International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 2, Issue 3, March 2013 ISSN 2319 - 4847

A REVIEW PAPER ON ALGORITHMS USED FOR TEXT CLASSIFICATION

Bhumika[1], Prof Sukhjit Singh Sehra[2], Prof Anand Nayyar[3]

[1]M.Tech (IT) Scholar
Guru Nanak Dev Engineering College, Ludhiana

[2] Department of Computer Science and Engineering
Guru Nanak Dev Engineering College, Ludhiana

[3] Department of Computer Applications & IT
KCL Institute of Mgmt. & Technology, Jalandhar

ABSTRACT

The textual revolution has seen a tremendous change in the availability of online information. Finding information for just about any need has never been more automatic. Text classification (also known as text categorization or topic spotting) is the task of automatically sorting a set of documents into categories from a predefined set. This task has several applications, including automated indexing of scientific articles, filing patents into patent directories, selective dissemination of information to information consumers, automated population of hierarchical catalogues of Web resources, spam filtering, identification of document genre. Automated text classification is attractive because it frees organizations from the need of manually organizing document bases, which can be too expensive, or simply not feasible given the time constraints of the application or the number of documents involved. The accuracy of modern text classification systems rivals that of trained human professionals, thanks to a combination of information retrieval (IR) technology and machine learning (ML) technology. The aim of this paper is to highlight the important algorithms that are employed in text documents classification, while at the same time making awareness of some of the interesting challenges that remain to be solved.

Keywords: Text categorization, information retrieval, Machine learning, patents, spam filtering.

1. INTRODUCTION

The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. Text categorization (also known as text classification or topic spotting) is the task of automatically sorting a set of documents into categories from a predefined set [9].The resources of unstructured and semi structured information include the world wide web, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail and blog repositories. Therefore, proper classification and knowledge discovery from these resources is an important area for research. Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the electronic documents. The main goal of text mining is to enable users to extract information from textual resources and deals with operations like retrieval, classification (supervised, unsupervised, unsupervised and semi supervised) and summarization. However how these documented can be properly annotated, presented and classified. So it consists of several challenges, like proper annotation to the documents, appropriate document representation, dimensionality reduction to handle algorithmic issues [1], and an appropriate classifier function to obtain good generalization and avoid over-fitting. Extraction, Integration and classification of electronic documents from different sources and knowledge discovery from these documents are important for the research communities. Today the web is the main source for the text documents, the amount of textual data available to us is consistently increasing, and approximately 80% of the information of an organization is stored in unstructured textual format [2], in the form of reports, email, views and news etc. The [3] shows that approximately 90% of the world's data is held in unstructured formats, so Information intensive business processes demand that we transcend from simple document retrieval to knowledge discovery. The need of automatically retrieval of useful knowledge from the huge amount of textual data in order to assist the human analysis is fully apparent [4]. Market trend based on the content of the online news articles, sentiments, and events are an emerging topic for research in data mining and text mining community [5]. For these purpose state-of-the-art approaches to text classifications are presented in [6], in which three problems were discussed: documents representation, classifier construction and classifier evaluation. So constructing a data structure that can represent the documents, and constructing a classifier that can be used to predicate the class

label of a document with high accuracy, are the key points in text classification. One of the purposes of research is to review the available and known work, so an attempt is made to collect what's known about the documents classification and representation. This paper covers the overview of syntactic and semantic matters, domain ontology, and tokenization concern and focused on the different machine learning techniques for text classification using the existing literature. The motivated perspective of the related research areas of text mining are: Information Extraction (IE) methods is aim to extract specific information from text documents. This is the first approach assumes that text mining essentially corresponds to information extraction.

Information Retrieval (IR) is the finding of documents which contain answers to questions. In order to achieve this goal statistical measures and methods are used for automatic processing of text data and comparison to the given question. Information retrieval in the broader sense deals with the entire range of information processing, from data retrieval to knowledge retrieval [7].Natural Language Processing (NLP) is to achieve a better understanding of natural language by use of computers and represent the documents semantically to improve the classification and informational retrieval process. Semantic analysis is the process of linguistically parsing sentences and paragraphs into key concepts, verbs and proper nouns. Using statistics-backed technology, these words are then compared to the taxonomy. Ontology is the explicit and abstract model representation of already defined finite sets of terms and concepts, involved in knowledge management, knowledge engineering, and intelligent information integration.

