. APPROACH

Our approach consists of three main steps. First step is to collect the dataset from bugzilla for train data and test data which contain a summary of Bug Reports. This summary consist of attributes and labels of bug reports either as Security Bug report (SBR) or Non security Bug report (NSBR). The second step is to Train the model with train data and Test the model through test data. Third step is the evaluation step which estimates the accuracy of the classification model. Below we intend to give a brief description of the steps involved in our approach.

A. Data Set Collection

Data Collection step prepares/collects the dataset. We obtain Bug Reports from bugzilla.mozilla.org. Bugzilla is an open source bug tracking system, which has several bug report, we make a summary of these reports. We have two types of dataset: one is train data to train the model and other is test data to test the model. Datasets contain eight attributes and category (SBR or NSBR) for each bug report.

Table II shows the dataset information for both data sets which are used in our experiment. Train data contains 1064 entries and Test data holds 1136 entries of bug reports. Fig. 1 and 2 are showing the graph view of distinct entries in train and test data.

Table III shows the eight attributes and category of the data set. These attributes are Id, Product, Component, Assignee, Status, Resolution, Changed and Summary. Product, Component, Assignee, Status and Resolution have same distinct entries in train data and test data. **Product** has 7 distinct entries (Bugzilla, Camino, Firefox, Toolkit, Mail News Core, Sea Monkey, Core), Component consists of 4 distinct entries (Backend, Security, UI Design, Networking), **Assignee** has 5 distinct entries (doug. turner, mail, mozilla, kaie, nobody) **Status** holds 3 distinct entries (Closed, Verified, Resolved), **Resolution** having 5 distinct entries (Expired, Duplicate, Fixed, Invalid, Works forms).

TABLE II
Dataset Information

DATASET INFORMATIONata Set	Train Data	Test Data	No. of Category
2200	1064	1136	2 (SBR and NSBR)

TABLE III
Attributes Information

Datasets	Attribute Name	No. of Distinct Entries	Entries
Train Data	Id	1064	Too big to show
	Product Component	7 4	Bugzilla, Camino, Firefox, Toolkit, Mail News Core, Sea Monkey, Core Backend, Security, UI Design, Networking
	Assignee	5	doug.turner, mail, mozilla, kaie, nobody
	Status	3	Closed, Verified, Resolved
	Resolution	5	Expired, Duplicate, Fixed, Invalid, Worksforme
	Changed	746	Too big to show
	Summary	1050a	Too big to show
Test Data	Id	1136	Too big to show
	Product	7	Bugzilla, Camino, Firefox, Toolkit, Mail News Core, Sea Monkey, Core
	Component	4	Backend, Security, UI Design, Networking
	Assignee	5	doug.turner, mail, mozilla, kaie, nobody
	Status	3	Closed, Verified, Resolved
	Resolution	5	Expired, Duplicate, Fixed, Invalid, Worksforme
	Changed	827	Too big to show
	Summary	1122	Too big to show

Category has two entries SBR (security bug report) and NSBR (non security bugreport).

Attributes Id, Changed and Summary have different distinct entries in both the data sets. **Id** has 1064 distinct entries in train data and 1136 distinct entries in test data, **Changed** consists of 746 distinct entries in train data and 827 distinct entries in test data, and **Summary** having 1050 distinct entries in train data and 1122 distinct entries in test data. Again in this case too Category has two entries SBR (security bug report) and NSBR (non security bugreport).

B. Training and Testing the Model

Training includes three sub-steps. First load train data, Second apply Gain ratio for attribute selection and third apply Naïve bayes classifier on selected attributes for classification of Bug Reports (BRs). Fig.3 shows the training data set.

Attribute selection play an important role in data mining. Asha Gowda, A.S. Manjunath & M.A.Jayaram [9] perform the comparative study of gain ratio and correlation based feature selection method for classifying Pima Indian Diabetic database and result shows that feature selection by CFS filter (correlation based feature selection) has marginal improvement when compared to information gain filter.

