A SURVEY ON TEXT CLASSIFICATION

KEYWORDS Text Classification, Data Mining.

Namekar Shirish Manohar
Department of Computer Engineering, PES's MCOE, Savitribai Phule Pune University,Pune, Maharashtra, India

Deipali V. Gore
Department of Computer Engineering, PES's MCOE, Savitribai Phule Pune University,Pune, Maharashtra, India

ABSTRACT *There are various application in various domains like data integration, marketing, medical diagnosis, news group filtering and documents organization etc. They are faced classification problems. This problem has been extensively studied in data mining, machine learning and informational retrieval. In this paper we will stipulate survey on text classifier.*

INTRODUCTION

The classification problem can be extensively studied in data mining and machine learning .Classification problem can be defined followed. There is training set data D={X1, X2,... Xn} such that each data and record is labeled values taken from set of k different values (1..k) which contains all possible data value pair as well as their probability distribution of appearance. It is subset of universal dataset due to lack of memory and some unavoidable reasons which is assumed to be exists the training data can be used for creating classification model which relates data values to class label. In the hard version of the classification problem, a particular label is explicitly assigned to the instance, whereas in the soft version of the classification problem, a probability value is assigned to the test instance.

A survey of a extensively variety of classification methods may be found in [1, 2], and a survey which is specific to the text domain may be found in [3]. A relative evaluation of different kinds of text classification methods may be found in [4]. A number of the techniques explained in this paper have also been converted into software and are publicly available in market.

In domain of Text mining, various applications that are faces problem of classification some of example domains can be used text classification.

Data Integration: A increase no of web portals are provide online shopping includes Amazon, ebay. A primary data integration task by this web portal is integration of data coming from multiple data providers into product catalog [8].

Opinion Mining: Customer reviews or opinions are often short text documents which can be mined to determine useful information from the review. Details on how classification can be used in order to perform opinion mining are discussed in [5] .

News filtering and Organization: There are large numbers of news web portals are in market. Which are electronic in nature in which a large volume of news articles are created In such cases, it is difficult to organize the news articles manually by human interval. Therefore, automated methods can be very useful for news categorization in a variety of web portals [6].

Document Organization and Retrieval: A many of supervised methods may be used for document organization. These include large digital libraries of documents like IEEE, web collections, scientific literature, or even social feeds. Taxonomy wise organized document collections can be especially useful for browsing and retrieval [19].

Email Classification and Spam Filtering: It is often desirable to classify email [23, 27, 85] In Email classification in order to determine either the subject or to determine junk email [113] in an automated way.

This is also considered to as spam filtering or email filtering.

A large number of techniques have been designed for text classification. Some key methods, which are commonly used for text classification, are as follows:

Bayesian Classifiers: Bayesian classifiers also called generative classifiers; Bayesian Classifiers is statistical classifier they attempt to predict class membership a probabilities. . The technique is then to classify text based on the posterior probability of the documents belonging to the different classes on the basis of the word presence in the documents.

SVM Classifiers: SVM Classifiers is algorithm for classification of both linear and nonlinear data. The key in such classifiers is to determine the optimal boundaries between the different classes for separation of data using essential training tuples called support vector and use them for the purposes of classification.

Neural Network Classifiers: Neural networks are used in a no of domains for the purposes of classification. In the context of text data, the main difference for neural network classifiers is to adapt these classifiers with the use of word features. We note that neural network classifiers are related to SVM classifiers; indeed, they both are in the category of discriminative classifiers, which are in contrast with the generative classifiers [102].

Decision Trees: Decision trees are designed with the use of a hierarchical division o f the underlying data space with the use of different text features. The hierarchical division of the data space is designed in order to create class partitions which are more skewed

in terms of their class distribution. For a given text instance, we determine the partition that it is most likely to belong to, and use it for the purposes of classification.

Pattern (Rule)-based Classifiers: In rule-based classifiers we determine the word patterns which are most likely to be related to the different classes. We construct a set of rules, in which the left hand side corresponds to a word pattern, and the right-hand side corresponds to a class label. These rules are used for the purposes of classification.

The domain of text categorization is so huge that it is impossible to cover all the different algorithms in detail in a single chapter or paper. Therefore, our aim is to provide the reader with an overview of the most important fundamental techniques, and also the direction to the different variations of these techniques.

CLASSIFICATION

Selection for Text Classification

First the most fundamental tasks that to be accomplished is that of document representation and feature selection. Feature selection is also used in other classification methods, it is most important in text classification because of it's the high dimensionality of text features. In general, text can be represented in two different ways. One is as a bag of words, in which a document is represented as a set of words with their associated frequency in the document. This representation should not dependent the sequence of words in the collection. The second method is to represent text directly as strings, in which each document is a sequence of words.

Most text classification methods use the bag-of-words representation because of its simplicity for classification purposes.

There is methods which are used for feature selection in text classification such as Gini index, Information gain, Mutual information, chi square distribution, Supervised Clustering for Dimensionality Reduction, Linear Discriminates Analysis.

