EFFICIENT FEATURE SUBSET SELECTION MODEL FOR HIGH DIMENSIONAL DATA

Chinnu C Georgel[1] and Abdul Ali[2]

[1]Department of Computer Science, Ilahia College of Engineering and Technology
chinnulaby@gmail.com
[2] Department of Computer Science, Ilahia College of Engineering and Technology
abdulali@icet.ac.in

ABSTRACT

This paper proposes a new method that intends on reducing the size of high dimensional dataset by identifying and removing irrelevant and redundant features. Dataset reduction is important in the case of machine learning and data mining. The measure of dependence is used to evaluate the relationship between feature and target concept and or between features for irrelevant and redundant feature removal. The proposed work initially removes all the irrelevant features and then a minimum spanning tree of relevant features is constructed using Prim's algorithm. Splitting the minimum spanning tree based on the dependency between features leads to the generation of forests. A representative feature from each of the forests is taken to form the final feature subset.

KEYWORDS

Feature subset selection, filter technique, feature clustering, feature reduction

1. INTRODUCTION

Data mining is the process of automatically discovering useful information from large data repositories [2]. The rapid advances in technologies led to accumulation of vast amount data. As the number of organizations grows day by day the breeding of new technologies and new styles of data also increases. It is difficult to handle them and store them in an efficient manner and of course retrieval of data is extremely challenging [3]. Nowadays we cannot use traditional methods to explore the data analysis because of the size of dataset. So it is very important to make study on data analysis. Most of the technologies are blended with data mining or we can say that data mining is vital and indispensible concept for every technology. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches [9][10].Data mining tasks are mainly divided into predictive and descriptive. Predictive refers to predict the particular attribute based on other attributes. Descriptive task is to derive patterns like correlations, trends, clusters, trajectories and anomalies. Association analysis is used to discover patterns from correlated data and the output of analysis is represented using implication rules. Cluster analysis is the method of partitioning the datasets into different clusters and each cluster data is strongly correlated with intra-manner and inter clusters shows strong repulsion. There are several clustering methods but it is difficult to give a crisp categorization. Anomaly detection is the opposite of cluster analysis. Anomaly detection is the process of identifying significantly different data from rest of data. Example: Fraud detection.

DOI: 10.5121/ijci.2016.5217

Choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility [9]. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches.The embedded methods incorporate feature selection as a part of the training process and therefore may be more efficient than the other three categories [23] decision trees or artificial neural networks are examples. 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 usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low. The hybrid methods are a combination of filter and wrapper methods [25] by using a filter method to reduce search space that will be considered by the subsequent wrapper.

2. RELATED WORKS

Dimensionality can be defined as the number of attributes or features that an object posses. Data objects compose of several features. Most of the features give unique characteristics to the object. As the number of features increases, processing will be difficult [11][3]. But lesser number of features will also, only give a vague idea about the object. So it is important to reduce the number of features into a number which represents the objects effectively [13][18].

Feature selection is an important concept of reducing dimensions [19][15]. Each feature in a dataset represents a particular characteristic of object. But sometimes some of them are irrelevant and or correlated. Finding such features and removing it is known as dimensionality reduction [20][19].Feature subset selection involves identifying and removing as much as irrelevant and redundant features as possible. This is because 1) irrelevant features do not contribute to the predictive accuracy [14], and 2) redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Most of the feature subset algorithms removes irrelevant features successfully but not able to handle redundant features [5],[18]. Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief [11], which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. This will not eliminate redundant features. But with irrelevant features redundant features will affect the speed and accuracy of machine learning algorithms. So the redundant features also have to be eliminated.

CFS [18], FCBF [20], and CMIM [15] are examples that take into consideration the redundant features. In CFS correlation based heuristic evaluation function is used. Feature subsets contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features are ignored and redundant features are screened out CFS uses symmetrical uncertainty to measure correlations. In FCBF predominant correlation is used. Fast filter method can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis. CMIM iteratively picks features which maximize their mutual information with the class to predict conditionally to the answer of any feature already picked. The proposed algorithm eliminates both irrelevant and redundant features efficiently using clustering based feature selection.

Cluster analysis is a method of classifying the given data. There are several cluster methods exists. It is difficult to say that one cluster method is efficient than other. Each of the cluster method overrides other based on at least one condition. Most of the cluster methods help to find irrelevant and correlated features. This implies that clustered data is efficient for removing relevant or irrelevant data. Hierarchical clustering has been adopted in word selection in the context of text classification (e.g., [22] and [1]). 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. [22] or on the distribution of class labels associated with each word.

3. PROPOSED METHOD

Irrelevant features, together with redundant features, will affect the correctness of the learning machines [6]. 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 hold features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.” [7].

