ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Efficient Feature Subset Selection using Kruskal's Process in Big Data

S.Saranya, S.L.Julian Austrina, K.Ravikumar

Post Graduate Student, Department of CSE, Rrase College of Engineering, Chennai, India

Assistant Professor, Department of CSE, Rrase College of Engineering Chennai, India

Professor, Department of CSE, Rrase College of Engineering, Chennai, India

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. 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.

KEYWORDS: Feature subset selection, filter method, feature selection, graph-based clustering

I. 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 the 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. Many feature selection methods have been adopted that has been classified and proposed for the machine learning process. The generality of the selected feature is limited and the computational complexity is complex. Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features. Along with the irrelevant features and redundant features also affect the speed and accuracy of learning algorithm. With the development of the information technology, the scales of data are increasing quickly. In recent years, the amount of data in our world has been increasing explosively, and analysing large data sets so called “Big Data” becomes a key basis of competition underpinning new waves of productivity growth, innovation and consumer surplus. Big data is becoming an integral part of solving the world's problems.

II. RELATED WORK

Feature subset selection is the process of removing the irrelevant and redundant feature as many as possible. Irrelevant feature doesn't contribute to the accuracy of the data and redundant feature doesn't help in achieving better predictor. Of many feature subset selection algorithms, some can effectively eliminate irrelevant feature but fail to handle redundant feature. FAST algorithm takes care of both elimination of irrelevant and the redundant feature effectively and efficiently. 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.

Hierarchical clustering also has been used to select features on spectral data. Van Dijck and Van Hulle proposed a hybrid filter/wrapper feature subset selection algorithm for regression. Krier et al. presented a methodology

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information. Their feature clustering method is similar to that of Van Dijck and Van Hulle except that the former forces every cluster to contain consecutive features only. Both methods employed agglomerative hierarchical clustering to remove redundant features.

III. PROPOSED ALGORITHM

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." A novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. This is achieved 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 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 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.

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.

To find the relevance of each attribute with the class label, Information gain is computed in this module. This is also said to be Mutual Information measure. Mutual information measures how much the distribution of the feature values and target classes differ from statistical independence. This is a nonlinear estimation of correlation between feature values or feature values and target classes. The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes, and has been used to evaluate the goodness of features for classification

The Information Gain is calculated as follows:

$$GAIN(X|Y) = H(X) - H(X|Y) \quad \text{eq. (1)}$$

$$= H(Y) - H(Y|X)$$

To calculate gain, entropy and conditional entropy values is computed

$$H(X) = -\sum_{x \in X} p(x) \log 2\, p(x) \quad \text{eq. (2)}$$

$$H(X|Y) = -\sum_{y \in X} p(y) \sum_{x \in X} p(x|y) \log 2\, p(x|y) \quad \text{eq. (3)}$$

Where,

p(x) and p(y) is the probability density function

p(x|y) is the conditional probability density function

The Symmetric Uncertainty (SU) is calculated using :

$$SU(X,Y) = (2 \times GAIN(X|Y)) / (H(X) + H(Y)) \quad \text{eq. (4)}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

IV. PSEUDO CODE

FAST - Fast Clustering - b**A**sed Feature **S**election Algori**T**hm

Let D(F1,F2.....Fm,C) is the given dataset and θ - the Threshold.
Step 1 : Calculate Information Gain (using eq(1)) to compute Symmetric Uncertainty(SU)
Step 2 : Check if (T-Relevance = SU(Fi,C)) is satisfied
if(T-Relevance > θ)
Select the Feature for Minimum Spanning Tree Construction (MST)
else
Ignore the Feature
Step 3 : Perform F-Correlation (Fi and Fj (Fi,Fj € ^ F ^ i ≠ j) on SU (F ′ i , F ′ j)for each selected pair from step 2
Step 4 : Construct Minimum Spanning Tree (MST) using Kruskal's Algorithm
minSpanTree = Kruskal's(G)
Where,
G is complete graph
Step 5 : Partition the tree that is constructed using step 4
Step 6 : Elect Representative from the featured selected using step 5
Step 7 : End

V. SIMULATION RESULTS

The simulation studies involves the process of forming the feature subset from the entire dataset. The subset formed is checked if it's being free from irrelevant and the redundant feature data. FAST algorithm obtains the average of best proportion of selected feature. FAST algorithm effectively filters out a mass of irrelevant features and reduces the possibility of improperly bringing the irrelevant features into the subsequent analysis. The algorithm removes a large number of redundant features by choosing a single representative feature from each cluster of redundant features. As a result , only a very small number of discriminative features are selected. This coincides with the desire happens of the data analysis.

VI. CONCLUSION AND FUTURE WORK

The simulation results showed that proposed 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. The future work, involved to explore different types of the filter methods and some formal properties of future space

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. L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.
5. R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.
6. 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.
7. 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.
8. 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

9. C. Cardie, "Using Decision Trees to Improve Case-Based Learning," Proc. 10th Int'l Conf. Machine Learning, pp. 25-32, 1993.
10. P. Chanda, Y. Cho, A. Zhang, and M. Ramanathan, "Mining of Attribute Interactions Using Information Theoretic Metrics," Proc. IEEE Int'l Conf. Data Mining Workshops, pp. 350-355, 2009.

BIOGRAPHY

Saranya is a post graduate student in the department of computer science engineering, Rrase college of engineering, Anna University. She received Bachelor of Engineering degree (B.E) in 2010 from Anna University, Chennai, India. Her research interests are Big Data, Information Security etc.