Feature Selection using Mutual Information for High- dimensional Data Sets

Arpita Nagpal
PhD. Scholar,
Computer Science Deptt.
ITM University
Gurgaon, India
arpitanagpal@itmindia.edu

Dr.Deepti Gaur
Associate Professor
Computer Science Deptt.
ITM University
Gurgaon, India
deeptigaur@itmindia.edu

Seema Gaur
Banasthali University
Banasthali
Rajasthan,India

***Abstract*- To reduce the dimensionality of dataset, redundant and irrelevant features need to be segregated from multidimensional dataset. To remove these features, one of the feature selection techniques needs to be used. Here, a feature selection technique to remove irrelevant features has been used. Correlation measures based on the concept of mutual information has been adopted to calculate the degree of association between features. In this paper authors are proposing a new algorithm to segregate features from high dimensional data by visualizing relevant features in the form of graph as a dataset.**

Keywords—Correlation; feature selection; minimum spanning tree; data set.

I. INTRODUCTION

Many real-world data sets consist of a very high dimensional feature space. Each attribute (feature) describing the object (measurements ,event) can be thought of as a different dimension. So, an Objects, is a point in a multi-dimensional space. Data is increasing day by data. To make use of this huge amount of data, Data Mining uses various methods, algorithms and techniques from many fields to extract the useful information. Generally, the data available has many attributes on which it is described. All these are not useful. Good feature subsets contain features highly correlated (predictive of) with the class, yet uncorrelated with (not predictive of) each other [4].One of the feature selection method is needed to remove the features which are irrelevant and redundant. So, the aim of reducing the dimensionality of the data can be fulfilled by removing the irrelevant and redundant attributes before actual algorithm is applied. Feature selectors are placed into two broader categories, Filter and Wrappers. In the filter approach unnecessary attributes are filtered out of the data before applying any other algorithm. This method does not depend on the learning algorithm to be applied. They are usually faster and computationally more efficient than wrapper [2]. Different types of filter approaches have been used like Correlation based Filter[4],Kira and Rendell,1992 attempts to rank features according to a relevancy score.

II. FEATURE SUBSET SELECTION TECHNIQUE

Accuracy of most of the algorithms depends upon the features selected. Irrelevant features can effect the results of the classification or recognition algorithms. So, Feature selection plays a key role in designing pattern recognition and machine learning systems. Feature extraction, is a process of extracting only those features from the original sample data set which are informative and can make the classification task more efficient. We would refer to feature selection, or variable selection, to the process of selecting the most relevant features (attributes) from an initial set of variables the dataset is representing [6].

The frame work of clustering of the graph based feature subset algorithm can be described as in figure 1

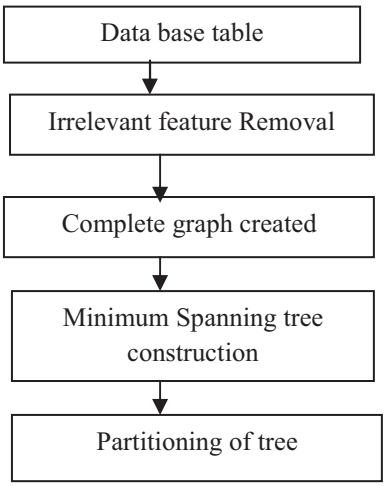

Fig. 1 Depicting the flow of Partitioning of high dimensional features.

Definition of relevance as given by John,Kohavi [6] :

A feature is said to be strongly relevant to sample S if there exist examples A and B in S that differ only in feature xi and have different labels.

The filter method relies on general characteristics of the training data to select some features without involving any learning algorithm. The wrapper method always needs one predetermined learning algorithm in feature selection and utilizes its performance to evaluate and determine the selected features. For each new subset of features, the wrapper method needs to learn a hypothesis (or a classifier). It always tries to find the features which are in compliance with the predetermined learning algorithm resulting in better learning performance, but it is more computationally expensive than the filter model. When the dimension of data increases, that is, the number of attributes (features) becomes huge, due to better computational efficiency of filter method; it is chosen [7].

The similarity between the two features or a feature and a class can be calculated by normalizing a possible relation among the entropies of both the features and the mutual information I(fi,fs) among them where fi and fs are two different features. This normalizing is called calculating the symmetric uncertainity.

One of the information Theoritical method called Mutual information is used to measure the similarity between two features or feature and a class.

Entropy is a measure of uncertainty of a random variable. Let X be a discrete random variable with alphabet Z and probability mass function p(x) = Pr{X = x}, x □ Z. The probability mass function is denoted by p(x). The entropy H(X) of a discrete random variable X is defined by:

$$H(X) = -\sum_{x\in Z} p(x)\log_2 p(x)$$

Entropy is always related to the probabilities of the random variable rather than their actual values.

