Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 10, Issue. 8, August 2021, pg.15 – 22

A Comparative Study of Parameters Measuring in Data Mining Function Using SVM

Dr. Meghna Utmal
Dept. of MCA, GGITS, Jabalpur, meghnautmal@gmail.com
DOI: 10.47760/ijcsmc.2021.v10i08.003

Abstract: ***Due to the vast amount of data available on the internet nowadays, it is necessary to categorise the data, and fast, accurate, and resilient algorithms for data analysis are required. Support vector machines (SVMs) are a form of machine learning technique that is commonly used to solve a variety of statistical learning issues. It's been designed as a reliable categorization tool, and it's especially useful when there's a lot of data. Machine learning is an area of artificial intelligence (AI) and computer science that focuses on using data and algorithms to mimic the way humans learn, with the goal of steadily improving accuracy. Algorithms are trained to create classifications by using statistical approaches. These should ideally have an impact on important growth measures. In this study, we found that employing the Support Vector Machine technique provides the best accuracy and efficiency for our dataset. Our work is based on the evaluation of parameters like accuracy, recall and precision.***

Keywords: ML, classification, text mining, SVM, accuracy.

1. INTRODUCTION

Analysis studies centered mostly on structured data in earlier days, were based on warehouse data, relational and transactional data. A large collection of documents and however, a significant part of the information available is stored in document databases, from text databases. Today there are several domains where we want to make use of information retrieval as we have abundant of news article, e-books etc. Text Word Today information retrieval has become a vast field to explore due to heavy information being available in form of different electronic records, journals etc. and the World Wide Web, databases are growing rapidly.

2. TEXT MINING

"In large textual datasets, text mining finds interesting regularities". The most relevant text mining task been done nowadays comprise of granular taxonomy creation, text clustering, text categorization and entity relationship modeling. Text research requires retrieval of data and lexical analysis to analyze distributions of word frequency. Handling and extracting information from text data in today's scenario is a difficult job as data is stored in unstructured format. Text mining has increasing attention in today's scenario to deal with unstructured or semi structured text[4, 3, 1, 2].

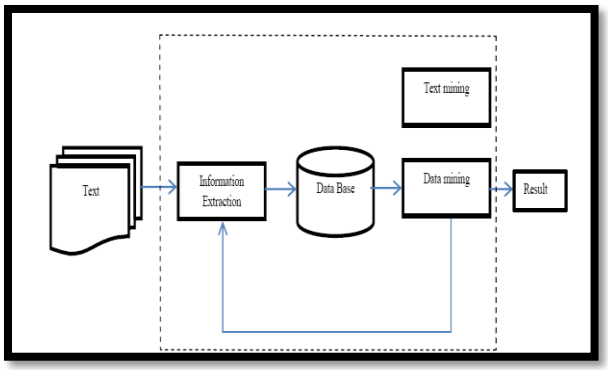

Fig.1: Architectural view of Text Mining

2.1 INFORMATION RETRIEVAL

In association with database we use Information retrieval (IR) for several years for drawing important information. Field of database system only focus on query and transaction processing while information retrieval is concerned with vast amount of text-based data present in documents.

The processing of information has found many uses because of the abundance of text information. There are several systems used for information retrieval like on-line document management systems, on-line library catalogue systems, and online search engines recently been developed [5].

In order to understand the quality of model it is mandatory to evaluate the performance of classification model. To refine the quality of model, understanding the quality of model and choosing the adequate model is an important task. [6]. Confusion matrix and receiver operating curves (ROC) are the two most widely used classification models.

