Text Classification through Statistical and Machine Learning Methods: A Survey

Krina Vasa
P.G Student of Computer Engineering
Marwadi Education Foundations Group of Institutions, Rajkot, India.

***Abstract* - With the instant growth of information, text classification has become a vital technique for handling and organizing text data. In general, Text classification plays an important role in information extraction, text summarization text retrieval, medical diagnosis, news group filtering, spam filtering, and sentiment analysis. This paper illustrates the text classification process using machine learning techniques and statistical techniques such as k-nearest neighbors, support vector machine, naive Bayesian method.**

***Keywords* - Text classification, k-nearest neighbor, naive bayesian, support vector machine**

I.INTRODUCTION

Text classification is the performance of separating a set of input documents into two or more classes where each document can be said to belong to one or multiple classes. In classification systems, group of words or terms are collected together and organized. Each of these terms will be associated with a particular concept.

System of classification have classically been hierarchical, that means more detail is gained the extra down the hierarchy one proceeds, while concepts are linked and planned around mutual characteristics. The documents to be classified may be texts, images, music, etc.

Text classification involves four phases by those modules text classification is done. For classification model many techniques are available like Decision tree, Rule based classification, Neural network, Support vector machine, k-Nearest neighbor, Naive bayesian etc.

- Text preprocessing
- Feature extraction
- Training classifier
- Classification model

Text classification also assists as a way of remembering and differentiating the types of data, making predictions about data of the same type, classifying the relationship between different texts, and providing precise category to the data. Text classification also useful in classifying the web data or online data.

The classification problem is one of the fundamental problems in the machine learning and data mining. The information is growing very fast and in very huge amount. So the classification is widely used to classify the text into different classes. With the explosion of information driven by the growth of the World Wide Web it is no longer feasible for a human observer to understand all the data coming in or even classify it into categories. With this growth of information classification is required for classify the data into appropriate category.

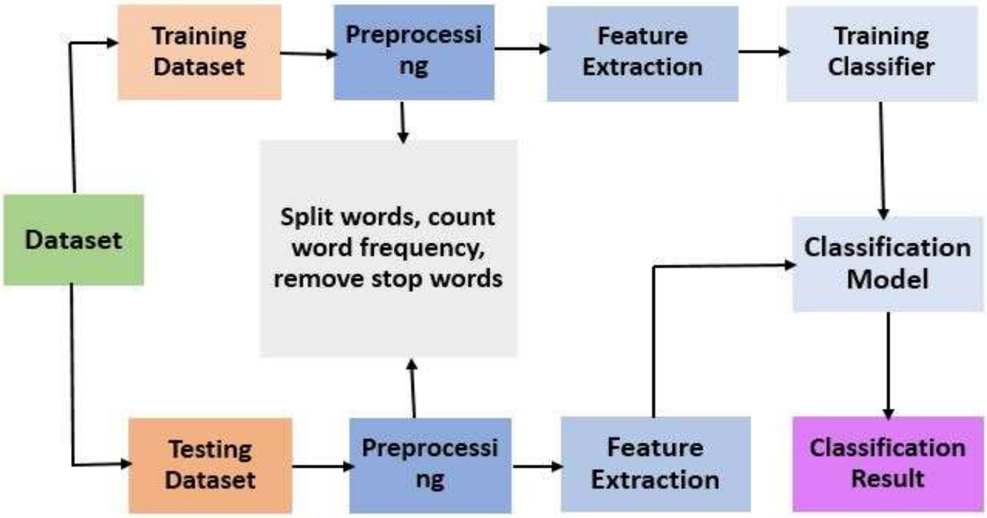

Figure: 1 Block diagram of text classification

II.REVIEW OF TEXT CLASSIFICATION TECHNIQUES

kNN: k-NEAREST NEIGHBOR

The most commonly and widely used distance function for the kNN classifier is the Euclidean distance formula and it is used to calculate the distance between the new unlabeled data point and the training data points. The main step in the classification stage of the KNN is to measure the distance in order to identify the nearest neighbors of the new input data point.

NAIVE BAYESIAN

Naive Bayesian is a simple classification method based on the Bayes rule with dimensionality of the inputs is high. In this Bayes rule applied to documents and classes. This model gives class labels to problem instances, represented as vectors of feature values. The value of a particular feature is autonomous of the value of other feature. Prior probability is known and posterior probability is checked in naive Bayesian technique. The class with higher posterior probability assigned to the document.

