Feature Subset Selection with Fast Algorithm Implementation

K.Suman[#1], S.Thirumagal[#2]

[#1]M.E, Computer Science and Engg, Chendhuran College of Engg and Tech, Anna University, Tamilnadu, India

[#2] Assistant Professor, Department of IT, Chendhuran College of Engg and Tech, Anna University, Tamilnadu, India

Abstract-- The Title "Feature Subset Selection with FAST Algorithm Implementation" has 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 FAST Algorithm in hands with Association Rule Mining.

Index Terms— FAST algorithm, Feature selection, Redundant data, Text classification, Clustering, Rule mining.

1. INTRODUCTION

In machine learning and statistics, feature selection also known as variable selection, attribute selection or variable subset selection. It 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 data points. The archetypal case is the use of feature selection in analysing DNA microarrays, where there are many thousands of features, and a few tens to hundreds of samples. Feature selection techniques provide three main benefits when constructing predictive models: improved model interpretability, shorter training times, and enhanced by reducing over fitting.

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods.

Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model.

Filter methods use a proxy measure instead of the error rate to score a feature subset. This measure is chosen to be fast to compute, whilst still capturing the usefulness of the feature set. Common measures include the Mutual Information, Pearson product-moment correlation coefficient, and inter/intra class distance. Filters are usually less computationally intensive than wrappers, but they produce a feature set which is not tuned to a specific type of predictive model. Many filters provide a feature ranking rather than an explicit best feature subset, and the cutoff point in the ranking is chosen via cross-validation.

Embedded methods are a catch-all group of techniques which perform feature selection as part of the model construction process. One other popular approach is the Recursive Feature Elimination algorithm, commonly used with Support Vector Machines to repeatedly construct a model and remove features with low weights. These approaches tend to be between filters and wrappers in terms of computational complexity.

Subset selection evaluates a subset of features as a group for suitability. Subset selection algorithms can be broken into Wrappers, Filters and Embedded. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset.

Wrappers can be computationally expensive and have a risk of over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in and specific to a model.

The Fast clustering Based Feature Selection algorithm (FAST) works into 2 steps. In the first step, features are divided into clusters by using graph theoretic clustering methods. In the second step, the most representative feature i.e., strongly related to target class is selected from each cluster to form the final subset of features. A feature in different clusters in relatively independent, the clustering based strategy of FAST has 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 four well known different types of classifiers. A good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other. Different from these algorithms, our proposed FAST algorithm employs clustering based method to features.

This paper discusses about feature selection, FAST algorithm, Text classification and so on. The next section is literature review.

2. LITERATURE REVIEW

This section literature review has provides an overview and a critical evaluation of a body of literature relating to a research problem. Literature review is the most important step in software development process. This paper presents [1], a temporal association rule mining approach named T-patterns, applied on highly challenging floating train data. The aim is to discover temporal associations between pairs of time stamped alarms, called events, which can predict the occurrence of severe failures within a complex bursty environment. The main advantage is Association Rule Mining Produces efficient cluster wise data maintenance and main drawback of this paper is cluster set is too large to handle. In this paper [2], an algorithm for mining highly coherent rules that takes the properties of propositional logic into consideration is proposed. The derived association rules may thus be more thoughtful and reliable. The advantage is reducing the Cluster set by applying Feature subset manipulations and drawback is resulting probability is low. In this paper [3], the basic objective of feature subset selection is to reduce the dimensionality of the problem while retaining the most discriminatory information necessary for accurate classification. Thus it is necessary to evaluate feature subsets for their ability to discriminate different classes of pattern. Now the fact that “two best features do not comprise the best feature subset of two features” demands evaluation of all possible subset of features to find out the best feature subset. If the number of features increases, the number of possible feature subsets grows exponentially leading to a combinatorial optimization problem. The advantage of this paper is optimal subset is created, so less time is required for manipulation and drawbacks are data security is an issue, high possibility for data loss. In the instance of this paper [4], we present here a new population-based feature selection method that utilizes dependency between features to guide the search. For the particular problem of feature selection, population-based methods aim to produce better”or fitter” future generations that contain more informative subsets of features. It is well-known that feature subset selection is a very challenging optimization problem, especially when dealing with datasets that contain large number of features. The main advantage of this paper is applying subset classification algorithm to simplify data and drawback is, it requires more time to manipulate data. In this paper [5], we propose a new hybrid ant colony optimization (ACO) algorithm for feature selection (FS), called ACOFS, using a neural network. A key aspect of this algorithm is the selection of a subset of salient features of reduced size. ACOFS uses a hybrid search technique that combines the advantages of wrapper and filter approaches. We evaluate the performance of ACOFS on eight benchmark classification datasets and one gene expression dataset, which have dimensions varying from 9 to 2000. The advantage along with this paper is hybrid methodology is applied to optimize data and other drawback is inefficient for large data sets. According to this paper which presents [6], a feature subset selection algorithm based on branch and bound techniques are developed to select the best subset of m features from an n-feature set. Existing procedures for feature subset selection, such as sequential selection and dynamic programming, do not guarantee optimality of the selected feature subset. Exhaustive search, on the other hand, is generally computationally unfeasible. The present algorithm is very efficient and it selects the best subset without exhaustive search. The advantage is, pointers are maintained to reach data efficiently and drawback which among inner looping produces lots and lots of confusions for large cluster.