1.1 Text Classification Process

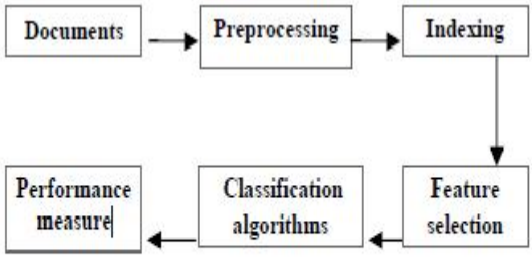


Fig. 1 Document Classification Process

1.1.1 Documents Collection

This is first step of classification process in which we are collecting the different types (format) of document like .html, .pdf, .doc, web content etc.

1.1.2 Pre-Processing

The first step of pre-processing which is used to presents the text documents into clear word format. The documents prepared for next step in text classification are represented by a greatamount of features. Commonly the steps taken are:

- *Tokenization:* A document is treated as a string, and then partitioned into a list of tokens.
- *Removing stop words*: Stop words such as "the", "a", "and", etc. are frequently occurring, so the insignificant words need to be removed.
- *Stemming word:* Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form, e.g. connection to connect, computing to compute.

1.1.3 Indexing

The documents representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector The Perhaps most commonly used document representation is called vector space model (SMART) vector space model, documents are represented by vectors of words. Some of limitations are: high dimensionality of the representation, loss of correlation with adjacent words and loss of semantic relationship that exist among the terms in a document. To overcome these problems, term weighting methods are used to assign appropriate weights to the term.

1.1.4 Feature Selection

After pre-processing and indexing the important step of text classification, is feature selection to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier. The main idea of Feature Selection (FS) is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Because of for text classification a major problem is the high dimensionality of the feature space. Many feature evaluation metrics have been notable among which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index.

1.1.5 Classification

The automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified by three ways, unsupervised, supervised and semi-supervised methods [1]. From last few years, the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Rocchio's.

1.1.6Performance Evaluations

This is Last stage of Text classification, in which the evaluations of text classifiers is typically conducted experimentally, rather than analytically. An important issue of Text categorization is how to measures the performance of the classifiers. Many measures have been used, like Precision and recall, fallout, error, accuracy etc. are given below *Precision wrt ci (Pri)* is defined as the as the probability that if a random document *dx* is classified under *ci*, this decision is correct. Analogously, *Recall wrt ci* (*Rei*) is defined as the conditional that, if a random document *dx* ought to be classified under *ci*, this decision is taken TPi–The number of document correctly assigned to this category.

FN - The number of document incorrectly assigned to this category

FPi - The number of document incorrectly rejected assigned to this category

TNi - The number of document correctly rejected assigned to this category

Fallout = FNi / FNi + TNi

Error = FNi +FPi / TPi + FNi +FPi +TNi

Accuracy = TPi + TNi

2. LITERATURE SURVEY

[1]. Vandana Korde et al (2012) discussed that the text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources which include unstructured and semi structured information. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization, Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the different types of the documents .

[2]. Zakaria Elberrichi, et al (2008) says that in this paper, a new approach is proposed for text categorization based on incorporating background knowledge (WordNet) into text representation with using the multivariate, which consists of extracting the K better features characterizing best the category compared to the others. The experimental results with both Reuters21578 and 20Newsgroups datasets show that incorporating background knowledge in order to capture relationships between words is especially effective in raising the macro-averaged F1 value. The main difficulty is that a word usually has multiple synonyms with somewhat different meanings and it is not easy to automatically find the correct synonyms to use.

[3]. William B. Cavnar et al (2010) says that the N-gram frequency method provides an inexpensive and highly effective way of classifying documents. It does so by using samples of the desired categories rather than resorting to more complicated and costly methods such as natural language parsing or assembling detailed lexicons. Essentially this approach defines a "categorization by example" method. Collecting samples and building profiles can even be handled in a largely automatic way. Also, this system is resistant to various OCR problems, since it depends on the statistical properties of N-gram occurrences and not on any particular occurrence of a word.