SVM: SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is an impressive approach for classifying high-dimensional data with the use of Structural Risk Minimization (SRM) principle. SVM has been conveyed as a discriminative classifier which is further accurate than most former classification prototypes. SVM acquires the optimal hyperplane that splits training data points from different classes by maximizing the classification margin. Similarly SVM applicable to data points with nonlinear decision surfaces by engaging a system identified as the kernel method that designs the input data to a higher dimensional feature space, where a linear separating hyperplane can be launch.

DECISION TREE

Decision tree shapes classification simulations in the usage of a tree structure. The aim is to construct a model that predicts the value of a target variable based on input variables. In these tree structures, leaves characterize class labels and branches denote features of class labels. It disturbances a dataset into smaller subsets while at the identical time a decision tree is incrementally established. The ultimate result is a tree with decision nodes and leaf nodes. Decision trees can knob both categorical and numerical data.

RULE BASED CLASSIFICATION

In rule based classification a set of rules are applied on data. In this classification method IF-THEN rules are used for classification. The IF (left part) part of the rule is known as rule antecedent or precondition. The THEN (right part) part of the rule is entitled rule consequent. The antecedent part of the state comprise of one or many attribute tests and those tests are logically ANDed. The consequent part involves class prediction.

NEURAL NETWORK

A neural network classifier is a network of units, where the input units show terms and output units speak for the category. Each unit receives set of inputs. There are two types of separators according to classes that are single linear perception and multiple perception. In single layer classes are not separated. Multiple layer used to find non-linear classification boundaries.

III.LITERATURE SURVEY

In research paper [2] authors presented a hybrid classification model that uses k-nearest neighbor and support vector machine techniques. This method is two stage approach based on the one-vs-near scheme was tested on big datasets. In the first stage, the kNN classifier is used to compute the category neighbor list which is learning phase. The kNN figures the distance between every centroid in the form of an ordered list which is used in second stage classifier. The second stage SVM uses the saved neighbor list to limit the dataset used for training the classifier for a single category.

In research paper [3], authors introduced a unique feature extraction technique and then classifies with linear support vector machine (SVM). In this approach, first, time domain and frequency domain analysis is done to the original data signal by translating and scaling co-efficient using Discrete Wavelet Transform (DWT) of mother wavelet Daubechies level 1 and then level 2 decomposition respectively. After that the subsequent transformed data is attended to a statistical method as Multi-Dimensional Scaling (MDS) to find out the similarities and dissimilarities to classify the precise class. Then for classifies the data SVM is applied on the data.

In research paper [4], authors developed a classification model grounded on Logical Analysis of Data (LAD). This paper deals with the problem of generating a fast and precise data classification, learning it from a possibly small set of records that are already classified. In this style, data should be encoded into binary form by means of a discretization process called binarization. This is done by using the training set for figuring exact values for each field, called cut-points in the case of numerical fields that split each field into binary attributes. The particular binary attributes constitute a support set, and are combined for generating logical rules called patterns. Patterns are used to classify each unclassified record, on the origin of the sign of a weighted sum of the patterns activated by that record.

In research paper [5], authors presented a novel fuzzy support vector machine (FSVM) tool or a variant of FSVM called modified fuzzy support vector machine (MFSVM). This variant is to classify the credit approval problem. In FSVM, each sample is given a fuzzy membership which denotes the attitude of corresponding point toward one class. The membership function which is a hyperbolic tangent kernel grips the impreciseness in training samples. In MFSVM, the victory of the classification lies in proper selection of the fuzzy membership function which is a function of center and radius of each class in feature space and is represented with kernel. The kernel used in MFSVM is hyperbolic tangent kernel. This kernel allows lower computational cost and higher rate of optimistic eigenvalues of kernel matrix which eases several limitations of other kernels.

In research paper [6], authors proposed an innovative and active approach to estimate the Bayesian probability. The Entropy-based Bayesian classifier, called EnBay, emphases on choosing the minimal set of long and not overlapped patterns that best complies with a conditional-independence model, based on an entropy-based evaluator. Additionally, the probability approximation is distinctly tailored to each group. This model works on two periods that are 1) partition the attribute set into a minimal number of large subsets so that their conditional dependence, given an arbitrary class, is minimized, 2) choose frequent item sets considered by conditionally independent attribute sets.

In research paper [7], authors described a cost-sensitive Naive Bayes method from a novel perspective of inferring the order relation between an instance's true classification probability of belonging to the class of interest and the cost-sensitive threshold. This method acquires and deduces the order relation from the training data and classifies the instance based on the inferred order relation.