High dimension data makes training and testing tasks difficult. Attributes selection is a method of selecting small subsets of attributes from a large dataset. The goal of attribute selection is to avoid selecting attributes which are not or less necessary. In this paper we are using Gain Ratio for attribute selection.

i) Gain Ratio:

Gain Ratio is a modification of the information gain. Gain Ratio is utilized as an iterative process where we select smaller sets of features in incremental fashion. These iterations terminate when there is only predefined number of features remaining. Gain ratio is used as one of disparity measures and the high gain ratio for selected feature implies that the feature will be useful for classification. It corrects the information gain by taking the split information. Split info is a value based on the column sums or sum of distinct entries of the attribute. Gain Ratio was firstly used in decision tree (C4.5), and applies normalization to information gain score by utilizing a split information value.

$$\text{Grain Ratio(Attribute Name)} = \frac{\text{Information Gain (Attribute Name)}}{\text{Split\_info(Attribute Name)}}$$

Anuj Sharma & Shubhamoy Dey [10] perform sentiment analysis and investigate performance of feature selection methods in term of recall, precision and accuracy, and found gain ratio gives the best result .Gain ratio is used to give the rank of the attributes and is used to avoid useless or unnecessary attributes. Attributes which have high gain ratio are selected for the classification.

ii) Naïve Bayes Classifier:

This Classifier simply computes the conditional probabilities of the different classes given the values of attribute and then selects the class with the highest conditional probability. A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (Naive) independence assumptions. 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.

Irina Rish [8] demonstrates that naïve bayes works best in two cases: completely independent features and functionally dependent features.

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:

$$p\left(\frac{c}{F1, \dots \dots \dots Fn}\right)$$

Using Bayes' theorem, we write:

$$p\left(\frac{c}{F1, \dots ..Fn}\right) = \frac{p(C)\, p\,\frac{(F1 \dots Fn)}{C}}{p(F1, \dots .\, Fn)}$$

$$posterior = \frac{prior * likelihood}{evidence}$$

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

$$p\left(\frac{c}{F1, \dots ..Fn}\right) = \frac{1}{z}p(C)\prod_{i=1}^{n} p\left(\frac{Fi}{c}\right)$$

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

Test Data just like Train data consists of 2 categories (SBR and NSBR) with larger dataset. This involves load test data and classifying the data through Naive Bayes.

C. Evaluation

Third and last step is the evaluation step which evaluates our model. In this step calculate the accuracy of Bug report classification in terms of correct classification and incorrect classification of those attributes which are selected through Gain ratio. We determine the probability of SBR and NSBR, through Naïve Bayes Classifier. If the probability of NSBR is high, category of the bug report is considered as NSBR and if the probability of SBR is high, category of the bug report is considered as SBR. We compare the classification of train data and test data which show the accuracy of the classification of bug report.

IV. EXPERIMENTAL DESIGN AND RESULT

The objective of this evaluation is to compare the classification accuracy when applying gain ratio to rank the attributes. We follow our approach which has mainly three steps: load data, train-test the model and evaluate the model.

A. Experimental Design

A train dataset with 1064 entries are classified with two categories SBR and NSBR, is used for train the model. Figure 3 shows the training dataset which include 1064 entries or total count, Total NSBR count is 997, total SBR count is 67 and no of attributes are 8 (Id, Product, Component, Assignee, Status, Resolution, Changed and Summary). Table III shows the information of attributes. Because of big number of distinct entries we are not considering the attributes like Id, Changed and Summary. These attributes are not necessary.

Now apply the gain ratio on training dataset to rank the attributes. Table IV shows the rank of the attributes and Fig 4 shows the graph view of the gain ratio of the attributes. Status has highest gain ratio, Component second highest, Assignee third highest, Product fourth highest and Resolution has lowest gain ratio. Now we make ten combinations of two attributes, like Status-Component, Product-Component, Product-Assignee, Product-Status, Product-Resolution, Component-Assignee, Component-Resolution, Assignee-Resolution, Assignee-Status and Status-Resolution.

And finally Applying Naïve Bayes classifier to each of above combination and classifying the bug report based on respective combination of the attributes. Through this process we trained our model. Fig 5 shows all combinations of the attributes of train data. Now we test the model to apply the same process for the test data, we do not apply the gain ratio for the test data.

