International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification

Pradnya Kumbhar[1], Manisha Mali[2]

[1, 2]Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune

Abstract: ***The rapid growth of World Wide Web has led to explosive growth of information. As most of information is stored in the form of texts, text mining has gained paramount importance. With the high availability of information from diverse sources, the task of automatic categorization of documents has become a vital method for managing, organizing vast amount of information and knowledge discovery. Text classification is the task of assigning predefined categories to documents. The major challenge of text classification is accuracy of classifier and high dimensionality of feature space. These problems can be overcome using Feature Selection. Feature selection is a process of identifying a subset of the most useful features from the original entire set of features. Feature selection (FS) is a strategy that aims at making text document classifiers more efficient and accurate. Feature selection methods provide us a way of reducing computation time, improving prediction performance, and a better understanding of the data. This paper surveys of text classification, several approaches of text classification, feature selection methods and applications of text classifications.***

Keywords: Feature Selection, Feature selection methods, Text Classification, Text Classification Algorithms, Text Mining

1. Introduction

The text mining studies are gaining huge importance recently because of the availability of the increasing number of the documents from a variety of sources which include unstructured and semi structured information. Almost 90% of the world's data generated is unstructured. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Operations like retrieval, text classification, text clustering, concept /entity extraction and summarization are some typical text mining tasks.

Text classification (TC) is an instance of text mining. Infeasibility of human beings to go through all the available documents to find the document of interest precipitated the rise of document classification. Automatically categorizing documents could provide people a significant ease in this realm. Text categorization is the task of classifying a given data instance into a pre-specified set of categories. In other words, given a set of categories, and a collection of text documents, text categorization or TC is the process of finding the correct topic for each document [7]. For example, automatically classify each incoming news story with a topic like "sports", "politics", or "art".

Text classification has two flavors: single label and multi-label text classification. In single label classification, document belongs to only one class i.e. exactly one category must be assigned to each document. In multi label classification document may belong to more than one class i.e. number of categories may be assigned to the same document. In this paper we consider only single label document classification.

Challenges of text classification include classifier accuracy and high dimensionality of feature space. This raises hurdles in applying sophisticated learning algorithms. It is important to use feature selection methods to reduce high dimensionality of data for effective text categorization. Feature Selection is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification. Feature selection mainly focuses on identifying relevant information without affecting the accuracy of the classifier. Feature extraction serves two main purposes. First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. Second, feature selection often increases classification accuracy by eliminating noise features.

In this paper, we will look at some of the widely used classification algorithms and feature selection methods. The rest of the paper is organized as follows. In Section 2 Text classification and process is presented followed by Feature Selection methods in Section 3. Section 4 presents brief description of various classification algorithms. Section 5 provides a brief overview of the related work. In Section 6, we present briefly the applications of Text Classification and finally in Section 7 we present the future scope and conclusion.

2. Classification and its Process

Text classification is a fundamental task in document processing, whose goal is to classify a set of documents into a fixed number of predefined categories. Text categorization is the task of assigning a Boolean value to each pair $\{d_j, c_i\} \in D \times C$, where D is a domain of documents and $C = \{c_1, c_2, .., c_i\}$ is a set of predefined categories. A value of T assigned to $\{d_j, c_i\}$ indicates a decision to file d_j under c_i, while a value of F indicates a decision not to file d_j under c_i. Fig1. represents various stages of text classification process.

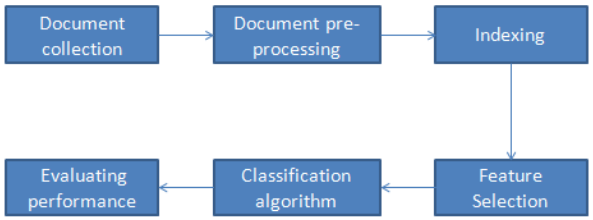

Figure 1: Stages of Text Classification

2.1 Document Collection

The first step of classification process includes collecting different types (format) of documents like .html, . pdf, .doc, etc.

2.2 Pre-processing

The real world data is inconsistent, incomplete and likely to contain errors, hence needs to be pre-processed. The pre-processing steps include tokenization, stop-word removal and stemming.

2.3 Indexing

One of the pre-processing techniques that are used to reduce 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, Boolean weighting model, Tf-idf weighting, Latent Semantic Indexing (LSI), LPI (Locality Preserving Indexing), etc.

