Feature Selection Techniques And Classification Algorithms

E.Rama Krishna (M.Tech)[1], Dr K.Anuradha (Professor) [2]
[1, 2] Dept of Computer Science and Engineering
[1, 2] Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, INDIA

Abstract- *World Wide Web has led to an immense growth of information. So more information is stored in text form. Text mining has gained more importance with the high availability of information from different sources, automatic categorization of documents has become key for managing, organizing the enormous amount of information and knowledge discovery. The task of assigning predefined categories to documents will be done by text classification. The accuracy of classifier and dimensionality of feature space are main challenges of text classification. These problems can be overcome using Feature Selection. Feature selection (FS) is a process of finding most useful features of a subset from the original set of features. FS aims for generating text document classifiers more efficient and accurate. Feature selection methods will improve prediction performance, reduce computation time and a better understanding of the data is prepared. In this paper text classification, several approaches to text classification and feature selection methods are described.*

Keywords- Feature Selection, Feature selection methods, Text Classification algorithms.

I. INTRODUCTION

The text mining has become important recently because of the availability of the collective number of the documents from a variety of sources which include unstructured and semi-structured information. Text mining is renowned as Text Data Mining or Knowledge-Discovery in Text (KDT), refers to the process of extracting non-trivial information and interesting and knowledge from unstructured text. Operations like retrieval, text classification, text clustering, extraction and summarization are some typical text mining tasks.

Text classification (TC) is an important part of text mining. Finding relevant document from infinity has raised document classification [1]. Automatically categorizing documents could provide people with a significant comfort. Text categorization is the task of classifying a given data example into a pre-specified set of categories. For example, automatically classifying each incoming news with specific topics like health, sports, politics or art.

Text classification has two flavours: single label and multi-label text classification. In this single label classification, the document belongs to only one class example texts i.e. only one category must be assigned to each document. In multi-label classification, the document may belong to more than one class such as texts, images, music, and videos. In this paper single label document classification is considered.

Challenges of text classification include classifier accuracy and high dimensionality of feature space. This increases difficulties in applying sophisticated algorithms. So feature selection methods are important to reduce high dimensionality of data for effective text categorization. FS mainly focuses on detecting relevant information without the accuracy of the classifier is affected. Feature extraction serves two main purposes. First applying a classifier more efficiently by decreasing the size of the actual vocabulary. Second, feature selection frequently increases classification accuracy by eliminating noise features.

II. DOCUMENT CLASSIFICATION PROCESS

Text classification is a fundamental task in document processing [2], whose goal is to classify a set of documents into a fixed number of predefined categories.

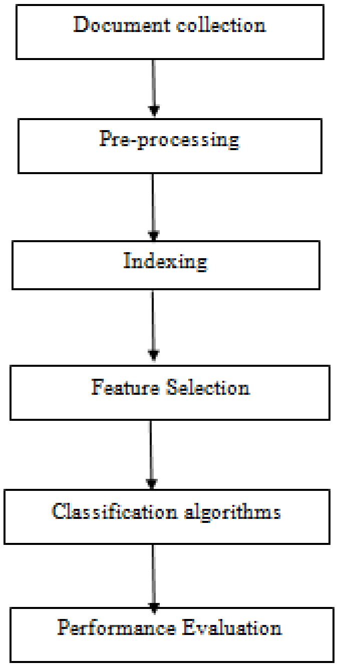

Figure 1: stages of classification.

2.1 Document Collection

The first step of classification process includes collecting different formats of documents like .html, .pdf, .doc, etc.

2.2 Indexing

It is one of the pre-processing techniques used for reducing the complexity of documents is document representation. The document is transformed from full-text version to a document vector. Most commonly used document representation models are vector space model, TF-IDF weighting.

2.2.1 Vector Space Model [VSM]

From [3] VSM is used in information filtering, information retrieval, indexing and relevancy rankings. The Vector Space model first extract features from the documents and assigns weights to them and finally compute document similarity. The standard practice is to take the set of all tokens that occur in the document and note their frequencies. This silently assumes position independence within the document, and also independence of terms with respect to each other. This assumption is naturally wrong (the term ‘surf’ has a different meaning within the context of surfing on the Web and surfing on the beach), but empirical NLP studies nevertheless report good results using it. This approximation of a document by their frequencies and a set of its terms is called the Bag of Words approximation.