In research paper [8], authors proposed a combined approach of k-nearest neighbor classification method and support vector machine classification algorithm. This integrated approach known as SVM-NN. In this method first the Bayesian vectorization tactic is used to convert training and training text documents into data points in numeric format. Then in training phase training data points are plotted into vector space of the SVM. Then support vectors of each class are recognized and the residual training data points are discarded. Now in classification stage new unlabeled data point in charted into the same vector space of support vectors which gained from the training phase. Then estimate average distance for each categories by using Euclidean distance formula and then after define the group of new unlabeled data point based on the shortest average distance between the support vector of the category and the new data point.

In research paper [9], authors projected a hybrid text classification using Bayesian-Support vector machine. In this first dataset is divided into training set and test set. In SVM input data is only in form of numerical form. So, testing data and training data are converted into numerical form by Bayesian vectorization method. By this the text data are transformed into numerical format. Then this data are input of support vector machine classifier. And output is classified data which is again used as a training dataset.

In research paper [10], developed Persian text classification using k-NN classifier with WordNet. There are mainly two steps of classification which are feature section and classification. In feature selection first stemmer was used to stemming the words in the document. By stemming similarity of documents having same words but in different forms is increased. After swapping stem of words in documents, frequency of these stems in the whole database and each document is achieved. Though frequency of stems is the most popular feature used in text classification, it is not a good discriminating feature. Many high frequency terms like "and", "or", "the" and so on are not suitable for distinguishing by information gain method. So high frequency words should be removed. Then used WordNet tool for precise classification. In the next step, frequency is calculated for documents in the training database. Then used PCA to reduce time needed to measure similarity between documents. Principal component analysis (PCA) transforms features into a new feature space and then eliminates features having low covariance. And then by use of k-NN classifier the text is classified.

In research paper [11], authors presented the SVM technique for classification. The brief of text categorization is the classification of natural text (or hypertext) documents into a fixed number of predefined categories based on their content. This problem arises in a number of different areas including email filtering, web searching, office automation and classification of news stories. SVM is a promising classification method settled based on the Structural Risk Minimization principle.

Authors of the paper [12] developed a privacy-preserving SVM classifier which is a SVM classifier with security. The privacy violation problem is a major problem in SVM so this classifier does not reveal the private content of support vector. The privacy-preserving SVM classifier (PPSVC) postprocesses the SVM classifier to transform it to a privacy-preserving SVM classifier which does not release the private content of the training data. For this privacy the Gaussian kernel function is used in PPSVC.

Authors of the paper [13] proposed SVM-based classification model. In this paper they presented different classification techniques like k-nearest neighbor, naive Bayesian, decision tree and SVM. In SVM there are mainly two points: (1) it is the case of linearly separable case, and the linearly inseparable case, in this the author used nonlinearity mapping to change the inseparable sample of low-dimensional sample space to high-dimensional feature space, then it will become linearly separable. (2)It is based on structural risk minimization theory to find the optimal separating hyperplane from the feature space. So learning machine can get global optimization, and the predictable risk of whole sample space would encounter a certain upper bound with a probability.

In research paper [14], authors developed localized support vector machine (LSVM) and profile support vector machine (PSVM). The LSVM is a nonlinear decision. Nonlinear support vector machine (SVM) is a global learning method and the problem of global learning technique is selection of model. So for overcome this problem the LSVM is developed. In LSVM build multiple linear SVMs but it is not possible with large amount of test data. To overcome this problem PSVM is developed. In this algorithm, the training data sets are partitioned into clusters based on their similarities and training a local SVM for each cluster. MagKmeans algorithm is used in PSVM.

In research paper [15], authors presented a rule extraction from support vector machines by active learning. This technique is concentrating on the problem areas, which for rule extraction are those spaces in the input space where the noise is the maximum. In this method initially, a preprocessing step is needed where the data missing values are remove and the encoding of nominal variables with weights of evidence(WOE) and a split of the data in training and test set. Following, the SVM model with RBF (Radial Basis Function) kernel is trained on the training data. Next, the generation of extra data instances. A simple, naive method is to generate extra data randomly throughout the input space. These all extra data is generated near to decision boundary. These extra-generated data instances are provided with a class label by the trained SVM model. Then measure the average distance to training data to support vectors. In final step, a rule induction technique C4.5 or RIPPER is applied on training data and extra generated data and then evaluated performance by its accuracy and fidelity.