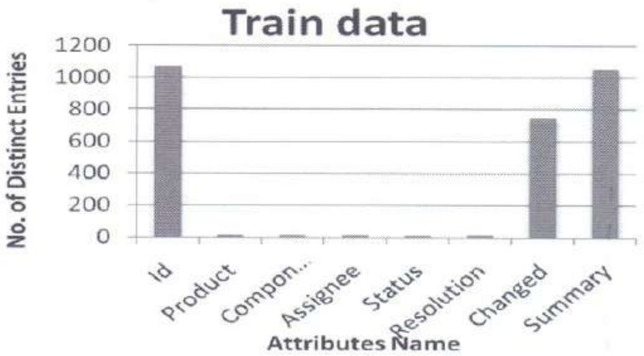

Fig. 1 shows the no. of distinct entries of the attributes for train data

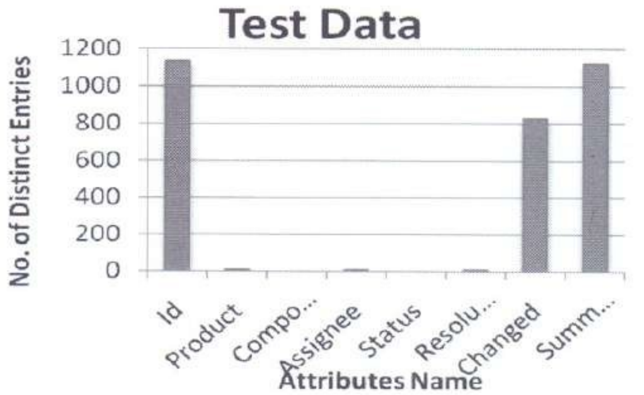

Fig. 2 shows the no. of distinct entries of the attributes for test data.

TABLE IV
Rank of the Attributres

Rank	Attributes Name	Gain Ratio
1	Status	0.67
2	Component	0.44
3	Assignee	0.29
4	Product	0.26
5	Resolution	0.24

B. *Experimental Result*

We did experiment to evaluate the accuracy of the classification of test data with all ten combinations and compare the accuracy of the ten combinations of the two attributes. In result we found that the classification accuracy of those attributes which have highest gain ratio is high as compare to other.

TABLE V
Classification result for Ranked Attributes

Attribute 1	Gain Ratio	Attribute 2	Gain Ratio	Correct Classification %	Incorrect Classification %
Status	0.67	Component	0.44	100%	0%
Product	0.26	Component	0.44	85.7%	14.3%
Product	0.26	Assignee	0.29	88.6%	11.4%
Product	0.26	Status	0.67	66.6%	33.4%
Product	0.26	Resolution	0.24	80%	20%
Component	0.44	Assignee	0.29	95%	5%
Component	0.44	Resolution	0.24	35%	65%
Assignee	0.29	Resolution	0.24	76%	24%
Assignee	0.29	Status	0.67	86.6%	13.4%
Status	0.67	Resolution	0.24	26.6 %	73.4%

