Empirical Studies on Machine Learning Based Text Classification Algorithms

Shweta C. Dharmadhikari[#1], Maya Ingle [*2], Parag Kulkarni [#3]

[#] Pune Institute of Computer Technology - EkLat Solutions
Pune , Maharashtra, India
[1]d.shweta18@gmail.com
[3]paragindia@gmail.com

**Indore Institute of Computer Applications , IIST,*
Indore, Madhya Pradesh , India
[2] maya_ingle@rediffmail.com

ABSTRACT

Automatic classification of text documents has become an important research issue now days. Proper classification of text documents requires information retrieval, machine learning and Natural language processing (NLP) techniques. Our aim is to focus on important approaches to automatic text classification based on machine learning techniques viz. supervised, unsupervised and semi supervised.
In this paper we present a review of various text classification approaches under machine learning paradigm. We expect our research efforts provide useful insights on the relationships among various text classification techniques as well as sheds light on the future research trend in this domain.

Keywords: Automatic text classification, Information Retrieval, NLP, Machine Learning.

I INTRODUCTION

The large availability of online text documents have provided us a very large amount of information. This available information must be organized systematically for its proper utilization. Systematic organization of information facilitates ease of storage, searching, and retrieval of relevant text content for needy application [9]. The Text Classification is an important technique for organizing text documents into classes [3][4]. Automatic Text classification is attractive because it relives the organizations from the need of manually organizing document bases, which is not only expensive, time consuming but also error prone [1].

Automatic classification of text is an important problem in many domains. It has many applications such as automated indexing of scientific articles, spam filtering, identification of document genre, authorship attribution, automated essay grading, survey coding, classification of news articles, etc.

This task falls at the crossroads of information retrieval, NLP and machine learning [8].Information Retrieval is the finding of the documents which contain answers to the questions. Statistical measures and methods are used to achieve the said goal[22] .Natural Language processing is used to get better understanding of natural language by representing documents semantically. This helps to improve classification results. Machine learning is concerned with the design and development of algorithms and techniques that allow computers to "learn" so as to improve the expected future performance.

DOI : 10.5121/acij.2011.2615

Algorithms used to train text categorization systems in information retrieval are often ad-hoc and poorly understood. In particular, very little is known about their generalization performance, that is, their behavior on documents outside the training data [14]. Machine learning techniques have the advantage that they are better understood from a theoretical standpoint, leading to performance guarantees and guidance in parameter settings

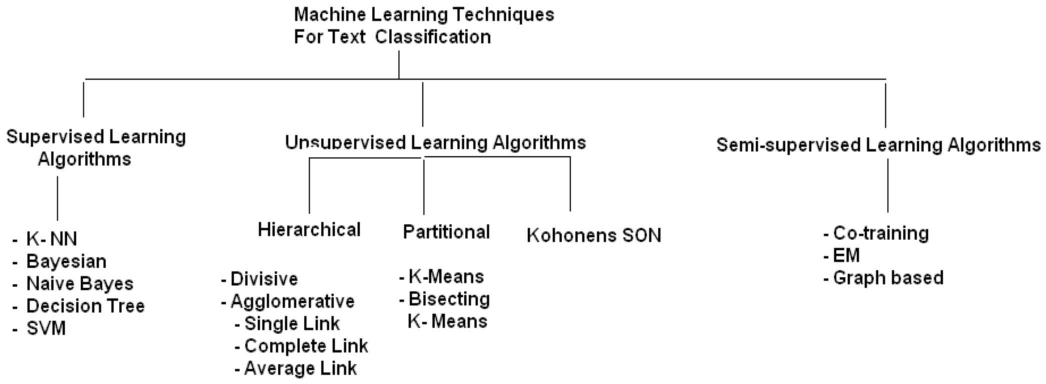

Fig.1 Classification Hierarchy of ML Algorithms for TC.

The major machine learning approaches falls under the category of supervised learning, unsupervised learning and semi supervised learning [1][7]. An increasing number of learning algorithms have been coming day by day, including Regrssion Models , Nearest Neighbor classification, Bayesian probabilistic approaches, Decision Trees , Neural Networks, On-line learning , Support Vector Machine(SVM), Co-training, Expectation Maximization , Graph based methods, Kohenen Self Organizing Maps, etc[3][7][8].