[4]. Andrew McCallumzy says that this paper has compared the theory and practice of two different first-order probabilistic classifiers, both of which make the naive Bayes assumption." The multinomial model is found to be almost uniformly better than the multi variate Bernoulli model. In empirical results on five real-world corpora we find that the multinomial model reduces error by an average of 27%, and sometimes by more than 50%.In future work we will investigate the role of document length in classification, looking for correspondence between variations in document length and the comparative performance of multi-variate Bernoulli and multinomial. We will also investigate event models that normalize the word occurrence counts in a document by document length, and work with more complex models that model document length explicitly on a per-class basis. We also plan experiments with varying amounts of training data because we hypothesize that that optimal vocabulary size may change with the size of the training set.

[5]. Fabrizio Sebastiani et al (2010) says that text categorization has evolved, from the neglected research niche it used to be until the late '80s, into a fully blossomed research field which has delivered efficient, effective, and overall workable solutions that have been used in tackling a wide variety of real-world application domains. Key to this success have been (i) the ever-increasing involvement of the machine learning community in text categorization, which has lately resulted in the use of the very latest machine learning technology within text categorization applications, and (ii) the availability of standard benchmarks (such as Reuters-21578 and OHSUMED), which has encouraged research by providing a setting in which different research efforts could be compared to each other, and in which the best methods

and algorithms could stand out. Currently, text categorization research is pointing in several interesting directions. One of them is the attempt at finding better representations for text; while the bag of words model is still the unsurpassed text representation model, researchers have not abandoned the belief that a text must be something more than a mere collection of tokens, and that the quest for models more sophisticated than the bag of words model is still worth pursuing.

[6]. Aurangzeb Khan et al (2010) says that this paper provides a review of machine learning approachesand documents representation techniques. Ananalysis of feature selection methods and classification algorithms were presented. It was verified from the study that information Gain and Chi square statistics are the most commonly used and well performed methods for feature selection; however many other FS methods are choices in pre-processing (stemming, etc.), indexing, dimensionality reduction and classifier parameter values etc. A performance compression in presented a controlled study on a large number of filter feature selection methods for text classification. Over 100 variants of fivemajor feature selection criteria were examined using four well-known classification algorithms: Naive Bayesian (NB) approach, Rocchio's-style classifier, k-NN method and SVM system. Two benchmark collections were chosen as the test beds: Reuters-21578 and small portion ofReuters Corpus Version 1 (RCV1), making the new results comparable to published results.

[7]. RON BEKKERMAN et al (2003) says that text categorization is a fundamental task in Information Retrieval, and much knowledge in this domain has been accumulated in the past 25 years. The "standard" approach to text categorization has so far been using a document representation in a word-based 'input space', i.e. as a vector in some high (or trimmed) dimensional Euclidean space where each dimension corresponds to a word. This method relies on classification algorithms that are trained in a supervised learning manner. Since the early days of text categorization, the theory and practice of classifier design has significantly advanced, and several strong learning algorithms have emerged. In contrast, despite numerous attempts to introduce more sophisticated techniques for document representation, like ones that are based on higher order word statistics or NLP, the simple minded independent word-based representation, known as bag-of words(BOW), remained very popular. Indeed, to-date the best multi-class, multi-labelled categorization results for the well-known Reuters-21578 dataset are based on the BOW representation.

[8]. Karuna Pande Joshi et al (Mar, 1997) says that this paper compares the various algorithms used for Data Mining and was submitted as part of project work for Advanced Algorithms course.

[9]. Fabrizio Sebastiani et al (2005) This paper will outline the fundamental traits of the technologies involved, of the applications that can feasibly be tackled through text classification, and of the tools and resources that are available to the researcher and developer wishing to take up these technologies for deploying real-world applications.

3. TASKS OF TEXT MINING ALGORITHMS

- Text categorization: assigning the documents with pre-defined categories (e.g decision trees induction).
- Text clustering: descriptive activity, which groups similar documents together (e.g. self-organizing maps).
- Concept mining: modelling and discovering of concepts, sometimes combines categorization and clustering approaches with concept/ logic based ideas in order to find concepts and their relations from text collections (e.g. formalconcept analysis approach for building ofconcept hierarchy).
- Information retrieval: retrieving the documents relevant to the user's query.
- Information extraction: question answering.