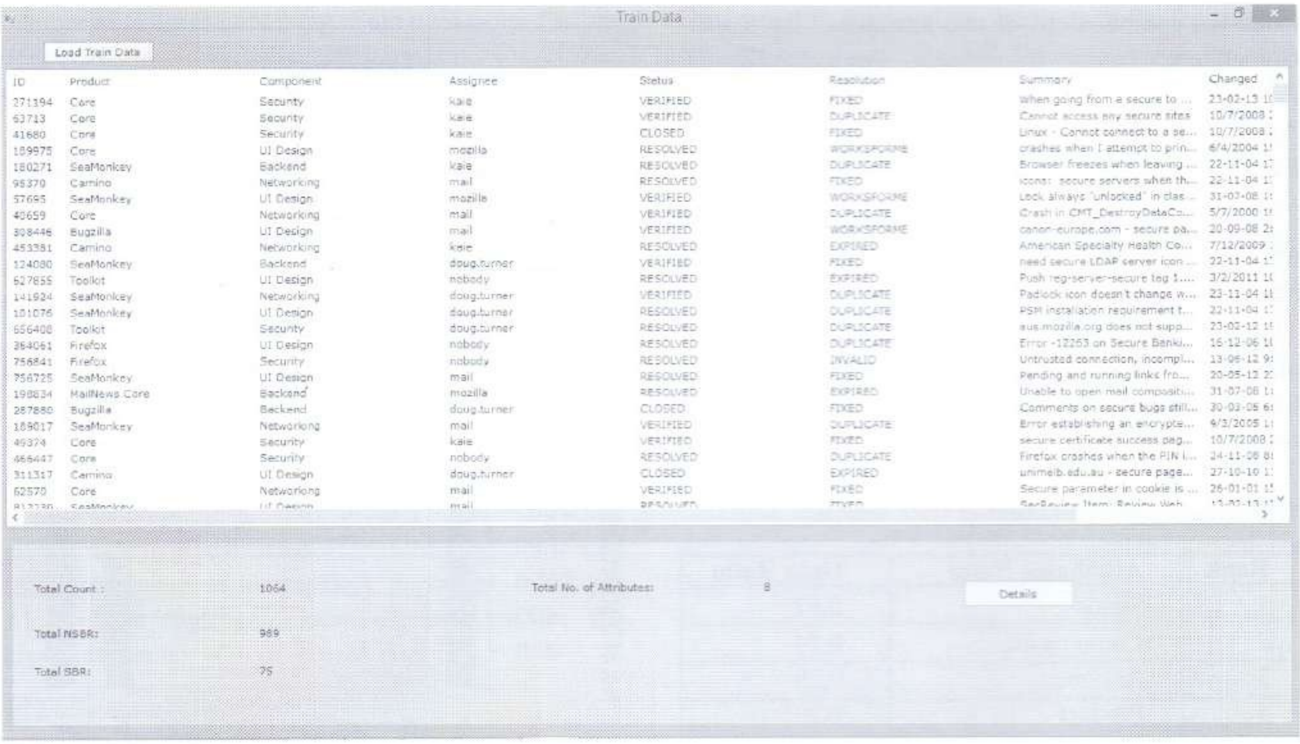

Fig. 3 Training Dataset

We found the classification accuracy of the combination of Status-Component is 100% means based on these two attributes we can classify our bug report 100% correctly which is highest than other combinations like Product- component which show 85.7 % correct classification and 14.3% incorrect classification, Product-Assignee

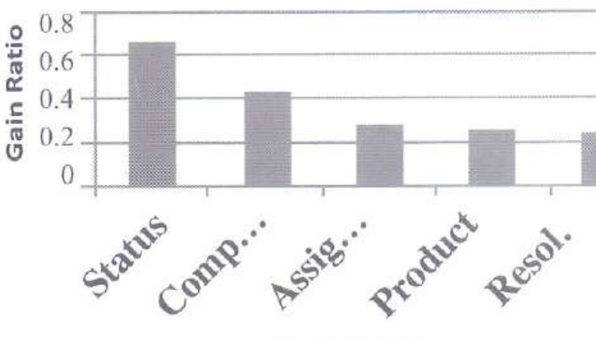

Fig. 4 Gain Ratio

show 88.6% correct classification and 11.4% incorrect classification, Product-Status show 66.6% correct classification and 33.4% incorrect classification ,Product-Resolution show 80% correct classification and 20% incorrect classification, Component- Assignee show 95% correct classification and 5% incorrect classification, Component-Resolution show 35% correct classification and 65% incorrect classification, Assignee-Resolution show 76% correct classification and 24% incorrect classification, Assignee-Status show 86.6% correct classification and 13.4% incorrect classification, and Status Resolution show 26.6 % correct classification and 73.4% incorrect classification.

This paper shows that if we apply gain ratio on attributes to rank them so, we can classify our data or bug report with small no of attributes and get the high accuracy. Table V shows the accuracy in terms of correct classification and incorrect classification in percentage and Fig. 6 is the graphical view of this result.

V. CONCLUSION AND FUTURE SCOPE OF WORK

In this work we perform the comparative study of the attributes which are ranked using gain ratio method. The dataset Bug Reports are collected from bugzilla.mozill.org for experiments. This experiment performs Classification based on the attributes values either belonging to SBR category or NSBR category. In this work, we compute the classification accuracy and compare the accuracy for all ten attributes. Our analysis suggests that high gain ratio containing attributes show highest classification accuracy. This paper shows that if we apply gain ratio on attributes, and ranked them so, we can classify our data or bug report with small no of attributes and get the high accuracy.