2.2.2 Term Frequency (TF)

Term frequency in the given document is the number of times a given term appears in that document. Every document is described as a vector consisting of words such as D = <term1, term2, term3 ….term n>

Where ‘D’ is Document and ‘Term’ means the word on that document and ‘n’ represents the number of words in the document. Importance of the term ’t’ within the particular document with “ni‟ being the number of occurrences of the considered term and the denominator is the number of occurrences of all terms.

$$TF = \frac{ni}{\sum_k n_k}$$

2.2.3 Inverse Document Frequency (IDF)

The inverse document frequency is of the general importance measure of the term in the corpus. [4] It assigns a smaller value to the words occurring in the most of the documents and higher values to those occurring in fewer documents. It is the logarithm of the number of all documents divided by the number of documents containing the term.

$\text{IDF}=\log \frac{|D|}{|(di \supset ti)|}$ or

$$\text{IDF} = \log \frac{|D|}{|\{di\ ti\ \in\ di\ \in\ D\}|}$$

|D| is no of documents in the corpus and | (di ⊃ ti) | is a number of documents where the term ‘ti’ appears.

2.2.4 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a numeric measure that is used to score the importance of a word in a document is based on how often it appeared in that existed document and a given collection of documents. Example if a word appears frequently in a document, then it should be important and we should give that word a high score. But if a word appears in many documents, then its unique score assigned to lower rank. The math formula for this measure:

TFIDF = TF * IDF

Where TF=$\frac{ni}{\sum_k n_k}$ IDF=$\log\frac{|D|}{|(di \supset ti)}$

2.3 Pre-processing:

In this pre-processing stage, incomplete, uneven data of the documents are removed by tokenization, filtering, lemmatization, stemming.

2.3.1 Tokenization:

Tokenization is breaking the order of characters into pieces of words/phrases called tokens and deleting certain characters such as punctuation marks. It is used in bank accounts, financial statements, medical records, criminal records, driver's licenses, loan applications, stock trades, and voter registrations [5].

2.3.2 Filtering:

Filtering is usually done on documents to remove some unwanted words. A common filtering is a stop-words removal [7]. Stop words have little meaning which often appears in the text without having much content information (e.g. prepositions, conjunctions, etc.). Example Email and Spam Filtering [6].

2.3.3 Lemmatization:

Lemmatization chops inflexions and relies on a lexical knowledge base like WordNet to obtain the correct base forms of words i.e. grouping together the various inflected forms of a word so they can be analysed as a single item. Specifying the Parts Of Speech (POS) of the documents for lemmatize the documents because POS is deadly and error-prone, in practice stemming methods are preferred.

2.3.4 Stemming:

Stemming is the task that considers the main essence of the word and chops off its inflexions. Stemming methods aim at obtaining stem (root) of unoriginal words [8]. The best-known and most popular stemming approach for English is the Porter stemming algorithm. Example running to run, cats to cat, ran to ran.

2.4 Feature Selection:

The important step of text classification, after pre-processing and indexing, is Feature selection. The main goal of FS is selecting a subset of features from the original ones without affecting classifier performance.

2.5 Classification Algorithms:

Various classification techniques have evolved from machine learning techniques such as Bayesian Classifier, Decision Tree, K-Nearest Neighbour, Neural networks and Support Vector Machines Support.

2.6 Performance Evaluation:

The evaluation of text classifiers is necessary to check whether the classifier is taking right categorization decisions. Various measures such as recall, precision, f measure, fallout, error, accuracy are used to test the performance of the classifier.

III. FEATURE SELECTION AND ITS METHODS

The accuracy of classifiers is based on Feature selection methods not only by classification algorithms. Selection of unrelated and incorrect features may result in the classifier to incorrect results. The solution to this problem is Feature Selection. FS selects a subset of features from the original features without affecting the classifier performance. It is also known as Attribute selection. Feature selection reduces the dimensionality of the dataset, increases the learning accuracy and improves the result.

Whether the training set is labelled or not, feature selection algorithms categorized into supervised [9], unsupervised [10, 11] and semi-supervised feature selection [12, 13]. Supervised feature selection methods can further be broadly categorized into filter methods, wrapper methods and embedded methods. Every feature selection algorithm uses any one of the three feature selection techniques.