4. TYPE OF TEXT MINING ALGORITHM

4.1 Classification Algorithm

The Classification problemcan be stated as a training data set consisting of records. Each record is identified by an unique record id, and consist of fields corresponding to the attributes. An attribute with a continuous domain is called a continuous attribute. An attribute with a finite domain of discrete values is called a categorical attribute. One of the categorical attribute is the classifying attribute or class and the value in its domain are called class labels.

4.1.1 Objective:

Classification is the process of discovering a model for the class in terms of the remaining attributes. The objective is to use the training data set to build a model of the class label based on the other attributes such that the model can be used to classify new data not from the training data set attributes. The objective is to use the training data set to build a model of the class label based on the other attributes such that the model can be used to classify new data not from the training data set.

4.1.2 Classification Model

The different type of classification models are as follows:

1. Decision tree.
2. Neural network.
3. Generic algorithm.

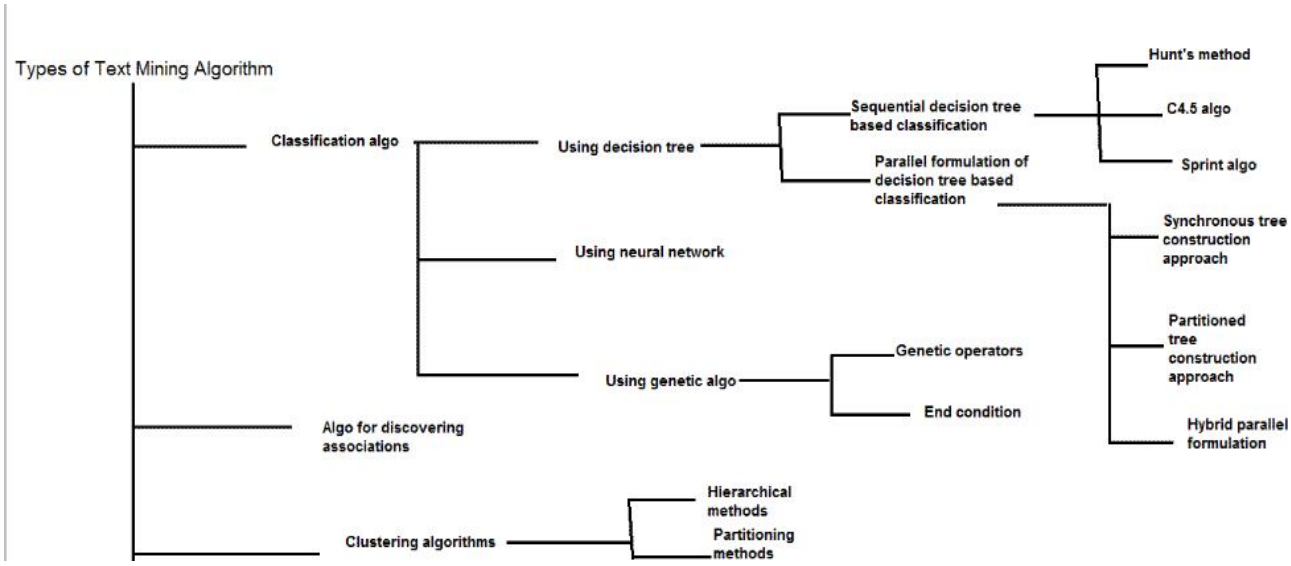
1. Classification using Decision Tree:
- Sequential decision tree based classification
- Parallel formulation of decision tree based classification.

4.1.2.1.1. Sequential Decision Tree based Classification:
A decision tree model consists of internal node and leaves. Each of the internal node has a decision associated with it and each of the leaves has a class label attached to it. A decision tree based classification consists of two steps.
1. Tree induction – A tree is induced from the given training set.
2. Tree pruning – The induced tree is made more concise and robust by removing any statistical dependencies on the specific training data set.

4.1.2.1.1.1. Hunt's method:
The following gives the recursive description of Hunt's method for constructing a decision tree from a set T of training cases with classes' denoted fC1, C2. Ckg.

Case 1 T contains cases all belonging to a single class Cj. The decision tree for T is a leaf identifying class Cj.

