Scientific Journal of Impact Factor (SJIF): 3.134

E-ISSN (O): 2348-4470
P-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 12,December -2015

FEATURE SUBSET SELECTION FOR HIGH DIMENSIONAL DATA BASED ON CLUSTERING

Prof. S.N.Zaware[1], Heena Shaikh[2], Sheefa Shaikh[3], Asmita Orpe[4], Pooja Rokade[5]

[1,2,3,4,5]*Computer Department, AISSMS IOIT Pune*

Abstract- *Feature selection is the process of evaluating and extracting desired data which can be grouped into subsets which retain the integrity of original data. A feature selection algorithm should be efficient and effective. Efficient means minimum time required and effective means quality of generated subset is not compromised. Our system proposes an algorithm which consists of following steps: Markov Blanket, Shannon Infogain, Minimum Spanning Tree, Tree Partition, Gaussian Distribution, Bayesian Probability. Applying these steps we get the desired subset from the clusters. Our system ensures to remove irrelevant data along with redundant data which most of the systems fail to perform.*

Keywords*-Markov Blanket, MST Creation, Gaussian Distribution, Shannon Infogain, Bayesian Probability, Fuzzy Logic*

I. INTRODUCTION

The basic idea of clustering is based on the fact that as the size of data set increases, the complexity of the cluster generation also increases (clusters are group of similar objects). So we have proposed a system that reduces the dataset by eliminating redundant and irrelevant data to enhance the quality of the cluster and speed up the cluster generation process. Previous algorithm could successfully remove the irrelevant data, but failed to remove the redundant data which degraded the quality of the cluster and provided ambiguous knowledge from that data. The system aims at increasing learning accuracy, and improving result comprehensibility.

Our system takes high dimensional dataset as input which consists of text and numeric data. The system preprocesses the data by applying certain steps such as special symbol removal, stemming, stop word detection and removal. After the data is preprocessed, Markov Blanket is applied on it which helps in removing irrelevant data. After this Shannon infogain is applied which helps in removing redundant data. It is followed by MST creation and partition. Further, Gaussian distribution is applied on each partition. Then interest ratio and Bayesian probability of features is calculated and final subset is generated from these clusters.

.

II. LITERATURE SURVEY

1. Qinbao Song, Jingjie Ni and Guangtao Wang, A Fast Clustering Based Feature Subset Selection Algorithm For High Dimensional Data, IEEE Transaction on knowledge and data Engineering, vol.25, no.1,2013.
   Objective: To select proper feature subset from the given data.
   Advantage: Feature subset selection, filter method, feature clustering, graph based clustering, kruskal's algorithm.

   Limitation: Prior algorithms are having issues related with efficiency as the process took much time.
2. I. Kononenko, estimating Attributes: Analysis and Extensions Of RELIF, Proc. European Conf. Machine Learning, pp. 171-182, 1994
   Objective: Relief which selects the relevant features by using a statistical method.
   Advantage: It requires only linear time in the number of given features. The Relief method is noise tolerant. It does not depend on heuristics and applicable even if the feature interact with each other.
   Limitation: It has non linear optimal feature set size.

3. Lydia Boudjeloud and Francois Poulet, Attribute Selection for High Dimensional Data Clustering, 2007
   Objective: Feature subset selection, filter method, feature clustering for feature extraction.
   Advantage: As it uses filter method for subset selection it is faster.
   Limitation: In high dimensional space finding clusters of data objects is challenging due to high dimensionality.

4. Luis Talavera, Feature Selection as a Preprocessing step for hierarchical clustering,2000
   Objective: The feature selection is done in a hierarchical manner by preprocessing step.
   Advantage: Proposed system removes the induced immaterial features.
   Limitation: Due to poor efficiency immaterial features get introduced.

5. M. Scherf and W. Brauer, Feature Selection by Means of Features Weighting Approach, Technical Report FKI-221-97, Institut fur Informatik, Technische Universitat Munchen, 1997.