3.1 Filter methods:

Ranking techniques are used as principal criteria in Filter method. The variables are assigned a score using a suitable ranking criterion and the variables having the score below some threshold value are removed. Filter methods are computationally cheaper, avoids overfitting but these methods ignore dependencies between the features. Hence, the selected subset might not be best and a redundant subset might be obtained. The RELIEF algorithm [14, 15] is another filter-based approach is used to rank the features. The basic filter feature selection algorithms are as follows:

3.1.1 Chi-square test

The chi-squared filter method test checks the independence between two events. The two events X, Y are defined to be independent if P(XY) = P(X)P(Y) or equivalently P(X/Y) = P(X) and P(Y/X) = P(Y). whether the occurrence of a specific term is tested by the feature selection and the occurrence of a specific class are independent. Thus we estimate the following quantity for each term and we rank them by their score:

$$X^2 = \sum (O-E)^2 / E$$

Where O represents the observed frequency. E is the expected frequency under the null hypothesis.

3.1.2 Euclidean Distance

In this feature selection technique, the correlation between features is calculated in terms of Euclidean distance. If sample feature says ‘a' contains ‘n', then these ‘n' number of features are compared with other ‘n-1' features by calculating the distance between them using the following equation.

$$d(a,b) = (\sum_{i}(a_i - b_i))^{\frac{1}{2}}$$

The distance between features remains unaffected even after the addition of new features.

3.1.3 Information Gain.

Information Gain (IG) tells us how important a given attribute of the feature vectors is. IG selects the terms having the highest information gain scores. Concretely, it measures the expected reduction in entropy (uncertainty associated with a random feature) defined as:

$$Entropy = \sum_{i=1}^{n} -p_i \ log_2 p_i$$

Where ‘n’ is the number of classes, and the Pi is the probability of S belongs to the class ‘i’. The gain of A and S is calculated as:

Gain (A) = $Entropy(s) - \sum_{k=1}^{m} \frac{|s_k|}{|s|} * Entropy(\quad)$

3.2 Wrapper methods:

Filter models select features independent of any specific classifiers. A major challenge of filter approach is ignoring the ignoring the selected feature subset of the induction algorithm [17, 16]. The optimal feature subset should depend on the specific biases and heuristics of the induction algorithm. By assumption of this, wrapper models utilize a specific classifier to evaluate the quality of selected features, and offer a simple and powerful way to address the problem of feature selection, regardless of the chosen learning machine [17, 18]. Given a predefined classifier, a typical wrapper model will perform the following steps:

- Step 1: searching a subset of features,
- Step 2: evaluating the selected subset of features by the performance of the classifier,
- Step 3: repeating Step 1 and Step 2 until the desired quality is reached

3.3 Embedded methods:

Embedded Models fixes feature selection with classifier construction, advantages are (1) wrapper models - they include the interaction with the classification model and (2) filter models - they are far less computationally intensive than wrapper methods [19, 20, and 21]. There are three types of embedded methods. The first pruning methods that utilizing all features to train a model and then attempt to eliminate some features by setting the corresponding coefficients to 0, while maintaining model performance such as recursive feature elimination using support vector machine (SVM) [22]. The second are models with a built-in mechanism for feature selection such as [23] Iterative Dichotomiser 3 (ID3) and C4.5 [24]. The third step is regularising models which minimize fitting errors and in time force of the coefficients to be small or to be exactly zero. Features with coefficients that are close to 0 are then eliminated.

IV. CLASSIFICATION ALGORITHMS

4.1 Naive Bayes

Naive Bayes is a simple classifier model based on Bayes theorem with strong independence assumptions between the features. Spam filtering is the best-known use of Bayes. It is a conditional probability model, easy to build and represented by a vector x = (x_1, x_2… x_j) representing some n features (independent variables), it assigns to this instance probabilities p (C_j | x_1, x_2… x_j) for each of K possible outcomes or classes. The conditional probability using Bayes theorem can be specified as:

$$p(x/C_j) = \frac{p(x/C_j)p(C_j)}{p(x)}$$

4.2 Decision Tree

Decision trees build classification in the form of tree structure. Where this classifier breaks dataset into small subsets for easy to understand. It can handle both categorical and numerical data. Decision trees are applied in some fields like missing data, for improving search engines. It is very useful while making simple decision trees but as the complications increases then its accuracy will decrease. Calculations will get very complex when many values are uncertain. The core algorithm for building decision trees is ID3 (Iterative Dichotomiser 3) by J. R. QuinlanID3 uses Information Gain to construct a decision tree.