Case 2 T contains cases that belong to a mixture of classes. A test is chosen, based on a single attribute that has one or more mutually exclusive outcomes fO1, O2, . . ., Ong. Note that in many implementations, n is chosen to be 2 and this leads to a binary decision tree. T is partitioned into subsets T1, T2, . . . , Tn, where Ti contains all the cases in T that have outcome Oi of the chosen test.The decision tree for T consists of a decision node identifying the test, and one branch for each possible outcome. The same tree building machinery is applied recursively to each subset of training cases.



Case 3 T contains no cases. The decision tree for T is a leaf, but the class to be associated with the leaf must be determined from information other than T.For example, C4.5 chooses this to be the most frequent class at the parent of this node.
Hunt's method works with the training data set. In case 2 of Hunt's method, a test based on a single attribute is chosen for expanding the current node. The choice of an attribute is normally based on the entropy gains of the attributes. The entropy of an attribute is calculated in terms of either entropy [Qui93] or Gini Index[BFOS84]. The best attribute is selected as a test for the node expansion.

4.1.2.1.1.2. C4.5 Algorithm:
The C4.5 algorithm generates a classification-decision tree for the given data-set by recursive partitioning of data. The decision is grown using Depth-firststrategy. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. For each continuous attribute, binary tests involving every distinct values of the attribute are considered. In order to gather the entropy gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted for the values of the continuous attribute and the entropy gains of the binary cut based on each distinct values are calculated in one scan of the sorted data. This process is repeated for each continuous attributes.[8]

4.1.2.1.1.3. SPRINT Algorithm:

In SPRINT, each continuous attribute is maintained in a sorted attribute list. In this list, each entry contains a value of the attribute and its corresponding record id. Once the best attribute to split a node in a classification tree is determined, each attribute list has to be split according to the split decision. A hash table, of the same order as the number of training cases, has the mapping between record ids and where each record belongs according to the split decision. Each entry in the attribute list ismoved to a classification tree node according to the information retrieved by probing the hash table. The sorted order is maintained as the entries are moved in pre-sorted order. Decision trees are usually built in two steps.

First, an initial tree is built till the leaf nodes belong to a single class only. Second, pruning is done to remove any over fitting to the training data. Typically, the time spent on pruning for a large dataset is a small fraction, less than 1% of the initial tree generation.

Advantages are they are inexpensive to construct, easy to interpret, and easy to integrate with the commercial database and they yield better accuracy. Disadvantages are it cannot handle larger data sets that are it suffers from memory limitations and it has low computational speed.

4.1.2.1.2. Parallel Formulation of Decision Tree based Classification

The goal of parallel formulation of decision tree based classification algorithms are scalability in both runtime and memory requirements. The parallel formulation overcome the memory limitation faced by the sequential algorithms, that is it should make it possible to handle larger data setswithout requiring redundant disk I/O. Also parallel formulation offer good speedup over serial algorithm.

Type of parallel formulations for the classification decision tree construction is

- Synchronous Tree Construction Approach
- Partitioned Tree Construction Approach
- Hybrid Parallel Formulation

4.1.2.1.2.1. Synchronous Tree Construction Approach

In this approach, all processors construct a decision tree synchronously by sending and receiving class distribution information of local data. Major steps for the approach are shown below:

1. Select a node to expand according to a decision tree expansion strategy (e.g. Depth-First or Breadth- First), and call that node as the current node. At the beginning, root node is selected as the current node.
2. For each data attribute, collect class distribution information of the local data at the current node.
3. Exchange the local class distribution information using global reduction [KGGK94] among processors.
4. Simultaneously compute the entropy gains of each attribute at each processor and select the best attribute for child node expansion.
5. Depending on the branching factor of the tree desired, create child nodes for the same number of partitions of attribute values, and split training cases accordingly.
6. Repeat above steps (1–5) until no more nodes are available for the expansion.

4.1.2.1.2.2. Partitioned Tree Construction Approach

In this approach, whenever feasible, different processors work on different parts of the classification tree. In particular, if more than one processors cooperate to expand a node, then these processors are partitioned to expand the successors of this node. Consider the case in which a group of processors Pn cooperate to expand node n.The algorithm consists of following steps:

Step 1 Processors in Pn cooperate to expand node n using the method described above. **Step 2** Once the node n is expanded in to successor nodes, n1, n2….nk, then the processor group Pn is also partitioned, and the successor nodes are assigned to processors as follows:

Case 1: If the number of successor nodes is greater than |Pn|,

1. Partition the successor nodes into |Pn| groups such that the total number of training cases corresponding to each node group is roughly equal. Assign each processor to one node group.

