ISSN: 2321-7782 (Online)

Volume 2, Issue 9, September 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Mining Feature Subset Using Cluster-Based Algorithm

K. Hepsiba[1]
Computer Science and Engineering
Gokula Krishna College of Engineering
Sullurupet – India

Y. Madusekhar[2]
Associate Professor, CSE
Gokula Krishna College of Engineering
Sullurupet – India

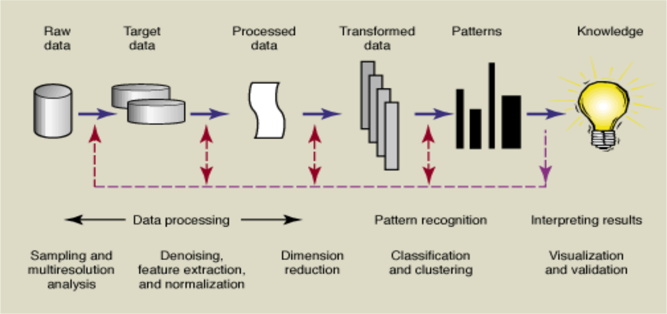
Abstract: Feature selection is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in classification process of data mining. Feature selection algorithm can be evaluated from mutually efficiency and effectiveness points of view. Our Proposed algorithm FAST is experimentally evaluated in this paper. FAST algorithm has three steps. In the first step irrelevant features are removed. In the second step, features are divided into clusters that posses features with redundant features. The third step selecting the most representative feature from each cluster that is closely related to the target class. This algorithm can be performed based on MST method for ensuring the efficiency.

Keywords—Data mining, Feature subset selection, feature clustering, MST construction.

I. Introduction

1. *Data Mining:*

Data mining starts with the raw data, which usually takes the form of simulation data, observed signals, or images. These data are preprocessed using various techniques such as sampling, multi resolution analysis, de noising, feature extraction, and normalization. Once the data are preprocessed or "transformed," pattern-recognition software is used to look for patterns. Patterns are defined as an ordering that contains some underlying structure. The results are processed back into a form-usually images or numbers-familiar to the scientific experts who then can examine and interpret the result.


Raw data
Target data
Processed data
Transformed data
Patterns
Knowledge
Data processing
Pattern recognition
Interpreting results
Sampling and multiresolution analysis
Denoising, feature extraction, and normalization
Dimension reduction
Classification and clustering
Visualization and validation

Sapphire, a data-mining infrastructure developed at Lawrence Livermore, is an iterative and interactive process designed to help scientific researchers uncover patterns in large data sets.

Fig1. Data mining Process

2. *Literature Survey*

Feature subset choice has been an active investigated topic since 1970s, and a large deal of research work has been published. Of the obtainable research work, most feature subset collection algorithms can effectively recognize the irrelevant features based on dissimilar assessment functions. Still very few of them can eliminate the redundant features and take the feature interface into consideration. On the basis of whether they can deal with immaterial features, unnecessary features and the feature interaction, the obtainable feature subset collection algorithms can be grouped into three categories: (a) the algorithms that can only handle immaterial features; (b) the algorithms that can handle both unrelated and unneeded features;

and (c) the algorithms that deals with unrelated and unneeded features while considering feature interface. Next, we give a brief analysis of the three categories correspondingly. Usually, the study work on feature subset selection has focused on search for relevant features. Feature weighting ranking algorithms weigh features independently and level them based on their importance to the target concept. Feature subset selection is the process of identifying and removing as many irrelevant and redundant features. Because (i) irrelevant features do not contribute to the predictive accuracy. (ii) Redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other features. The many feature subset selection algorithms can effectively eliminate irrelevant features but fail to handle redundant features however some of others can eliminate the irrelevant at the same time as taking care of the redundant features. The proposed FAST algorithm falls into the removing of redundant features. Conventionally, feature subset selection research has focused on searching for relevant features. In Relief, each feature according to its ability to classify instances under different targets based on distance-based criteria function. However, Relief is not removing redundant features. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but accuracy of learning algorithms, and thus should be eliminated as well. CFS and FCBF are examples that take into consideration the redundant features. CFS is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, so far uncorrelated with each other. FCBF is a fast filter method which can identify relevant features as well as redundancy among relevant features without correlation analysis. Different from these algorithms, the proposed FAST Algorithm employs clustering based method to select features.