2.4 Feature Selection

The important step of text classification, after pre-processing and indexing is Feature selection. The main goal of Feature selection is to select a subset of features from the original features without affecting the classifier performance. Section 3 describes feature selection and its methods in detail.

2.5 Classification Algorithms

Automatic classification has been observed to have an active attention, and is being extensively studied from the past few years. Various classification techniques have evolved from machine learning techniques such as Bayesian classifier, K-Nearest Neighbor (KNN), Decision tree to soft computing methodologies like neural Networks, Genetic Algorithms, Fuzzy logic, Support vector machine(SVM), etc. Section 4 describes the classification algorithms in detail.

2.6 Performance Measure

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

2.7 Training phase

It is also called as Model Construction or Learning Phase.
The set of documents used for learning phase is called training set. It describes a set of predetermined labeled classes. Each document in the training set is assumed to belong to a predefined class (labeled documents). This set is used to train the classifier to take appropriate categorization decisions. Here, the classifier learns from training data, hence called learning phase. The model is represented as classification rules, decision trees, or mathematical formulae [2] [3].

2.8 Testing phase

After training phase, then comes the testing phase which is also called Mode Usage or Classification Phase. It is used for classifying future or unlabeled documents. The known label of test document is compared with the classified result to estimate the accuracy of the classifier.

3. Feature Selection and its Methods

The accuracy of the classifier not only depends on the classification algorithm but also on the feature selection method. Selection of irrelevant and inappropriate features may confuse the classifier and lead to incorrect results. The solution to this problem is Feature Selection i.e. feature selection is must in order to improve efficiency and accuracy of classifier.

Feature selection selects subset of features from original set of features by removing the irrelevant and redundant features from the original dataset. It is also known as Attribute selection. Feature selection reduces the dimensionality of the dataset, increases the learning accuracy and improves result comprehensibility.

The two search algorithms ‘forward selection’ and ‘backward eliminations’ are used to select and eliminate the appropriate feature. Feature selection is a three step process namely search, evaluate and stop.

Feature selection methods are also classified as attribute evaluation algorithms and subset evaluation algorithms. In first method, features are ranked individually and then a weight is assigned to each feature according to each feature’s degree of relevance to the target feature. The second approach in contrast, selects feature subsets and then ranks them based on certain evaluation criteria. Attribute evaluation methods do not measure correlation between feature are hence likely to yield subsets with redundant features. Subset evaluation methods are more efficient in removing redundant features. Different kinds of feature selection algorithms have been proposed. The feature selection techniques are broadly categorized into three types: 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 principle criteria in Filter method. The variables are assigned a score using a suitable ranking criterion and the variables having score below some threshold value are removed. Filter methods are computationally cheaper, avoids over fitting but these methods ignore dependencies between the features. Hence, the selected subset might not be optimal and a redundant subset might be obtained. The RELIEF algorithm [5,6] is another filter based approach wherein a feature relevance criterion 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). More specifically in feature selection it is used to test whether the occurrence of a specific term 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:

$$\chi^2(D,t,c) = \sum_{et\in\{1,0\}} \sum_{ec\in\{1,0\}} \frac{(N_{etec} - E_{etec})^2}{E_{etec}} \quad (1)$$

High scores on χ_2 indicate that the null hypothesis (H_0) of independence should be rejected and thus that the occurrence of the term and class are dependent. If they are dependent then we select the feature for the text classification.

3.1.2 Euclidean Distance

In this feature selection technique, the correlation between features is calculated in terms of Euclidean distance. If sample feature say '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.

$$d(a,b) = \left\{ \sum_i (a_i - b_i)^2 \right\}^{1/2} \quad (2)$$

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

3.1.3 Correlation criteria

Pearson correlation coefficient is simplest criteria and is defined by the following equation:

$$R(i) = \frac{\mathrm{cov}(x_i, Y)}{\sqrt{\mathrm{var}(x_i) * \mathrm{var}(Y)}} \quad (3)$$

Where, x_i is i^{th} variable, Y is the output class, var() is the variance and cov() denotes covariance. The disadvantage is that correlation ranking can only detect linear dependencies between variable and target.