2. Shuffle the training data such that each processor has data items that belong to the nodes it is responsible for.

3. Now the expansion of the sub trees rooted at a node group proceeds completely independently at each processor as in the serial algorithm.

Case 2: Otherwise (if the number of successor nodes is less than |Pn|),

1. Assign a subset of processors to each node such that number of processors assigned to a node is proportional to the number of the training cases corresponding to the node.
2. Shuffle the training cases such that each subset of processors has training cases that belong to the nodes it is responsible for.
3. Processor subsets assigned to different nodes develop subtrees independently. Processor subsets that contain only one processor use the sequential algorithm to expand the part of the classification tree rooted at the node assigned to them. Processor subsets that contain more than one processor proceed by following the above steps recursively.

At the beginning, all processors work together to expand the root node of the classification tree. At the end, the whole classification tree is constructed by combining subtrees of each processor.

First (at the top of the figure), all four processors cooperate to expand theroot node just like they do in the synchronous tree construction approach. Next, the set of four processors is partitioned in three parts. The leftmost child is assigned to processors 0 and 1, while the other nodes are assigned to processors 2 and 3, respectively. Now these sets of processors proceed independently to expand these assigned nodes. In particular, processors 2 and processor 3 proceed to expand their part of the tree using the serial algorithm. The group containing processors 0 and 1 splits the leftmost child node into three nodes. These three new nodes are partitioned in two parts the leftmost node is assigned to processor 0, while the other two are assigned to processor 1. From now on, processors 0 and 1 also independently work on their respective sub trees.

Advantages:

- The advantage of this approach is that once a processor becomes solely responsible for a node, it can develop a sub-tree of the classification tree independently without any communication overhead.

Disadvantages:

- The first disadvantage is that it requires data movement after each node expansion until one processor becomes responsible for an entire sub-tree. The communication cost is expensive in the expansion of the upper part of the classification tree
- The second disadvantage is poor loadbalancing inherent in the algorithm. Assignment of nodes to processors is done based on the number of training cases in the successor nodes. However, the number of training cases associated with a node does not necessarily correspond to the amount of work needed to process the subtrees rooted at the node.

4.1.2.1.2.3. Hybrid Parallel Formulation
The hybrid parallel formulation has elements of both schemes. The Synchronous Tree Construction Approach incurs high communication overhead as the frontier gets larger. The Partitioned Tree Construction Approach incurs cost of load balancing after each step. The hybrid scheme keeps continuing with the first approach as long as the communication cost incurred by the firstformulation is not too high. Once this cost becomes high, the processors as well as the current frontier of the classification tree are partitioned into two parts. The description assumes that the number of processors is a power of 2, and that these processors are connected in a hypercube configuration. The algorithm can be appropriately modified if P is not a power of 2. Also this algorithm canbe mapped on to any parallel architecture by simply embedding a virtual hypercube in the architecture.

4.1.2.2. Classification Using Neural Network:
An **artificial neural network**, often just named a **neural network**, is a mathematical model inspired by biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases a neural network is an adaptive system changing its structure during a learning phase. Neural networks are used for modeling complex relationships between inputs and outputs or to find patterns in data.

In Supervised learning, we are given a set ofexample pairs (x,y), $x \in X$, $y \in Y$ and the aim is to find a function f in the allowed class of functions that matches the examples. In other words, we wish to infer the mapping implied by the data. he cost function is related to the mismatch between our mapping and the data and it implicitly contains prior knowledge about the problem domain. Tasks that fall within the paradigm of supervised learning are pattern recognition (also known as classification) and regression (also known as function approximation).The supervised learningparadigm is also applicable to sequential data (e.g., for speech and gesture recognition). With respect to the above specification the following assumptions have been considered.

(1)Multi-Layer Perceptions is the simple feed forward neural network is actually called a Multilayer perception (MLP). An MLP is a network of perceptions. The neurons are placed in layers with outputs always flowing toward the output layer. If only one layer exists, it is called aperception. If multiple layers exist, it is an MLP.