Based on these aspects a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. The algorithm works in four steps. First step is to calculate the symmetric uncertainty of features between target classes and other features. In this paper the algorithm calculates symmetric uncertainty not only considering the co-occurrence of feature but also the change of feature when other changes. Then the features whose symmetric uncertainty values less than a threshold is considered irrelevant and are removed. For the remaining features minimum spanning tree is constructed using prim’s algorithm. The third step splits the minimum spanning tree into several forests. Each forest is considered as a single cluster. As we know that the elements within the cluster is strongly correlated. So we can select the representative features from it. Fourth step takes representative features from every cluster.

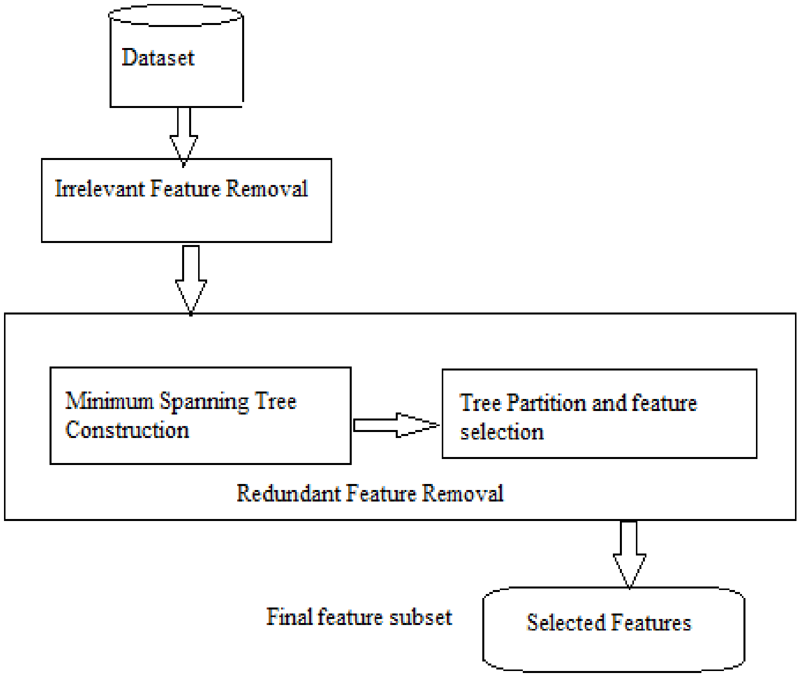
Dataset
Irrelevant Feature Removal
Minimum Spanning Tree Construction
Tree Partition and feature selection
Redundant Feature Removal
Final feature subset
Selected Features

Figure 1.Framework of the proposed feature subset selection algorithm.

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. Most of the information contained in redundant features is already present in other features. As a result, unneeded features will affect the accuracy. Symmetric uncertainty is a nonlinear estimation of correlation between feature values. The symmetric uncertainty is derived from mutual information by normalizing it to the entropies of feature values. The symmetric uncertainty can be defined as the measure of correlation between either two features or a feature and target concept.

$$SU(X,Y) = \frac{(2x\text{Information Gain}(X,Y))}{H(X) + H(Y)}$$

Where, H(X) is the entropy of a discrete random variable X. Suppose p(x) is the prior probabilities for all values of X, H(X) is defined by

$$H(X) = -\sum p(x) \log2\, p(x)$$

Gain(X/Y) is the amount by which the entropy of Y decreases. It reflects the additional information about Y provided by X and is called the information gain [23] which is given by

$$\begin{aligned} Gain(X/Y) &= H(X) - H(X/Y) \\ &= H(Y) - H(X/Y) \end{aligned}$$

Where H(X) is the conditional entropy which quantifies the remaining entropy (i.e., uncertainty) of a random variable X given that the value of another random variable Y is known.

Information gain is a symmetrical measure. That is the amount of information gained about X after observing Y is equal to the amount of information gained about Y after observing X. This ensures that the order of two variables will not affect the value of the measure. Symmetric uncertainty treats a pair of variables symmetrically, it compensates for information gain's bias toward variables with more values and normalizes its value to the range [0,1].

Given SU(X,Y) the symmetric uncertainty of variables X and Y , the relevance T-Relevance between a feature and the target concept C, the correlation F-Correlation between a pair of features, the feature redundance F-Redundancy and the representative feature R-Feature of a feature cluster can be defined as follows.

Definition 1 (T-Relevance): 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 value we say that Fi is a strong T-Relevance feature.

Definition 2(F-Correlation): The correlation between any pair of features Fi and Fj is called the FCorrelation of Fi and Fj, and denoted by SU(Fi,Fj).

Definition 3 (F-Redundancy): Let S {F1, F2, . . . , Fi, . . . Fk } be a cluster of features. if SU(Fj,C) ≥ SU(Fi,C) ^ SU(Fi,Fj)> SU(Fi,C)is always corrected for each Fi, then Fi are redundant features with respect to the given Fj (i.e., each Fi is a F-Redundancy).

