Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 7, July 2016, pg.234 – 238

FAST Clustering Based Feature Subset Selection Algorithm for High Dimensional Data

Vishnu Tore[1], Prof. P.M.Chawan[2]

[1]Department of Computer and IT, VJTI, India

[2]Department of Computer and IT, VJTI, India

[1] vishnutore319@gmail.com; [2] pmchawan@vjti.org.in

Abstract— Data mining is the extraction of hidden predictive results from large databases. Feature selection is defined as process of selecting subset of relevant feature. The method for using a feature selection technique is that data may contain many redundant or irrelevant features. 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. Our proposed FAST algorithm falls into the second group.
In early days, feature subset selection research has focused on searching for relevant features. This paper provides the implementation of this algorithm on high dimensional data.

Keywords- Data mining, Feature selection, FAST algorithm, relevant features, redundant features

1. INTRODUCTION

1.1 DATA MINING

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

Till now there are different feature subset selections algorithms have been proposed for machine learning applications. They can be categorized into four categories: the Embedded, Filter, Wrapper, and Hybrid approaches. The wrapper method used to determine the quality of the selected subsets, accuracy of this algorithm is ordinarily high. But the computational complexity is more. To reduce dimensionality and to obtained quality features FAST algorithm is used by removing irrelevant data with most suitable way. To obtain the goal of FAST clustering algorithm, it works in two parts. In the first part, it divides the input dataset into clusters using graph methods. For this purpose Minimum

Spanning Method (MST) is used. In second part, the subsets that are having great accuracy or relative with the required search are selected from cluster and form a feature subset. Feature subset identifies and removes as many irrelevant and redundant features as possible. All the related feature set and unrelated feature set are clustered in different sets. This technique is useful in taking take care of related and unrelated feature sets for prediction as per the requirement of target.

1.2 Feature Selection

Feature 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 data points) [2].
Research on feature selection has been done for last several decades and is still in focus. Reviews and books on feature selection can be found in [3, 4, and 5]. Recent papers such as [6, 7, 8, 9, 10] address some of the existing issues of feature selection. Feature subset selection is an effectual way for dimensionality reduction, elimination of inappropriate data, rising learning accurateness, and recovering result unambiguousness. Numerous feature subset selection methods have been planned and considered for machine learning applications. They can be separated into four major categories such as: the Wrapper, Embedded, and Filter and Hybrid methods. In particular, we accept the minimum spanning tree based clustering algorithms, for the reason that they do not imagine that data points are clustered around centers or separated by means of a normal geometric curve and have been extensively used in tradition.

2. RELATED WORK

The following papers explain different views regarding the feature selection and its techniques. Also they elaborate the different feature selection methods.

Qinbao Song, Jingjie Ni and Guangtao Wang [1] presented paper stating that FAST clustering algorithm is more powerful and efficient related to quality and time. The paper also explained about Wrapper algorithm, Filter algorithm, Hybrid techniques and its drawbacks.

L. Yu and H. Liu [2] proposed an approach that data mining for high dimensional data is big problem. This is focused on removing irrelevant data. It uses correlation-based approach for feature selection. Which helps to reduce features which are having zero linear correlation and reduce redundancy amongst selected features.

M. Dash, H. Liu, and H. Motoda [3] these people have major concentration on consistency measure for feature selection. This paper have elaborated the consistency with properties and greatness of this with other present techniques for selection of features.

In [4] the paper contains main idea of working of FAST clustering based feature subset algorithm. It also cover idea that if more than one feature are joint and if matching with target feature then declare it as relevant. It also elaborates the distributed clustering and time complexity of Prim's algorithm.

In [5] Dingcheng Feng, Feng Chen_, and Wenli Xu have explained how to calculate optimized feature set on the basis of classification and clustering. In this paper the drawbacks of backward greedy elimination algorithm by leave one out strategy

3. EXISTING METHOD

The embedded techniques 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. Old machine learning algorithms like decision trees or artificial neural networks are examples of embedded methods. 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 generally high. However, the generality of the selected features is limited and the computational complexity is large. The next is filter methods which are independent of learning algorithms, with good generality. Their computational complexity is lower than previous one, but the accuracy of the learning algorithms is not guaranteed. The last methods known as 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.

4. PROPOSED METHOOD

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. 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.

4.1. 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.

4.2. 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. 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.

4.3. 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.

4.4. TIME COMPLEXITY

The major amount of work for this algorithm involves the computation of SU values for TR relevance and F-Correlation, which has linear complexity in terms of the number of instances in a given data set. 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.