4.3 K-Nearest Neighbour

K-Nearest Neighbour (KNN) is a non-parametric method used for classification. The output of KNN classification is a class membership. One object can be classified by the majority of its neighbours where the object being assigned to the class most common among its k-nearest neighbours and it is measured by a distance function. The distance measure used depends on the application and nature of data. For text documents, cosine similarity is widely used whereas Euclidean distance is commonly used for relational data. K-NN is a type of instance-based learning called lazy learning. It is used in areas like credit card rating, bank loan issues, voting.

4.4 Neural networks

Neural networks from [25] can be used to model complex relationships between inputs and outputs to find patterns in data. When we use neural networks as a tool the data warehousing firms are gathering information from datasets. A neural network classifier is a network of unit s, where the input units usually represent terms, the output unit represents the category. Text document classification can be done by its term weights assigned to the input units; the activation of the units is forward through the network[26].it is used in complex pattern matching, character recognition and forecasting.

4.5 Support Vector Machines

Support Vector Machines (SVM) are used for classification and regression, supervised learning model defined by identifying a separating hyperplane between classes. SVM outputs an optimal hyperplane. SVM in text categorization and it had the highest classification precision. The main aim of the algorithm is finding the hyperplane which gives the largest minimum distance to the training example. This distance is called ‘margin’. It is used in bioinformatics, handwritten text recognition, permutation test.

V. PERFORMANCE MEASURES

There are various methods to determine the effectiveness or the performance of the algorithms. The metrics Precision, Recall, and F-measure is most often used. Precision [27, 28 and 29] is determined as the conditional probability that a random document d is classified under a category (ci).

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

The recall is an ability that a random document (dx) should be classified under category the (ci), this decision is taken.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

Where True Positive (TP) - when the classifier correctly classifies a positive test case into the positive class;

True Negative (TN) –when the classifier correctly classifies a negative test case into the negative class;

False Positive (FP) – when the classifier incorrectly classifies a negative test case into the positive class;

False Negative (FN) - when the classifier incorrectly classifies a positive test case into the negative class;

Precision and recall are often combined in order to get a better picture of the performance of the classifier given as F-Measure [30]

$$\text{F-Measure} = \frac{2*\mathrm{Recall}*\mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}}$$

VI. COMPARATIVE OBSERVATION

The performance of a classification algorithm is most affected by the quality of data source. Irrelevant and redundant features of data not only increase the cost of the mining process and reduce the Quality of the result in some cases [31]. Decision tree is less effective than all algorithms which are described above but Easy to understand and generate rule with Reduce problem complexity Each algorithm has its own advantages and disadvantages and most common method in most cases is support vector machine has better accurate result which Work well on numeric or textual data and Easy to implement and computation and also Work for both linear and nonlinear data and next neural network is accurate, naïve Bayes is last because of Perform very poorly when features are highly correlated and its indexing is gain break

VII. APPLICATIONS OF TEXT MINING

Text Mining has wide range of applications in our daily life often unknowingly. For example, the emails sent to us may be filtered through a text mining tool before being delivered to us [32]. Broadly these applications are categories in four main areas: business, medicine, law, and society. These applications are not limited to these areas. Text classification also has wide variety of applications in the domain of text mining some of which are listed below:

a) News filtering and organization
b) Document organization and retrieval
c) Opinion mining
d) Email classification and spam filtering
e) Text filtering
f) Hierarchical categorization of web pages

VIII. CONCLUSION

Text classification and Feature selection both are general domains of research covering Data mining, NLP and Machine Learning. These techniques have gained significant importance owing to the high growth rate of internet. This survey paper confines literature survey of feature selection methods and different classifiers. We can conclude that among three approaches to Feature selection method, filter methods should be used if we want results in less time for large datasets. If we want the results to be accurate and optimal, then wrapper method should be used. In embedded model it might be possible that the features which are relevant are already removed in the Filter approach, so, even if we go for wrapper those useful features cannot be added. It is deemed that no single representation scheme and classifier can be used as a general model for any application. However, all the discussed classifiers can only predict the class of unknown document; they do not provide degree of relevance of a particular document to a particular class. Also the data needs to be certain, precise and accurate. These can be overcome by using soft computing methodologies aim to exploit the tolerance for imprecision, uncertainty, partial truth and approximation. As a part of future scope, soft computing methodologies like Fuzzy logic, Evolutionary algorithms can be used for feature selection as well as classification algorithm.

REFERENCES

[1] G. Doquire, M. Verleysen, ―Mutual information-based feature selection for multilabel classification,‖ Neurocomputing, Elsevier, June 2013.

[2] Pradnya Kumbhar, Manisha Mali, - A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification,2015.