In general, text can be represented in two different ways. One is as a bag of words, in which a document is represented as a set of words with their associated frequency in the document. This representation should not dependent the sequence of words in the collection. The second method is to represent text directly as strings, in which each document is a sequence of words.

Most text classification methods use the bag-of-words representation because of its simplicity for classification purposes.

There is methods which are used for feature selection in text classification such as Gini index, Information gain, Mutual information, chi square distribution, Supervised Clustering for Dimensionality Reduction, Linear Discriminates Analysis

Stop-word removal and stemming is used in supervised and unsupervised applications for feature selection In stop-word removal, we determine the common words in the documents which are not specific or discriminatory to the different classes. In stemming, different forms of the same word are consolidated into a single word. For example, singular, plural and different tenses are consolidated into a single word. We understand that these methods are not for only the classification problem but also used in unsupervised applications such as clustering and indexing.

A wide variety of feature selection methods are discussed in [10, 11. Many of these feature selection methods have been compared with one another, and the experimental results are presented in [10].

Decision Tree Classifiers

A decision tree [11] is important a hierarchical decomposition of the (training) data space, in this a predicate or a condition on the attribute that should value which is used such that to divide the data space hierarchically.

The subdivision of the data space is performed recursively in the decision tree, until the leaf nodes contain a minimum number of records,

The majority class label in the leaf node is used for the purposes of classification. For a given test instance, we apply predicates at the nodes such that order to traverse a path of the tree in top-down fashion.

There is many predicates. It may not be necessary to use individual terms for partitioning, but one may measure the similarity of documents to correlated sets of terms. These correlated sets of terms may be used to further partition the document collection, based on the similarity of the document to them. The different kinds of splits are as follows:

Single Attribute Splits: In Stop-word removal and stemming is used in supervised and unsupervised applications for feature selection In stop-word rule-based classifiers find out the word patterns which are most likely to be related to the different classes. There is a set of rules, in which the left hand side corresponds to a word pattern, and the right-hand side corresponds to a class label. These rules are used for the purposes of classification.

Similarity-based multi-attribute split: we use meta-documents and use the similarity of the documents to these words clusters in order to perform the split. For the selected word cluster, the documents are further partitioned into groups by rank ordering the documents by similarity value, and splitting at a particular threshold. We select the word-cluster for which rank-ordering by similarity provides the best separation between the different classes\

Single Attribute Splits: For this split is to use discriminants such as the Fisher discriminant for performing the split. Such discriminants provide the directions in the data along which the classes are best separated. The documents are projected on this discriminant vector for rank ordering, and then split at a particular coordinate. The choice of split point is picked in order to maximize the discrimination between the different classes. The work in [18] uses a discriminant-based split, though this is done indirectly because of the use of a feature transformation to the discriminant representation, before building the classifier. Some of the existing implementation of classifiers may be found in [13, 12]. These are a rule-based classifier, which can be consider either as a decision tree or a rule-based classifier. Most of small variations of standard packages such as ID3 and C4.5 are available for decision tree implementations in the text literature , in order to model to text classification.

Many of these classifiers are designed as foundation for review with other learning models [14].

A well known implementation of the decision tree classifier is based on the C4.5 taxonomy of algorithms [106] is presented in [13]. More specifically, the work in [13] uses the successor to the C4.5 algorithm, which is also known as the C5 algorithm. This algorithm uses single-attribute splits at each node, where the feature with the highest information gain is used for the purpose of the split. Decision trees have also been used in conjunction with boosting techniques. An adaptive boosting technique [48] is used in order to improve the accuracy of classification.

Rule-based Classifierss

Decision trees are also generally related to rule-based classifiers. In rule-based classifiers, the data space is modeled with a set of rules, in which the left hand side is a condition on the underlying feature set, and the right hand side is the class label. The rule set is essentially the model which is generated from the training data. For a given test instance, we determine the set of rules for which the test instance satisfies the condition on the left hand side of the rule. We determine the predicted class label as a function of the class labels of the rules which are satisfied by the test instance.

A number of criteria can be used in order to generate the rules from the training data. Two of the most common conditions which are used for rule generation are those of support and confidence. These conditions are common to all rule-based pattern classifiers [13] and may be defined as follows:

Support: This quantifies the absolute number of instances in the training data set which are relevant to the rule. For example, in a corpus containing 1000,000 documents, a rule in which both the left-hand set and right-hand side are satisfied by 50,0000 documents more important than a rule which is satisfied by 20 documents. Essentially, this quantifies the statistical volume which is associated with the rule. However, it does not encode the strength of the rule.

Confidence: This quantifies the conditional probability that the right hand side of the rule is satisfied, if the left-hand side is satisfied. This is a more direct measure of the strength of the underlying rule.

We note that the afore-mentioned measures are not the only measures which are possible, but are widely used in the data mining and machine learning literature [88] for both textual and non-textual data, because of their intuitive nature and simplicity of interpretation. One criticism of the above measures is that they do not normalize for the a-priori presence of different terms and features, and are therefore prone to misinterpretation when the feature distribution or class-distribution in the underlying data set is skewed.

