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

IJCSMC, Vol. 4, Issue. 7, July 2015, pg.161 – 166

RESEARCH ARTICLE

Improve the Efficiency of High Dimensional Data by using FAST and Feature Subset Selection Algorithm

Panga Gurivi Reddy[1], K. Ishthaq Ahamed[2]

[1]P.G Student, Department of Computer Science& Engineering, G. PullaReddy Engineering College, Kurnool, Andhra Pradesh, India

[2]Associate Professor, Department of Computer Science & Engineering, G. Pulla Reddy Engineering College, Kurnool, Andhra Pradesh, India

[1] gurivireddypanga513@gmail.com, [2] ishthaq1@yahoo.com

***Abstract*----Feature Selection involves recognizing a subset of the majority helpful features that produces attuned results as the unique set of features. The aim here is to select some of the features to form a feature sub set. Feature selection has been effective technique to deal with irrelevant features, and hence improving comprehensibility. Existing Feature selection algorithm removes only irrelevant features. We adopted a new clustering algorithm FAST removes both irrelevant and redundant features, and hence improve the time complexity. The FAST algorithm mainly works in two steps, in primary step, features are divided into clusters, based on the graph-theoretic clustering methods, and in secondary step the most representative feature is selected form a target feature set to form a final subset of features. This survey mainly focus on how FAST Clustering producing a subset of useful and independent features for a high dimensional data."**

***Key words*--- "Feature Selection Algorithm", "FAST Clustering", "Graph-Theoretic Clustering", "Irrelevant data", "Redundant data".**

I. INTRODUCTION

A. General Background

Data mining, the extraction of hidden predictive information from large data bases, is a powerful new technology, the data mining tools allowing business to provide knowledge-driven decision, and the automated prospective analysis of past events provided by retrospective tools typical of decision support systems. Data mining techniques are the result of a long process of research and product development.

Data mining is ready for application in the business community because it is supported by three technologies are

A. Massive data collection
B. Powerful multiprocessor computers
C. Data mining algorithms

Data mining tasks are specified by its functionalities that task are classified into two forms:

Descriptive mining tasks: Port or the general properties of the data.

Predictive mining tasks: Perform the implication on the current data order to craft prediction.

B. Data mining Functionalities:

- Characterization and Discrimination
- Mining frequent patterns
- Association and Correlation
- Classification and Prediction
- Cluster Analysis
- Outlier Analysis
- Evolution Analysis

C. Feature Selection:

Feature Selection is the process of selecting a subset of relevant features for use in model construction. The main aim of the feature selection technique is that the data contains many redundant or irrelevant feature sets. The Redundant features are provide no more information than the currently selected features, and the irrelevant features provide no more information in any context. Feature extraction creates new subset of features from original set of features. Feature selection is an effectual way for dimensionality reduction, elimination of inappropriate data, raising learning accuracy. The widely used feature subset selection methods are wrapper method, embedded, and filter and hybrid methods.

In particular, we accept the minimum spanning tree based clustering algorithms, for the reason that 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.

II. EXISTING SYSTEM

The embedded methods incorporate feature selection as a part of the training process and are usually specific to the given learning algorithms. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of a selected subsets. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are provide good generality, their computational cost is low, but the accuracy of the learning algorithms is not guaranteed.

The hybrid methods are a combination of the filter and wrapper methods by using a filter method to reduce a 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 particular learning algorithm with similar time complexity of the filter methods.

A. Drawbacks of existing system:

- Lacks Speed
- Security Issues

- Performance Related Issues
- The generality of the selected features is limited and the computational complexity is large

III. PROPOSED SYSTEM

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. Some of the feature subset selection algorithms are effectively eliminate irrelevant features but fail to handle redundant features yet some others can eliminate the irrelevant while taking care of the redundant features. We proposed FAST algorithm falls into second group. Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief which weights each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. Relief-F extends relief, enabling this method to work with noisy and incomplete datasets and to deal with multiclass problems, but still cannot identifying redundant features. We proposed FAST Clustering algorithm efficiently deals with both irrelevant and redundant features, and hence increase time complexity and accuracy.

A. Advantages of proposed system

- Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.
- We proposed FAST algorithm effectively deal with both irrelevant and redundant features, and obtain a good feature subset.
- Generally, all the six algorithms achieve significant reduction of the dimensionality by selecting only a small portion of the original features.
- The Friedman test tells that all the feature selection algorithms are equivalent in terms of runtime.

IV. FAST ALGORITHM

```
Inputs: D (F1, F2 …Fi, C)-The given data se
θ -The T-Relevance threshold
Output: S-Selected feature subset
//----part1: Irrelevant feature removal---
Step 1: For i=1 to m do
Step 2: T-Relevance = SU (Fi, c)
Step 3: If T-Relevance>θ then
Step 4: S =SU {Fi}
//part2: Minimum spanning tree Construction-----
Step 5: G=NULL; // G is a complete graph
Step 6: For each pair of features {F’i, F’j} CS do
Step 7: F-Correlation =SU (F’i, F’j)
Step 8:Add F’i and/or F’j to G with F-Correlation as the weight of the corresponding edge;
Step 9: minSpanTree=Prim (G); //using Prim Algorithm to generate the minimum spanning tree
//---part3: Tree partition and Representation feature selection---
Step 10: Forest= minSpanTree
Step 11: For each edge Eij∈ Forest d0
Step 12: If SU (F’I, F’j) <SU (F’I, C) <SU (F’I, F’j) <SU (F’j, C)
Step 13: Forest= Forest-Eij
Step 14: S= ϕ
Step 15: For each tree Ti € Forest do
Step 16: $F^{j}_{r}$= argmax F’k € Ti SU (F’k, C)
Step 17: S= SU {Fjr};
Step 18: Return S
```

