Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 09, September - 2017

Comparison of SVM and Naïve Bayes Text Classification Algorithms using WEKA

Nitin Rajvanshi,
Research Scholar
Dept. of Computer Sci. & Engg,
M.B.M. Engineering College, Jodhpur

Dr. K. R. Chowdhary,
Director,
Jodhpur Institute of Engineering &
Technology - SETG, Visiting faculty of IITJ,
Formerly, Professor & Head, Dept of CSE,
J.N.V. University, Jodhpur, India.

***Abstract*— Due to the growing amount of textual data, automatic methods for managing the data are needed. Automated text classification has been considered as a vital method to manage and process a large amount of documents in digital formats that are continuously increasing at an exponential rate. In general, text classification plays an important role in information extraction, summarization and text retrieval. This paper illustrates the text classification process using SVM and Naïve Bayes techniques. It automatically assigns documents to a set of classes based on the textual content of the document. In this paper after feature selection of text, machine learning algorithms Naïve Bayesian, Support Vector Machine(SVM) are applied. Evaluation and Comparison of algorithms is depicted. Topic-based text categorization classifies documents according to their topics. Performed through WEKA tool**

Index Terms*— *Machine Learning, Feature Selection, Stop words, Naïve Bayesian, Support Vector Machine (SVM), WEKA

1 .INTRODUCTION

Texts can also be written in many classes or species, for instance: scientific articles, news reports, movie reviews, and advertisements. Text Classification is the task of classifying a document under a predefined category. More formally, if d_i is a document of the entire set of documents D and $\{C_1, C_2,.,C_n\}$ is the set of all the classes, then text classification assigns one class C_j to a document d_i. Text Classification process is defined as eight stage process namely: Read Document, Tokenize Text, Stemming, Stopwords Removal, Vector Representation of text, Feature Selection or Feature Transformation and Learning Algorithms. The level of difficulty of text classification tasks naturally varies. As the number of distinct classes increases, so does the difficulty. WEKA tool is used here for comparative analysis of SVM and Naïve Bayes classification algorithms.

2. TEXT CLASSIFICATION

Generally, in the text classification task, a document is expressed as a vector of many dimensions,
x = (x1, x2,...,xl). Each feature of a document vector has two values: whether a certain word appears in the document and the real value that is weighted by a suitable method, for example, TF-IDF.
For example, the following two documents, "All-star game will held in Jodhpur" (document 1)

And "Chess is the champion of games" (document 2) are expressed as x1, x2 (Fig. 1) using the four word features "all-star", "Jodhpur", "chess", "game".

	"all-star"	"Jodhpur"	"chess"	"game"
document 1 (x1)	1	1	0	1
document 2 (x2)	0	0	1	1

Figure 1: Vector representation of two documents.

In the example above, the document is expressed by a 4 dimensional feature vector. However, it is desirable to use at least 10,000 features, or as many as possible, to classify various documents at a high accuracy. However, when most machine learning techniques are used, having many features causes overlearning and a very long calculation time. In order to avoid these problems, several feature selection methods have been proposed to cut down the features from 100 to 10,000 by using various evaluation standards such as word appearance frequency, document frequency, mutual information, and information profit. On the other hand, a class label y is given, which stands for which class the document belongs to. The number of classes can be two or more. A two-class case, which solves whether a document belongs to a class or not, is the easiest case and is called a "binary-class problem." A three-class or more case is called a "multi-class problem." Also, the problem can be divided into two cases; namely, the case where a document has only one label and the case where a document has two or more labels, called "multi-label." Generally, multi-class or multi-label classification problems are solved by combining many binary-class classifiers.

3. TEXT CLASSIFICATION ALGORITHMS

Machine Learning and Natural Language processing techniques can be used for the categorization. Some of the existing categorization methods include decision trees, decision rules, k-nearest neighbor, Bayesian approach, neural networks, regression-based methods, vector-based method etc. In this paper a comparative analysis of Naïve Bayes (NB) and Support Vector Machines(SVM) is done.