Mutual information, is a measure of the amount of It is the decrease in the uncertainty of one random variable information that one random variable contains about another random variable. due to the knowledge of the other. The mutual information I(X; Y) is the relative entropy between the joint distribution and the product distribution [5]

$$I(X,Y)=\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (1)$$

If X and Y are discrete random variables following Equations 2 and 4 give the entropy of Y before and after observing X.

$$H(Y)=-\sum_{y\in Y} p(y)\log_2 p(y) \quad (2)$$

$$H(Y|X)=-\sum_{x\in X} p(x)\sum_{y\in Y} p(y|x)\log_2 p(y|x) \quad (3)$$

$$=-\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x) \quad (4)$$

The amount by which the entropy of Y decreases reflects the additional information about Y provided by X and is called the information gain (Quinlan,1993).

Information gain is given by

$$Gain(X,Y)=\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

$$=\sum_{x,y} p(x,y)\log\frac{p(y\mid x)}{p(y)} \quad (5)$$

$$=$$

$$-\sum_{x,y} p(x,y)\log p(y)+\sum_{x,y} p(x,y)\log\ p(y|x)$$

$$=-\sum_{x} p(y)\log p(y)-(-\sum_{x,y} p(x,y)\log p(y|x)$$

$$= H(Y) - H(Y|X) \quad (6)$$

The Information gain is the reduction in uncertainity of Y due to knowledge of attribute X. It is a symmetrical measure that is, the amount of information gained about Y after observing X is equal to the amount of information gained about X after observing Y.

By Symmetry, Gain (X,Y) = H(X) – H(X|Y)

It is observed that the information gain is biased in favour of features with more values that is, attributes with large number of values will appear to gain more information than those with fewer values even if they are not much informative. Symmetrical uncertainty (Press et al1988), compensates for information gain's bias toward attributes with more values and normalizes its value to the range[0,1]:

$$\text{symmetrical uncertainty, } SU(X,Y) = 2*\left[\frac{gain(X,Y)}{H(Y)+H(X)}\right]$$

Using the symmetric uncertainity SU(X,Y) we can calculate the relevant features SU(xi,C) which is the relevance of each feature(xi) with its corresponding class label(C).

The Correlation between each pair of feature xi and xi is denoted by SU(xi,xj).

.

III. ALGORITHM

Input to this algorithm is the dataset D with m features X={ x1,x2,x3,……xm }and class C.

Initially, compute the relevant features SU(xi,C) for each feature. If value of SU is greater than 0 then, keep those features as relevant ones. A relevant feature subset FS={ x1',x2',x3',……..xh'} where h<m is formed.

Now, calculate the correlation SU(x'i,x'j) value for each set of features x'i and x'j where $i \neq j$.

Then construct a graph taking each relevant feature as a vertex. There are h features which are relevant. Each relevant feature acts as a node in the graph. The value of each node in graph is the value of SU calculated above for each feature with its class label.

It is a weighted complete graph with h(h-1)/2 edges and correlation value as weights on edges.

Using Kruskal's Spanning tree algorithm, the number of edges from the complete graph are reduced to (h-1). Now, to partition the tree one of the edge whose weight is greater than the weight of both of its corresponding node's values is removed. This divides the complete feature set into two groups of the features.

Require: D(x1,x2,x3,……xm,C) given data set along with class label of each

Output: fs- selected feature point subset.

Procedure:

1. For i=1: number of features of data do
2. Z=MI(xi,C)
3. Relevance=SU(xi,C)
4. If relevance > 0 then,
5. FS=FS U {XI}
6. For each pair of features {xi,xj} □ FS do
7. Correlation= SU(xi,xj)
8. Construct a 2D FSx FS matrix with value of correlation between each feature.
9. SpanTree=kruskal(M)
10. If SU(x1,x2)< SU (x1,C) && SU(x1,x2)< SU(x2,C)
11. Then, SpanTree=SpanTree-edge(x1,x2)
12. Return SpanTree.

End procedure

IV. Experimental Analysis

Data sets: We have choosen 3 data sets from tunedit.org and archive.ics.uci.edu. Base-hock dataset consist of 256 attributes (features), number of classes is 2 (0 and 1).AR10P dataset has 2400 attributes and it is dived into 10 classes. Pie10P dataset has 2421 features and has 10 classes. These three datasets have been applied on the algorithm. The result obtained has been analyzed and depicted in table 1. It is given on the basis of total number of features which were already there in dataset, number of relevant attributes found after applying the algorithm on this dataset; number of edges of the complete graph and number edges after applying kruskal algorithm. It is depicted in the following Table 1.