[3] Radim Rehuek. Plagiarism Detection through Vector Space Models Applied to a Digital Library.

[4] S.W. Mohod Deptt, Dr C.A.Dhote Feature Selection Technique for Text Document Classification: An Alternative Approach.

[5] Jonathan J Webster and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In Proceedings of the 14th conference on Computational linguistics-Volume 4. Association for Computational Linguistics, 1106–1110.

[6] Catarina Silva and Bernardete Ribeiro. 2003. The importance of stop word removal on recall values in text categorization. In Neural Networks, 2003. Proceedings of the International Joint Conference on, Vol. 3. IEEE, 1661–1666.

[7] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of Twitter. (2014).

[8] Martin F Porter. 1980. An algorithm for suffix stripping. Program: electronic library and information systems 14, 3 (1980), 130–137.

[9] J. Weston, A. Elisseff, B. Schoelkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. Journal of Machine Learning Research, 3:1439–1461, 2003.

[10] P. Mitra, C. A. Murthy, and S. Pal. Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24:301– 312, 2002.

[11] J.G. Dy and C.E. Brodley. Feature selection for unsupervised learning. The Journal of Machine Learning Research, 5:845–889, 2004.

[12] Z. Zhao and H. Liu. Semi-supervised feature selection via spectral analysis. In Proceedings of SIAM International Conference on Data Mining, 2007.

[13] Z. Xu, R. Jin, J. Ye, M. Lyu, and I. King. Discriminative semi-supervised feature selection via manifold regularization. In IJCAI' 09: Proceedings of the 21st International Joint Conference on Artificial Intelligence, 2009.

[14] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. Morgan Kaufmann, 1998.

[15] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics pages 2313–2351, 2007.

[16] M.A. Hall and L.A. Smith. Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In Proceedings of the Twelfth International

Florida Artificial Intelligence Research Society Conference, volume 235, page 239, 1999.

[17] R. Kohavi and G.H. John. Wrappers for feature subset selection. Artificial intelligence, 97(1-2):273–324, 1997.

[18] I. Inza, P. Larra˜naga, R. Blanco, and A. J. Cerrolaza. Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial intelligence in medicine, 31(2):91–103, 2004.

[19] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4):491, 2005.

[20] S. Ma and J. Huang. Penalized feature selection and classification in bioinformatics. Briefings in bioinformatics, 9(5):392–403, 2008.

[21] Y. Saeys, I. Inza, and P. Larra˜naga. A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19):2507–2517, 2007.

[22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3):389–422, 2002.

[23] J. R.Quinlan. Induction of decision trees. Machine learning, 1(1):81–106, 1986.

[24] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[25] Payal R. Undhad, Dharmesh J. Bhalodiya. Text Classification and Classifiers: A Comparative Study.

[26] Ramasundram, S.P.Victor, "Text Categorization by Backpropagation Network," International Journal of Computer Applications, vol. 8, no. 6, 2010.

[27] G.Angulakshmi, Dr.R.Manicka Chezian, " Three-Level Feature Extraction For Sentiment Classification", International Journal of Innovative Research in Computer and Communication Engineering, Volume 2, Issue 8, Page.No5501-5507, August 2014

[28] Jafar Ababneh, Omar Almomani Wael Hadi, Nidhal Kamel Taha El-Omari, and Ali Al-Ibrahim, "Vector Space Models to Classify Arabic Text", International Journal of Computer Trends and Technology (IJCTT), Volume 7, Issue 4, Page.No 219-223, January 2014.

[29] Zakaria Elberrichi, Abdelattif Rahmoun, Mohamed Amine Bentaalah, "Using WordNet for Text Categorization", International Arab Journal of Information Technology, Volume 5, Issue 1, Page.No16-24, January 2008.

[30] Shweta C. Dharmadhikari, Maya Ingle, Parag Kulkarni, "A Comparative Analysis of Supervised Multi-label Text Classification Methods", International Journal of Engineering Research and Applications (IJERA) Vol. 1, Issue 4, Page.No 1952-1961, March 2012.

[31] Dr.K.Prabha S.Brindha Dr.S.Sukumaran, "A SURVEY ON CLASSIFICATION TECHNIQUES FOR TEXT MINING," IEEE, 2016.

[32] M. Mali, M. Antique, ―Applications of Text Classification using Text Mining,‖ International Journal of Engineering Trends and Technology (IJETT) – Volume 13, Number 5, July 2014.