3.1 Naïve Bayes Algorithm

The naïve Bayesian model is a probabilistic approach. It is based on the assumption of conditional independence among attributes. Given a training set containing attribute values and corresponding target values (classes), the naïve Bayesian

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 09, September - 2017

classifier predicts the class of an unseen (new) instance, based on previously observed probabilities of the feature terms occurring in that instance.

Let C indicate the set of pre-defined classes to which the text may belong. Let B indicate the training set of pre-processed documents, while Bc is a pre-labeled subset of B that contains text of some class c Є C. Let F be the final feature set generated during the Feature Extraction and Enrichment Phase. The probabilities of occurrence each of the features in the feature set F for each class, was computed by making one pass over the document training set. First, the naïve Bayesian classifier [1] computes the prior probabilities of each class c Є C as indicated by Equation (1).

for each class c Є C do

$$P(C) = \frac{|B_C|}{|B|} \qquad (1)$$

Now, in order to classify a new text entry e, the probability of it belonging to each class is predicted as shown in Equation (2).

In Equation (2) the P(fi/c) terms indicate the statistical probability of occurrence of the ith feature term in a text entry of category c.

$$Pe(C) = P(C)\prod_{i=1}^{|F|} P(^{fi}/_{fc}) \qquad (2)$$

The document is then assigned to the class with the highest probability as given by the following equation 3:

$$maxpr = \arg max\ \big(pe(c)\big) \qquad (3)$$

Where, c Є C

3.2 Support Vector Machine (SVM) Algorithm

Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. For categorical variables a dummy variable is created with case values as either 0 or 1. Thus, a categorical dependent variable consisting of three levels, say (A, B, C), is represented by a set of three dummy variables:

A: {1 0 0}, B: {0 1 0}, C: {0 0 1}

To construct an optimal hyperplane, SVM employs an iterative training algorithm, which is used to minimize an error function. According to the form of the error function, SVM models can be classified into four distinct groups:

- Classification SVM Type 1 (also known as C-SVM classification)
- Classification SVM Type 2 (also known as nu-SVM classification)
- Regression SVM Type 1 (also known as epsilon-SVM regression)
- Regression SVM Type 2 (also known as nu-SVM regression)

There are number of kernels that can be used in Support Vector Machines models. These include linear, polynomial, radial basis function (RBF) and sigmoid.

Kernel Functions

$$K(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} \mathbf{X}_i \cdot \mathbf{X}_j & \text{Linear} \\ (\gamma \mathbf{X}_i \cdot \mathbf{X}_j + C)^d & \text{Polynomial} \\ \exp(-\gamma |\mathbf{X}_i - \mathbf{X}_j|^2) & \text{RBF} \\ \tanh(\gamma \mathbf{X}_i \cdot \mathbf{X}_j + C) & \text{Sigmoid} \end{cases}$$

Where $K(\mathbf{X}_i, \mathbf{X}_j) = \phi(\mathbf{X}_i) \bullet \phi(\mathbf{X}_j)$

that is, the kernel function, represents a dot product of input data points mapped into the higher dimensional feature space by transformation ϕ.Gamma is an adjustable parameter of certain kernel functions.

The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis. In order to use Radial Basis Function (RBF), it needed to specify the hidden unit activation function, the number of processing units, a criteria for modelling the given a training finding the algorithm for finding out the parameters.

4. EXPERIMENTAL DETAILS

4.1. Data

Here , the data is of Car Dataset available from UCI repository(https://archive.ics.uci.edu/ml/machine-learning-databases/car/car.data).

4.1.1 Description of Data

The Car Dataset relates CAR to the six input attributes: buying, maint (for maintainence), doors, persons, lug_boot (boot space) , safety.

Attribute Information:

Class Values: 4
unacc, acc, good, vgood
Attributes: 6
buying: vhigh, high, med, low.
maint: vhigh, high, med, low.
doors: 2, 3, 4, 5more.
persons: 2, 4, more.
lug_boot: small, med, big.
safety: low, med, high.
Total number of instances of data were 1728.