This paper surveys the major text classification algorithms under these three categories . Figure 1. shows main algorithms which will be discussed in this paper. This paper is organized as : Section II provides information about Supervised machine learning algorithms, Section III provides information about unsupervised machine learning algorithms, Section IV informs about Semi Supervised machine learning algorithms, Section V enlightens some of the hybrid algorithms. Subsequent sections convey conclusion & future work followed by references.

II SUPERVISED TEXT CLASSIFICATION ALGORITHMS

These algorithms use the training data, where each document is labeled by zero or more categories, to learn a classifier which classifies new texts. A document is considered as a positive example for all categories with which it is labeled, and as a negative example to all others. The task of a training algorithm for a text classifier is to find a weight vector which best classifies new text documents [21].

A K-Nearest Neighbor classifier

It is a well known pattern recognition algorithm. Given a test document, the kNN algorithm finds the k nearest neighbors

among the training documents and uses the categories of the k nearest neighbors to weight the category candidates [2]. The similarity score of each neighbor document to the test document is used as the weight of the categories of the neighbor document. This algorithm is based on the assumption that the characteristics of members of the same class should be similar. Thus observations located close together in covariate space are members of the same class.

It is suitable for data streams. It does not build a classifier in advance.

Merits: This method is effective, simple, non-parametric and easy to implement.

Demerits: The major drawback of this method is that it becomes slow when size of training set grows The presence of irrelevant features severely degrades its accuracy.

B Naïve Bayes Method

The Bayesian method that makes independence assumption is termed as Naïve Bayes classifier [2].

It predicts by reading a set of examples in attribute value-representation and then by using the Bayes theorm to estimate the posterior probabilities of all qualifications. The independence assumptions of features make the features order irrelevant and presence of one feature does not affect other features in classification task.[22]

Merits: This method requires a small amount of training data to estimate the parameters necessary for classification. The classifiers based on this algorithm exhibited high accuracy and speed when applied to large databases.

Demerits: This method works well only if assumed features are independent; when dependency arises then it gives low performance.

C Decision Trees

The decision tree categorizes the training documents by constructing well-defined true/false queries in the form of tree structure. In this leaves represent the corresponding category of the text documents and branches represent conjunctions of features that lead to these categories [22].

Merits: This method works on data of any type. It is fastest even in the presence of large amounts of attributes.

Demerits: The major risk of implementation of decision tree is it over fits the training data with the occurrence of an alternative tree.

D Decision Rules Classification

This method uses the rule-based inference to classify documents to their annotated categories [22]. These classifiers are useful for analyzing non-standard data. It constructs a rule set that describe the profile for each category. Rules are in the form of “If condition Then conclusion”, where condition portion is filled by features of the category, and conclusion portion is represented with the categories name or another rule to be tested.

Merits: This method is capable to perform semantic analysis.

Demerits: The major drawback of this method is the need of involvement of human experts to construct or update the rule set.

E Support Vector Machines

It is a statistical based learning algorithm [22]. This algorithm addresses the general problem of learning to discriminate between positive and negative members of a given class of n-dimensional vectors. It is based on the Structural Risk Minimization principle from computational learning theory.

The SVM need both positive and negative training set which are uncommon for other classification methods. The performance of the SVM classification remains unchanged even if documents that do not belong to the support vectors are removed from the set of training data; this is one of its major advantages.

Merits: Amongst existing supervised learning algorithms for TC SVM has been recognized as one of the most effective text classification methods [7][21] as it is able to manage large spaces of features and high generalization ability.

Demerits: But this makes SVM algorithm relatively more complex which in turn demands high time and memory consumptions during training stage and classification stage.

III UNSUPERVISED TEXT CLASSIFICATION ALGORITHMS / TEXT CLUSTERING

In unsupervised clustering, we have unlabelled collection of documents. The aim is to cluster the documents without additional knowledge or intervention such that documents within a cluster are more similar than documents between clusters. These are categorized into two major groups as partitioned and hierarchical.

