RNI Regd No: APENG/2011/47294 ISSN No: 2320-3706(PRINT)
International Journal & Magazine of Engineering, Technology, Management and Research
YUVAENGINEERS
Transforming Young Engineers for Better Tomorrow

Feature Extraction for Text Classification Methodology Based on Frequent term Measures

S.Venkata Ramana
Assistant Professor,
Department of Computer Science,
Malla Reddy Engineering College for Women,
Maisammaguda, Hyderabad.

Shabanaunnisa Begum
Assistant Professor,
Department of Computer Science,
Malla Reddy Engineering College for Women,
Maisammaguda, Hyderabad.

Abstract:

This system is intended to show the things occurred in between the searches happened in the place of client, and server. The users clearly know about the process of how to sending a request for the particular thing, and how to get a response for that request or how the system shows the results explicitly. But no one knows about the internal process of searching records from a large database. This system clearly shows how an internal process of the searching process works. In text classification, the dimensionality of the feature vector is usually huge to manage. The Problems need to be handled are as follows: a) the current problem of the existing feature clustering methods b) The desired number of extracted features has to be specified in advance c) When calculating similarities, the variance of the underlying cluster is not considered d) How to reduce the dimensionality of feature vectors for text classification and run faster?. These Problems are handled by means of applying the K-Means Algorithm in hands with Association Rule Mining.

Keywords:

Document clustering, Clustering algorithm, Frequent Concepts based Clustering.

INTRODUCTION:

With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches.

The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high.

However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality.Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods, by using a filter method to reduce search space that will be considered by the subsequent wrapper.

They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. The wrapper methods are computationally expensive and tend to overt on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large.

With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms. Pereira Baker and Dillon employed the distributional clustering of words to reduce the dimensionality of text data. In cluster analysis, graph-theoretic methods have been well studied and used in many applications.

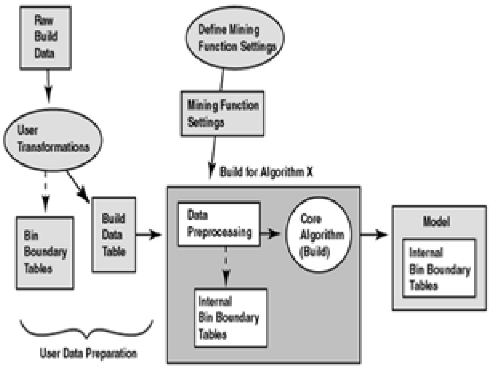


Fig:- Data Mining – Hybrid Methods

Their results have, sometimes, the best agreement with human performance. The general graph-theoretic clustering is simple: Compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter (according to some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. In our study, we apply graph theoretic clustering methods to features. In particular, we adopt the minimum spanning tree (MST) based clustering algorithms, because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice. Based on the MST method, we propose a Fast clustering-based feature Selection algorithm (FAST).

The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features. Features in different clusters are relatively independent, the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. The experimental results show that, compared with other five different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of the four well-known different types of classifiers. good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other.

For all the above mentioned terms this system is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis, and iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked. Different from these algorithms, our proposed FAST algorithm employs clustering based method to choose features. Several algorithms which illustrates how to maintain the data into the database and how to retreive it faster, but the problem here is no one cares about the database maintenance with ease manner and safe methodology. The systems like Distortion and Blocking algorithm, which creates an individual area for each and every word from the already selected transactional database, those are collectively called as dataset, which will be suitable for a set of particular words, but it will be problematic for the set of records, once the user get confused then they can never get the data back. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

A FAST algorithm research has focused on searching for relevant features. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

LITERATURE SURVEY:

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.
The major part of the project development sector considers and fully survey all the required needs for developing the project.