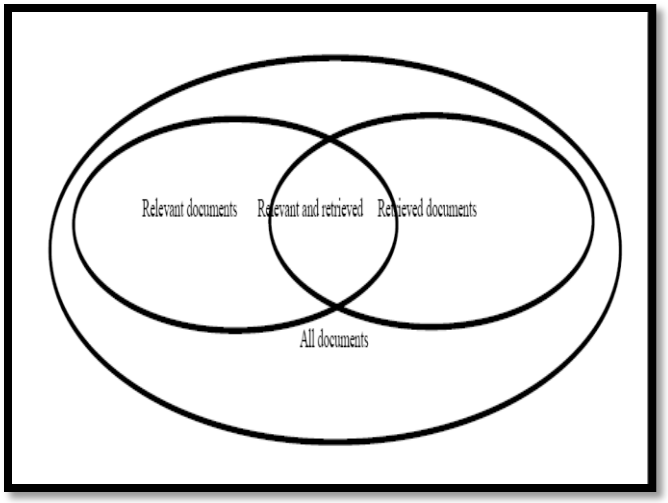

Fig.2: Relationship between all Documents

3. DATA MINING CLASSIFICATION

The present research work focuses on the Association Rule Mining and Classification individually but our present paper focuses solely on classification technique. In order to claim existence of group of labeled instance for class of objects we use predictive method of supervised machine learning. Figure given below depicts the process of classification. We feed a set of attributes as input and the model applies and output is attained. In order to predict the new class of instance we use classifier.

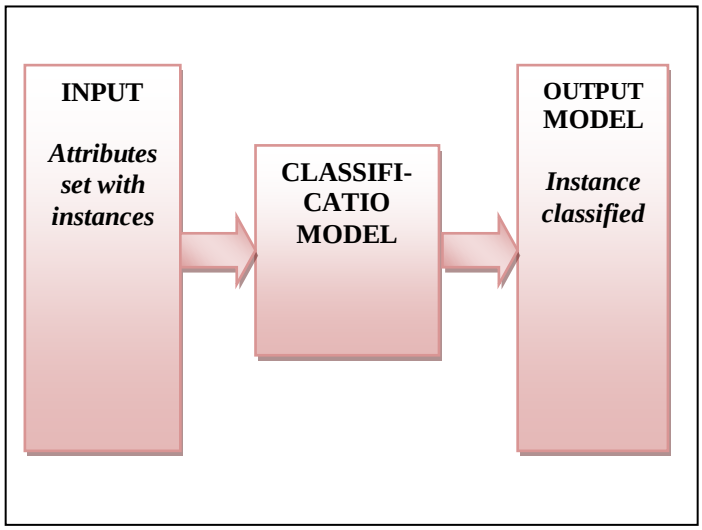

Fig. 3: Process of Classification

3.1 CLASSIFICATION MODEL EVALUATION

Testing of algorithms is an important measure used in every data mining. The widely used methods in evaluating the effects of classification algorithms used (ROC) are learning curves and confusion matrix.

The number of correct and incorrect predictions are displayed with the help of and their results made by the model. Table 1 displays the uncertainty matrix for a classifier.

Table 1: A Typical Confusion Matrix with Two Classes

		Classes predicted	
		True class	False class
Current classes	True class	True Pozitive	Fals Pozitive
	False class	True Negative	Fals Negative

The description of the above mentioned Confusion matrix with two classes is as follows:

Table 2: Comparison table

Name	Formula	Explanation
True Positive Rate (TP rate)	TP / (TP + FP)	The closer to 1, the better. TP rate = 1 when FP = 0. (No false positives)
True Negative Rate (TN rate)	TN / (TN + FN)	The closer to 1, the better. TN rate = 1 when FN = 0. (No false negatives)
False Positive Rate (FP rate)	FP / (FP + TN)	The closer to 0, the better. FP rate = 0 when FP = 0. (No false positives)
False Negative Rate (FN rate)	FN / (FN + TP)	The closer to 0, the better. FN rate = 0 when FN = 0. (No false negatives)

There are various parameters for calculating performance evaluation like accuracy, kappa, correlation, weighted mean precision and recall. The details of the measures are summarized as following,

(i) Accuracy is one of the measures which evaluate the performance of the models. Correctly classified examples are defined by accuracy. It is calculated by taking % of correct predictions over total number of examples. It can be express through the given formula

$$\textbf{Accuracy} = (TP + TN) / (P + N)$$

Where TP = True positive,
TN= True Negative,
P= Positive and N= Negative

(ii) Precision is the number of the predicted positive values that were correct.

$$\textbf{Precision} = TP / (TP + FP)$$

Where TP= True Positive,
FP= False Positive

(iii) The average of precision of every class is calculated by the weighted mean.

$$\textbf{Weighted Mean Precision} = \frac{1}{|I|}\sum_{j=1}^{|I|}\frac{|aj \cap bj|}{|bj|}$$

Where a_j and b_j are instances of true positives and true negatives outcomes.