Hierarchical algorithms produce nested partitions of data by splitting (divisive approach) or merging (agglomerative approach) clusters based on the similarity among them.

Partitional clustering algorithms group the data into un-nested non-overlapping partitions that usually locally optimize a clustering criterion.

A Hierarchical Clustering Techniques

Hierarchical clustering algorithms produce a cluster hierarchy in the form of tree structure named a dendrogram[11][21]. The root of the tree consists of single cluster containing all observations, and the leaves correspond to individual observations.

Merits: The major advantage of these techniques lies in its simplicity and ability to capture the information properly.

Demerits : It does not give discrete clusters, and we get different clustering for different experiment set.

This technique having following two subtypes:

B Divisive Hierarchical Clustering

Divisive algorithms start with one cluster at a time. During each iteration split the most appropriate cluster until a stopping criterion such as a requested number *k* of clusters is achieved.

In this technique in each step the cluster with the largest diameter is split, i.e. the cluster containing the most distant pair of documents. As we use document similarity instead of distance

as a proximity measure, the cluster to be split is the one containing the least similar pair of documents. Within this cluster the
document with the least average similarity to the other documents is removed to form a new singleton cluster.
The algorithm proceeds by iteratively assigning the documents in the cluster being split to the new cluster if they have greater average similarity to the documents in the new cluster.

C Agglomerative Hierarchical Clustering

Agglomerative clustering algorithms start with each document in a separate cluster and at each iteration merge the most similar clusters until the stopping criterion is met. They are mainly categorized as single-link, complete-link and average-link depending on the method they define inter-cluster similarity [16][21].

a) Single-link: The single-link method defines the similarity of two clusters *Ci* and *Cj* as the similarity of the two most similar documents. But
b) Complete-link: The complete-link method defines the similarity of two clusters *Ci* and *Cj* as the similarity of the two least similar documents.

c) Average-link: The average-link method defines the similarity of two clusters *Ci* and *Cj* as the average of the pairwise similarities of the documents from each cluster:

The aim of text clustering is to group text documents such that intra-group similarities are high and inter-group similarities are low. Document clustering has many application areas.
In Information Retrieval (IR), it has been used to improve precision and recall, and as an efficient method to find similar documents. More recently, document clustering has been used in automatically generating hierarchical groupings of documents and in document browsing to organize the results returned by a search engine.

D Partitional Clustering Techniques

In this clustering technique classes are mutually exclusive. Each object is the member of with which it is most similar. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. In this given the number of clusters *k*, an initial partition is constructed; next the clustering solution is refined iteratively by moving documents from one cluster to another [21].
Merits: It requires only one pass through dataset and therefore it is faster.
Demerits: In this resulting clusters are not independent of the order in which the documents are processed, with the first clusters formed usually being larger than those created later in the clustering run.

K-Means Clustering: In this each of *k* clusters can be represented by the mean of the documents assigned to that cluster, which is called the *centroid* of that cluster. There
are two versions of *k*-means algorithm. The first version is the *batch* version and is also known as Forgy's algorithm [11][12]. It consists of the following two-step major iterations:

- Reassign all the documents to their nearest centroids
- Recompute centroids of newly assembled groups
  (i) Before the iterations start, firstly *k* documents are selected as the initial centroids.
  (ii) Iterations continue until a stopping criterion such as no reassignments occur is achieved

Merits: The main advantage of k-means is its speediness and simplicity.
Demerits: Its random process makes this an indeterminate method.

Bisecting K-Means: It is actually a divisive clustering algorithm that achieves a hierarchy of clusters by repeatedly applying the basic *k*-means algorithm,
In each step of bisecting *k*-means a cluster is selected to be split and it is split into two by applying basic *k*-means for *k* = 2. The largest cluster that is the cluster containing the maximum number of documents, or the cluster with the least overall similarity can be chosen to be split[21].