II. Feature Selection

The important reason of the feature extraction algorithm is that they discover out or extract only targeted features out of many features. They don’t measure the irrelevant and redundant data because irrelevant and redundant data affects the competence and effectiveness of the algorithm. In existing algorithm that uses a variety of different techniques to decide relevant features or removing irrelevant or redundant features , when it removes the irrelevant features it does not consider the interface of different features. Proposed algorithm not only removes the irrelevant features but also focus on interaction of the features.

In machine learning and statistics, feature selection, also known as attribute selection or variable subset selection, is the process of selecting a subset of relevant features for the model construction. The main use of feature selection technique is that the data contains many redundant or irrelevant features. Redundant features 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 training data, 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). Feature selection techniques provide three main benefits when constructing predictive models.

- Improved model interpretability,
- Shorter training times,
- Enhanced generalization by reducing over fitting.

1. *Architectural diagram*

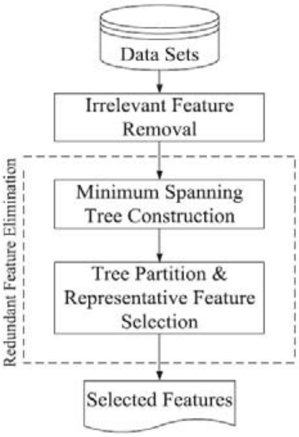

Data Sets
Irrelevant Feature Removal
Minimum Spanning Tree Construction
Tree Partition & Representative Feature Selection
Redundant Feature Elimination
Selected Features

Fig 2. Architecture diagram

FAST algorithm can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset, which is 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 subset. 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 (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; and (iii) the selection of representative features from the clusters. The methodology used here is a five step methodology. The steps are data set management, finding symmetric uncertainty, irrelevant feature removal, minimum spanning tree construction and its clustering and the final step is target matching.

III. FAST ALGORITHM

Proposed FAST algorithm mainly consists of 3 steps: 1) Removing irrelevant features, 2) constructing an MST and 3) Partitioning the MST and select exchanging features. In first step, we calculate the T-Relevance. In second step, we compute F-Correlation value for each pair of feature. In the third step, we remove the edges, whose weight is less than both of the T-Relevance and SU(Fj,C) from the MST. After removing all the surplus edges, a forest is obtained. Every tree Tj Forest represents a cluster that's denoted as V (Tj), that is that the vertex set of Tj moreover. As illustrated on top, the features in every cluster are redundant. The steps for FAST algorithm as follows.

```
Input
D (F1, F2, …..Fm)      -  Data set
C                      -  Target class
TH                             -  Threshold value
Output
S -selected feature subset.
Step 1:  Irrelevant Feature Removal
for i =1 to m do
T-Relevance =SU (Fi,C)
if T-Relevance > TH then S =SU{Fi}

step 2:  Clustering
```

Construct a Graph G such that for each feature in D calculate F-Correlation =SU (Fi, Fi+1) Construct a tree with clusters.

Step 3: Redundant feature removal

With each tree construct a forest F Select a representative feature Fi^R from each cluster

Step 4: Subset selection

With edges of Forest select the subset S such that

S = S U { Fi^R }, where Fi^R is the relevant feature subset selected.

IV. Algorithm Implementation

T-Relevance(Irrelevant Feature Removal)

The relevance between the feature Fi 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, then Fi is a strong T-Relevance feature.

$$SU(X,Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)}$$

After finding the relevance value, the redundant attributes will be removed with respect to the threshold value.

F-Correlation

The correlation between any pair of features Fi and Fj is called the F-Correlation 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.

Construction of Minimum spanning Tree

Prim's algorithm is a greedy algorithm that finds a minimum spanning tree for a connected weighted graph. It finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized. If the graph is not connected, then it finds a minimum spanning forest (a minimum spanning tree for each connected component).

Description:

- Create a forest F (a set of trees), where each vertex in the graph is a separate tree.
- Create a set S containing all the edges in the graph .
- While S is nonempty and F is not yet spanning.
- Remove an edge with minimum weight from S.
- If that edge connects two different trees, then add it to the forest, combining two trees into a single tree

At the termination of the algorithm, the forest forms a minimum spanning forest of the graph. If the graph is connected, the forest has a single component and forms a minimum spanning tree. The sample tree is as follows