For every project Literature survey is the most important sector in software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

The Text classification contains many problems, which has been widely studied in the data mining, machine learning, database, and information retrieval communities with applications in a number of diverse domains, such as target marketing, medical diagnosis, news group filtering, and document organization. The text classification technique assumes categorical values for the labels, though it is also possible to use continuous values as labels. The latter is referred to as the regression modeling problem. The problem of text classification is closely related to that of classification of records with set-valued features; however, this model assumes that only information about the presence or absence of words is used in a document. In reality, the frequency of words also plays a helpful role in the classification process, and the typical domain-size of text data (the entire size) is much greater than a typical set-valued classification problem.

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Rakish Agawam introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {mathrm{onions, potatoes}} Rightarrow{mathrm{burger}} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection, Continuous production and bioinformatics. As opposed to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

FAST Algorithm is a classic algorithm for frequent item set mining and association rule learning over transactional databases. This FAST algorithm inbuiltly contains an algorithm called Apriori, which proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction.

The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or datapoints).

Feature selection techniques provide three main benefits when constructing predictive models:

•Improved model interpretability,
•Shorter training times,
•Enhanced generalisation by reducing overfitting.

Feature selection is also useful as part of the data analysis process, as shows which features are important for prediction, and how these features are related.

EXISTING SYSTEM:

In the past approach there are several algorithm which illustrates how to maintain the data into the database and how to retreive it faster, but the problem here is no one cares about the database maintenance with ease manner and safe methodology. A Distortion algorithm, which creates an individual area for each and every word from the already selected transactional database, those are collectively called as dataset, which will be suitable for a set of particular words, but it will be problematic for the set of records. A Blocking algorithm make propagation to the above problem, and reduce the problems occurred in the existing distortion algorithm, but here also having the problem called data overflow, once the user get confused then they can never get the data back.

The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods.

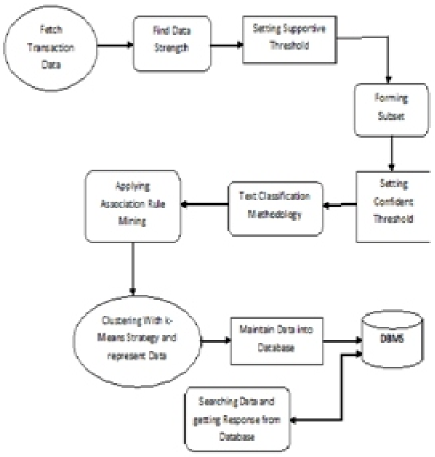
Disadvantages:

•Lacks speed
•Security Issues
•Performance Related Issues
•The generality of the selected features is limited and the computational complexity is large.
•Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

So the focus of our new system is to enhance the throughput for any basis to eliminate the data security lacks therein and make a newer system prominent handler for handling data in an efficient manner.

PROPOSED SYSTEM:

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features.


Fetch Transaction Data
Find Data Strength
Setting Supportive Threshold
Forming Subset
Setting Confident Threshold
Text Classification Methodology
Applying Association Rule Mining
Clustering With k-Means Strategy and represent Data
Maintain Data into Database
DBMS
Searching Data and getting Response from Database

Our proposed FAST algorithm falls into the second group. Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function.

However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

Advantages:

1. Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with each other.
2.The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.
In our proposed FAST algorithm, it involves
(i)the construction of the minimum spanning tree (MST) from a weighted complete graph;
(ii)the partitioning of the MST into a forest with each tree representing a cluster;
(iii)the selection of representative features from the clusters.

IMPLEMENTATION

User Module:

Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first. Once the user wants to access the system, the user needs to prove the originality and permission rights regarding to access the provisions present into the system. So that the user has to initially submit the original identity with proper username and password for accessing the features in the system. After submitting the details the user can easily access any of the features present into the system, but the user need to keep one thing in mind carefully, if any one of the details found to be wrong at any circumstances the administrator can block the access rights to the user without any intimation.