Probabilistic and Naive Bayes Classifiers

We note that the Probabilistic Navie bays classifier is mixture model for generating document. Navie bays is simple method for classification and its most popular method. This is gives model distribution of feature in using probability.

TABLE – 1
COMPARATIVE STUDY

SR.NO	Comparative study		
	Paper	Technique	Conclusion
1	Panagiotis apadimitriou and Panayiotis Tsaparas,Taxonomy-AwareCatalog Inte- gration, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGI- NEERING, VOL.25, NO. 7, JULY2013.	The Base Classi cation,Taxonomy Aware Processing Step	It is better classifi cation accuracy than Navie bayes Classi cation.
2	R. Agrawal and R. Srikant, On Integrating Catalogs,Proc. 10th Intl Conf World Wide Web(WWW), pp. 603-612, 2001.	Enhanced Naive Bayes Class cafition	It is better Classification accuracy than Navie bayes Classi cafication but it lower classifisi cation accuracy than TACI approach and requires a standalone tune set to and optimal value of parameter which control in nse of source categorization information on classi cation .
3	S. Sarawagi, S. Chakrabarti, and S. Godbole, Cross- Training:Learning Probabilistic Map- pings between Topics, Proc. Ninth ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining (KDD), 2003..	semi- supervised learning in the presence of multiple label sets	They assume the existence of some training data labelled using both the source and master taxonomy.
4	D. Zhang and W.S. Lee, Web Taxonomy Integration through Co-Bootstrapping, Proc. 27th Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 410-417, 2004.	Boosting machine learning method for classi fication	No burden of tune set to optimal value of parameter which controls influence of source cate- gorization information on classi fication

5	D. Zhang and W.S. Lee, Web Taxonomy Integration Using Support Vector Machines, Proc. 13th Intl Conf. World Wide Web (WWW), pp. 472-481, 2004	Support vector machine(SVM) machine learning method for classi - fication	In contrast to NB, SVM is disciminative classi cation i.e. SVM does not posit a generative model.
5	A. Nandi and P.A. Bernstein, Hamster: Using Search Click- logs for Schema and Taxonomy Matching, Proc. VLDB Endowment vol. 2, no. 1,pp. 181-192,2009.	It matching taxonomies based on query term distributions and it performs the mapping at the taxonomy level, mapping categories from the source to the target.	It performs the mapping at the taxonomy level, mapping cate- gories from the source to the target.

CONCLUSIONS

The classification problem is one problem in the machine learning and data mining concept. The area of these sets is large so, text mining techniques need to be designed to effectively. Techniques for classification such as decision trees, rules, Bays methods, nearest neighbor classifiers, SVM classifiers, and neural networks have been extended to the case of text data.

The vast development and use of web and social network technologies like market places, facebook have lead to a tremendous interest in the classification of text documents containing links or other meta-information. Recent research scientist has shown that the incorporation of linkage information into the classification process can significantly improve the quality of the underlying results.

REFERENCE

[1]Aiello, M. A., and Leuzzi, F. (2010), "Waste Tyrerubberized concrete: Properties at fresh and hardened state." Journal of Waste Management, ELSEVIER, 30,1696-1704. | [1] R. Duda, P. Hart, W. Stork. Pattern Classification, Wiley Interscience,2000. | [2]M. James. Classification Algorithms, Wiley Interscience, 1985. | [3]F. Sebastiani. Machine Learning in Automated Text Categorization,ACM Computing Surveys, 34(1), 2002. | [4]Y. Yang, L. Liu. A re-examination of text categorization methods,ACM SIGIR Conference, 1999. | [5]B. Liu, L. Zhang. A Survey of Opinion Mining and Sentiment Analysis.Book Chapter in Mining Text Data, Ed. C. Aggarwal, C. Zhai,Springer, 2011. | [6] K. Lang. Newsweeder: Learning to filter netnews. ICML Conference,1995. | [7]S. Chakrabarti, B. Dom. R. Agrawal, P. Raghavan. Using taxonomy,discriminants and signatures for navigating in text databases,VLDB Conference, 1997. | [8]Ariel Fuxman, Panagiotis, TACI: Taxonomy-Aware Catalo Integration,IEEEtran, vol. 25,July2013. | [9]Y. Yang, J. O. Pederson. A comparative study on feature selection in text categorization, ACM SIGIR Conference, 1995. | [10]Y. Yang. Noise Reduction in a Statistical Approach to Text Categorization, ACM SIGIR Conference, 1995. | [11]J. R. Quinlan, Induction of Decision Trees, Machine Learning, 1(1), pp 81–106, 1986. | [12]D. Lewis, J. Catlett. Heterogeneous uncertainty sampling for supervised learning. ICML Conference, 1994. | [13]Y. Li, A. Jain. Classification of text documents. The ComputerJournal, 41(8), pp. 537–546, 1998. | [14]T. Joachims. Text categorization with support vector machines: learning with many relevant features. ECML Conference, 1998. | [15]J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61. | [16]S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device |