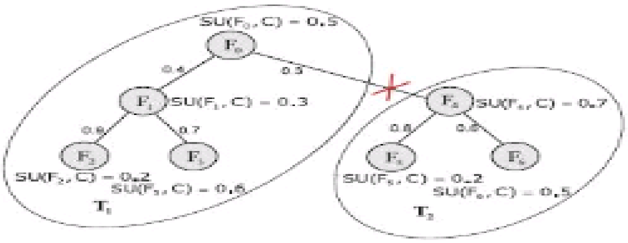


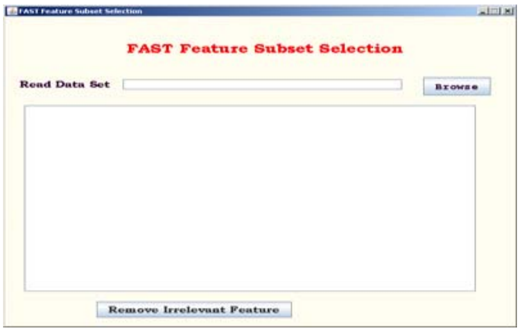
Fig 3. Example of minimum spanning tree

Feature selection

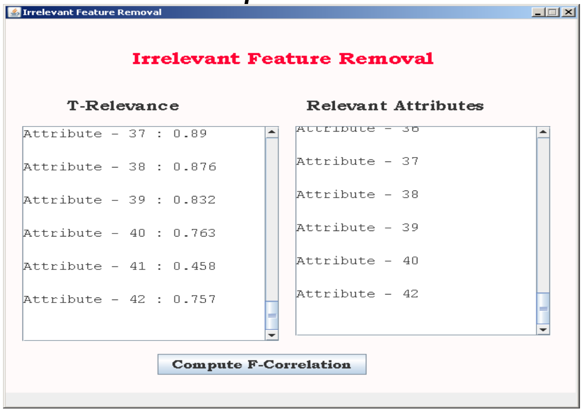
The features in each cluster are redundant, so for each cluster V(Tj) we choose a representative feature Fj^R whose T-Relevance (Fj^R, C) is the greatest. All Fj^R (j=1….|Forest|) comprise the final feature subset ∪ Fj^R

V. Experimental Results

A. *Loading of data set:*


FAST Feature Subset Selection
FAST Feature Subset Selection
Read Data Set
Browse
Remove Irrelevant Feature

B. *Calculation of T-Relevance and F-Correlation*

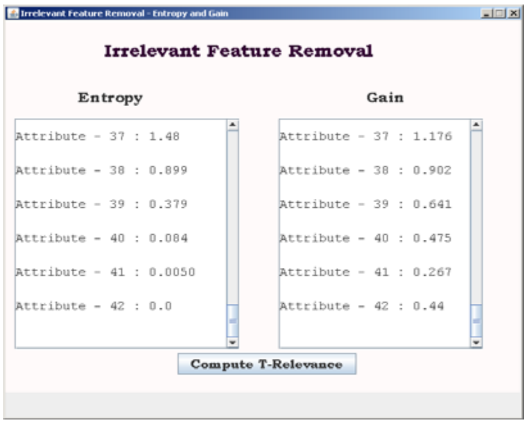


T-Relevance

```
Attribute - 37 : 0.89
Attribute - 38 : 0.876
Attribute - 39 : 0.832
Attribute - 40 : 0.763
Attribute - 41 : 0.458
Attribute - 42 : 0.757
```

Relevant Attributes

```
Attribute - 36
Attribute - 37
Attribute - 38
Attribute - 39
Attribute - 40
Attribute - 42
```


Irrelevant Feature Removal - Entropy and Gain
Irrelevant Feature Removal
Compute T-Relevance

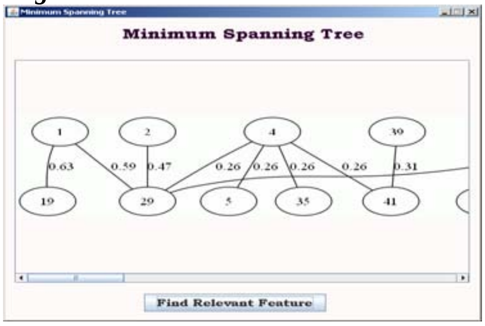
Entropy

```
Attribute - 37 : 1.48
Attribute - 38 : 0.899
Attribute - 39 : 0.379
Attribute - 40 : 0.084
Attribute - 41 : 0.0050
Attribute - 42 : 0.0
```

