IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-3, Issue-8 E-ISSN: 2347-2693

A Comparative Study of Spam Detection in Social Networks Using Bayesian Classifier and Correlation Based Feature Subset Selection

Sanjeev Dhawan[1], Kulvinder Singh[2] and Meena Devi[3*]
[1, 2] *Faculty of Computer Science & Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra- 136119, Haryana, India*
[3*] *Dept. of Computer Engineering) Research Scholar, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra-136119, Haryana, India*

Received: Jul /09/2015 Revised: Jul/22/2015 Accepted: Aug/20/2015 Published: Aug/30/ 2015

***Abstract*—** The article gives an overview of some of the most popular machine learning methods (Naïve Bayesian classifier, naïve Bayesian k-cross validation, naïve Bayesian info gain, Bayesian classification and Bayesian net with correlation based feature subset selection) and of their applicability to the problem of spam-filtering. Brief descriptions of the algorithms are presented, which are meant to be understandable by a reader not familiar with them before. Classification and clustering techniques in data mining are useful for a wide variety of real time applications dealing with large amount of data. Some of the application areas of data mining are text classification, medical diagnosis, intrusion detection systems etc. The Naive Bayesian Classifier technique is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayesian can often outperform more sophisticated classification methods. The approach is called "naïve" because it assumes the independence between the various attribute values. Naïve Bayesian classification can be viewed as both a descriptive and a predictive type of algorithm. The probabilities are descriptive are used to predict the class membership for a untrained data.

***Keywords*—** Bayesian Classifier, Feature Subset Selection, Naïve Bayesian Classifier, Correlation Based FSS, Info Gain, K-cross validation, Spam, Non-Spam

I. INTRODUCTION

Classification techniques analyze and categorize the data into known classes. Each data sample is labeled with a known class label. Clustering is a process of grouping objects resulting into set of clusters such that similar objects are members of the same cluster and dissimilar objects belongs to different clusters.[1] In classification the classes are pre-defined. Training sample data are used to create a model, where each training sample is assigned a predefined label. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Other than collection and managing data, data mining also includes analysis and prediction. In this paper we will try to understand the logic behind Bayesian classification. The Naive Bayesian Classifier technique is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayesian can often outperform more sophisticated classification methods.

II. Naïve Bayesian Classifier

The Naive Bayesian classifier is a straightforward and frequently used method for supervised learning. It provides a flexible way for dealing with any number of attributes or classes, and is based on probability theory. It is the asymptotically fastest learning algorithm that examines all its training input. It has been demonstrated to perform surprisingly well in a very wide variety of problems in spite of the simplistic nature of the model. Furthermore, small amounts of bad data, or ''noise,'' do not perturb the results by much.[2] However, as mentioned above, the central assumption in Naive Bayesian classification is that given a particular class membership, the probabilities of particular attributes having particular values are independent of each other. However, this assumption is often violated in reality. For example, in demographic data, many attributes have obvious dependencies, such as age and income. A plausible assumption of independence is computationally problematic. This is best described by redundant attributes. If we posit two independent features, and a third which is redundant (i.e., perfectly correlated) with the first, the first attribute will have twice as much influence on the expression as the second has, which is a strength not reflected in reality. The increased strength of the first attribute increases the possibility of unwanted bias in the classification. Even with this independence assumption, Naive Bayesian classification still works well in practice. However, some researchers have shown that although irrelevant features should theoretically not hurt the accuracy of Naive Bayesian, they do degrade performance in practice. This paper illustrates that if those redundant or irrelevant attributes are eliminated, the performance of Naïve Bayesian Classifier can significantly increase.

III. NAÏVE BAYESIAN K-CROSS VALIDATION

For k-fold cross-validation, data is split into k groups (e.g. 10). Then select one of those groups and use the model (built from your training data) to predict the 'labels' of this testing group. Once you have your model built and cross-validated, then it can be used to predict data that don't currently have labels.[5] The cross-validation is used to prevent over fitting. In K cross validation only 1 of the 10 groups is not used. Let's say you had 100 samples. You split it into groups 1-10, 11-20, ... 91-100. You would first train on all the groups from 11-100 and predict the test group 1-10. Then you would repeat the same analysis on 1-10 and 21-100 as the training and 11-20 as the testing group and so orth. The results typically averaged at the end.