(2) Back Propagation algorithm is learning
Technique that adjusts weights in neural network by propagating weight changes backward from the sink to the source nodes.[12]

Advantages of Neural Network:

- Artificial neural networks make no assumptions about the nature of the distribution of the data and are not therefore, biased in their analysis. Instead of making assumptions about the underlying population, neural networks with at least one middle layer use the data to develop an internal representation of the relationship between the variables.
- Since time-series data are dynamic in nature, it is necessary to have non-linear tools in order to discern relationships among time-series data. Neural networks are best at discovering nonlinear relationships.
- Neural networks perform well with missing or incomplete data. Whereas traditional regression analysis is not adaptive, typically processing all olderdata together with new data, neural networks adapt their weights as new input data becomes available.

Disadvantages of Neural Network:

- No estimation or prediction errors are calculated with an artificial neural network
- Artificial neural networks are “black boxes,” for it is impossible to figure out how relations in hidden layers are estimated.

Tasks of Neural Network:

The tasks to which artificial neural networks are applied tend to fall within the following broad categories:

- Function approximation, or regression analysis, including time series prediction and modeling.
- Classification, including pattern and Sequence.
- Recognition, novelty detection and sequential decision making.
- Data processing, including filtering, clustering, blind source separation and compression.

4.1.2.3. Classification using Genetic Algorithm:
Genetic algorithms are heuristic optimization methods whose mechanisms are analogous to biological evolution .In Genetic Algorithm, the solutions are called individuals or chromosomes. After the initial population is generated randomly, selection and variation function are executed in a loop until some termination criterion is reached. Each run of the loop is called a generation. The selection operator is intended to improve the average quality of the population by giving individuals of higher quality a higher probability to be copied into the next generation. The quality of an individual is measured by a fitness function.

4.1.2.3.1. Genetic Operators
The genetic algorithm uses crossover and mutation operators to generate the offspring of the existing population. Before genetic operators are applied, parents have been selected for evolution to the nextgeneration. The crossover and mutation algorithm is used to produce next generation. The probability of deploying crossover and mutation operators canbe changed by user. In all of next generation, WTSD has used as the fitness function.

4.1.2.3.2. End Condition
GA needs an End Condition to end the generation process. If there is no sufficient improvement in two or more consecutive generations; stop the GA process. In other cases, time limitation can be used as a criterion for ending the process.

4.2 Algorithm for Discovering Associations:
4.2.1 Objective:
In order to discover associations present in the data. The problem was formulated originally in the context of the transaction data at supermarket. This market basket data, as it is popularly known, consists of transactions made by each customer. Each transaction contains items bought by the customer. The goal is to see if occurrence of certain items

in a transaction can be used to deduce occurrence of other items, or in other words, to find associative relationships between items. Traditionally, association modelsare used to discover business trends by analyzing customer transactions. However, they can also be used effectively to predict Web page accesses for personalization. For example, assume that after mining the Web access log, Company X discovered an association rule "A and B implies C," with 80% confidence, where A, B, and C are Web page accesses. If a user has visited pages A and B, there is an 80% chance that he/she will visit page C in the same session. Page C may or may not have a direct link from A or B. This information can be used to create a dynamic link to page C from pages A or B so that the user can "click-through" to page C directly. This kind of information is particularly valuable for a Web server supporting an ecommerce site to link the different product pages dynamically, based on the customer interaction.

4.2.2.2. Sequential Algorithm for finding Association:

The concept of association rules can be generalized and made more useful by observing another fact about transactions. All transactions have a timestamp associated with them; i.e. thetime at which the transaction occurred. If this information can be put to use, one can find relationships such as if a customer bought book today, then he/she is likely to buy a book in a few days' time. The usefulness of this kind of rules gave birth to the problem of discovering sequential patterns or sequential associations. In general, a sequential pattern is a sequence of item-sets with various timing constraints imposed on the occurrences of items appearing in the pattern. Example Consider the instance that A, B, C, D are the set of transactions such that (A) (C, B) (D) encodes a relationship that event D occurs after an event-set (C, B), which in turn occurs after event A.Prediction of events or identification of sequential rules that characterize different parts of the data, are some example applications of sequential patterns. Such patterns are not only important because they represent more powerful and predictive relationships, but they are also important from the algorithmic point of view. Bringing in the sequential relationships increases the combinatorial complexity of the problem enormously. The reason is that, the maximum number of sequences having k events is O(mk2k1), where m is the total number of distinct events in the input data. In contrast, there are only (m C k) size-k item-sets possible while discovering non-sequential associations from m distinct items.

