Feature Subset Selection for High Dimensional Data Based On Clustering

Asmita Orpe[1] Heena Shaikh[2] Pooja Rokade [3] Sheefa Shaikh[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Savitribai Phule Pune University, Pune

Abstract— Feature selection is the process of examining, evaluating and extracting required data which can be clustered into subsets which contain and retain the integrity of original data. A feature selection algorithm should be adept and productive. Adept means minimum time required and productive means quality of generated subset is not compromised. Our system proposes an algorithm which consists of following steps: Markov Blanket, Shannon Info gain, 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 eliminate. Irrelevant features are the extraneous features or data objects, whereas the redundant ones are the repetitious features. These data objects tend to consume memory and do not contribute in generating accurate results.

Key words: Bayesian Probability, Fuzzy Logic, Gaussian Distribution, Markov Blanket, MST Creation, Shannon Info gain

I. INTRODUCTION

Feature selection has been an active and fruitful field of research and development for decades in statistical pattern recognition. In theory, more features should provide more discriminating power, but in practice, with a limited amount of training data, excessive features will not only significantly slows down the learning process but also produce ambiguous results. 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 pre-processes the data by applying certain steps such as special symbol removal, stemming, stop word detection and removal. After the data is pre-processed, 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.[1]

II. PREVIOUS WORK

A. 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.

B. 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.

C. M. Scherf and W. Brauer, Feature Selection by Means of Features Weighting Approach, Technical Report FKI-22197, 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.

D. 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.

E. 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.

F. M.A. Hall, Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, Proc. 17th Intl Conf. Machine Learning, pp. 359-366, 2000.

Objective: This paper describes a fast, correlation-based filter algorithm that can be applied to continuous and discrete problems. Experiments using the new method as a

preprocessing step for naive Bayes, instance-based learning, decision trees, locally weighted regression, and model trees show it to be an effective feature selector as it reduces the data in dimensionality by more than sixty percent in most cases without negatively affecting accuracy.
Advantage: This paper presents a new approach to feature selection, called CFS, (Correlation-based Feature Selection) that uses a correlation based heuristic to evaluate the worth of features. The algorithm is simple, fast to execute and extends easily to continuous class problems by applying suitable correlation measures.
Limitation: The computation time of this method is intensive.

III. EXPERIENTIAL SETUP

Our proposed system is implemented using the following resources.

A. Hardware Resources:

- Processor: Dual Core of 2.2 GHZ,
- Hard Disk: 100 GB,
- RAM : 2GB

B. Software Resources:

- Platform: JAVA
- Technology: JDK 1.6 and Above
- IDE: Netbeans 6.9.1
- Data base : MySQL 5.0

These resources are easily available. Database like MySQL is open source and can be downloaded and then unzip the setup file and execute the downloaded MSI File. Then choose typical type of setup and then install. Choose detailed configuration and then select developer machine. Select multifunctional database. Select decision support system/ OLAP. Then enable both strict mode and TCP/IP networking. Select standard character set and select windows as service. Finally set password, select all options and click execute and the click finish.

For NetBeans, first click on setup, then double click windows exe file. Then NetBeans IDE installer will be launched. The wizard page will display default packages. Click customize button and select all packages and click ok. Accept license agreement. Then click unblock if an alert appears. After complete installation click finish.

IV. ARCHITECTURE

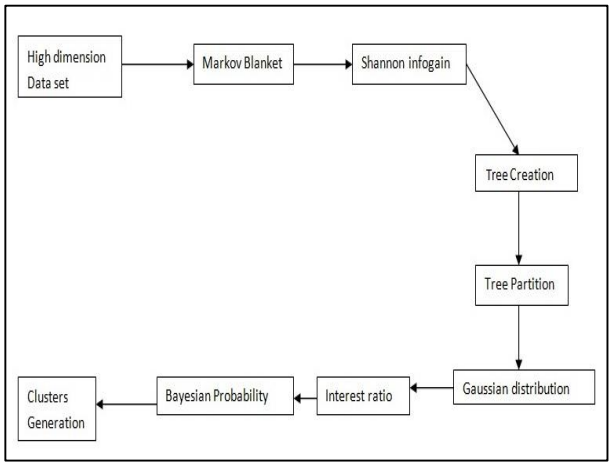

Fig. 1: Architecture diagram

V. WORKING

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

A. Pre-processing:

Initially the high dimensional data taken as input is pre-processed by applying the steps such as:
1) Special symbol removal: In this the special symbols such as , ? ! etc are replaced by blank spaces.
2) Stop word removal: The stop words such as and, or, the, if, at etc are removed.
3) Stemming: in this the prefixes and suffixes of the words are removed. E.g. playing becomes play.