TABLE I. Result obtained after applying the Data sets on Algorithm

Data Set Name	*Base-hock*	*AR10P*	*PIE10P*
Number of Features(x)	256	2400	2421
Number of Rows(I)	256	130	210
Number of Classes(T)	2	10	10
Number of Relevant features found	97	2400	2420
Number Edges before applying Kruskal	4,657	2878800	3014110
Number of edges after applying Kruskal	96	2399	2419
Domain	Text	Image,Face	Image,Face

The symmetric uncertainty value between each feature and its corresponding class value is calculated to get whether a feature is relevant or not. These values are depicted in figure 2. Figure 3 given below shows that the number of relevant attributes found from the Base-hock data set. Out of total 256 attributes in database table. Number of relevant attributes came out to be 97. Then correlation is calculated between each of the 97 attributes and all values found are given as weight on the edge of the graph. This forms weighted complete graph with 4,657 edges. Figure 4 displays that the number of edges formed before applying kruskal's which is 4,657. It also shows a weighted matrix. It has the value of correlation (weight) between each feature. These values are used as weight on edge of the weighted graph. Then, Kruskal's algorithm is applied and number of edges get's reduced to 96. This is depicted in figure 5 below.

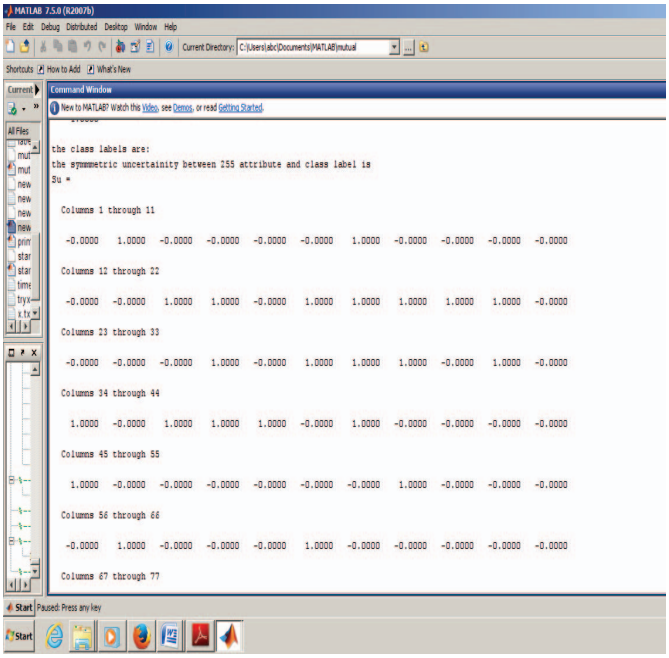

Fig. 2. Shows the symmetric uncertainity values between each attribute and class label.

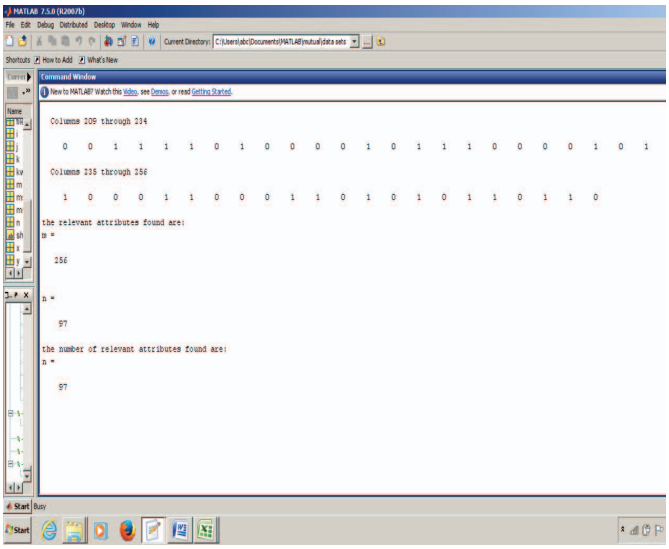

Fig.3. Showing number of relevant attributes found are: 97 after running on base-hock data

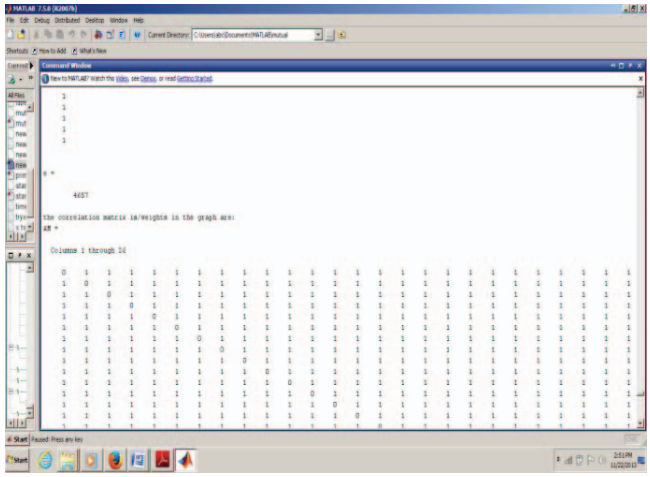

Fig. 4. On base-hock data, correlation matrix is depicted.