3.1.4 Information Gain

Information gain tells us how important a given attribute of the feature vectors is. IG feature selection method selects the terms having the highest information gain scores. Information gain (IG) measures the amount of information in bits about the class prediction, if the only information available is the presence of a feature and the corresponding class distribution. Concretely, it measures the expected reduction in entropy (uncertainty associated with a random feature) defined as:

$$Entropy = \sum_{i=1}^{n} - p_i \log_2 p_i \quad (4)$$

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

$$Gain(A) = Entropy(S) - \sum_{k=1}^{m} \frac{S_k}{S} \times Entropy(Sk) \quad (5)$$

Where, S_k is the subset of S.

3.1.5 Mutual Information

Information theoretic ranking criteria [ ] uses the measure of dependency between two variables. To describe MI we must start with Shannon's definition for entropy given as:

$$H(X) = -\sum_i P(y)\log P(y) \quad (6)$$

Above equation represents the uncertainty (information content) in output Y. Suppose we observe a variable X then the conditional entropy is given by:

$$H(Y \mid X) = -\sum_x \sum_y P(x,y)\log P(y \mid x) \quad (7)$$

Above equation implies that by observing a variable X, the uncertainty in the output Y is reduced. The decrease in uncertainty is given as:

$$I(Y,X) = H(Y) - H(Y \mid X) \quad (8)$$

This gives the MI between Y and X meaning that if X and Y are independent then MI will be zero and greater than zero if they are dependent. This implies that one variable can provide information about the other thus proving dependency. The definitions provided above are given for discrete variables and the same can be obtained for continuous variables by replacing the summations with integrations.

3.1.6 Correlation based Feature Selection (CFS)

Correlation-based Feature Selection algorithm selects attributes by using a heuristic which measures the usefulness of individual features for predicting the class label along with the level of inter-correlation among them. The highly correlated and irrelevant features are avoided. The equation used to filter out the irrelevant, redundant feature which leads the poor prediction of the class is defined as:

$$F_s = \frac{N * r_a}{N + N(N-1)r_u} \quad (9)$$

3.1.7 Fast Correlation based Feature Selection

FCBF (Fast Correlation Based Filter) [4] is a multivariate (see section 1) feature selection method which starts with full set of features, uses symmetrical uncertainty to calculate dependences of features and finds best subset using backward selection technique with sequential search strategy. The

FCBF algorithm consists of two stages: the first one is a relevance analysis that orders the input variables depending on a relevance score, which is computed as the symmetric uncertainty with respect to the target output. This stage is also used to discard irrelevant variables, whose ranking score is below a predefined threshold. The second stage is a redundancy analysis, which selects predominant features from the relevant set obtained in the first stage. This selection is an iterative process that removes those variables which form an approximate Markov blanket. Symmetrical Uncertainty (SU) is a normalized information theoretic measure which uses entropy and conditional entropy values to calculate dependen cies of features. An SU value of 1 indicates that using one feature other feature's value can be totally predicted and value 0 indicates two features are totally independent.

3.2 Wrapper methods

Wrapper methods are better in defining optimal features rather than simply relevant features. They do this by using heuristics of the learning algorithm and the training set. Backward elimination is used by the wrapper method to remove the insignificant features from the subset. The SVM-RFE is one of the feature selection algorithms which use the Wrapper method. The Wrapper method needs some predefined learning algorithm to identify the relevant feature. It has interaction with classification algorithm. The over fitting of feature is avoided using the cross validation. Though wrapper methods are computationally expensive and take more time compared to the filter method, they give more accurate results than filter model. In filter model, optimal features can be obtained rather than simply relevant features. Another advantage is it maintains dependencies between features and feature subsets. Wrapper methods are broadly classified as sequential selection algorithms and heuristic search algorithms as follows:

3.2.1 Sequential Selection Algorithms

The Sequential Feature Selection (SFS) [22][24][25] algorithm starts with an empty set and adds one feature for the first step which gives the highest value for the objective function. After the first step, the remaining features are added individually to the current subset and the new subset is evaluated. The individual features that give maximum classification accuracy are permanently included in the subset. The process is repeated until we get required number of features. This algorithm is called a naive SFS algorithm since the dependency between the features is not taken into consideration.

A Sequential Backward Selection (SBS)[21][23] algorithm is exactly reverse of SFS algorithm. Initially, the algorithm starts from the entire set of variables and removes one irrelevant feature at a time whose removal gives the lowest decrease in predictor performance. The Sequential Floating Forward Selection (SFFS) algorithm is more flexible than the naive SFS because it introduces an additional backtracking step. The algorithm starts same as the SFS algorithm which adds one feature at a time based on the objective function. SFFS algorithm then applies one step of SBS algorithm which excludes one feature at a time from the subset obtained in the first step and evaluates the new subsets. If excluding a feature increases the value of the objective function then that feature is removed and algorithm switches back to the first step with the new reduced subset or else the algorithm is repeated from the top. The entire process is repeated until the required numbers of features are obtained or required performance is reached. SFS and SFFS produce nested subsets since forward inclusion was unconditional.