3. PROPOSED APPROACH

In the Proposed Approach, the feature subset selection algorithm the data is to be viewed as the process of identifying and eliminating as many immaterial and unneeded features as probable. This is because inappropriate features do not supply to the projecting accuracy and unneeded features do not redound to getting a better predictor for that they provide mostly information which is already present in other features. In past systems users can store the data and retrieve the data from server without any knowledge about how it is maintained and how it gets processed, but in the present system 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; this is the present advantage in our system. In text classification, the dimensionality of the feature vector is usually huge to manage. In the present system we eliminate the redundant data finding the representative data, reduce the dimensions of the feature sets, find the best set of vectors which best separate the patterns and two ways of doing feature reduction, feature selection and feature extraction. Good feature subsets contain features highly correlated with predictive of the class, yet uncorrelated with each other. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. In future the same system can be

developed with the help of mobile applications to support the same procedure in mobile environments so that the users can easily manipulate their data in any place.

4. SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

Generally algorithms shows a result for exploring a single thing that is either be a performance, or speed, or accuracy, and so on. But here is a requirement of proving the above mentioned all in a single algorithm. The system has to start with a concept of collecting transactional database, and process the step with the rule of association, further implementation proceeds with an implementation of rule mining; then finally end with the process of sensitive rule mining. For all our goal is to apply various association rules to generate quicker results that may enhance algorithm process.

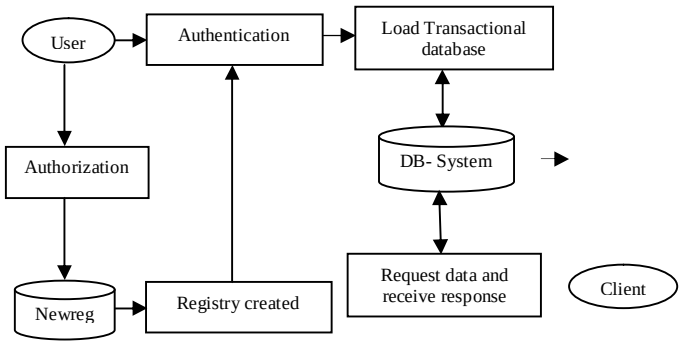


Fig.a.System Architecture

4.2 MODULES

A. User Module
B. Circulated Clustering
C. Text Detachment
D. Association Rule Mining
E. Text Organization
F. Data Representation

A. USER MODULE

In this 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.

B. CIRCULATED CLUSTERING

Cluster is nothing but it is a combination of various features including text subsets, it has been used to cluster words into groups based either on their contribution in particular grammatical relations with other words. Here the distributed clustering focuses on the cluster with various text subsets. In this module the system can manage the cluster with various classifications of data.