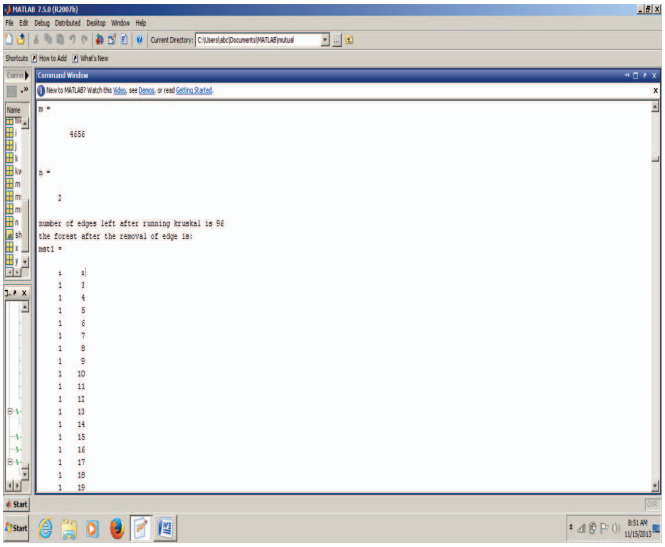

Fig.5: Depicting the edges after applying kruskal

V. Conclusion

In this paper, we have presented a technique to reduce the dimensions of data using filter approach. Mutual information is used as a feature selection technique. Relevant features from the high dimensional data set are viewed as a complete weighted graph. Correlation value using mutual information has been calculated between features and it act as weight on the edges of the graph. The number of edges in the complete graph has been greatly reduced using Minimum Spanning tree algorithm. Reducing these edges can help to easily group the features for any further use. A criterion has been used to further cut the tree from one of the edge and features get partitioned into two groups. It can be further partitioned using different cut algorithms.

REFERENCES

[1] Quinbao Song,Ni,Wang," A fast clustering based feature Subset Selection Algorithm for high Dimensional Data", IEEE Transaction on knowledge and data Engineering, Vol.25,No.1,January 2013.

[2] Huang,Cai,Xu," A filter Approach to feature selection based on mutual information", 5th IEEE int. Conf.,2006.

[3] Tangsen Zhan, Zhou," Clustering Algorithm on High-dimension Data Partitional Mended Attribute", 9th International Conference on Fuzzy Systems and Knowledge Discovery 2012.

[4] Mark A. Hall, Smith, "Feature Selection for machine learning: Comparing a correlation based Filter Approach to the Wrapper", American Association for Artificial Intelligence, 1998

[5] Cover and Thomas, "Elements of Information Theory, Chapter 2", 1991 John Wiley & Sons

[6] G.H. John, R. Kohavi and K. Pfleger, Irrelevant features and the subset selection problem. In: Proceedings 11th International Conference on Machine Learning New Brunswick, NJ, , Morgan Kaufmann, San Mateo, CA, pp. 121–129, (1994).

[7] Lei Yu, Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003

[8] Reena Srivastava, MM Gore "A Micro-Clustering based Method for Multi-Class Classification in Multi-Relational Databases", MNNIT Allahabad, 2008.

[9] Sotoca,Pla, "Supervised feature selection by clustering using conditional mutual information based distances", Elsevier Science direct, Pattern Recognition 43 pp.2061-2081, 2010.

[10] Sanmay Das, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection", Proc. 18th Imt'l Conf. Machine Learning, pp.74-81, 2001

[11] Caiming Zhong, Duoqian Miao, Pasi Fränti, "Minimum spanning tree based split-and-merge: A hierarchical clustering method", Elsiver Science Direct, Information Sciences 181 (2011) 3397–3410

[12] O. Grygorash, Yan Zhou, Z. Jorgensen, "Minimum Spanning Tree Based Clustering Algorithms", Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06) 2006

[13] Gavin Brown, Adam Pocock, Ming-Jie Zhao, Mikel Luj, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection",Journal of machine learning pp. 26-66 2012