ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A REVIEW ON CLUSTERING-BASED FEATURE SUBSET SELECTION ALGORITHM FOR HIGH DIMENSIONAL DATA

Madhuri B Patil*, Anil Rao
Department of Computer Science and Engineering,IET Alwar,Rajasthan,India

ABSTRACT

In HD dataset, feature selection involves identifying the subset of good features by using clustering approach. Feature selection involves removal of irrelevant and redundant features which are the essential data preprocessing activities for effective data mining. A Clustering based approach for good feature selection evaluated from both the efficiency and effectiveness points of view. Efficiency relates the time required to find a subset of good features while the effectiveness is related to the quality of the subset of features. The feature selection algorithm for high dimensional data produces the more compatible results as the original entire set of results based on search strategies, evaluation criteria, and data mining tasks. It reveals unattempted combinations, and provides guidelines in selection of feature selection algorithms. FAST algorithm for feature subset selection works in two steps. The first step involves distribution of feature subsets into clusters by using graph-theoretic clustering methods and the second step involves selection of most useful ,efficient features that is strongly related to the target classes which form the subset of good features. In FAST algorithm to increase the the efficiency we adopted efficient Minimum Spanning Tree clustering method. Based on some of these criteria, a clustering-based feature selection algorithm for HD data is proposed and experimentally evaluated in this paper.

KEYWORDS: Minimum Spanning Tree, Good Feature subset selection, Clustering

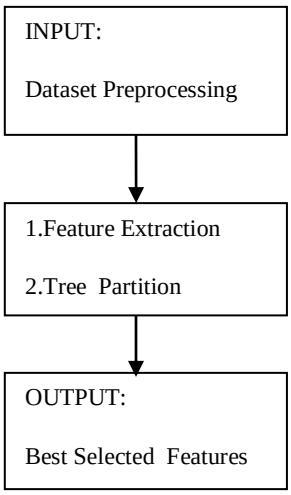
INTRODUCTION

Data mining is the process that analyzes and converts mountains of data into nuggets. Data preprocessing is essential for successful data mining. Feature selection is one of the important, efficient and frequently used techniques in data preprocessing for successful data mining. The main aim of this project is choosing a subset of good features with respect to the target concepts. Good feature subset selection is an efficient approach that minimizes dimensionality, removes irrelevant and redundant data increases learning accuracy, and improves result comprehensibility. Based on some criteria always an optimal feature subset is to be evaluated. There are many filter selection methods are available by which feature selection process is evaluated. Clustering is collection of similar data in which the datasets are divided into number of clusters. The features in each cluster are relatively independent on the features in another cluster. Many feature selection algorithms effectively handle irrelevant features but fail to avoid redundant features, while some other algorithms effectively handles both. FAST feature subset selection algorithm used to minimize the time complexity and it increases the learning accuracy. Data model for the system has some rules for comparing data. The methods used in the existing system are wrapper, filter, embedded and hybrid but these methods had some bugs such as large computational complexity, less accuracy, expensive, limited feature subset selection, and removal of redundant features which are overcome by our proposed system. Experiments are carried out to compare FAST and other representative feature selection algorithms such as FCBF, ReliefF, CFS, Consist, and FOCUS-SF regarding with four well known classifiers namely the instance based IB1, the rule based RIPPER, Probability based Naive Bayes and the tree based c4.5. In the proposed system architecture, dataset preprocessing involves irrelavent and redundant features removal. It brings immediate effects and results for the applications as well as increases the speed of data mining algorithms.After dataset preprocessing ,the best features are extracted to construct a tree.Here, we will use k-means clustering

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

to partition tree.Here, we will use k-means clustering to partition the clusters.Each cluster will contain feature subsets and the features in different clusters will be relatively independent. One of the popular cluster analysis in data mining is the K-means clustering. It partitions n-observations into K-clusters where each cluster related to the cluster with the nearest mean which serve as a prototype for that particular cluster. K-means clustering results into the partition of datasets into different voronoi cells or the clusters. Finally, using PRECISION and RECALL the best selected features are retrieved. PRECISION is also known as positive predictive value which is the fraction of relevant retrieved instances. RECALL is the Fraction of retrieved relevant instances.

Figure:



Proposed System Architecture

MATERIALS AND METHODS

All the algorithms developed before this mainly focused on identifying or searching relevant features. One of them is Relief, which according to the distance based criteria function finds out each feature according to its ability to classify the instances under different targets. However, it is ineffective in identifying or removing redundant features which are highly correlated. Relief-F is the extension of Relief which deals with multiclass problems and works with incomplete and noisy datasets but still it was ineffective in identifying redundant features. Redundant and irrelevant features affects the accuracy and speed of learning algorithms hence needs to handle effectively. Other algorithms, CFS, FCBF, CMIM takes into account redundant features.CFS stands for Correlation-Based Feature Selection, which is achieved by hypothesis that a best feature subset is one that contains features that are highly correlated with the target concepts but they are uncorrelated with each other. It reduces the complications occurred in selecting the feature subset selection& increases the accuracy in classification. FCBF stands for Fast Correlation Based Filter solution. As its name indicates, it is a fast filter process which identifies redundant as well as relevant features without any correlation based pairwise analysis. Other algorithm, CMIM stands for Conditional Mutual Information Maximization which picks each feature iteratively. Apart from all these algorithms, FAST algorithm uses clusters to select the features. In the area of text classification, hierarchical clustering has been used. It is also used to select the features which are on the spectral data. The FAST algorithm uses methods depend upon minimum spanning tree to divide the features among clusters. In FAST algorithm, the limitation to specific type of data has been overcomed. FAST evaluate its results into two steps. In first step, relevance information is calculated using suitable method to construct a tree. It involves dividation of tree into different clusters using graph theoretic clustering methods. In the second step, most likely features are selected from each cluster to form the subset of good features. FAST is quite different from other hierarchical clustering. FAST algorithm checks whether or not method is useful in practice and allows other researchers to confirm their results with 35 publically available datasets. In the proposed algorithm, we extend the FAST, to maximize the accuracy for high dimensional data using Precision and Recall