3.2.2 Heuristic Search Algorithms

Heuristic search algorithms include Genetic algorithms(GA)[6][9][11]12][14][33], Ant Colony Optimization(ACO)[27][28][29], Particle Swarm Optimization(PSO)[30][31],etc.A genetic algorithm is a search technique used in computing to find true or approximate solution to optimization and search problems. Genetic algorithms are based on the Darwinian principle of survival of the fittest theory. ACO is based on the shortest paths found by real ants in their search for food sources. ACO approaches suffer from inadequate rules of pheromone update and heuristic information. They do not consider random phenomenon of ants during subset formations. PSO approach do not employ crossover and mutation operators, hence is efficient over GA but requires several mathematical operators. Such mathematical operations requires various user-specified parameters and dealing with these parameters, deciding their optimal values might be difficult for users.

Although these ACO and PSO algorithms execute almost identically to GA, GA has received much attention due to its simplicity and powerful search capability upon the exponential search spaces.

3.3 Embedded methods

In embedded method[6][26][34], a feature selection method is incorporated into a learning algorithm and optimized for it. It is also called the hybrid model which is combination of filter and wrapper method. Embedded methods [8] reduce the computation time taken up for reclassifying different subsets which is done in wrapper methods. The KP-SVM is the example for embedded method.

The problem of nesting effect of SFS and SFFS was overcome by developing an adaptive version of SFFS called Adaptive Sequential Forward floating Selection (ASFFS) algorithm. In ASFFS algorithm, two parameters ‘r’ and ‘o’ are used where ‘r’ specifies number of features to be added while parameter ‘o’ specifies number of features to be excluded from the set so as to obtain less redundant subset than the SFFS algorithm.

The Plus L is a generalization of SFS and the Minus R is the generalization of SBE algorithm. If L>R then the algorithm start with SFS algorithm i.e. start from empty set and add the necessary features to the resultant set, else the algorithm start with the SBE algorithm i.e. start from entire set and start eliminating the irrelevant features and produce the resultant set. The Plus-L-Minus-r search method also tries to avoid nesting. In this method, the parameters L and r have to be chosen arbitrarily. It consumes less time than wrapper method but gives less accurate results than wrapper model as

some of the important features may be lost by the filter model.

4. Classification Algorithms

Several classification algorithms have been developed and used for classifying the documents, few of which are discussed below.

4.1 Bayesian Classifier

Bayesian classifier also called Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $x = (x_1, x_{2\ldots}\ x_n)$ representing some n features (independent variables), it assigns to this instance probabilities $p(C_k \mid x_1, x_2, \ldots\ldots\ldots, x_n)$ for each of *K* possible outcomes or *classes*. The conditional probability using Bayes theorem can be specified as:

$$P(C_k \mid x) = \frac{P(x \mid C_k)P(C_k)}{P(x)} \quad (10)$$

The denominator is effectively constant, so only the numerator is of concern. According to Naïve conditional independence assumptions, each feature Fi is conditionally independent of other feature F_j for $j \neq i$, given the category C. The joint model can be expressed as:

$$P(C_k \mid x_1, x_2, \ldots x_n) \propto P(C_k)\prod_{i=1}^{n} p(x_i \mid C_k) \quad (11)$$

The goal of text classification is to find the best class for the document. The best class in Naive Bayes classification is the one having maximum a posteriori and is defined by a function that assigns a class label y = Ck ; for some k as follows:

$$y = \arg\max P(C_k \prod_{i=1}^{n} p(x_i \mid C_k)) \quad (12)$$

a) Advantages

1) Naive Bayes classifier is quite efficient since it is less computationally intensive (in both CPU and memory).
2) It necessitates a small amount of training data.

b) Limitations

1) Assumption of conditional independence results in degraded performance.
2) As Bayesian classifier processes binary feature vectors, they have to abandon possibly relevant information.

