Chapter 5. Machine Learning Classifiers[7]

Recently, Machine Learning Classifiers have gained popularity in spam classification area. A number of machine learning algorithms have been developed and tested on different corpuses. Machine Learning can also be named as experience-based learning that works by learning a pattern of the old examples with the use of different statistical rule to generate a robust filtering model to classify new examples. These classifiers always suffer from problems of bad sampling, weak statistical rules and bad learning etc. that result in too many errors in the classification that minimise the classification performance. In this study, six different Machine Learning Classifiers are tested on feature selected from three different publically available corpuses for evaluating a best classifier.

5.1 Evaluation of Best Machine Learning Classifier

The description and experiment design of this study are given below:

5.1.1 Aim of this study

A comparative study is done for evaluating best Machine Learning Classifier.

5.1.2 Corpora for this study

This study incorporates three different corpuses taken from three different sources, discussed in Chapter 3. Our main tests are performed on "Enron email" corpus and further "SpamAssassin" and "LingSpam" corpuses are employed for validation of the results observed from first corpus.

5.1.3 Feature Subset Search Techniques

Greedy Feature Subset search method is preferred for this study that gives a subset of less number of most informative features.

[7] Trivedi, Shrawan Kumar, and Shubhamoy Dey. "Effect of feature selection methods on machine learning classifiers for detecting email spams."*Proceedings of the 2013 Research in Adaptive and Convergent Systems*. ACM, 2013.

5.1.4 Classifiers for this study

Bayesian, Naïve Bayes (*NB*), Support Vector Machine (*SVM*), Genetic Algorithms (*GA*), Simple Decision Tree (*J48*), and Random Forest (*RF*) are the preferred classifiers of this study.

5.1.5 System Design

After pre-processing, Greedy feature subset search method is used for selecting the most informative features. Apart from all, 49 features for Enron, 36 features for SpamAssassin and 60 features for LingSpam are chosen for testing the concerned classifiers. All corpuses are split for performing the training and testing, where 66% instances are taken for training on the selected features and remaining 34% kept for testing.

5.1.6 Evaluation Metrics

Classification accuracy, F-Value and False Positive Rate are the preferred evaluation metrics of this study.

5.1.7 Results and Analysis

This study examines a comparative analysis of various Machine Learning Classifiers that are tested on three different corpuses (Enron, SpamAssassin, and LingSpam). Percentage Accuracy, F-Value and False Positive rate are incorporated for analysis purpose where the results of Accuracy and F-Value are more or less same. For clear understanding, the analysis is divided into three sections. First section analyses Machine Learning Classifiers. Second section shows the analysis of False Positive Rate for evaluating the sensitivity of the accurate classification and the last section observes the combine Effect of all the metrics.

5.1.7.1 Analysis with Percentage Accuracy and F-Value

Results of the classifiers tested on the Enron corpus is shown in Table 29 and Figure 30 that demonstrates, Random Forest (RF) and Support Vector Machine (SVM) both are performing

excellent among the other classifiers with the classification accuracy 98.4% and 97.8% respectively. In addition, Genetic Classifier comes in the third position with classification accuracy of 96.4% whereas Bayesian and Naïve Bayes, are predicted to be fourth best performer with classification accuracy of 93.1% and 93.9% respectively.

On the other side, Classifiers tested on the SpamAssassin corpus (Table 29 and Figure 30) validates the results of Enron corpus. In this case, again RF and SVM have given best classification accuracy amongst other classifier, i.e. 93.8% and 94.3% respectively. Probabilistic Classifiers, i.e. Bayesian and Naïve Bayes, are the second best performer with classification accuracy 93% and 94% respectively. Rule based Classifier i.e. genetic classifier is proven to be the worst amongst all with classification accuracy 87.5%.

When same tests are performed on LingSpam corpus then the results are slightly different from the results of Enron and SpamAssassn corpus. In this case, again SVM is predicted to be an excellent classifier together with Bayesian and Naïve Bayes where the classification accuracy is 97.8% for all classifiers. This time, RF is the second best performer with 96.6% classification accuracy. In addition, Genetic classifier continually disappoints with its worst performance (93.3% classification accuracy).

Table 29. Accuracy and F-Value for All Corpuses

In Percentage							
Corpuses	**Metrics**	Genetic	Bayesian	NB	SVM	J48	RF
Enron	Acc	96.4	97.1	96.6	**97.8**	97.9	**98.4**
	F-Value	96.4	97.1	96.7	**97.8**	97.9	**98.4**
SpamAssassin	Acc	87.6	93	94	**94.2**	92.1	93.8
	F-Value	87.5	93.1	93.9	**94.3**	92.2	93.9
LingSpam	Acc	93.2	97.7	97.8	**97.8**	95.7	96.5
	F-Value	93.3	97.8	97.8	**97.8**	95.7	96.6