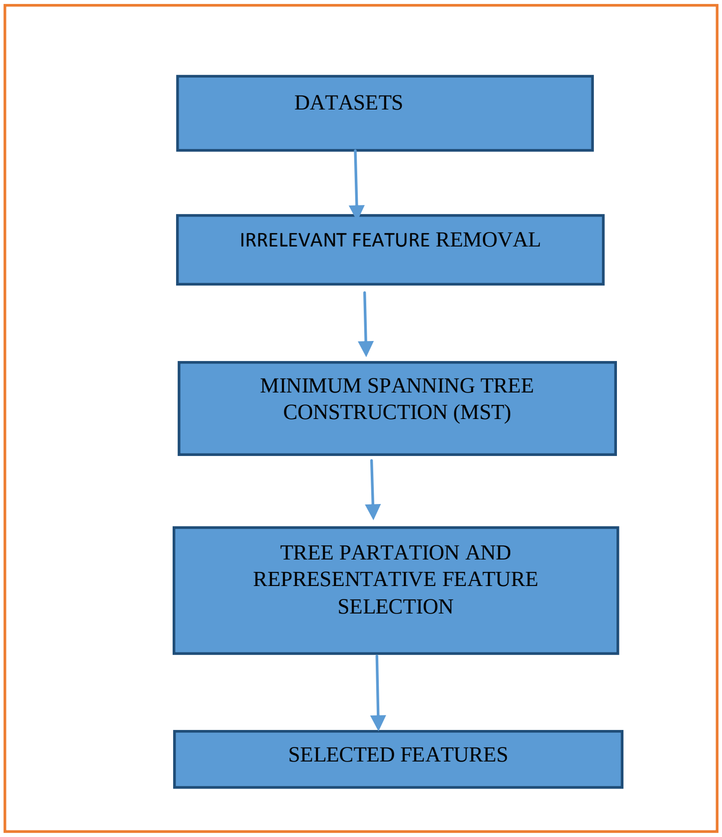
DATASETS
IRRELEVANT FEATURE REMOVAL
MINIMUM SPANNING TREE CONSTRUCTION (MST)
TREE PARTATION AND REPRESENTATIVE FEATURE SELECTION
SELECTED FEATURES

Figure1.1: System Flow

V. RELATED WORK

Some of the feature Subset Selection algorithms eliminate irrelevant features but fail to handle redundant features. Yet some other eliminate the irrelevant while taking care of the redundant features. FAST (fast-clustering based feature selection algorithm) algorithm falls into the second group. The Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function.

CFS is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target concept, yet uncorrelated with each other. FCBF is a fast filter method which identifies both irrelevant and redundant features without pair wise correlation analysis. Apart from these algorithms, FAST algorithm employs clustering-based method to choose features.

VI. 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. We proposed FAST algorithm effectively deals with both irrelevant and redundant features, and hence produces a good feature subset.

REFERENCES

[1].H. Almuallim and T.G. Dietterich, “Algorithms for Identifying Relevant Features,” Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.
[2] H. Almuallim and T.G. Dietterich, “Learning Boolean Concepts in the Presence of Many Irrelevant Features,” Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305, 1994.
[3].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.
[4]D.A. Bell and H. Wang, “A Formalism for Relevance and Its Application in Feature Subset Selection,” Machine Learning, vol. 41, no. 2, pp. 175-195, 2000.
[5].J. Biesiada and W. Duch, “Features Election for High-Dimensional data a Pearson Redundancy Based Filter,” Advances in Soft Computing, vol. 45, pp. 242-249, 2008.
[6].R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, “On Feature Selection through Clustering,” Proc. IEEE Fifth Int’l Conf. Data Mining, pp. 581-584, 2005.
[7].C. Cardie, “Using Decision Trees to Improve Case-Based Learning,” Proc. 10th Int’l Conf. Machine Learning, pp. 25-32, 1993.
[8].S. Chikhi and S. Benhammada, “ReliefMSS: A Variation on a Feature Ranking Relief Algorithm,” Int’l J. Business Intelligence and Data Mining, vol. 4, nos. 3/4, pp. 375-390, 2009.
[9].W. Cohen, “Fast Effective Rule Induction,” Proc. 12th Int’l Conf. Machine Learning (ICML ’95), pp. 115-123, 1995.
[10].M. Dash and H. Liu, “Feature Selection for Classification,” Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.
[11].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.

BIOGRAPHY

Panga Gurivi Reddy is a M.tech student in Computer Science & Engineering at G. Pulla Reddy Engineering College, Kurnool, Andhra pradesh, India. He recived B.Tech Degree in Computer Science & Engineering in AVR & SVR College of Engineering and Technology, Nandyal, Kurnool, Andhra Pradesh ,India.

K.Ishthaq Ahamed is currently working as Associate professor at G. Pulla Reddy Engineering College, Kurnool, Andhra pradesh, India. His research area in Artificial Intelligence.