International Journal of Pure and Applied Mathematics
Volume 116 No. 22 2017, 425-433
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

Comparison of Classification Algorithms in Text Mining

[1]Ananthi Sheshasaayee and [2]G. Thailambal
[1]Quaid-e-Milleth College for Women,
Chennai, India.
ananthi.research@gmail.com
[2]SCSVMV University, Kancheepuram,
Chennai, India.
thaila.research@gmail.com

Abstract

Web Mining is searching useful data from the World Wide Web repository which is divided into Content Mining, Usage Mining and Structure Mining in which Content Mining uses text, images, Audio and Video to extract useful information which is Unstructured. Web Mining is sub process of Data Mining which involves Anomaly detection, Classification, Clustering, Association Rule Mining, Regression and Summarization. This discovers patterns in large data sets involving many disciplines such as Artificial Intelligence, Machine Learning, Statistics and Database Systems. Machine Learning is the emerging technology to make the machines to predict values for new data inputs according to the previous data inputs trained with some Algorithms. Among all, the Classification is in supervised Learning of Machine learning where a training set of correctly predicted observation is available. In this paper three algorithms Naïve Bayes, Random Forest and Support Vector Machine used in Rapid Miner with 500 example dataset. Accuracy and Classification error of three algorithms compared and a chart is displayed which shows that Support vector machine is 97.40% accurate than other two algorithms.

Key Words: Machine learning, classification, random forest, naïve bayes, support vector machine, precision, recall.

1. Introduction

Today web is the main source of information. Google is the highly searchable Search Engine. Nearly 57,000 searches per second on a day according to Internet Live Stats, 2016. Web Content Mining is finding useful information from web pages. Analysing the web page by extracting its unstructured information helps us to understand its usability, predicting the future requirement of the user and many more. Web content Mining uses Text, Images, Audio and videos for extraction of information from Web. Among all text mining is very popular since most of the search purely uses only text documents. Text Mining helps to search related patterns from web Repository. The task which is very difficult in text mining is extracting useful information from unstructured text as there is no proper format of text in web. Statistical and Machine Learning algorithms used for Web Content Mining is to find relevancy of web page contents. [1]

2. Machine Learning

Machine Learning is automatically learn to make predictions on current data based on past history. It is divided into Supervised and Unsupervised Learning. Supervised Learning is when for every observation i = 1, 2, 3, 4… n and a vector of measurement x_i but not associated response $\mathbf{y_i}$. Unsupervised learning has inputs but no supervising Outputs to learn Relationships and structure of data. Predicting a continuous quantitative Output value is referred as Regression Problem. Predicting a non-numerical, Qualitative value or categorical Output value is Classification. Observing only Input Variables and No Output variables and grouping those input variables depending on their characteristics called Clustering. Input variables are referred as Predictors, Independent, Features or variables X. Output variables are referred as Response or Dependent variable Y. The relationship between Y (Response) and X (Predictors) and it is written as

$$Y = f(x) + E \quad (1)$$

'f' is fixed or unknown function and E is a random error term independent of X and mean zero. To estimate the function f apply a statistical learning method to the training data. Accuracy of f depends on two quantities Reducible error or Irreducible error. If the error can be reduced by increasing the accuracy then it is reducible error. If it cannot be reduced in any case then it is Irreducible error. When a given method yields a small training MSE but a large test MSE it is over fitting of data. Always training MSE smaller than the testing MSE. Variance is the amount by which 'f' would change if estimated it using a different training data set. [2].

The first increasing volume of readily available digital texts makes natural language into a fertile area and becomes most important data format of Machine language application. They include fundamental language processing and linguistic problems, identify a word's part of speech on its meaning, finding

relations between words. In Machine Learning system the performance should drastically improve the experience.

The Knowledge in Machine Learning System is represented as Symbolic Declarative as decision trees, numeric format as Support Vector Machine and Naïve Bayes or in Model based ways such as Neural Network and Hidden Markov Model. A well stated Machine Learning problem needs its input and output to be Quantified or Categorized. The Categorization of output is easier if it is based on objective factors, as it is in document class by topic.

3. Classification

Classification and Prediction are the two important methods of data analysis [3]. The first step of classification process is collecting the documents in different extensions. The collected documents to be converted into a pre-processed document with techniques like tokenization, Stop Word Removal, Stemming. This makes the document to get reduced from its original size and easy to handle. Then Indexing is done using Vector space model, Semantic representation, and Ontological representation, N-Grams, Boolean weighting and many more. To improve the Efficiency, Scalability and Accuracy of text feature Selection is made using Term Frequency, Chi-Square, Information Gain and Genetic Algorithm Optimization. Classification is done after selecting feature using some machine learning algorithms Bayesian Classifier, Decision Tree, K-Nearest Neighbor, Support Vector Machines and Neural Networks. When classification is done it should be evaluated experimentally through evaluation techniques such as Precision, Recall, Accuracy and many more [4].True Positive is Number of correctly Classified positive examples. False Positive is incorrectly classified Positive examples. False Negative is incorrectly classified Negative examples. Precision measures how many of the classifier is correct. Recall measures how few correct verdicts classifier has missed. In application where positive and negative examples equally treated Standard Accuracy is calculated. Selection of text representation features can make a difference between successful and Unsuccessful applications [12]