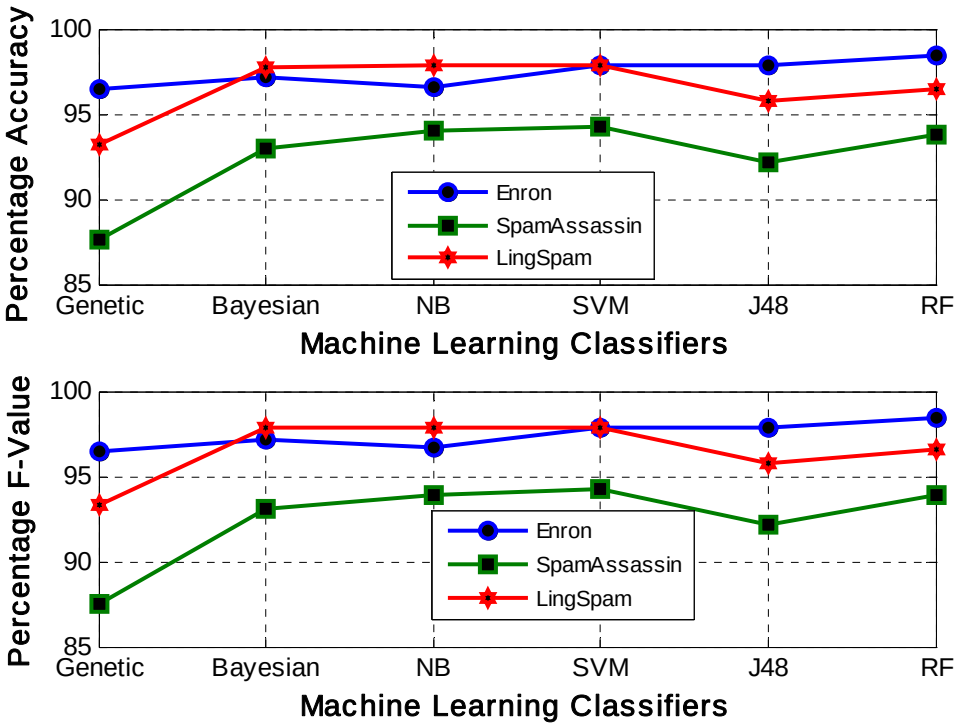

Fig 30. Accuracy and F-Value for All Corpuses

Table 30. False Positive Rate for All Corpuses

In Percentage Corpuses	Genetic	Bayesian	NB	SVM	J48	RF
Enron	22.5	1.8	4.4	2.6	2.7	2.7
SpamAssassin	2.4	3.6	4.5	1.0	2.2	1.7
LingSpam	10.0	0.0	0.0	1.9	3.8	3.8

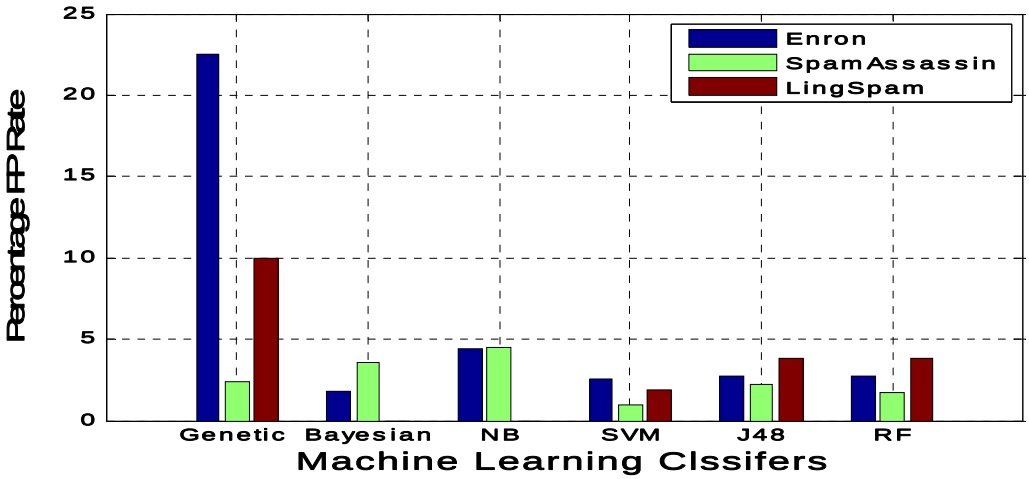

Fig 31. False Positive Rate for All Corpuses

5.1.7.2 Analysis with False Positive Rate

Although, some machine learning classifiers give the best classification accuracy but measurement of the sensitivity of classifiers accuracy is necessary because the possibility of misclassification of the positive instances. Legitimate Emails or Ham emails are sometimes seen important and misclassification of these emails as Spam create serious problems. This problem can be well tackled by obtaining the False Positive Rate (FP Rate) that reveals rate of misclassified legitimate (Positive).

From Table 30 and Figure 31, it is clear that, Function based Classifiers, i.e. SVM and Bayesian classifier, have given more sensitive analysis. In this case the FP Rate is less in all three corpuses and come 0.0%, 3.6%, and 1.8% for Bayesian classifier and 1.9%, 1.0% and 2.6% for SVM tested on LingSpam, SpamAssassin and Enron corpuses respectively.

5.1.7.3 Analysis with All Metrics Together

The observations from the different metrics suggest that Support Vector Machine (SVM) is the excellent classifier that gives not only excellent classification accuracy, but also shows exceptional sensitivity towards accurate classification with low false positive rate. RF and Bayesian are also seen promising whose results were very close to the best one.

5.2 Discussion

Best classification accuracy of the classifiers with the help of the minimum number of best feature subset is always a concern in the classification research for cost-sensitive evaluation. This study has done a comparative analysis of various Machine Learning Classifiers. Results demonstrate three different observations. First one deals with Machine Learning Classifiers where Function based Classifier i.e. SVM has given the best result in classification accuracy and is also proven to be the best sensitive classifier with less False Positive Rate. In addition, RF and Probabilistic Classifiers were the second best performer and close to the best one. Rule Based Classifier was the worst performer in this research.

This chapter has observed the excellent machine learning classifier i.e. RF and SVM. In this study, Greedy stepwise feature search method was preferred for selecting most informative feature. In the next chapter, various machine learning classifiers in conjunction with different feature search methods are tested to obtain best machine learning and feature search method pair.