IV. NAÏVE BAYESIAN INFO GAIN

The information gain of a given attribute X with respect to the class attribute Y is the reduction in uncertainty about the value of Y when we know the value of X[3].The uncertainty about the value of Y is measured by its entropy, H(Y). The uncertainty about the value of Y when we know the value of X is given by the conditional entropy of Y given X, H (Y|X) as shown in below:

IG = H (Y) – H (Y|X) = H (X) – H (X|Y)

IG is a symmetrical measure [11]. The information gained about Y after observing X is equal to the information gained about X after observing Y.

V. BAYESIAN CLASSIFIER

The Bayesian classifier is a simple but effective learning algorithm which can be used to classify the incoming messages into several classes (ω1, ω2…ωn). In fact, it is capable of much more than just that. The Bayesian classifier is used in document classification, voice recognition and even in facial recognition [9]. It is a simple probabilistic classifier (mathematical mapping system) which requires the following:

1. The prior probability that a given event belongs to a specific class
2. The likelihood function of a given feature set describing a class P(x|ω1)

Once these data are available, the classifier divides the sample space into disjoint regions (Ω1, Ω2…Ωn). When there are only two classes (in our case: spam and not-spam), the classifier also provides a decision function δ(x) such that

$\delta(x) = \omega_1$ if $x \in \Omega_1$

$\delta(x) = \omega_2$ if $x \in \Omega_2$

Initially, the classifier needs to be trained on labeled features to allow it to build up the likelihood functions and the priori probabilities. After the classifier is put to work, as it comes across newer values for the features, it automatically adjusts the likelihood functions and the decision boundaries appropriately.

Bayesian theorem provides a way of calculating the posterior probability, $P(c \mid x)$, from $P(c)$, $P(x)$, and $P(x \mid c)$. Naive Bayesian classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

Likelihood Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\, P(c)}{P(x)}$$

Posterior Probability Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \ldots\ldots\ldots \times P(x_n \mid c) \times P(c)$$

- $P(c \mid x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x \mid c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

VI. CORRELATION BASED FSS

CFS algorithm relies on a heuristic for evaluating the worth or merit of a subset of features. This heuristic takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them. The hypotheses on which the heuristic is based can be stated:

Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.

Features are relevant if their values vary systematically with category membership. In other words, a feature is useful if it is correlated with or predictive of the class; otherwise it is irrelevant. Empirical evidence from the feature selection literature shows that, along with irrelevant features, redundant information should be eliminated as well [6].

A feature is said to be redundant if one or more of the other features are highly correlated with it. The above definitions for relevance and redundancy lead to the idea that best features for a given classification are those that are highly correlated with one of the classes and have an insignificant correlation with the rest of the features in the set.

If the correlation between each of the components in a test and the outside variable is known, and the inter-correlation between each pair of components is given, then the correlation between a composite consisting of the summed components and the outside variable can be predicted from

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k - (k-1)\overline{r_{ii}}}} \quad (1)$$

Where

r_{zc} = correlation between the summed components and the outside variable.

k = number of components (features).

$\overline{r_{zi}}$ = average of the correlations between the components and the outside variable.

$\overline{r_{ii}}$ = average inter-correlation between components.

Equation 1 represents the Pearson's correlation coefficient, where all the variables have been standardized. The numerator can be thought of as giving an indication of how predictive of the class a group of features are; the denominator of how much redundancy there is among them [7]. Thus, equation 1 shows that the correlation between a composite and an outside variable is a function of the number of component variables in the composite and the magnitude of the inter-correlations among them, together with the magnitude of the correlations between the components and the outside variable. Some conclusions can be extracted from (1):

- The higher the correlations between the components and the outside variable, the higher the correlation between the composite and the outside variable.
- As the number of components in the composite increases, the correlation between the composite and the outside variable increases.
- The lower the inter-correlation among the components, the higher the correlation between the composite and the outside variable.