Definition 4 (R-Feature).A feature is a representative feature of the cluster S (i.e., Fi is R-Feature) if and only if, Fi = argmax SU(Fj, C). This means the feature, which has the strongest TRelevance, can act as a R-Feature for all the features in the cluster.

According to the above definitions, feature subset selection can be the process that identifies and retains the strong T-Relevance features and selects R-Features from feature clusters. The behind heuristics are that

1. irrelevant features have no/weak correlation with target concept;
2. redundant features are assembled in a cluster and a representative feature can be taken out of the cluster.

Proposed Algorithm

Inputs: D (F1, F2, …,Fm, C) – the given dataset
T-Relevance threshold.
Output: S - Final feature subset

```
1. for i=1 to m do
2. T-Relevance = SU(Fi,C)
3. if T-Relevance > T-Relevance threshold then
4. S = S ∪ {Fi};
5. G = NULL; //G is a complete graph
6. for each pair of features {Fi, Fj } ⊂ S do
```

```
7. F-Correlation = SU (Fi’, Fj’)
8. While calculating F-Correlation for each feature check change of a feature with respect to another
8. Add Features Fi and / or Fj to G with F-Correlation as the weight of the corresponding edge
9. minSpanTree = prim(G); // Generate minimum spanning tree using Prim’s algorithm
10. Forest = minSpanTree
11. For each edge Eij∈ Forest do
12. if SU(Fi, Fj) <SU(Fi, C) ^ SU(Fi, Fj) < SU(Fj, C) then
13. Forest = Forest - Eij
14. Select representative features from each tree (one with maximum symmetric uncertainty value)
17. S = S ∪ {Representative features};
18. return S
```

Consider a data set D consists of m features f= {f1, f2…fm} and class C, the T-Relevance SU (fi, C) value for each feature is calculated in the first step. The features whose T-Relevance value greater than a predefined threshold value forms the relevant feature subset F= {F1, F2 …Fk}.In the second step F-correlation values for each pair of features is calculated considering the whether a feature changes when other feature changes along with checking co-occurrence of the features. A graph G=(V,E) is constructed with each target-relevant feature as node and the F-Correlation value as the weight of the edge. Graph G reflects the correlations among all the target-relevant features .Then a minimum spanning tree is constructed which connects all vertices such that the sum of the weights of the edges is the minimum, using the well-known Prim algorithm [21].After constructing minimum spanning tree edged with weights smaller than both of the T-Relevance SU(Fi, C) and SU(Fj, C), are deleted from the MST. This deletion will result in many disconnected trees T1,T2 etc. . Each tree T represents a cluster and features in a cluster are strongly correlated so are redundant. Finally a representative feature is selected from each disconnected tree to form the final feature subset.

4. EXPERIMENTAL RESULTS

We present the experimental results in terms of proportion of selected features.

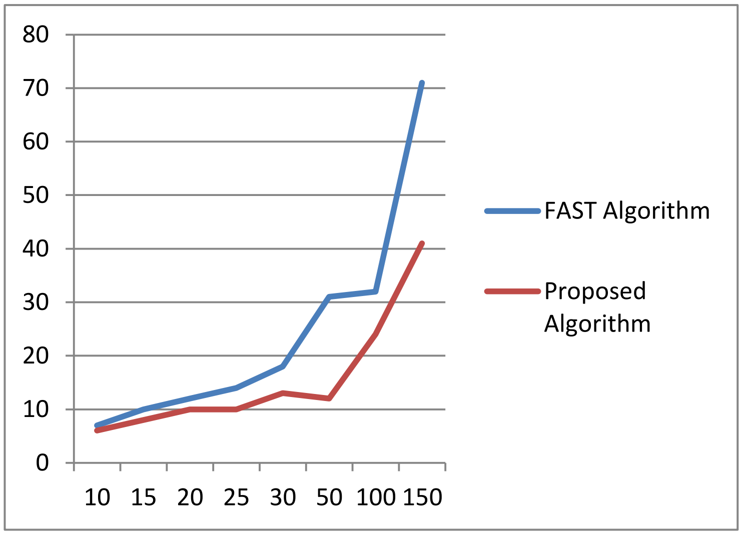

Figure 2. Comparison of New algorithm and FAST algorithm

Both algorithms FAST and Proposed algorithm achieve significant reduction of dimensionality by selecting only a small portion of the original features. But the experimental results shows that the proposed algorithm achieves more reduction in dimensionality. We get only the most important features which is capable to classify it accurately.

5. Conclusion

In this paper we have presented a feature selection algorithm for high dimensional data in which a clustering-based approach is used. A dependency measure is used to find out the correlation between features .The model involves the following steps removing irrelevant features, constructing minimum spanning tree, partitioning the minimum spanning tree and selecting representative features. In the algorithm partitioning of minimum spanning tree results in clusters and each cluster is treated as a single feature. Thus the dimensionality is drastically reduced.