In research paper [16], authors proposed a text classification based on multi-word with support vector machine. In this method first the multi-word extraction based on the syntactical structure of the noun multi-word phrases implemented. For the reduction on computation, pattern identification is proposed to be extracted from sentences and then use the extracted replication patterns for regular expression matching to extract the multi-words. By use of the multi-words for representation, two approaches were established based on the different semantic level of the multi-words: the first is the decomposition strategy using general ideas for representation and the second is combination strategy using subtopics of the general i for representation. Moreover, information gain method was working as a measure to remove the multi-word from the feature set to learning the robustness of the classification performance. Then a series of text classification tasks were done with SVM in linear and non-linear kernels, respectively, to analyze the effect of different kernel functions on classification performance.

IV.CONCLUSION

Text classification is the ultimate problem in data mining and machine learning. After look over the papers, we can say that there are so many techniques in text classification. Support vector machine, k-nearest neighbor and nave Bayesian method are widely used techniques in text classification. The hybrid approach of these techniques also very useful in text classification. In today's world the demand of support vector machine is increased because of its kernel functions. Kernel functions plot the data into higher dimensional spaces and then data could convert effortlessly separated. The hybrid approach of linear SVM and k-NN provides better accuracy but with the use of kernel function gives better performance than linear SVM. In future work, we will use the combine approach of Bayesian vectorization module, fuzzy support vector machine and k-nearest neighbor.

V.REFERENCES

[1] C. C. Aggarwal and C. Zhai, Mining Text Data, 2012.

[2] M. Kepa, J. Szymanski, "Two stage SVM and kNN text documents classifier," In: Pattern Recognition and Machine Intelligence, Kryszkiewicz M. (Ed.), Lecture Notes in Computer Science, Vol. 9124, pp. 279-289, 2015.

[3] R. C. Barik and B. Naik, "A Novel Extraction and Classification Technique for Machine Learning using Time Series and Statistical Approach," Computational Intelligence in Data Mining, vol. 3, pp. 217-228, 2015.

[4] R. Bruni and G. Bianchi, "Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis," IEEE Trans. Knowl. Data Eng., vol. 27, no. 9, pp. 2349-2361, 2015.

[5] A. Chaudhuri, "Modified fuzzy support vector machine for credit approval classification," IOS Press and Authors, vol. 27, no. 2, pp. 189-211, 2014.

[6] E. Baralis, L. Cagliero, and P. Garza, "EnBay: A novel pattern-based Bayesian classifier," Tkde, vol. 25, no. 12, pp. 2780-2795, 2013.

[7] X. Fang, "Inference-Based Naive Bayes: Turning Naive Bayes Cost-Sensitive," vol. 25, no. 10, pp. 2302-2314, 2013.

[8] C. H. Wan, L. H. Lee, R. Rajkumar, and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine," Expert Syst. Appl., vol. 39, no. 15, pp. 11880-11888, 2012.

[9] L. H. Lee, R. Rajkumar, and D. Isa, "Automatic folder allocation system using Bayesian-support vector machines hybrid classification approach," Appl. Intell., vol. 36, no. 2, pp. 295-307, 2012.

[10] M. Parchami, B. Akhtar, and M. Dezfoulian, "Persian text classification based on K-NN using wordnet," Adv. Res. Appl. …, vol. 7345, pp. 283-291, 2012.

[11] M. Wang H. Zhang, and R. Ding, "Research of text categorization based on SVM," International Conference on Infomatics, Cybermetics and Computer Engineering ICCE, Vol. 2, pp. 69-77, 2011.

[12] K. Lin, and M. Chen, "On the Design and Analysis of the privacy-preserving SVM classifier," IEEE Trans. Knowl. Data Eng., vol. 23, no. 11, pp. 1704-1717, 2011.

[13] Z. Liu, X. Lv, K. Liu, and S. Shi, "Study on SVM compared with the other text classification methods," 2nd Int. Work. Educ. Technol. Comput. Sci. ETCS 2010, vol. 1, pp. 219-222, 2010.

[14] H. Cheng, P. Tan, and R. Jin, "Efficient algorithm for localized support vector machine," Knowl. Data Eng. …, vol. 22, no. 4, pp. 537-549, 2010.

[15] D. Martens, B. B. Baesens, and T. Van Gestel, "Decompositional Rule Extraction from Support Vector Machines by Active Learning," IEEE Trans. Knowl. Data Eng., vol. 21, no. 2, pp. 178-191, 2009.

[16] W. Zhang, T. Yoshida, and X. Tang, "Text classification based on multi-word with support vector machine," Knowledge-Based Syst., vol. 21, no. 8, pp. 879-886, 2008.

[17] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Min. Knowl. Discov. vol. 2, no. 2, pp. 121-167, 1998.