4.5 FLOW CHART

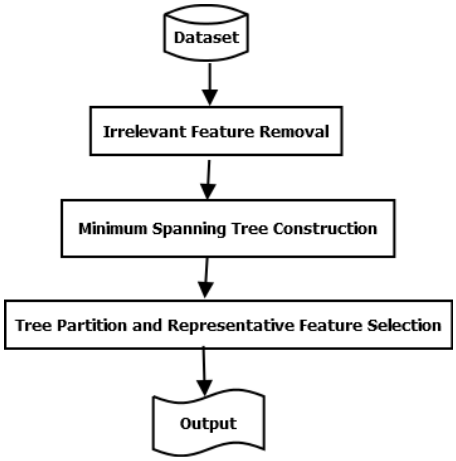
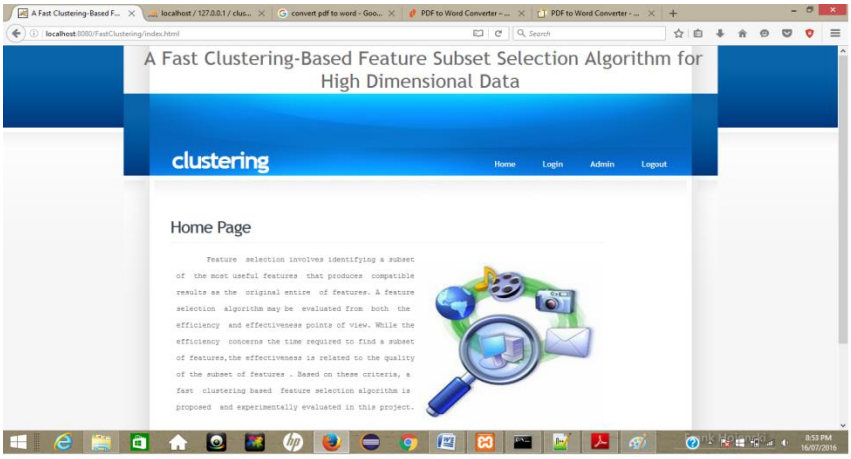
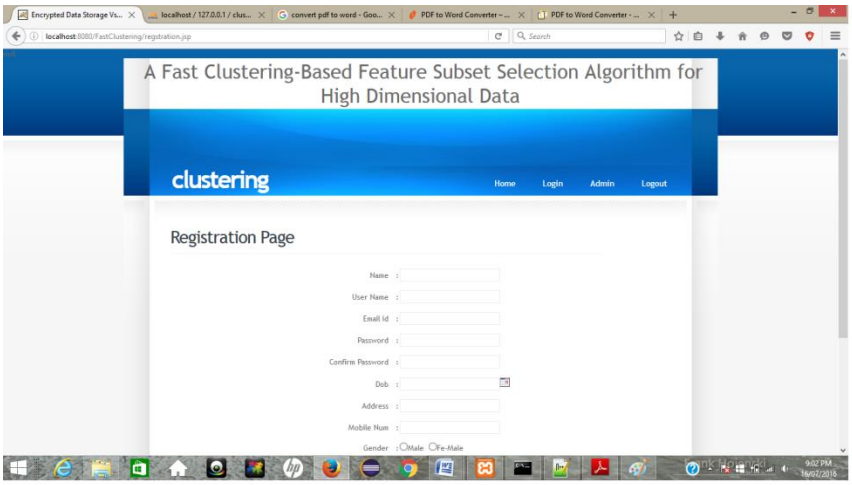
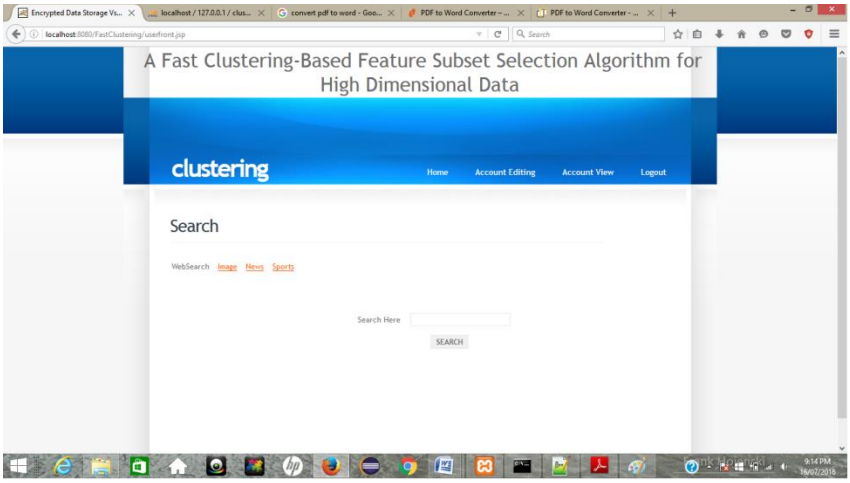


Figure 1 Flow chart for feature selection

5. IMPLEMENTATION RESULTS







6. CONCLUSION

This paper explains about the data mining functionalities and also about the feature subset selection. In this we have explained different methods proposed for feature subset selection. The proposed method is used to extract the features based on clustering. This also provides the implementation details of the proposed algorithm. The implementation details include the modules User Module, Distributed Clustering, Subset Selection Algorithm.

REFERENCES

[1] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast clustering based feature subset selection algorithm for high dimensional data", In proceedings of the IEEE Transactions n Knowledge and data engineering, 2013.
[2] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.
[3] M. Dash, H. Liu, and H. Motoda, "Consistency Based Feature Selection," Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109, 2000.
[4] A New Clustering Based Algorithm for Feature Subset Selection (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5272-5275.
[5] Dingcheng Feng, Feng Chen_, and Wenli Xu "Efficient Leave-One-Out Strategy for Supervised Feature Selection" TSINGHUA SCIENCE AND TECHNOLOGY ISSN 1007-0214 09/10 pp629 635 Volume 18, Number 6, December 2013.
[6] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int"l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.
[7] H. Liu, H. Motoda, and L. Yu, "Selective Sampling Approach to Active Feature Selection," Artificial Intelligence, vol. 159, nos. 1/2, pp. 49-74, 2004.
[8] Qinbao Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," IEEE Transaction on Knowledge and Data, Engineering, Vol. 25, No. 1, January 2013.
[9] M. Dash, H. Liu, and H. Motoda, "Consistency Based Feature Selection," Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109, 2000.
[10] Jesna Jose,"Fast for Feature Subset Selection Over Dataset" International Journal of Science and Research (IJSR), Volume 3 Issue 6, June 2014.