E Kohonen's Self Organizing Network

It uses a special type of neural network called Kohonen's self-organizing network. The novelty of the method is that it automatically detects the number of classes present in the given set of text documents and then it places each document in its appropriate class [20][21].
The method initially uses the Kohonen's self-organizing network to explore the location of possible groups in the feature space. Then it checks whether some of these groups can be merged on the basis of a suitable threshold resulting in expected clustering. These clusters represent the various groups or classes of texts present in the set of given text documents. Then these groups are labeled on the basis of frequency of the class titles found in the documents of each group.
Merits: These algorithms can make it easy for humans to see relationships between vast amounts of data. These are commonly used in the applications for visualization aid.
Demerits: These are more complex and computationally extensive.

IV SEMI SUPERVISED TEXT CLASSIFICATION ALGORITHMS

These algorithms make use of unlabeled data along with few labeled data to classify new unlabeled text document. In text classification most of the times there is limited labeled data, and in most cases it can be expensive to generate that labeled data so semi-supervised algorithms gives good solution in such a situations. Its framework is applicable to both classification and clustering [13]. Some of the important algorithms discussed here are as:

A Co-training

The co-training setting applies when a dataset has a natural division of its features. Co-training requires two *views* of the data. It assumes that each example is described using two different feature sets that provide different, complementary information about the instance. These feature sets are conditionally independent [5][17].

Co-training first learns a separate classifier for each view using any labeled examples. The most confident predictions of each classifier on the unlabeled data are then used to iteratively construct additional labeled training data [18].

Co-training has been used to classify web pages using the text on the page as one view and the anchor text of hyperlinks on other pages that point to the page as the other view. The co-training models utilize both classifiers to determine the likelihood that a page will contain data relevant to the search criteria.

Merits: Its major advantage is it's simplicity and applicability to almost all existing classifiers.

Demerits: In many applications natural feature split may not available as per the requirement of this method in such cases its performance is low.

Co-training was used on FlipDog.com; a job search site by the U.S. Department of Labor. It has been used in many other applications, including statistical parsing and visual detection.

B EM based

It stands for Expectation Maximization algorithm. These algorithms are used to train classifiers by estimating parameters of a generative model through iterative Expectation-Maximization (EM) techniques[17][18]. In this, Text documents are represented with a bag-of-words model, which leads to a generative classification model based on a mixture of multinomial.

The basic EM procedure works well when data conform to the generative assumptions of the model. However these assumptions are often violated in practice and poor performance can result. But when model probability and classification are well-correlated, the use of deterministic annealing finds more probable and more accurate classifiers.

Merits :

1. This model is an extremely simplistic representation of the complexities of written text.

2. The Expectation-Maximization shows that with sufficiently large amounts of unlabeled data generated by the model class in question, a more probable model can be found than if using just the labeled data alone.

Demerits :

EM suffers from the problem of local maxima. It does not guarantee global maxima in model probability space.

C Graph based

These methods are generally non-parametric and transductive. In this a learning algorithm works on a closed data set and test set is revealed at the time of training. Here one assumes that the data is embedded within a low-dimensional manifold which is expressed by a graph. Each data sample is represented by vertex within a weighted graph with the weights providing a measure of similarity between vertices.

But many graph based SSL algorithms assume binary classification tasks and require the use of sub-optimal approaches. Modified versions of Graph based SSL are based on optimizing a loss function composed of KL-divergence terms between probability distributions defined for each graph vertex.

Merits: These methods not only provide solution to binary TC but also for multiclass TC. Measure for uncertainty [5][19] is also provided by these techniques.

Demerits: These methods require excessive computation.

V OTHER METHODS

Depending upon the need of application, in order to create more accurate text classifier from learner one can make use of hybrid learning algorithms. These algorithms combine the steps from few of above mentioned algorithms.

Eg. Naïve bias and EM algorithm when applied together yields more performance effective classifier. Decision tree along with Neural Network constructs the networks by directly mapping decision nodes or rules to the neural units and compresses the network by removing unimportant and redundant units and connections.

Conclusion and future work

This review paper is focused on the existing literature and explored major classification techniques along with their respective merits and demerits.

From most of the literature it is clear that performance of classification algorithm in text classification is greatly influenced by the quality of data source, feature representation techniques as the irrelevant and redundant features of data degrades the accuracy and performance of the classifier.