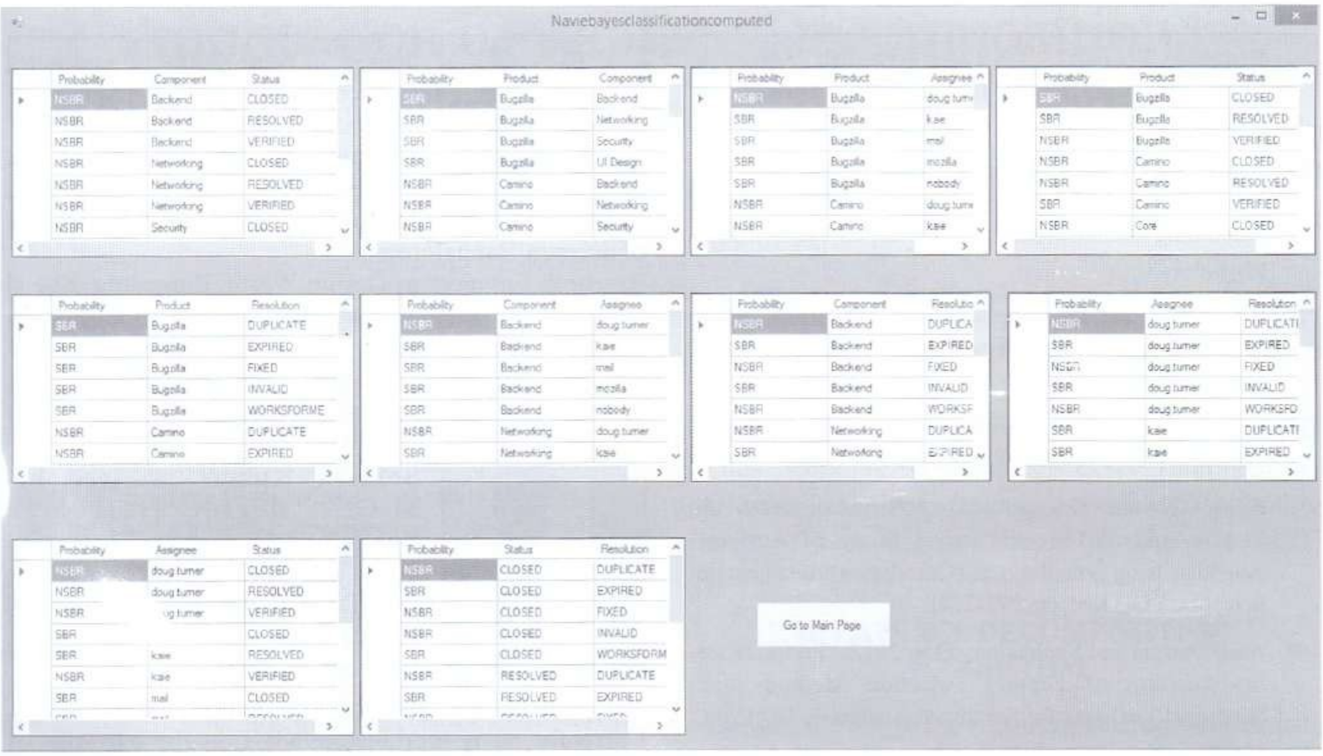

Fig. 5 Classification of ten combinations of five attributes

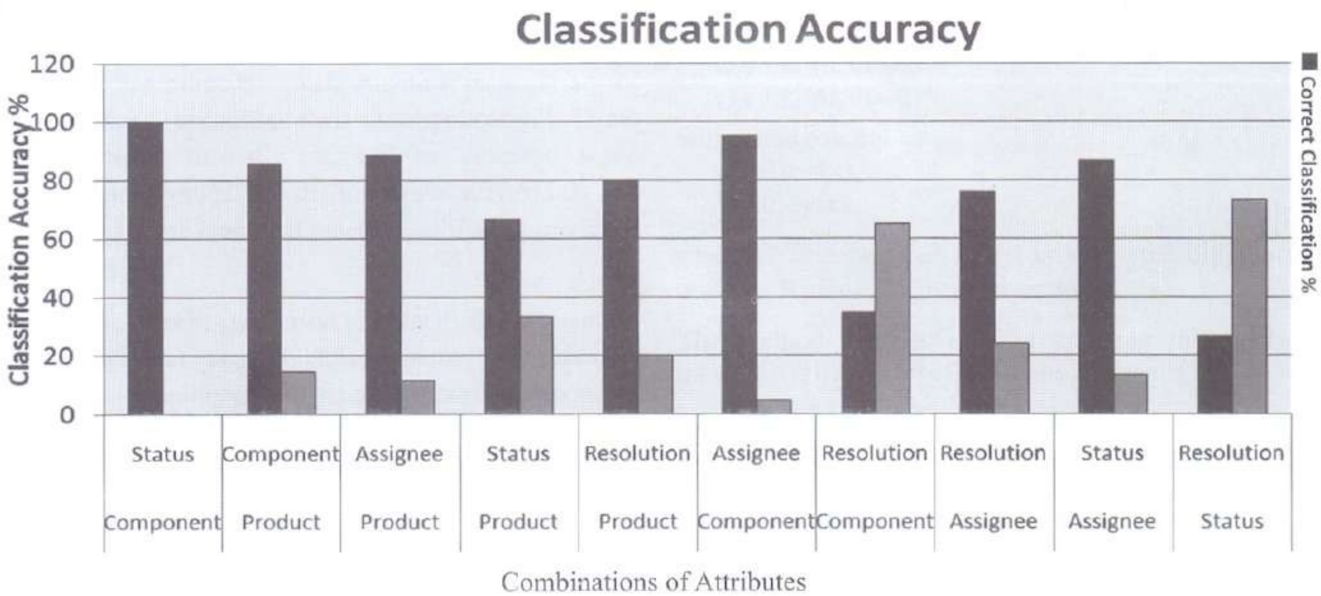

Fig. 6 Classification Accuracy of all ten combinations of the attributes

For Future enhancement different classification methods may be used with gain ratio and compare the accuracy. Also other attribute selection method can be used with naïve bayes classifier or other classification techniques and compare the performance and accuracy.

REFERENCES

[1] J.Han and M. Kamber, 2001, "Data Mining Concepts and Techniques", Morgan Kaufmann.

[2] Marc Boulle, 2007," Compression Based Averaging of Selective Naïve Bayes Classifiers", Journal of Machine Learning Research, volume 8, pp. 1659-1685.

[3] M Gegick, P. Rotella, and T.Xie, 2010," Identifying Security Bug Reports via Text mining: An Industrial Case Study.", IEEE, pp. 11-20.

[4] R. Praveena Priyadarsini and S. Sivakumari, 2011, "Gain Ratio Based Feature Selection Method for Privacy Preservation", ICTACT Journal on Softcomputing, pp. 201-205.

[5] Sangeeta Lal and Ashish Sureka, 2012,"Comaprison of

Seven Bug Report Types: A case-study of Google Chrome Browser Project", 19th Asia-Pacific Software Engineering Conference, IEEE Computer Society, pp. 517-526.

[6] Astha Chharia and R.K. Gupta, 2013, "Enhancing Naïve Bayes Performance with Modified Absolute Discount Smoothing Method in Spam Classification", IJARCSSE, pp. 424-429