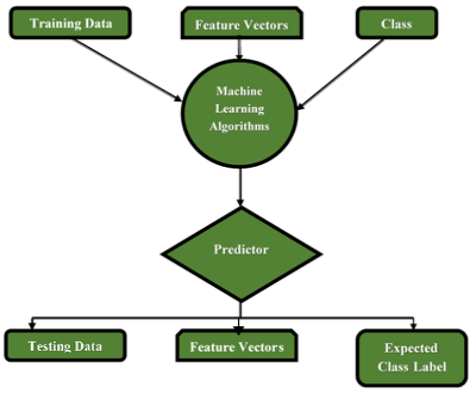

Fig. 1: Machine Learning Model

Figure 1 explains how this general model is used for finding a new class label for the given testing data by training some data with its features. Training Data is given as input and feature vectors selected and a class label is assigned to it. Then apply any of the machine Learning algorithms and with the testing data the new feature vectors are selected. A new class label is identified depending on the feature vector. Many Real time application in Classification is E-mail, Handwriting Recognition, Face Recognition, Speech Recognition, Information retrieval, Intrusion Detection, Anomaly Detection, Epileptic Seizure Detection, Finance, Music, and Signal processing [5]

4. Classification Algorithms

Random Forest

Ensemble Learning algorithms are accurate and robust to noise since it is a combination of more than one classifier. It performs well than single Classifier. Breiman in 2001 suggested this classifier with many advantages such as efficient, more input variables handled, importance of variables, robust to noise and also outliers and it is lighter than other ensemble algorithms [6]. Random forests helps in ranking the variables in regression or classification.

Table 1: Random Forest Matrix with Precision and Recall Values

	True Negative	True Positive	Class Precision
Pred. Negative	214	36	85.60%
Pred. Positive	24	226	90.40%
Class Recall	89.92%	86.26%	

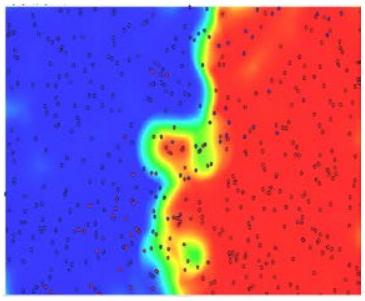

Fig. 2: Random Forest Density graph

Support Vector Machine

Support Vector Machine is a supervised machine learning algorithm used in Classification or Regression. Vikramjit Mitra et.al. Proposes LS-SVM based system which has accuracy of 99.9% used with the Gaussian Radial Basis function (GRBF) kernel. This paper uses Document titles in Library System instead of using document content to increase the performance time and relevancy for Semantic Classification [10]. The simplest SVM is a binary classifier, which is mapping to a class and can identify an instance belonging to the class or not. To produce a SVM classifier for class C, the SVM must be given a set of training samples including positive and negative samples. Positive

samples belong to C and negative samples do not. After text pre-processing, all samples can be translated to n-dimensional vectors. SVM tries to find a separating hyper-plane with maximum margin to separate the positive and negative examples from the training samples. [11]. SVM has a technique called the kernel trick. These are functions which takes low dimensional input space and transform it to a higher dimensional space. It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then find out the process to separate the data based on the labels or outputs. [7]

Table 2: SVM Matrix with Precision and Recall Values

	True Negative	True Positive	Class Precision
Pred. Negative	233	8	96.68%
Pred. Positive	5	254	98.07%
Class Recall	97.90%	96.95%	

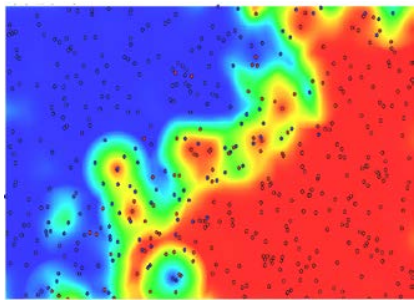

Fig. 3: SVM Density Graph

Naïve Bayes

Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes classifier is simple classifier with is based on Bayes Theorem of conditional probability and strong independence assumptions. This classifier emphasizes on measure of probability that whether the document A belongs to class B or not. It is based on independent feature model. It is based on the assumption that occurrence or non-occurrence of a particular attribute is unrelated to the occurrence or non-occurrence of a particular attribute. The advantage of Bayesian classifier is that it requires small training data set for classification. It is easier for implementation, fast to classify and more efficient. It is non sensitive to irrelevant features. It is used in personal email sorting, document categorization, email spam detection and sentiment detection [8].

