A Fast Clustering Features Based on Sub Selection Algorithm in Big Data Using Fidoop

M.Keerthana[1], S.Priyanka[2], N.Ramya[3], C.Porkodi[4]
Ganadipathy Tulsis Jain Engineering College
keekeerthi4115@gmail.com
Ganadipathy Tulsis Jain Engineering College
priyankasampath11@gmail.com
Ganadipathy Tulsis Jain Engineering College
ramyan1794@gmail.com
Assistant professor, Ganadipathy Tulsis Jain Engineering College
porkodi_cse@gtec.ac.in

ABSTRACT

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, a fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated in this paper. 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 a 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. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) using the Kruskal's Algorithm clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study.

Key words: Frequent Item Sets, Frequent items ultra metric tree(FIU tree), Hadoop cluster, Load balance, MapReduce .

1. INTRODUCTION

It is widely recognized that a large number of features can adversely affect the performance of inductive learning algorithms, and clustering is not an exception. However, while there exists a large body of literature devoted to this problem for supervised learning task, feature selection for clustering has been rarely addressed. The problem appears to be a It is widely recognized that a large number of features can adversely affect the performance of inductive learning algorithms, and clustering is not an exception. However, while there exists a large body of literature devoted to this problem for supervised learning task, feature selection for clustering has been rarely addressed. The problem appears to be a difficult one given that it inherits all the uncertainties that surround this type of inductive learning. Particularly, that there is not a single performance measure widely accepted for this task and the lack of supervision available.as many irrelevant and redundant features as possible. This is because 1) irrelevant features do not contribute to the predictive accuracy, and 2) redundant features do not Feature subset selection can be viewed as the process of identifying and removing edound 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 difficult one given that it inherits all the uncertainties that surround this type of inductive learning. Particularly, that there is not a single performance measure widely accepted for this task and the lack of supervision available.

2. EXISTING SYSTEM

features. A well as much of the irrelevant and redundant information are 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. We achieve this through a new feature selection framework which composed of the two connected components of irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final sub set. The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the clusters.

 COPYRIGHT © 2016 IJREST, ALL RIGHT RESERVED

3. PROPOSED SYSTEM

Quite different from these hierarchical clustering-based algorithms, our proposed FAST algorithm uses minimum spanning tree-based method to cluster features. Meanwhile, it does not assume that data points are grouped around centers or separated by a regular geometric curve. Moreover, our proposed FAST does not limit to some specific types of data.
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. We achieve this through a new feature selection framework which composed of the two connected components of irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final sub set. The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the clusters.

4. METHODOLOGY

1. Network Analysis
2. Spatial Data Analysis
3. Classification, Clustering
4. Outlier Detection
5. Sentiment Analysis, Text Analytic
6. Cluster Formation using Big data
7. Data Stored in HDFS

I. Network Analysis

Load the data into the process. The data has to be preprocessed for removing missing values, noise and outliers. Then the given dataset must be converted into the arff format which is the standard format for WEKA toolkit. From the arff format, only the attributes and the values are extracted and stored into the database. By considering the last column of the dataset as the class attribute and select the distinct class labels from that and classify the entire dataset with respect to class labels.

II. Spatial Data Analysis

Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation.

III. Classification,Clustering

The relevance between the feature Fi € F and the target concept C is referred to as the T-Relevance of Fi and C, and denoted by SU (Fi,C). If SU (Fi,C) is greater than a predetermined threshold , we say that Fi is a strong T-Relevance feature.

IV. Outlier Detection

The correlation between any pair of features Fi and Fj (Fi,Fj € ^ F ^ i ≠ j) is called the F-Correlation of Fi and Fj, and denoted by SU(Fi, Fj). The equation symmetric uncertainty which is used for finding the relevance between the attribute and the class is again applied to find the similarity between two attributes with respect to each label.

V. Sentiment Analysis,Text Analytic

After building the MST, in the third step, we first remove the edgeswhose weights are smaller than both of the T-Relevance SU (Fi`, C) and SU (Fj`, C), from the MST.After removing all the unnecessary edges, a forest is obtained. Each tree Tj € Forest represents a cluster that is denoted as V (Tj), which is the vertex set of Tj as well. As illustrated above, the features in each cluster are redundant, so for each cluster V (Tj) we choose a representative feature Fj R whose T-Relevance SU (Fj R,C) is the greatest.

VI. Data Stored in HDFS

An HDFS cluster is comprised of a NameNode which manages the cluster metadata and DataNodes that store the data. Files and directories are represented on the NameNode by inodes. Inodes record attributes like permissions, modification and access times, or namespace and disk space quotas.The file content is split into large blocks (typically 128 megabytes), and each block of the file is independently replicated at multiple DataNodes. The blocks are stored on the local file system on the datanodes. The Namenode actively monitors the number of replicas of a block. When a replica of a block is lost due to a DataNode failure or disk failure, the NameNode creates another replica of the block.

5. SYSTEM ARCHITECTURE

 COPYRIGHT © 2016 IJREST, ALL RIGHT RESERVED

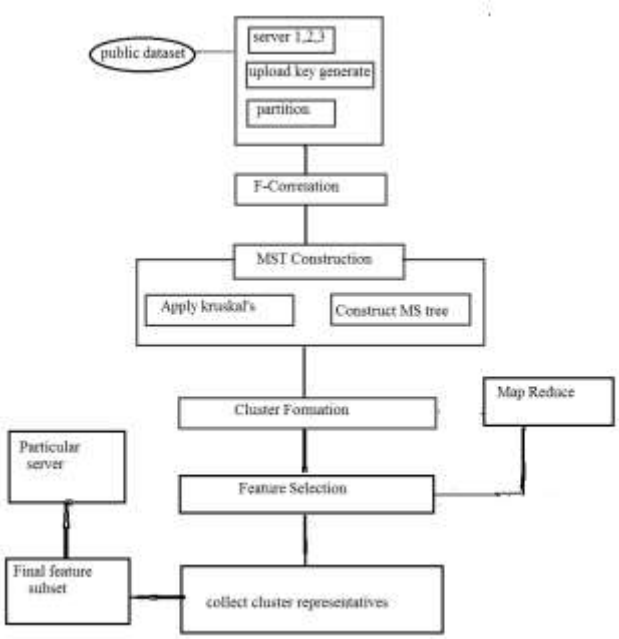

Fig. System Architecture

6. CONCLUSION

In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

REFERENCES

[1]YalingXun,JifuZhang,andXiaoQin,SeniorMember,IEEE,"A Fast Clustering Feature Based On Subselection Algorithm in Bigdata using Fidoop,"2015.

[2]J.Choi,C.Choi,K.Yim,J.Kim,and P. Kim,"Intelligent reconfigurable method of cloud computing resources for multimedia data delivery,"Informatica, vol. 24, no. 3, pp. 381–394, 2013.

[3]Rini T.Kaushik,Klara Nahrstedt,"FP-tax:Tree structure based generalised association rule mining,"pp..85-94,2012.

[4]R.Rahulurgaonkar,Bhuvanurgaonkar,MichaelJ.Neely,Anand Sivasubramanian,"Optimal power cost management using stored energy in data centeres.,pp.399-403,2011.

[5]J.Dean and S.Ghemawat, "MapReduce: A flexibledata processingtool,"commun.ACM,volume.53,number.1,pp.72-77,jan.2010.

 COPYRIGHT © 2016 IJREST, ALL RIGHT RESERVED