B. Markov Blanket:

It is the blanket or cover of other nodes around a specific node and contains enough knowledge to predict the behaviour 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

C. 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:

$$H=-\sum_{i=1}^{N} pi(x) \log pi(x) \quad \text{....(e.q. 1)}$$

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.

D. 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.

E. 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.

F. 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/σ√2□) exp^(-(x-μ)2/2σ2)....(e.q. 2)
where,
μ- mean of Shannon infogain values
σ- Std deviation
σ2- variance.
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.

G. 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 ....(e.q. 3)
The final subset of features is derived from the refined clusters.

VI. Result and Analysis

Some experimental evaluations are performed to show the effectiveness of the system. And these experiments are conducted on windows based java machine with universally used IDE Netbeans. Also the numbers of retrieved clusters from the data set is used to set benchmark for performance evaluation. Numbers of relevant retrieved clusters from the dataset is used to show the effectiveness of the system. Below are the definition of the used measuring techniques i.e. precision and recall. Precision: it is a ratio of numbers of proper clusters retrieved to the sum of total numbers of relevant and irrelevant novelty clusters retrieved. Relative effectiveness of the system is well expressed by using precision parameters. Recall: it is a ratio of total numbers of relevant clusters retrieved to the total numbers of relevant clusters not retrieved. Absolute accuracy of the system is well narrated by using recall parameter. Numbers of scenarios presents where one measuring parameter dominates the other. By taking such parameters into consideration we used two measuring parameters such as precision and recall. For more clarity let we assign:
X = the number of relevant clusters retrieved
Y= the number of relevant clusters are not retrieved
Z = The number of irrelevant clusters are retrieved.
So, Precision = (X/ (X+ Z))*100….(e.q. 4)
Recall = (X/ (X+ Y))*100….(e.q. 5)

By observing the precision graph it is clear that the average precision obtained by using fast clustering method is approximately71.8From Recall graph it shows 82.3recall for the cluster detection method.

By comparing these two graphs we can conclude that the cluster extraction by fast clustering techniques method gives high recall value compare to the precision value and this the good sign of any cluster formation methods.

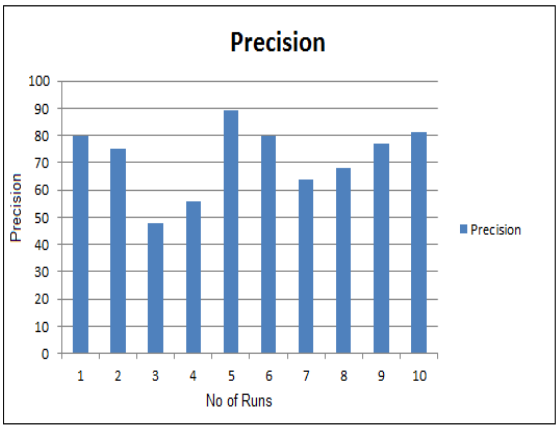

Fig. 2: Precision graph

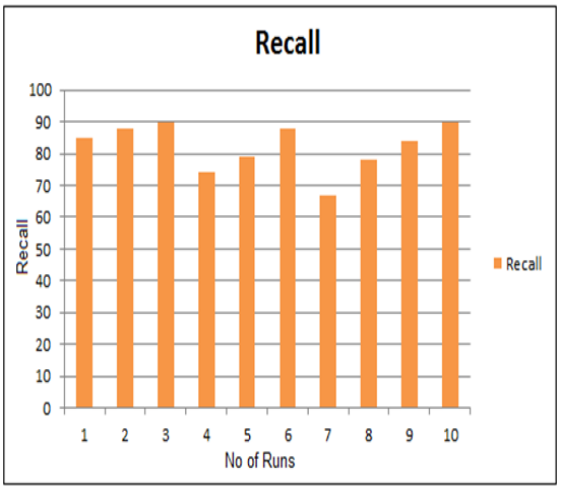

Fig. 3: Recall graph

VII. Conclusion

Our proposed system is intended to reduce the high dimensional data i.e. which consists of combination of text and numerics to extract desired feature subsets in the form of cluster by removing redundant and irrelevant data through a series of methodologies like Markov Blanket, Shannon Infogain, MST, Gaussian Distribution and Bayesian Probability. We can insert data _les in our system from external and internal hard disk. The intended areas for the application of the project is Weather data, news data, stock exchange, etc as our system is integrated in trend generation system, market analysis. With further improvements and developments in this system, we can also use image data along with text and numerics. Further it can be applied on data generated from CCTV cameras, or any image related application.

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] 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