(iv) Expected accuracy is compared with observed accuracy via Kappa measure on random chance basis. The kappa statistic evaluates classifier among themselves.

Kappa = (observed accuracy - expected accuracy) / (1 - expected accuracy)

(v) Returns the correlation coefficient between the label and prediction attributes.

$$\textbf{Corr (x, y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}}$$

Where x_i , y_i denotes the data point at various intervals; x and y are the mean values.

(vi) It is expressed in % and is defined as records retrieved: total number of records in database**.**

Recall= a / (a+b)

Where,

a = no. of relevant retrieved records

b = no. of relevant records not retrieved

(vii) The no. of sample classified incorrectly is described by the classification error Ei and is evaluated by the formula:

$$E_i = \frac{f}{n} \cdot 100$$

Where f = incorrectly classified sample cases

n= total number of sample cases.

The above parameters are used in text classification that has been compared with the traditional methods with the standard parameters of performance measurement as depicted below in Table2:

Table 3: Classification methods and algorithms

Classification methods	Algorithms
- Classification and decision trees - Bayesian classifiers - Artificial neural network - K-nearest neighbor classifier - Regresion - Classifiers based on association rules - Rough set	- ID3, C4.5 şi C5.0, CART, SPRINT, THAID, CHAID - Naive Bayes, BayesNet - Single-Layer Perceptron, Multy-Layer Perceptron, RBFNetwork, SVM - K-NN, PEBLS - Linear Regresion, SimpleLogistic - RIPPER, CN2, Holte's 1R, C4.5

4. THE SUPPORT VECTOR MACHINE (SVM)

Suggested by Vapnik the concept of SVM has attracted research community for machine learning [8]. On the basis of many studies it was proclaimed that SVM (support vector machines) deliver higher accuracy performance than any other techniques.

SVMs belong to a group of similar supervised learning strategies used to identify and regress[8]. SVM belong to the family of linear classification. SVM maximizes the geometric margin and

reduces the empirical classification error denoted by a special property and system risk is minimized by SVM. There is maping of input vector into larger dimensional space where there is a separation of hyperplane being constructed. [9]. The hyperplanes separate the data on both sides of the hyperplane. There is one general concept with this demarcation that more is the distance or greater is the margin between the hyperplane more distinctly is the data being separated, the better the classifier's generalization error would be. [8, 10].

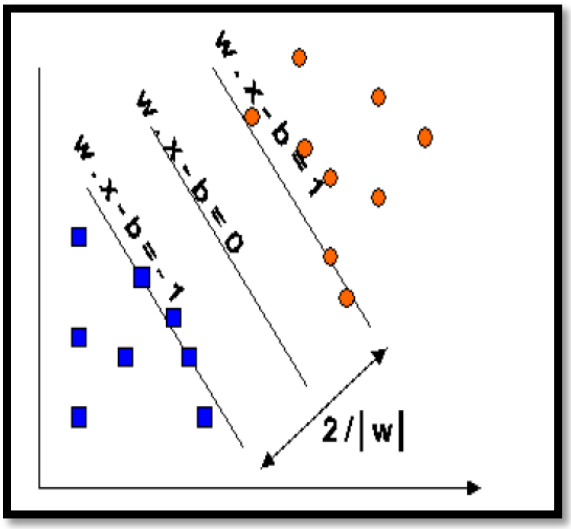

Fig. 4: Maximum Margin Hyperplanes

Now here is a comparison of Parameters with respect to Naïve Bayes, & Naïve Bayes Classification with term weight modification using SVM.