Distributed Clustering:

The Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Pereira et al. or on the distribution of class labels associated with each word by Baker and McCallum . As distributional clustering of words are agglomerative in nature, and result in suboptimal word clusters and high computational cost, proposed a new information-theoretic divisive algorithm for word clustering and applied it to text classification. Proposed to cluster features using a special metric of distance, and then makes use of the of the resulting cluster hierarchy to choose the most relevant attributes.

Unfortunately, the cluster evaluation measure based on distance does not identify a feature subset that allows the classifiers to improve their original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower. Clustering spatially distributed data is well motivated and especially challenging when communication to a central processing unit is discouraged, e.g., due to power constraints. The former procedures suffers from the fact that only a small subset of the system are responsible for relaying the messages, and thus cause rapid consumption of the energy of these system.

The latter procedures use the residual energy of each block in order to decide about whether it will elect itself as a leader of a cluster or not. Distributed clustering schemes are developed in this paper for both deterministic and probabilistic approaches to unsupervised learning. The centralized problem is solved in a distributed fashion by recasting it to a set of smaller local clustering problems with consensus constraints on the cluster parameters. The resulting iterative schemes do not exchange local data among nodes, and rely only on single-hop communications.

Subset Selection Algorithm:

The Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. Keeping these in mind, we develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

Subset selection is a method for selecting a subset of columns from a real matrix, so that the subset represents the entire matrix well and is far from being rank deficient. This system begins by extending a deterministic subset selection algorithm to matrices that have more columns than rows. Then investigate a two-stage subset selection algorithm that utilizes a randomized stage to pick a smaller number of candidate columns, which is forwarded for to the deterministic stage for subset selection.

After this the approach performs extensive numerical experiments to compare the accuracy of this algorithm with the best known deterministic algorithm. And also introduce an iterative algorithm that systematically determines the number of candidate columns picked in the randomized stage, and provides a recommendation for a specific value.

Association Rule Mining:

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Rakesh Agrawal introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {mathrm {onions, potatoes}} Right arrow{mathrm{burger}} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection, Continuous production and bioinformatics.

As opposed to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions. Association rule mining is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Based on the concept of strong association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in real world applications like supermarkets, stores and etc.

Text Classification Process:

The Text classification contains many problems, which has been widely studied in the data mining, machine learning, database, and information retrieval communities with applications in a number of diverse domains, such as target marketing, medical diagnosis, news group filtering, and document organization. The text classification technique assumes categorical values for the labels, though it is also possible to use continuous values as labels. The latter is referred to as the regression modeling problem. The problem of text classification is closely related to that of classification of records with set-valued features; however, this model assumes that only information about the presence or absence of words is used in a document. In general, text classification includes topic based text classification and text genre-based classification. Topic-based text categorization classifies documents according to their topics.

Texts can also be written in many genres, for instance: scientific articles, news reports, movie reviews, and advertisements. Genre is defined on the way a text was created, the way it was edited, the register of language it uses, and the kind of audience to whom it is addressed. Previous work on genre classification recognized that this task differs from topic-based categorization. Typically, most data for genre classification are collected from the web, through newsgroups, bulletin boards, and broadcast or printed news. They are multi-source, and consequently have different formats, different preferred vocabularies and often significantly different writing styles even for documents within one genre. Namely, the data are heterogenous. In reality, the frequency of words also plays a helpful role in the classification process, and the typical domain-size of text data (the entire size) is much greater than a typical set-valued classification problem.

Feature Selection Algorithm:

Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or datapoints). Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. For example, physician may make a decision based on the selected features whether a dangerous surgery is necessary for treatment or not. Feature selection in supervised learning has been well studied, where the main goal is to find a feature subset that produces higher classification accuracy.