4.2 Decision Tree Induction

Decision trees are non-parametric supervised learning method that predicts the value of target variable by learning simple decision rules inferred from the data features. In other words, decision tree classification is the learning of decision trees from labeled training documents. A decision tree is a flowchart like tree structures, where each internal node denotes a test on document, each branch represents an outcome of the test, and each leaf node holds a class label. It is a top-down method which recursively constructs a decision tree classifier. Some of the most well- known decision tree algorithms are ID3 and its successor C4.5 and C5, CART, MARS, etc.

a) Advantages

1) Robust to noisy data and capable of learning disjunctive expressions.
2) Decision trees are simple to understand and interpret and require little data preparation.

b) Limitations

1) Decision-tree learning algorithms are based on heuristic algorithms where decisions are made at each node locally, hence doesn't guarantee to return the globally optimal decision tree.
2) Decision-tree learners can create over-complex trees that do not generalize the data well, which is also called over fitting.

4.3 K-Nearest Neighbor

K-NN is a non-parametric method used for classification. The output of KNN classification is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its 'k' nearest neighbors, measured by 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, also called lazy learning, where the function is only approximated locally and all computation is deferred until classification.

a) Advantages

1) Simplicity i.e. simplest of all machine learning algorithms. It has reasonable similarity measures and does not need any resources for training.
2) K-NN is also very flexible. It can work with any arbitrarily shaped decision boundaries.

b) Limitations

1) It uses all features in computing distance and costs very much time for classifying objects.
2) KNN algorithm is sensitive to the local structure of the data.
3) KNN consumes more time for classifying objects when large number of training examples are given.

4.4 Support Vector Machines

Support Vector machines are non-probabilistic binary linear classifiers, supervised learning model defined by a separating hyper plane. In other words, given labeled training data, the algorithm outputs an optimal hyper plane which categorizes new examples. The operation of the SVM algorithm is based on finding the hyper plane that gives the largest minimum distance to the training examples. This distance is called "margin" within SVM's theory. The best hyper plane is the one that represents largest margin or separation between two classes. So we choose the hyper plane so that the distance

from it to the nearest data point on each side is maximized. If such a hyper plane exists, it is known as the "maximum-margin hyper plane" and the linear classifier it defines is known as a "maximum margin classifier"; or the "perceptron of optimal stability".

a) Advantages

1) Effective in high dimensional spaces.
2) Superior runtime-behavior during the classification of new documents because only one dot product per new document has to be computed
3) Versatile: different Kernel functions can be specified for the decision function.

b) Limitations

1) A document could be assigned to several categories because the similarity is typically calculated individually for each category.
2) It works only in real-valued space. For a categorical attribute, we need to convert its categorical values to numeric values.
3) It allows only two classes, i.e., binary classification.
4) The hyperplane produced by SVM is hard to understand by users. It is difficult to picture where the hyperplane is in a high-dimensional space.

4.5 Neural Networks

Neural networks [22][23] have emerged as an important tool for classification. For classifying a document in a neural network, its feature weights are loaded into the input nodes; the activation of the nodes is propagated forward through the network, and the final values on output nodes determine the categorization decisions. The neural networks are trained by back propagation, whereby the training documents are loaded into the input nodes. If misclassification occurs, then the error is propagated back through the network, modifying the link weights in order to minimize the error. The simplest kind of a neural network is a "perceptron" which has only two layers: the input and the output nodes. More sophisticated neural networks consist of hidden layer(s) between the input and output nodes. These feed-forward -nets consist of at least three layers and use back propagation as learning mechanism.

a) Advantages

1) High flexibility: Neural networks are nonlinear models, which makes them flexible in modeling real world complex relationships.
2) Neural networks are able to estimate the posterior probabilities, which provide the basis for establishing classification rule and performing statistical analysis.
3) They can handle noisy or contradictory data very well.

b) Limitations

1) With increase in the number of input and hidden nodes, the parameters needed for neural network also increases this result in over fitting of the data.
2) Very high computing costs.
3) Extremely difficult to understand for an average user.

5. Related work

There has been an immense amount of work in the area of feature selection that has been presented in the past few decades for mining the optimal features subset from the high-dimensional feature spaces. Along with it, lot of research has been done by the researchers on classification algorithms to obtain more accurate, and efficient classification results. Various authors have worked on various techniques out of which some are listed below.

Sheng-yi Jiang et al. [13] proposed a novel approach that measures the correlation between a continuous feature and a discrete feature. They also proposed an efficient filter feature selection algorithm based on correlation analysis that removed irrelevant features according to the correlation between features and the class feature, and eliminated redundant features among the relevant ones. The proposed algorithm achieves high degree of dimensionality reduction and decrease classification error rates with the selected features compared to FCBF, FarVPKNN and NDEM.