We have compared the performance of the proposed algorithm with FAST algorithm and found that the new algorithm achieve significant reduction of dimensionality by selecting only a small portion of the original features.

For the future work, we plan to explore different types of correlation measures.

Acknowledgements

I wish to thank the Management, the Principal and Head of the Department (CSE) of ICET for the support and help in completing the work.

References

[1] Qinbao Song, Jingjie Ni, and Guangtao Wang, “A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data” IEEE Transactions on knowledge and data engineering, Vol. 25, No. 1, January 2013

[2] Data Mining: Concepts and Techniques 2nd ed. Jiawei Han and Micheline Kamber

[3] 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.

[4] Top 10 algorithms in data mining XindongWu • Vipin Kumar • J. Ross Quinlan • Joydeep Ghosh • Qiang Yang Hiroshi Motoda • Geoffrey J. McLachlan • Angus Ng • Bing Liu • Philip S. Yu •Zhi-Hua Zhou • Michael Steinbach • David J. Hand • Dan Steinberg

[5] R. agarwal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc, 20th Int’l Conf. Very large Data Bases, pp.487-499,1994.

[6] 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.

[7] R. Kohavi and G.H. John, “Wrappers for Feature Subset Selection,” Artificial Intelligence, vol. 97,nos.1/2, pp. 273-324, 1997.

[8] M.A. Hall and L.A. Smith, “Feature Selection for Machine Learning:Comparing a Correlation-Based Filter Approach to the Wrapper,” Proc.12th Int’l Florida Artificial Intelligence Research Soc. Conf., pp. 235-239, 1999.

[9] T.M. Mitchell, “Generalization as Search,” Artificial Intelligence, vol. 18, no. 2, pp. 203-226, 1982.I.Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” J. Machine Learning Research, vol 3, pp. 1157-1182, 2003.

[10] K. Kira and L.A. Rendell, “The Feature Selection Problem: Traditional Methods and a New Algorithm,” Proc. 10th Nat’l Conf.Artificial Intelligence, pp. 129-134, 1992.

[11] 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.

[12] 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.

[13] G.H. John, R. Kohavi, and K. Pfleger, “Irrelevant Features and theSubset Selection Problem,” Proc. 11th Int’l Conf. Machine Learning, pp.121-129, 1994

[14] F. Fleuret, “Fast Binary Feature Selection with Conditional Mutual Information,” J. Machine LearningResearch, vol. 5, pp. 1531-1555, 2004.

[15] A.Y. Ng, “On Feature Selection: Learning with Exponentially Many Irrelevant Features as Training Examples,” Proc. 15th Int’l Conf. Machine Learning, pp. 404-412, 1998.

[16] L. Yu and H. Liu, “Feature Selection for High-Dimensional Data: A FastCorrelation-Based Filter Solution,” Proc. 20th Int’l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.

[17] M.A. Hall, “Correlation-Based Feature Subset Selection for Machine Learning,” PhD dissertation, Univ. of Waikato, 1999.

[18] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufman, 1993.

[19] 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.

[20] R.C. Prim, “Shortest Connection Networks and Some Generalizations,” Bell System Technical J., vol. 36, pp. 1389-1401, 1957

[21] F. Pereira, N. Tishby, and L. Lee, “Distributional Clustering of English Words,” Proc. 31st Ann. Meeting on Assoc. for Computational Linguistics, pp. 183-190, 1993. I.Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” J. Machine Learning Research, vol 3, pp. 1157-1182, 2003

[22] .U. Fayyad and K. Irani, “Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning,” Proc. 13th Int’l Joint Conf. Artificial Intelligence, pp. 1022-1027, 1993.

[23] J. Yu, S.S.R. Abidi, and P.H. Artes, “A Hybrid Feature Selection Strategy for Image Defining Features: Towards Interpretation of Optic Nerve Images,” Proc. Int’l Conf. Machine Learning and Cybernetics, vol. 8, pp. 5127-5132, 2005.

AUTHORS

Chinnu C. George, is a PG Scholar in Ilahia college of Engineering and Technology, Muvattupuzha, Kerala. She received B-Tech Degree in Computer Science from Viswajyothi College of Enginnering and Technology (M G University) Vazhakulam, Kerala in 2007. Her research interest includedatamining and Natural Language Processing.

Abdul Ali, Assistant Professor, is an Assistant Profersor of Computer Science Department, in Ilahia college of Engineering, Muvattupuzha, Kerala. He received M-Tech Degree in Computer and Information Technology from Center for information technology and engineering University Campus,M S University, Tirunelveli, Tamilnadu in 2010. He received B-Tech Degree in Computer Science from M G University College Thodupuzha, Kerala in 2007 .His research interest include image processing and networking.