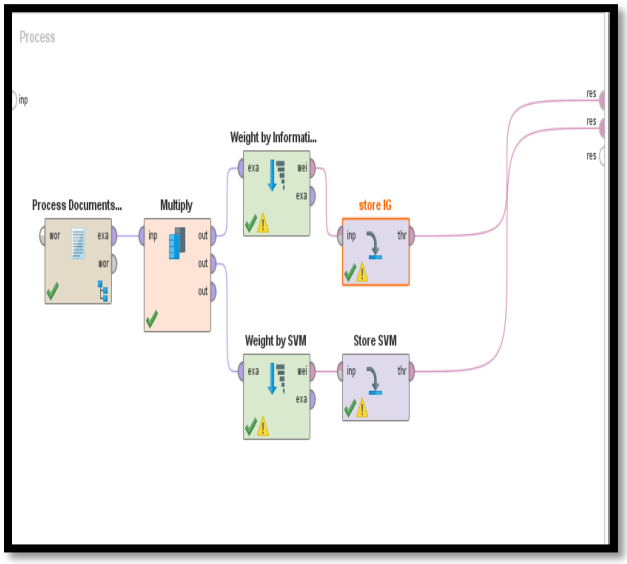

Fig. 5: Process window of Term weight adjustment by SVM

The design window for Term weight adjustment by Naïve Bayes Classification with term weight modification using SVM & Naïve Bayes Classification with ARM is shown in the Fig. 5 .We find out the result and compare the parameters with respect to Naïve Bayes, & Naïve Bayes Classification with term weight modification using SVM . The comparison is shown in table 1.

Table 4: Comparison of Parameters with respect to Naïve Bayes, & Naïve Bayes Classification with term weight modification using SVM

Methods/Parameters	**Accuracy(%)**	**Weighted mean recall (%)**	**Weighted mean precision (%)**
Naïve Bayes Classification	80.83	78.33	70
Naïve Bayes Classification with term weight modification using SVM	94.17	91.67	90

The Naïve Bayes Classification with term weight modification using SVM method gives highest accuracy with compare to the traditional method of text classification. The reason is that it has the highest number of true positives thus increasing the accuracy. The Naïve Bayes method has total of 80.83%, Naïve Bayes Classification with term weight modification using SVM 94.17 % of accuracy. The association between different frequent words is an efficient reason to improve the accuracy of the system. The parameters like weighted mean recall and weighted mean precision have increased in Naïve Bayes Classification with term weight modification using SVM approach which is shown in Table 1. Based on the three valuation parameters we find that this method outperforms in most of the metrics.

5. CONCLUSION

In the last decade, several data mining techniques have been suggested. Mining of association rule, sequential pattern mining, maximum pattern mining etc is included in this method. Accessing the efficiency of classification models is the main work proposed in the paper. Commonly used technique for supervised learning is Classification. This is the mechanism by which a collection of features and models representing the classes or concepts of data are defined. The theory and application of different model evaluation mechanisms in data mining are also explained in this paper. Performance analysis is largely based on a matrix of uncertainty. Various statistical measures are also defined here, such as accuracy, ROC region, etc., used for performance analysis. To streamline and enhance the efficiency, we applied & Naïve Bayes classification on various data sets.

REFERENCES

[1]. H. Yu, J. Han, and K. Chang. PEBL: web page classification without negative examples. IEEE Transactions on Knowledge and Data Engineering, 16(1):70–81, 2004.

[2]. X. Zhu, X. Wu, and Y. Yang. Dynamic classifier selectionfor effective mining from noisy data streams. Proceedings of the 4th international conference on Data Mining,(ICDM'04), pages 305–312, 2004.

[3]. Z. Jiawei Han and Micheline Kamber Data mining concepts and techniques book referred third edition 2010

[4]. Z. Oxford publication Arun kumar pujari book referred second edition 2011.

[5]. Data Mining: Concepts and Techniques Second Edition Jiawei Han University of Illinois at

Urbana-Champaign Micheline Kamber 2006.

[6]. X. Li and B. Liu. Learning from Positive and Unlabeled Examples with Different Data Distributions. Proceedings of European Conference on Machine Learning (ECML-05), pages 218–229, 2005.

[7]. Gorunesu, F., Data Mining. Concepte, modele şi tehnici, Editura Albastră, 2006, Cluj-Napoca.

[8]. V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.

[9]. Chih-Wei Hsu, Chih-Chung Chang, and Chih- Jen Lin. "A Practical Guide to Support Vector Classification". Deptt of Computer Sci. National Taiwan Uni, Taipei, 106, Taiwan http://www.csie.ntu.edu.tw/~cjlin 2007.

[10]. Boser, B. E., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages. 144 - 152. ACM