Results and discussion

We will firstly discuss about the disadvantages or the limitations that this feature subset selection algorithm incurs:

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

i)Discarding irrelavent features before Minimum Spanning tree has been developed leads to less accuracy of good feature selection.
ii)Other procedures for good feature selection requires more computation cost.
To remove these bugs or these disadvantages is the plan of this research paper.The solution to the above is as follows for the feature subset selection of high dimensional data:As we have already discussed, there are lots of methods of feature selection.To minimize the cost of computation and the problem of less accuracy we proposed the PRECISION and RECALL technique for good feature selection after minimum spanning tree construction, which will provide good accuracy and time complexity. Using k-means clustering , the features are divided into different clusters and good features are selected from that clusters.The block diagram represented above describes this method. This method provides better accuracy and time complexity.

MODULES

The modules may be change or vary by requirement, it is not necessary that they may be same in the future too. We have introduced some of the modules here ,they are as follows:

1. *User*: Here, the User needs to be authenticated or genuine to access the details. Before accessing the details the user needs to have his own account or he should get register first.
2. *Clustering Method* :Here, the proper method for clustering is used which will cluster different words into groups. Words in the same group are relatively independent of words in the other group.
3. *Algorithm for Feature subset Selection:* By using the clustering algorithm, we have to develop the algorithm which will increase not only the effectiveness of feature subset selection but also it will provide better accuracy and time complexity.
4. *Complexities:* Time and space complexity for FAST involves computation of many processes like SU,TR and F.

CONCLUSION AND FUTURE SCOPE

A good feature subset involves identifying the most useful features that will produce most compatible results as compared to the target classes. The algorithm is evaluated from both efficiency and effectiveness point of view. Here ,we presented an approach for feature selection for High dimensional data.The algorithm involves three major steps:

i)It involves removing redundant and irrelevant data for identifying the subset of good features.
ii) Features are divided into different clusters based on minimum spanning tree construction.
iii) Partitioning of MST and best features are selected from different clusters of MST.
It increases accuracy and time complexity. That means the Performance Of the algorithm when compared with other algorithms is increased upto7 to 9%. This performance may increase or decrease in future.

REFERENCES

[1] Sergio Francisco da Silva, Marcela Xavier Ribeiro,Joao do E.S. Batista Neto ,Caetano Traina -JrAgma J.M. Traina ,"Improving the ranking quality of medical image retrieval using a genetic feature selection method" , Decision Support Systems, 2011
[2] Yu L. and Liu H., Redundancy based feature selection for microarray data,In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 737-742, 2004
[3] Xing E., Jordan M. and Karp R., Feature selection for high-dimensional genomic microarray data, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 601-608, 2001.
[4] 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 YEAR 2013.
[5] Molina L.C., Belanche L. and Nebot A., Feature selection algorithms: A survey and experimental evaluation, in Proc. IEEE Int. Conf. Data Mining, pp 306-313, 2002
[6] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Selection through Clustering, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp
[7] L. Hawarah, A. Simonet, and M. Simonet, "A probabilistic approach to classify incomplete objects using decision trees," in DEXA, ser. Lecture Notes in Computer Science, vol. 3180. Zaragoza, Spain: Springer, 30 Aug.-3 Sep. 2004, pp. 549–558.
[8] Krier C., Francois D., Rossi F. and Verleysen M., Feature clustering and mutual information for the selection of variables in

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

spectral data, In Proc European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning, pp 157-162, 2007.

[9] Yu L. and Liu H., Efficient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Reseach 10(5), pp 1205-1224,2004

[10] L.Ladha, T.Deepa, "Feature Selection Methods AndAlgorithms", International Journal on ComputerScience and Engineering (IJCSE), Vol. 3 No. 5 May2011

[11] Raman B. and Ioerger T.R., Instance-Based Filter for Feature Selection,Journal of Machine Learning Research, 1, pp 1-23, 2002.

[12] Oliveira A.L. and Vincentelli A.S., Constructive induction using a nongreedystrategy for feature selection, In Proceedings of Ninth International Conference onMachine Learning, pp 355-360, 1992.

[13] Park H. and Kwon H., 'Extended Relief Algorithms in Instance-Based Feature Filtering', In Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology, pp 123-128, 2007.

[14] Sunita Beniwal and Jitendar Arora (2012), 'Classification and Feature Selection Techniques in Data Mining', International Journal of Engineering Research and Technology, Volume 1 Issue 6, August 2012

[15] 'Fleuret. F (2004), 'Fast Binary Feature Selection with Conditional Mutual Information', Journal of Machine Learning Research 5(2004) 1531–1555.

[16] Huan Liu, Hiroshi Motoda (2002), and Lei Yu, 'Feature Selection with Selective Sampling', Proceedings of 19th International Conference on Machine Learning, pp. 395-402, 2002.

AUTHOR BIBLOGRAPHY

Madhuri B. Patil
Has received degree in BE computer Engineering from Pune University,India in 2010.She is currently a master student in the Department of Computer Science and Engg,RTU Kota University.
Her research focuses on feature subset selection.
mail: madhur299@gmail.com

Anil Rao
He is currently working as an Assistant Professor in IET Alwar,Rajasthan
Email: anil.alw@gmail.com