It was verified from the study that among the unsupervised techniques, *k*-means and bisecting *k*-means perform the best in terms of time complexity and the quality of the clusters produced. On the other hand, among the supervised techniques support vector machines achieve the highest performance while naive bayes performs the worst. Among the Semi-supervised techniques Graph based algorithm performs well as compare to co-training and Expectation-Maximization, but hybrid algorithm of EM and naïve bayes gives better results.

In many cases hybrid algorithms outperforms other and is gaining more research attention.

However there is need to make more generalized text classification systems which will efficiently able to utilize huge amount of unlabelled data .Such systems should also consider the nature of the input text documents [eg. Single or multi-label ie whether text document can belong to single or multiple class labels.]

The first author is doing research in Semi- supervised learning methods and proposes a new semi supervised Methodology for Multi-Label Text classification problem where a new text document can be assigned to more relevant category.

REFERENCES

[1] Fabrizio Sebastiani , Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 1–47.

[2] Kavi Narayana Murthy Advances in Automatic text categorization.DRTC Workshop on Semantic Web, Bangalore, India, 8-10 December, 2003.

[3] Boyapati,V. Improving hierarchical text classification using unlabeled data. Proceedings of SIGIR, (2002).

[4] Maribor, Slovenia. Text Categorization for Multi-label Documents and Many Categories. Twentieth IEEE International symposium on Computer-Based Medical Systems. June 20 2007.

[5] Amarnag Subramanya, Jeff Bilmes , Soft-Supervised Learning for text classification.Proceedings of the Conference on Empirical Methods in Natural Language Processing. Pages: 1090-1099.2008.

[6] Ganesh R.,Deepa P.,Byron Dom. A structure-sensitive framework for text categorization.CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005.

[7] Z. Wang, X. Sun, D. Zhang. An optimal Text categorization algorithm based on SVM.

[8] Goncalves, Paulo Quaresma. The impact of NLP techniques in the text multilabel classification problem.

[9] L.Tang,S. Rajan,V.K. Narayanan. Large Scale Multi-Label Classification via MetaLabeler. In *Proceedings of the Data Mining and Learning 2009.*

[10] Yu-Chuan Chang, S. Ming Chen. Multi-Label Text Classification based on a new linear classifier learning method and a category sensitive refinement method. ScienceDirect Expert Systems with Applications. Volume 34, Issue 3, April 2008. 1948-1953.

[11] Hartigan, J., *Clustering Algorithms*, John Wiley & Sons, New York, NY, 1975.

[12]Berkhin, P., “Survey of Clustering Data Mining Techniques”, Research paper, Accrue Software, http://www.accrue.com/products/researchpapers.html, 2002.

[13] N.M. Pise, Dr. Parag Kulkarni. A survey of Semi-Supervised Learning Methods”. IEEE International conference on Computational Intelligence and Security. 2008.30-34.

[14] Ido Dagan, Y. Karov, Dan Roth. Mistake- driven learning in text categorization.

[15] Arturo Montejo-Raez . Automatic Text Categorization of documents in the High Energy Physics domain. Thesis submitted in 15 December, 2005.

[16] Tom M. Mitchell, The Discipline of Machine Learning **,** CMU-ML-06-108, July 2006.

[17]Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning,39, 103–134.

[18] Nigam, K., McCallum, A., & Mitchell. T. (2006). Semi-supervised Text Classification Using EM. MIT Press.

[19] Zheng-Jun Zha, T. Mei.Graph-based Semi-Supervised Learning with Multi-label text classification. IEEE International conference on Multimedia.June 23 2008.1321-1324.

[20] N. Choudhary. An Unsupervised Text Classification Method using Kohonen’s Self Organizing Network.

[21] Arzucan Ozgur. Supervised and unsupervised machine learning techniques for text document categorization.Thesis submitted in Department of Computer Science, Bogaziki University. 2004.

[22] A.Khan,B.Baharudin,Lan Hong Lee. A Review of Machine Learning Algorithms for Text-Documents Classification. Journal Of Advances in Information Technology, Vol. 1 , No. 1, Feb.2010.