Uysal et al. [4] proposed s a novel filter based probabilistic feature selection method, called distinguishing feature selector (DFS), for text classification. The proposed method obtained promising results in terms of classification accuracy, processing time and dimension reduction as compared to well-known filter approaches such as chi square, information gain, Gini index and deviation from Poisson distribution.

G. Docquire et al. [7] introduces a new methodology to perform feature selection in multi-label classification problems using the multivariate mutual information criterion combined with a problem transformation and a pruning strategy. The proposed method considered joint relevance between features and obtained better results than chi-square technique. A. Kadhim et al. [19] used tf-idf weighting for extracting features considering the cosine similarity of text documents. The system reduced the feature set tocertain extenr but required larger execution time.

Besides filter based approaches, a number of algorithms using sequential search strategy [20] [21] [22] [23] [24] [25] have been proposed. S. Guan et al. [22], uses SFS strategy where significant features are added sequentially to the Neural Network (NN) during training. The feature addition process depends on the improvement of NN performance. Kabir et al. [24] [25] also proposed approaches based on SFS-based FS. The key idea of these approaches is to provide the correlation information of input features to the NN classifiers during training. In other studies, S. Abe et al. [20], C. Gasca et al. [21] and C. Hsu et al. [23] considered SBS in FS process, where the least salient features are deleted in stepwise fashion during the training of Neural Network.

Apart from the sequential search-based FS algorithms, global search-based algorithms (i.e., meta-heuristic algorithms) start searching in a full space rather than partial space for finding high-quality solution [26]. ACO algorithm is used for feature

selection in the systems proposed by R. Sivagaminathan et al. [27], M. Aghdam et al. [28], L. Ke et al. [29]. The basic idea in ACO is formulated by the observation of real ants in their search for the shortest paths to food sources. L. Ke et al. [29] proposed method using ACO algorithm that measures the heuristic information by using filter tools, i.e. statistical analysis of features. Wrapper or filter approach has generally been used in order to evaluate the constructed subsets.

Wang et al. [30] proposed a FS approach based on rough sets and PSO algorithm. Multi Swarm PSO method for selecting effective emotional features form reviews was proposed by Z. Liu et al. [31] that obtained promising results by effectively reducing redundant features.

GA is dominantly a nice tool used in several studies [6] [9] [14] for FS. A. Ozcift et al. [9] proposed a novel approach for differential diagnosis of erythemato-squamous diseases based on Genetic Algorithm (GA) wrapped Bayesian Network (BN) Feature Selection (FS). GA makes a heuristic search to find most relevant feature model that increase accuracy of BN algorithm with the use of a 10-fold cross-validation strategy. R. Wekilala et al. [12], M. Pedergnan et al. [33] have used Genetic algorithms for optimal Feature Selection in image datasets. Feature Selection method based on GA for classification as well as clustering was proposed by S. Hong et al. [32]. C. Tsai et al. [17] introduced a system that performs Feature Selection as well as instance selection based on GA to examine classification performance over different domain datasets.

Recently, hybrid approaches were also proposed. A. Ghareb et al. [6] proposed a hybrid feature selection technique that combines the advantages of filter feature selection methods with an enhanced GA (EGA) in a wrapper approach. A hybrid wrapper-filter feature selection technique was proposed by J. Apolloni et al. [34]. M. Kabir et al. [26] presented a new hybrid genetic algorithm (HGA) for Feature selection, where a new local search operation was embedded in HGA to fine-tune the search in FS process.

6. 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 [47]. 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