C. TEXT DETACHMENT

The text detachment is the filtration process, which filters the combination of various subsets present in the cluster into matching clusters belonging to the text head with the help of k-Means algorithm. A novel algorithm which can efficiently and effectively deal with both inappropriate and superfluous features, and acquire a good feature subset.

.

D. ASSOCIATION RULE MINING

Association rule mining is the best method for discovering interesting relations between variables in large databases or data warehouse. It is intended to identify strong rules discovered in databases using different measures of interestingness. With the help of this rule mining the system can manipulate and associates the text cluster into the respective heads based on the internal features of data.

E. TEXT ORGANIZATION

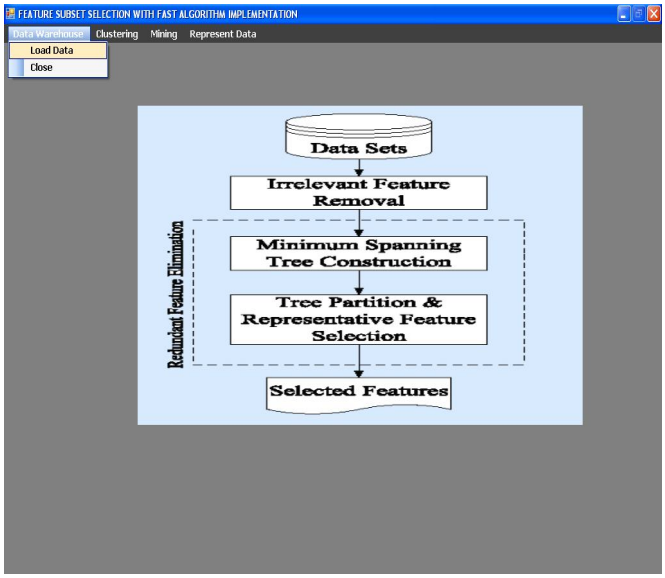
The Text organization contains many problems, which has been widely studied in the data mining, machine learning, database, and information retrieval communities with applications in a number of diverse domains, such as target marketing, medical diagnosis, news group filtering, and document organization. The text organization technique assumes categorical values for the labels, though it is possible to use continuous values as labels. The latter is referred to as the regression modeling problem. The problem of text organization 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 typical domain size of text data (the entire size) is much greater than a typical set valued classification problem.

F. DATA REPRESENTATION

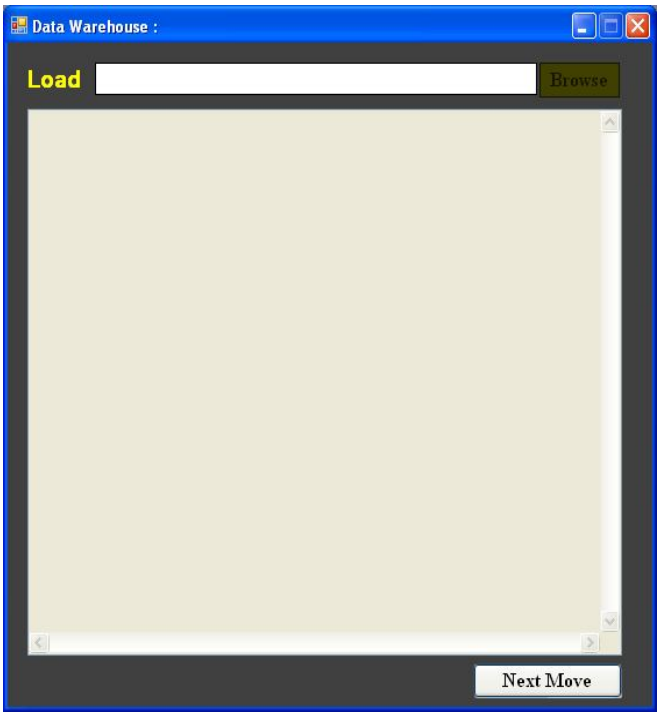
With the help of text classification and association rule mining the cluster is assembled with proper subset and correct header representations, in this stage the system can easily find out the text representation with maximum threshold value.