Time Complexity:

The major amount of work for FAST Algorithm involves the computation of Subset values for items relevance and Correlation, which has linear complexity in terms of the number of instances in a given data set. The time complexity of an algorithm quantifies the amount of time taken by an algorithm to run as a function of the length of the string representing the input. The time complexity of an algorithm is commonly expressed using big O notation, which excludes coefficients and lower order terms. When expressed this way, the time complexity is said to be described asymptotically, i.e., as the input size goes to infinity. For example, if the time required by an algorithm on all inputs of size n is at most 5n3 + 3n, the asymptotic time complexity is O(n3). Time complexity is commonly estimated by counting the number of elementary operations performed by the algorithm, where an elementary operation takes a fixed amount of time to perform. Thus the amount of time taken and the number of elementary operations performed by the algorithm differ by at most a constant factor. Since an algorithm's performance time may vary with different inputs of the same size, one commonly uses the worst-case time complexity of an algorithm, denoted as T(n), which is defined as the maximum amount of time taken on any input of size n. Time complexities are classified by the nature of the function T(n). For instance, an algorithm with T(n) = O(n) is called a linear time algorithm, and an algorithm with T(n) = O(2n) is said to be an exponential time algorithm. The first part of the algorithm has a linear time complexity in terms of the number of features m. Assuming features are selected as relevant ones in the first part, when k ¼ only one feature is selected.

CONCLUSION:

The area of document clustering has many issues which need to be solved. In this work, few issues e.g. high dimensionality and accuracy are focused but there are still many issues that can be taken into consideration for further research which are as follows: 1. The proposed algorithm can be modified to soft clustering. 2. Efficiency of the proposed work can be improved by adding more issues. 3. Each concept represents a topic enclosed in the document. This fact could be used to generate titles for a document or a group of document by post processing the set of concepts assigned to a document.

Future work:

In future we apply the algorithms called association rule mining with FAST process to achieve greater performance of header based data identity. The application of these algorithms results the following strategies (i) removing irrelevant features from generated feature itemset, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features. The informative rule set will generates significant features for a given database that can be generated more efficiently than the other approaches like Subset formation and c-Means applications. The efficiency improvement results from that the generation of the informative rule set needs fewer candidates and database accesses than that of the association rule set rather than large memory usage. So that the future implementation will definitely results higher end data maintenance and better performance to the users.

REFERENCES:

[1] J. Han and M. Kimber. 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.

[2] Jain, A.K, Murty, M.N., and Flynn P.J. 1999. Data clustering: a review. ACM Computing Surveys, pp. 31, 3, 264-323.

[3] M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. KDD Workshop on Text Mining‟00.

[4] P. Berkhin. 2004. Survey of clustering data mining techniques [Online]. Available: http://www.accrue.com/products/rp_cluster_review.pdf.

[5] Xu Rui. 2005. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 16(3):pp. 634-678.

[6] Miller G. 1995. Wordnet: A lexical database for English. CACM, 38(11), pp. 39–41.

[7] L. Zhuang, and H. Dai. 2004. A Maximal Frequent Itemset Approach for Document Clustering. Computer and Information Technology, CIT. The Fourth International Conference, pp. 970 – 977.

[8] R. C. Dubes and A. K. Jain. 1998. Algorithms for Clustering Data. Prentice Hall college Div, Englewood Cliffs, NJ, March.

[9] D. Koller and M. Sahami. 1997. Hierarchically classifying documents using very few words. In Proceedings of (ICML) 97, 14th International Conference on Machine Learning, pp. 170–178, Nashville, US.

[10] B.C.M.Fung, K.Wan, M.Ester. 2003. Hierarchical Document Clustering Using Frequent Itemsets", SDM‟03.

[11] Green, S. J. 1999. Building hypertext links by computing semantic similarity. T KDE, 11(5), pp. 50–57.

[12] Sedding, J., & Kazakov, D. 2004. Wordnet-based text document clustering. 3rd Workshop on Robust Methods in Analysis of Natural Language Data, pp. 104–113.

[13] Y. LI, and S.M. Chung. 2005. Text Document Clustering Based on Frequent Word Sequences. In Proceedings of the. CIKM, 2005. Bremen, Germany, October 31-November 5.