7. Conclusion

Text classification and Feature selection both are widespread domains of research encompassing Data mining, NLP and Machine Learning. These techniques have gained significant importance owing to the high growth rate of internet. In this survey paper, various feature selection techniques along with their advantages and limitations have been discussed. Various approaches for text categorization, their advantages and limitations also have been discussed in the paper. This survey paper circumscribes literature survey feature selection methods and different classifiers. From the survey done, we can conclude that among three the approaches to Feature selection method, filter methods should be used if we want results in lesser time and for large datasets. If we want the results to be accurate and optimal, wrapper method like GA 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. Performance of different algorithms varies according to the data collection and requirements. 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 difficulties can be overcome by using soft computing methodologies that 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] N. Azam, J. Yao, “Comparison of term frequency and document frequency based feature selection metrics in text categorization,” Expert Systems with Applications, Elsevier 39, 4760–4768,2012.
[2] A. Mesleh, “Feature sub-set selection metrics for Arabic text classification,” Pattern Recognition Letters, 32(14), 1922–1929, August 2011.
[3] H. Ogura, H. Amano, M. Kondo, “Feature selection with a measure of deviations from Poisson in text categorization,” Expert Systems with Applications, 36(3),6826–6832, 2009
[4] Uysal, A. K., & Gunal, S., “A novel probabilistic feature selection method for text classification,” Knowledge-Based Systems, 36, 226–235, 2012.
[5] P. Somol , P. Pudil , J. Novovicova, P. Paclik, “Adaptive Floating search methods in feature selection,”, Pattern Recognition Letters 20, 1157-1163, 1999.
[6] A. Ghareb , A. Bakar, A. Hamdan, “Hybrid feature selection based on enhanced genetic algorithm for text categorization,” Expert SystemsWith Applications, Elsevier, 2015.
[7] G. Doquire , M. Verleysen, “Mutual information-based feature selection for multilabel classification,” Neuro-computing, Elsevier, June 2013.

[8] M. Zhu, J. Song, "An Embedded Backward Feature Selection Method for MCLP Classification Algorithm," Information Technology and Quantitative Management, Elsevier, 2013.
[9] A. Ozcift, A. Gulten, "Genetic algorithm wrapped Bayesian network feature selection applied to differential diagnosis of erythemato-squamous diseases," Digital Signal Processing, Elsevier, July 2013.
[10] J. Cadenas, M. Carmen Garrido, R. Martínez, "Feature subset selection Filter-Wrapper based on low quality data," Expert Systems with Applications, Elsevier,2015.
[11] C. Lin, H. Chen, Y. Wua, "Study of image retrieval and classification based on adaptive features using genetic algorithm feature selection," Expert Systems with Applications, Elsevier,2014.
[12] R.A. Welikala, M.M. Fraz, J. Dehmeshki, A. Hoppe, V. Tah, S. Mann,T.H. Williamson, S.A. Barmana, "Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy," Computerized Medical Imaging and Graphics, Elsevier, March 2015.
[13] S. Jiang, L. Wang, "Efficient Feature Selection Based on Correlation Measure between Continuous and Discrete Features," 2015.
[14] P. Alirezazadeh, A. Fathi, F. Abdali-Mohammadi, "A Genetic Algorithm-Based Feature Selection for Kinship Verification," DOI 10.1109 / LSP.2015.2490805, IEEE Signal Processing Letters, 2015.
[15] Y. Lei, H. Liu. , "Feature selection for high-dimensional data: A fast correlation-based filter solution," ICML. Vol. 3. 2003.
[16] P. Pudil, J. Novovicova, J. Kittler, "Floating search methods in feature selection'" Pattern Recog Letters, 15:1119–25, November 1994.
[17] C. Tsai, W. Eberle, C. Chu, "Genetic algorithms in feature and instance selection," Knowledge-Based Systems, November 2012.
[18] H. Peng, F. Long, C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8), 1226–1238, August 2005.
[19] A. Kadhim, "Feature Extraction for Co-Occurrence-Based Cosine Similarity Score of Text Documents," IEEE Conference, 2014.
[20] S. Abe, "Modified backward feature selection by cross validation," Proceedings of the European Symposium on Artificial Neural Networks, 163–168, April 2005.
[21] E. Gasca, J. Sanchez, R. Alonso, "Eliminating redundancy and irrelevance using a new MLP-based feature selection method," Pattern Recognition 39, 313–315, 2006.
[22] S. Guan, J. Liu, Y. Qi, "An incremental approach to contribution-based feature selection," Journal of Intelligence Systems 13 (1), 2004.
[23] C. Hsu, H. Huang, D. Schuschel, "The ANNIGMA-wrapper approach to fast feature selection for neural nets," IEEE Transaction son Systems, Man, and Cybernetics—Part B:Cybernetics32(2)207–212, April 2002.
[24] M.M. Kabir, M.M. Islam, K. Murase, "A new wrapper feature selection approach using neural network," in: Proceedings of the Joint Fourth International Conference on Soft Computing and Intelligent Systems and Ninth International Symposium on Advanced Intelligent Systems (SCIS&ISIS2008), Japan, pp. 1953–1958, 2008.
[25] M.M. Kabir, M.M. Islam, K. Murase, "A new wrapper feature selection approach using neural network," Neurocomputing 73, 3273–3283, May 2010.
[26] M.M. Kabir, M.M. Islam, K. Murase, "A new local search based hybrid genetic algorithm for feature selection," Neurocomputing 74, 2194-2928, May 2011.
[27] R.K. Sivagaminathan, S. Ramakrishnan, "A hybrid approach for feature subset selection using neural networks and ant colony optimization," Expert Systems with Applications 33, 49–60, 2007.
[28] M.H. Aghdam, N.G. Aghaee, M.E. Basiri, "Text feature selection using ant colony optimization," Expert Systems with Applications 36, 6843–6853, 2009.
[29] L. Ke, Z. Feng, Z. Ren, "An efficient ant colony optimization approach to attribute reduction in rough set theory," Pattern Recognition Letters 29, 1351–1357, March 2008.
[30] X. Wang, J. Yang, X. Teng, W. Xia, R. Jensen, "Feature selection based on rough sets and particle swarm optimization," Pattern Recognition Letters 28 (4), 459–471, November 2006.
[31] Z. Liu, S. Liu, L. Liu, J. Sun, X. Peng, T. Wang, "Sentiment recognition of online course reviews using multi-swarm optimization-based selected features," Neuro-Computing, Elsevier, December 2015.
[32] S. Hong, W. Lee, M. Han, "The Feature Selection Method based on Genetic Algorithm for efficient Text Clustering and Text Classification ," Int. J. Advance Soft Computing Applications, Vol. 7, No. 1, March 2015.
[33] M. Pedergnan, "A Novel Technique for Optimal Feature Selection in Attribute Profiles Based on Genetic Algorithms," IEEE Transactions on Geoscience and Remote Sensing, Vol. 51, No. 6, June 2013.
[34] J. Apolloni, G. Leguizamon, E. Alba, "Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments," Applied Soft Computing, Elsevier, October 2015.
[35] M. Gupta, N. Aggrawal, "Classification Techniques Analysis," NCCI National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, pp. 128-131, March 2010.
[36] G. Zhang, "Neural Networks for Classification: A Survey", IEEE Transactions on systems, man and cybernetics-Part C, Applications and Reviews, Vol. 30, NO. 4, November 2000.
[37] L. Manevitz, M. Yousef, "One-class document classification via Neural Networks," Neurocomputing 70, 1466–1481, Elsevier, 2007.