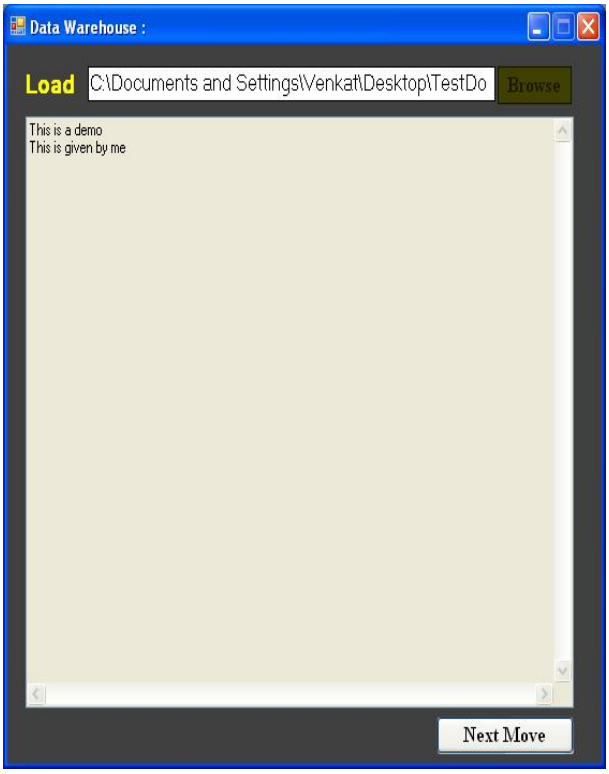
4.3 SCREEN SHOTS

A. Main form

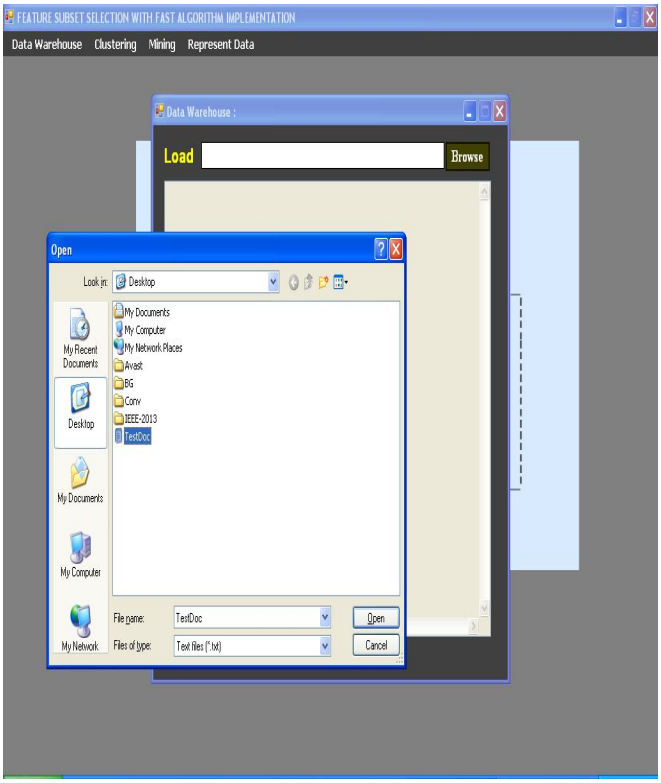


B. Load Transactional Data

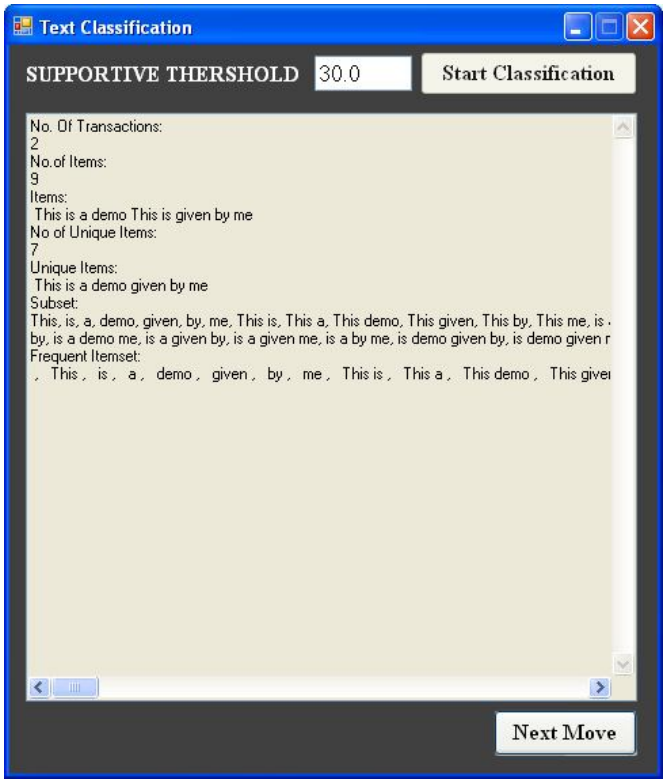




C. Select the Data



D. Text Classification



5. CONCLUSION AND FUTURE ENHANCEMENT

For the entire Fast algorithm in hands with association rule implementation gives flexible results to users, like removing irrelevant features from the Original Subset, and constructing a minimum spanning tree from the relative subset whatever present in the data store. By partitioning the minimum spanning tree we can easily identify the text representation from the features. Association Rule Mining gives ultimate data set with header representation as well as FAST algorithm with applied K-Means strategy provides efficient data management and faster performance. The revealing regulation set is significantly smaller than the association rule set, in particular when the minimum support is small. The proposed work has characterized the associations between the revealing regulation set and the non-redundant association rule set, and discovered that the enlightening regulation set is a subset of the non-redundant association rule set.

ACKNOWLEDGMENT

We would like to sincerely thank Assistant Prof. S.Thirumagal for his advice and guidance at the start of this article. His guidance has also been essential during some steps of this article and his quick invaluable insights have always been very helpful. His hard working and passion for research also has set an example that we would like to follow. We really appreciate his interest and enthusiasm during this article. Finally we thank the Editor in chief, the Associate Editor and anonymous Referees for their comments.

REFERENCES