Objective: Selecting a set of features which is optimal for given optimization task using the robust and flexible filter technique like EUBAFES.
Advantage: It computes binary features weights and therefore solution in the feature selection sense and also gives detailed information of relevance by continuous weights.
Limitations: This methodology is used to detect only different types of relevant features but cannot determine redundant features from the dataset.

III. WORKING

Our system consists of the following steps to reduce high dimensional data and acquire relevant feature subset from clusters. The steps are:

3.1 Preprocessing: Initially the high dimensional data taken as input is preprocessed by applying the steps such as:

3.1.1 Special symbol removal: In this the special symbols such as , ? ! etc are replaced by blank spaces.
3.1.2 Stop word removal: The stop words such as and, or, the, if, at etc are removed.
3.1.3 Stemming: in this the prefixes and suffixes of the words are removed. E.g. playing becomes play.

3.2 Markov Blanket: It is the blanket or cover of other nodes around a specific node and contains enough knowledge to predict the behavior of that node. Suppose A is the node, then the blanket consists of all of A's direct children and direct parents and also it's children's direct parents. Node A is not affected by nodes outside the blanket. This makes A conditionally independent of all nodes in the model. It removes features which are really unnecessary.

3.3 Shannon infogain: It is used to quantify information, i.e. numerical value is assigned to each feature which is directly proportional to the importance of the feature. The Shannon infogain values range from 0 to 1. Shannon infogain formula:

$$H=-\sum_{i=1}^{N} p_i(x) \log p_i(x) \quad \text{where,}$$

N- total no. of relevant features
'H' is also called the entropy. More the entropy, more the information gain. Redundancy is reduced by this step.

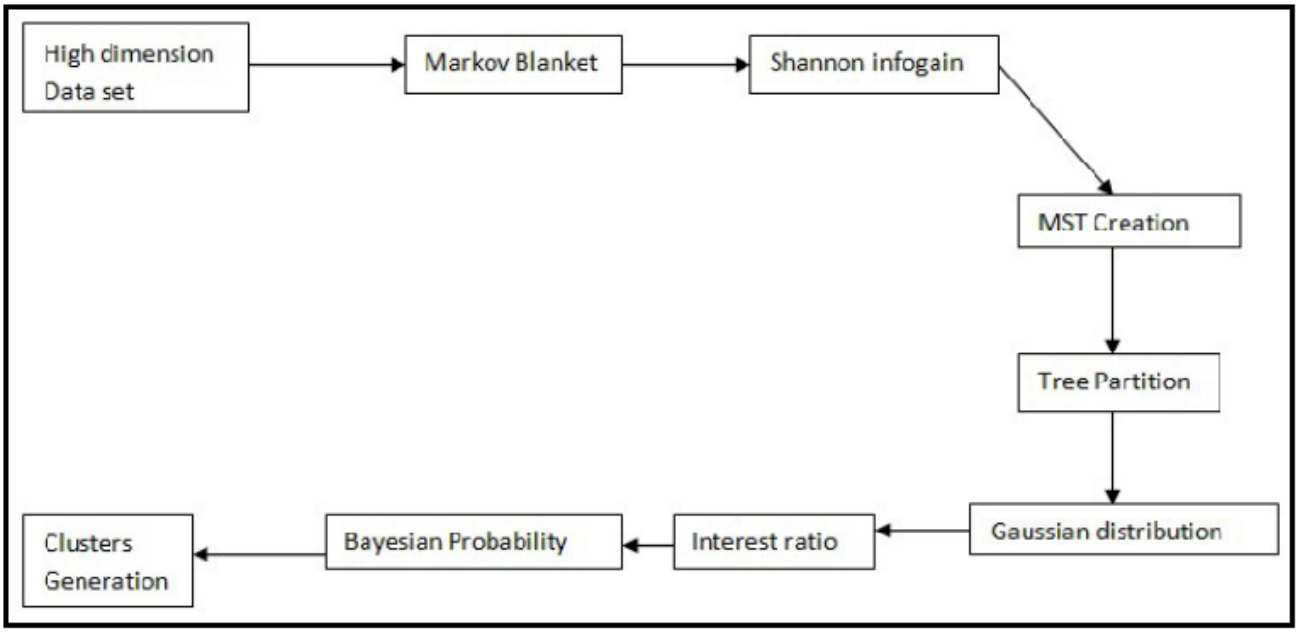
3.4 MST Creation: A minimum spanning tree is created from the Shannon infogain values. The highest infogain value is considered as the root. If there are one or more features with highest infogain values, the 1 is considered as the root. The nodes in the tree represent the features.