[38] G. Zhao, Y. Wu, F. Chen, J. Zhang, J. Bai, "Effective feature selection using feature vector graph for classification," Neurocomputing, Elsevier,2015.
[39] P. Bermejo, J. Gamez, J. Puerta, "Speeding up incremental wrapper feature subset selection with Naive Bayes classifier," Knowledge-Based Systems, Elsevier, 2014.
[40] H. Chen, C. Huang, X. Yu, X. Xu, X. Sun, G. Wangd, S. Wang, "An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach," Expert Systems with Applications, Elsevier, 2013.
[41] U. Singh, S. Hasan, "Survey paper on document classification and classifiers," International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 2, Mar-Apr 2015.
[42] A. Saxena, V. Dubey, "A Survey on Feature Selection Algorithms," International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 3 Issue: 4, 2015.
[43] S.Vanaja, K. Ramesh Kumar, "Analysis of Feature Selection Algorithms on Classification: A Survey," International Journal of Computer Applications (0975 – 8887) Volume 96– No.17, June 2014.
[44] V. Korde, "Text Classification and Classifiers: A SURVEY," International Journal of Artificial Intelligence & 19. Applications (IJAIA), Vol.3, No.2, March 2012.
[45] M. Bali, D. Gore," A Survey on Text Classification with Different Types of Classification Methods," International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 5, May 2015.
[46] G. Chandrashekar , F. Sahin, "A survey on feature selection methods, " Computers and Electrical Engineering, Elsevier, 2014.
[47] M. Mali, M. Atique, "Applications of Text Classification using Text Mining," International Journal of Engineering Trends and Technology (IJETT) – Volume 13, Number 5, July 2014.

Author Profile

Pradnya Kumbhar received the B.E. degree in Information Technology from Shivaji University in 2014. She is currently pursuing Masters in Computer Engineering at Savitribai Phule Pune University. Her current research interests include Text mining and Data Mining.