Table 3: Naïve Bayes Matrix with Precision and Recall Values

	True Negative	True Positive	Class Precision
Pred. Negative	216	27	88.89%
Pred. Positive	22	235	91.44%
Class Recall	90.76%	89.69%	

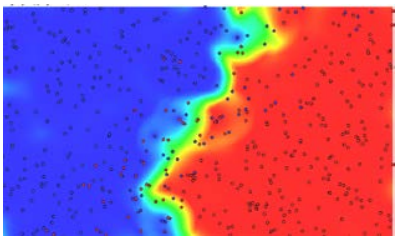

Fig. 4: Naïve Bayes Density Graph

Table 4: Combination of Algorithms Used by Different Authors

Author Name & Title	Dataset used	Performance Measures	Algorithm	Pre-processing
Keyword extraction using Naïve Bayes, Yaxin Usun	Set of trained documents	Tf-idf score, Distance of the word to beginning of text, paragraph and sentence	Naïve Bayes	Yes
Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study, Nazeeh Ghatasheh	1000 instances	Sensitivity, Precision, F-Measure	Random Forest	No
A Novel text mining approach based on TF-IDF and Support Vector Machine for news classification, Seyyed Mohammad Hossein Dadgar	2BBC and 20 Newsgroup sets	Precision	TF-IDF, SVM	Yes

5. Comparison of Algorithms

Applying different Classification Algorithm for a sample Dataset with 500 examples which contains a class Positive or Negative. Comparing the three algorithms shows that among them SVM is predicting 97% accuracy, Naïve Bayes is predicting 90% and Random forest predicts 88% for example set of 500 in a dataset. The Precision, Recall Matrix and Density graph is given for each algorithm. Rapid Miner tool is used to implement these algorithms.

Table 5: Comparison of SVM, NB and Random Forest

Algorithm	Accuracy	Classification Error
SVM	97.40%	2.60%
Naïve Bayes	90.20%	9.80%
Random Forest	88%	12%

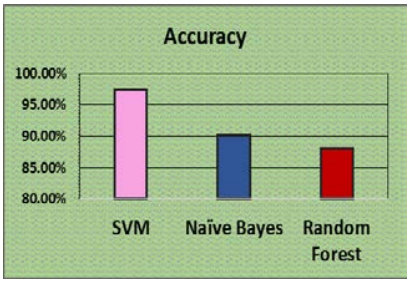

Fig. 5: Comparison of Three Algorithms with its Accuracy

Fig. 5 shows that SVM is better than Naïve Bayes and Random forest in its accuracy.

6.Conclusion

The Machine learning helps in doing research in a better way. The three algorithms in machine learning were the best algorithms and so many researchers combined these algorithms to improve the accuracy. The Density graph of three algorithms differs and in which SVM density is more than the other two proves that Accuracy is more when density is more. Many more dataset examples can be implemented with this Rapid miner with the different algorithms for getting faster and accurate results.

References

[1] Bharamagoudar G.R., Totad S.G., Prasad Reddy P.V.G.D., Literature Survey on Web Mining, IOSR Journal of Computer Engineering 5(4) (2012), 31-36.

[2] James G., Witten D., Hastie T., An Introduction to Statistical Learning: With Applications in R, Springer (2014).

[3] Vijayarani S., Muthulakshmi M., Comparative Study on Classification Meta Algorithms, International Journal of Innovative Research in Computer and Communication Engineering 1(8) (2013), 1768-1774.

[4] Korde V., Mahender C.N., Text classification and classifiers: A survey, International Journal of Artificial Intelligence & Applications 3(2) (2012).

[5] Das S., Dey A., Pal A., Roy N., Applications of Artificial Intelligence in Machine Learning: Review and Prospect, International Journal of Computer Applications 115(9) (2015).

[6] Rodriguez-Galiano V.F., An assessment of the effectiveness of a Random forest classifier for land-cover classification, ISPRS Journal of Photogrammetry and Remote Sensing 67 (2012), 93–104.

[7] Aparicio R., Acuna E., Using Ontologies To Improve Document Classification with Transductive Support Vector Machines, International Journal of Data Mining & Knowledge Management Process 3(3) (2013).

[8] Rajeswari R.P., Kavitha Juliet, Dr.Aradhana, Text Classification for Student Data Set using Naive Bayes Classifier and KNN Classifier, International Journal of Computer Trends and Technology 43(1) (2017), 8-12.

[9] Maroco J., Silva D., Rodrigues A., Guerreiro M., Santana I., de Mendonça A., Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression,

neural networks, support vector machines, classification trees and random forests, BMC research notes 4(1) (2011).

[10] Mitra V., Wang C.J., Banerjee S., Text classification: A least square support vector machine approach, Applied Soft Computing 7(3) (2007), 908-914.

[11] Fu J., Huang C., Lee S., A multi-class svm classification system based on methods of self-learning and error filtering, Taiwan, Republic of China (2008).

[12] Sokolova M., Szpakowicz S., Machine learning applications in mega-text processing, Handbook of research on machine learning applications and trends: algorithms, methods and techniques (2009), 325-347.