4.2 WEKA parameters for comparison

This paper presents comparison of SVM and Naïve Bays algorithm on following WEKA parameters

a. Correctly Classified Instances (CCI): Based on the training set how many instances are classified as positive i.e. True Positive.

b. Incorrectly Classified Instances (ICI): Based on the training set how many instances are classified as negative i.e. True Negative.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 09, September - 2017

c. Kappa Statistic (KS): The Kappa statistic (or simply *kappa*) is intended to measure agreement between two raters, and that possible value range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to -1 (complete disagreement)

d. Mean Absolute Error: It is used to measure how close prediction is to actual outcome. e. Root Mean Squared Error (MAE): It is frequently used measure of the differences between values predicted by a model and the values actually observed from the thing being modeled.

Following are results from WEKA:

4.3 Results and Analysis

There are 6 attributes namely- buying capacity, maintenance, number of doors, seating Capacity, boot space, safety and class and there were 1728 total instances of the dataset.

Table 1
Result from WEKA for Car Dataset

Algorithm	*CCI (%)*	*ICI(%)*	*KS*	*MAE*	*RMSE*
Naïve Bayes	85.53	14.47	0.6665	0.1137	0.2262
Radial Biased Function (RBF-SVM)	94.21	5.79	0.8752	0.676	0.1571

From the above results obtained for dataset, it is clearly shown that RBF (SVM) outperforms the Naïve Bayes algorithm. The Kappa statistic for RBF is close to perfect agreement (i.e. 0.8752). It has got 94.21% that is more as compared Naïve Bayes for correctly classified Instances. Moreover, RBF has got lesser mean absolute error and root mean square error.

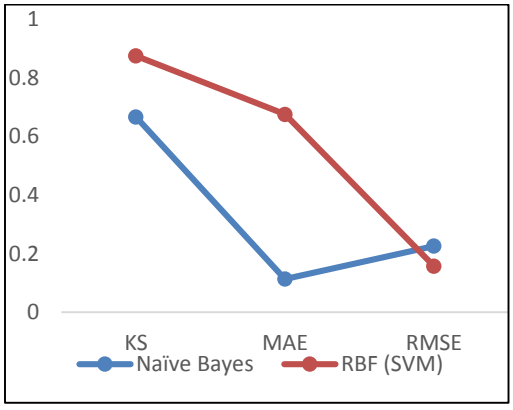

Figure 2: Graph of KS, MAE and RMSE for Dataset

5. CONCLUSION

In this paper two machine learning algorithms SVM (RBF) and Naïve Bayes have been evaluated and compared in WEKA. It can be inferred that none of the algorithm is perfect.

Since here a single dataset was considered, so its variation with different parameters were compared and evaluated. By increasing the datasets, results can be finer. Further, it can be said that different text classification algorithms work efficiently and may show different behavior for different datasets. The accuracy of predictive model is affected by the attributes of the data chosen.

REFERENCES

[1] Daniel T. Larose, "Data Mining Methods and Models", John Wiley & Sons, INC Publication, Hoboken, New Jersey (2006).
[2] Ryan Potter, "Comparison of Classification Algorithms Applied to Breast Cancer Diagnosis and Prognosis", Wiley Expert Systems, 24(1), 17-31, (2007).
[3] Yoav Freund and Llew Mason, "The Alternative Decision Tree Learning Algorithm" International Conference on Machine Learning, 124-133, (1999).
[4] Xindog Wu, Vipin Kumar et al., "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, 14(1), 1-37 (2008).
[5] Sebastiani, "Machine Learning in automated text categorization", ACM Computer Surveys, Vol. 34, March 2002
[6] Bharat Deshmukh, Ajay S. Patil, B.V.Pawar, "' Comparison of classification algorithms using WEKA on various Datasets", International Journal of Computer Science and Information Technology.
[7] Adrian G. Bors, I. Pitas, "Introduction to RBF Network", Online Symposium for Electronics Engineers, 1(1), 1-7 (2001).