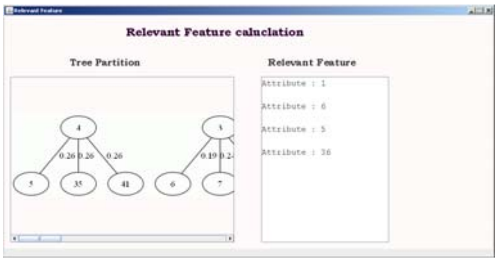
Gain

```
Attribute - 37 : 1.176
Attribute - 38 : 0.902
Attribute - 39 : 0.641
Attribute - 40 : 0.475
Attribute - 41 : 0.267
Attribute - 42 : 0.44
```

C. *Construction of Minimum spanning tree:*



D. *Selection of relevant feature*



```
Attribute : 1
Attribute : 6
Attribute : 5
Attribute : 36
```

VI. Conclusion

Feature selection is an effective and efficient process for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Feature selection is applied to reduce the number of features in many applications where data has hundreds or thousands of features. Existing feature selection methods mainly focus on finding relevant features. For the future work, we can study some formal properties of feature space. In feature we are going to classify the high dimensional data.

References

1. Biesiada J. and Duch W., Feature selection for high dimensional data Pearson redundancy based filter, Advances in Soft Computing, 45, pp 242C249,2008.
2. Dash M., Liu H. and Motoda H., Consistency based feature Selection, In Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109, 2000.
3. Bhaskar Adepu and Kiran Kumar Bejjanki., “A Novel Approach for Minimum Spanning Tree based Clustering Algorithm”
4. Yu L. and Liu H., “Feature selection for high-dimensional data: a fast correlation-based filter solution”, in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856- 863, 2003.
5. Yu L. and Liu H.,” Efficient feature selection via analysis of relevance and redundancy”, Journal of Machine Learning Research, 10(5), pp 1205-1224, 2004.
6. Zhao Z. and Liu H. (2009), ‘Searching for Interacting Features in Subset Selection’, Journal of Intelligent Data Analysis, 13(2), pp 207-228, 2009.
7. Huan Liu and Lei Yu, ‘Towards Integrating Feature Selection Algorithms for Classification and Clustering’, IEEE Transactions on Knowledge and Data Engineering Vol. 17 No. 4, 2005.
8. Sha. C, Qiu X. and Zhou A (2008), ‘Feature Selection Based on a New Dependency Measure’, Fifth International Conference on Fuzzy.
9. Sunita Beniwal and Jitendar Arora (2012), ‘Classification and Feature Selection Techniques in Data Mining’, International Journal of Engineering Research and Technology, Volume 1 Issue 6, August 2012.
10. K. Kira and L.A. Rendell (1992), ‘The Feature Selection Problem: Traditional Methods and a New Algorithm’, Proceedings of 10th National Conference on Artificial Intelligence, pp. 129-134, 1992.
11. I. Kononenko (1994), ‘Estimating Attributes: Analysis and Extension of RELIEF’, Proceedings of Sixth European Conference on Machine Learning, pp. 171-182, 1994.
12. Huan Liu, Manoranjan Dash and Hiroshi Motoda, ‘Feature Selection Using Consistency Measure’, Second International Conference, DS’99 Tokyo, Japan, December 6–8, 1999 Proceedings, pp 319-320
13. M.A. Hall (2000), ‘ Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning’, Proceedings of 17Th International Conference on Machine Learning, pp. 359-366, 2000.
14. Huan Liu and Rudy Setiono, ‘A Probabilistic approach to Feature Selection – A Filter approach’, 2000.
15. Huan Liu, Hiroshi Motoda (2002), and Lei Yu, ‘Feature Selection with Selective Sampling’, Proceedings of 19th International Conference on Machine Learning, pp. 395-402, 2002.
16. Huan Liu and Lei Yu (2003), ‘Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution’, Proceedings of 20th International Conference on Machine Learning, pp. 856-863, 2003.
17. Fleuret. F (2004), ‘Fast Binary Feature Selection with Conditional Mutual Information’, Journal of Machine Learning Research 5 (2004) 1531–1555.

AUTHOR(S) PROFILE

K HEPSIBA Pursing M.Tech Computer Science & Engineering in Gokula Krishna College of Engineering Sullurpeta affiliated to JNTU, Anantapur and received Master of Computer Application from Sri Padmavathi Mahila Visva Vidyalayam UniversitySPMVVU) in 2000. Her field of interest is Data WareHousing and Mining.

Y.MADHU SEKHAR Received the degree of MASTER OF COMPUTER SCIENCE in JNTU, ANANTHAPUR. He is currently working as an Associate Professor in Gokula Krishna College of Engineering, Sullurpet, Andhra Pradesh, India. His field of interest is Data Ware Housing and Mining.