4.3. Clustering Algorithm:

4.3.1. Objective:

Clustering is a division of data into groups of similar objects Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modelling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis.

4.3.2. Clustering Algorithms:

Clustering Algorithms are classified into following two methods:

4.3.2.1. Hierarchical Methods:

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels ofgranularity. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down).

An **agglomerative** clustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters. A **divisive** clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until astopping criterion (frequently, the requestednumber k of clusters) is achieved.

Advantages are

1) Embedded flexibility regarding the level of granularity.
2) Ease of handling of any forms of similarity or distance.
3) Consequently, applicability to any attribute types.

Disadvantages are

1) Vagueness of termination criteria
2) The fact that most hierarchical algorithms do not revisit once constructed (intermediate) clusters with the purpose of their improvement

4.3.2.2 Partitioning Methods:

In data partitioning algorithms, which divide data into several subsets. Since checking all possible subset systems is computationally infeasible, certain greedy heuristics are used in the form of iterative optimization. Specifically, this means differentrelocation schemes that iteratively reassign points between the k clusters. Unlike traditional hierarchical methods, in which clusters are not revisited after being constructed, relocation algorithms gradually improve clusters. With appropriate data, this results in high quality clusters. One approach to data partitioning is to take a conceptual point of view that identifies thecluster with a certain model whose unknownparameters have to be found. More specifically, probabilistic models assume that the data comes from a mixture of several populations whose distributions and priors we want to find. One advantage of probabilistic methods is the interpretability of the constructed clusters. Having concise cluster representation also allows inexpensive computation of intra-clusters measures of it that give rise to a global objective function.

REFERENCES

[1].Vandana Korde et al Text classification and classifiers:" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012".
[2]."Zakaria Elberrichi,Abdelattif Rahmoun, and Mohamed Amine Bentaalah""Using WordNet for Text Categorization""The International Arab Journal of Information Technology, Vol. 5, No. 1, January 2008".
[3]."William B. Cavnar and John M. Trenkle""N-Gram-Based Text Categorization""vol.5 IJCSS 2010"
[4]."Andrew McCallumzy and Kamal Nigamy""A Comparison of Event Models for Naive Bayes Text Classification".
[5]."FabrizioSebastian""Text Categorization""Vol 5 2010".
[6]."Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan""A Review of Machine Learning Algorithms for Text-Documents Classification""JOURNAL OF ADVANCES IN INFORMATION TECHNOLOGY, VOL. 1, NO. 1, FEBRUARY 2010".
[7] "RON KERMAN "Distributional Clustering Categorization Haifa Jan 2003."

AUTHORS

Bhumika

She is pursuing M.Tech in the Dept. of Information Technology, Guru Nanak Dev Engineering College Ludhiana, India. She received Diploma from Govt. Polytechnic College, Kandaghat and B.Tech from KC College of Engg. & IT. She had actively participated in various national seminars. Her Field of interest is Database Management System, Data Mining, Programming Languages etc. She is currently working at KC Polytechnic College, Nawanshahr.

Sukhjit Sehra

He is a professor working in the Dept. of Computer Science Engineering. Guru Nanak Dev Engineering College Ludhiana, India. He has published various international and national papers on Data Mining.

Anand Nayyar

Anand Nayyar (B.Com, MCA, M.Phil, M.Tech, MBA). He is a certified professional in various International Certifications like A+, CCNA, MCSE, MCTS, MCITP, RHCE, CEH, OCP, AutoCad, Project Management, Google Certified Searcher and many more. He has published more than 180 Research Papers in various National and International Conferences cum Journals with Impact Factor. He has published 10 Books on various subjects of Computer Science. He is a member of more than 100 Journals as Editorial Board cum Review Board. He has been awarded with various prestigious recognitions like Shiksha Ratan Puraskar, Best Citizen of India, Best Reviewer, Researcher and Outstanding Teacher. He is member of more than 40 International and National Research Associations. He is currently working as Assistant Professor in Department of Computer Applications & IT at KCL Institute of Management and Technology, Jalandhar, Punjab.