3.5 Tree Partition: The MST is partitioned into five parts according to fuzzy logic levels (*very low, low, medium, high, very high*). Further fuzzy logic levels are applied to each of the five parts, which are further disintegrated into five parts. This cycle continues until all the features are grouped into clusters of similar features. Because of this raw clusters are generated.

3.6 Gaussian distribution: It is used to find the distribution of density of all the features in a cluster. Gaussian distribution is applied to all the clusters. After the Gaussian distribution is applied the features are distributed in the form of 'bell shaped curve'. In Gaussian function, μ is the mean of all Shannon infogain values in a cluster. σ is used to calculate the amount of dispersion or how the values are spread throughout in a cluster.
Gaussian function: $p(x)=(1/\sigma\sqrt{2\pi})\exp\hat{}(-(x-\mu)^2/2\sigma^2)$ where,
μ- mean of Shannon infogain values
σ- std deviation
σ^2- variance
The std deviation is calculated as: σ= Shannon infogain values - μ
The highest density area or the peak or the 'bell curve' is selected as the main cluster which consists of the highest density features.

3.7 Interest ratio and Bayesian Probability: The interest ratio of features of lower density is calculated and used in calculating Bayesian probability, so as to allot lower density features to appropriate clusters. This refines the raw clusters.
posterior probability=(likelihood* prior probability)/ evidence
The final subset of features is derived from the refined clusters.

IV. SYSTEM ARCHITECTURE



4.1 system architecture

V. CONCLUSION

Our proposed system is intended to reduce the high dimensional data which consists of combination of text and numeric data to extract desired feature subsets in the form of clusters by removing redundant and irrelevant data through a series of methodologies mentioned above. This approach can be applied in various fields like, stock exchange, weather forecast or any organizational database.

REFERENCES

[1] Qinbao Song, Jingjie Ni and Guangtao Wang , "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE Transaction on knowledge and data Engineering, vol.25, no.1 2013.

[2] Kira K. and Rendell L.A., "The feature selection problem: Traditional methods and a new algorithm", In Proceedings Of Ninth National Conference On Artificial Intelligence, pp 129-134 1992

[3] Mark A. Hall, Feature Selection for Discrete and Numeric Class Machine Learning 2000.

[4] Jianchao Han, Ricardo Sanchez, Xiaohua Hu, T.Y. Lin, "Feature Selection Based on Relative Attribute Dependency: An Experimental Study", 1993

[5] Scherf M. and Brauer W., " Feature Selection By Means of a Feature Weighting Approach", Technical Report FKI-221-97, Institut fur Informatik, Technische Universitat Munchen, 1997.

[6] I. Kononenko, "Estimating Attributes: Analysis and Extensions Of RELIF", Proc. European Conf. Machine Learning, pp. 171-182, 1994

[7] Kale Sarika Prakash, P.M.J Prathap, "A Survey on Iceberg Query Evaluation Strategies", International Journal of Modern Trends In Engineering and Research, e-ISSN No.2349-9745, July 2015

[8] Lydia Boudjeloud and Francois Poulet, "Attribute Selection for High Dimensional Data Clustering", 2007

[9] Luis Talavera, " Feature Selection as a Preprocessing step for hierarchical clustering", 2000