[1] Sammouri W., "Temporal association rule mining for the preventive diagnosis of onboard subsystems within floating train data framework", pp 1351-1356, June 2012.

[2] Chun-Hao Chen., "A High Coherent Association Rule Mining Algorithm", In Proceedings of the IEEE international Conference on Technologies and applications of artificial intelligence, pp 1-4, Nov 2012.

[3] Chakraborty.B.., "Bio-inspired algorithms for optimal feature subset selection" In Proceedings of the Fifth IEEE international Conference on Computers and Devices for Communication , pp 1-7, Dec 2012..

[4] Ahmed Al-Ani, Rami N. Khushaba., "A Population Based Feature Subset Selection Algorithm Guided by Fuzzy Feature Dependency", Volume 322, pp 430-438, Dec 2012.

[5] Md. Monirul Kabir., "A new hybrid ant colony optimization algorithm for feature selection", Volume 39, issue 3, pp 3747-3763, Feb 2012.

[6] P.M. Narendra, K. Fukunaga, "A Branch and Bound Algorithm for Feature Subset Selection", Volume C-26, issue 9, pp 917-922, Aug 2006.

AUTHORS PROFILE

K.Suman is currently a PG scholar in Computer Science Engineering from the Department of Computer Science and Engineering at Chendhuran College of Engineering and Technology, Pudukkottai. He received his Bachelor Degree in Information Technology from Sudharsan Engineering College, Pudukkottai and Tamilnadu. His Research areas include Data mining, grid computing and wireless sensor networks.

S.Thirumagal is currently working as an Asst. Prof. from the Department of Information Technology at Chendhuran College of Engineering and Technology, Pudukkottai. She received his Bachelor Degree from Shanmuganathan Engineering College, Pudukkottai. She received his master degree from Infant Jesus college of Engineering, Thoothukudi and Tamilnadu. She Published 1 National Conference and 1 International Journal. Her main research interests lie in the area of Data mining and Data warehousing.