VII CLASSIFICATION RESULTS

Classifier	TP Rate	FP Rate	Precision	Recall
Naïve Bayes	0.793	0.152	0.842	0.793
Naïve Bayes 20 Folds	0.692	0.046	0.959	0.692
NB Info Gain FSS	0.8	0.196	0.808	0.8
Bayes Net	0.9	0.123	0.9	0.9
Bayes Net + CFS	0.924	0.096	0.925	0.924

Table 1 Comparison of Performance of Various Algorithms

In this above table comparision of performance of various algorithm has been shown and from the above table it is found that performance of Bayesian Net with Correlation Based Feature Subset Selection is best among all these algorithm with respect to TP Rate,FP Rate, Precision and Recall

VII. CONCLUSION AND FUTURE SCOPE

Feature subset selection (FSS) plays a vital act in the fields of data excavating and contraption learning. A good FSS algorithm can efficiently remove irrelevant and redundant features and seize into report feature interaction. This also clears the understanding of the data and additionally enhances the presentation of a learner by enhancing the generalization capacity and the interpretability of the discovering mode.An alternative way employing a classifier on a corpus of e-mail memos from countless users and a collective dataset.

In this work, we have worked on improving SPAM detection based on feature subset selection of Spam data set. The Feature Subset selection methods such as Info Gain Attribute selection and Correlation based Attribute Selection can be perceived as the main enhancement to Naïve Bayesian/ probabilistic methods. We have analyzed the Probabilistic SPAM Filters and attained more than 92% of success in filtering SPAM.

However, many open issues still remain open such as the system deals only with content as it has been translated to plain text or HTML. Since some spam is sent where most of the messages are inbuilt in image, it would be worth looking at ways in which images and other attachments could be examined by the system. These could include algorithms which extract text from the attachment, or more complex analysis of the information contained within the attachment. We can also work on a technique to recognize web junk e-mail according to finding these boosting pages in place of web spam page itself. We will begin from a small set of spam seed pages to get a hold of boosting pages. Then web junk e-mail pages are supposed to be identified making use of boosting pages. We can also work on a better larger dataset; the system should be tested over a longer period than the one-year one available in the public domain.

ACKNOWLEDGEMENT

I would like to acknowledge Dr. Sanjeev Dhawan, Assistant Professor, University Institute of Engineering and Technology (U.I.E.T), Kurukshetra University, Kurukshetra for introducing the present topic and for his inspiring guidance, valuable suggestions and support throughout the work.

REFERENCES

[1] Rushdi Shams and Robert Mercer," Classifying Spam Emails using Text and Readability Features," IEEE 13th International Conference on Data Mining (ICDM), **2013,** pp. **657-666.**

[2] Chotirat "ANN" Ratana Mahatana and Dimitrios Gunppulos," Feature Selection For the Naïve Bayesian Classifier Using Decision Trees," Applied Artificial Intelligence, Volume-17, **2003,** pp. **475-487.**

[3] Mehdi Naseriparsa, Amir-Masoud Bidgoli, Touraj Varaee,"A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithms," International Journal of Computer Applications (0975-8887), Volume 69, No-17, May **2013.**

[4] Aakriti Aggarwal and Ankur Gupta, "Detection of DDoS Attack Using UCLA Dataset on Different Classifiers, International Journal of Computer Science and Engineering, Volume-03, Issue-08, August **2015,** pp. **33-37.**

[5] Ioannis Kanaris, Konstantinos Kanaris, Ioannis Houvardas, And Efstathios Stamatatos, "Words Vs. Character N-Grams For Anti-Spam Filtering," International Journal on Artificial Intelligence Tools, **2006,** pp.**1-20.**

[6] Mehdi Naseriparsa, Amir-Masoud Bidgoli and Touraj Varaee," A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithms" International Journal of Computer Applications (0975 – 8887),Volume 69, Issue- 17,May **2013**

[7] Sanjeev Dhawan and Meena Devi, "Spam Detection in Social Networks Using Correlation Based Feature Subset Selection," International Journal of Computer Applications Technology and Research, Volume 4, Issue-8, August **2015**, pp. **629-632.**

[8] Dipali Bhosale and Roshani Ade," Feature Selection based Classification using Naive Bayesian, J48 and Support Vector Machine," International Journal of Computer Applications (0975 – 8887) Volume 99– No.16, August **2014.**

[9] Anjana Kumari," Study on Naive Bayesian Classifier and its relation to Information Gain," International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2, Issue- 3, March **2014**, pp.**601 – 603.**

AUTHORS PROFILE

Meena Devi has done her bachelor of technology degree in Computer Science and Engineering with first division in year 2013 and currently persuing her Master of Technology degree in Computer Engineering from Kurukshetra University, Kurukshetra. Her areas of interest are WEKA, java.