[7] R.S. Anu Gowsalya and S. Miruna Joe Amali, 2014, "Naive Bayes Based Network Traffic Classification Using Correlation Information", IJARCSSE, pp. 8-14.

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

[9] Asha Gowda Karegowda, A.S. Manjunath & M.A.Jayaram 2010,"Comparative Study of Attribute Selection using Gain Ratio and Correlation based Feature Selection", IJITKM, pp. 271-277.

[10] Anuj Sharma and Shubhamoy Dey, 2012, "Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis", IJCA, pp. 15-20.

Smita Mishra received her B.Tech degree in year 2006 in computer science from Bhopal Institute of Technology and Science, Bhopal, affiliated to RGPV, Bhopal. In between 2007 and 2011 she worked as project assistant in Indian Institute of Science, Bangalore and as Research Engineer in Design Worth, Bangalore. She is currently pursuing her M.Tech. Degree in Computer Science & engineering from Noida Institute of Engineering & Technology. Her research interest includes software testing and dataming.

Somesh Kumar received his M.C.A. degree from MJP Rohilkhand University, Bareilly in 2000, and M.E. and Ph.D. degrees from Dr. B. R. Ambedkar University, Agra in 2006 and 2011 respectively. Between 2000 and 2011, he served with the SGI and Apeejay Groups. Currently he is working at NIET in the capacity of Professor & Head of IT Department. Dr. Somesh has published a number of research papers in Elsevier, Springer, Inderscience etc