Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/XRGDH

IMPROVING CLUSTER BASED FEATURE SELECTION USING MODIFIED MINIMUM SPANNING TREE

Dr. A. SURESH
Department of Computer Science Sona College of Arts and Science, Salem, India

A. KALEEMULLAH
Department of Computer Science, Mazharul Uloom College, Ambur, India

ABSTRACT

The issue of feature subset selection for data of high dimensionality in classifying opinions has been investigated in this paper. Feature selection schemes select the crucial feature subset producing similar or superior classification outcomes compared to the original feature set obtained. Despite their higher efficiencies, wrapper-based feature selection schemes have high overheads of computation. However, the issue is the problem is Non-deterministic Polynomial (NP) hard. This work suggests the clustering scheme based on MST optimized by the Group Search Optimization (GSO) for efficient selection of features. The Amazon camera review dataset is used for evaluating the suggested scheme. Experiments are conducted to evaluate the proposed method and compared with the MRMR feature selection, FCM clustering, and MST based clustering.

Keywords: *Sentiment analysis, Feature Selection, Minimum Spanning Tree (MST), Data Analytics, Clustering and Group Search Optimization (GSO).*

1 INTRODUCTION

Feature selection in machine learning is also referred to as changeable subset selection. This refers to the method of choosing a subset of pertinent features for building a prototype. When correlated models are being built, the following advantages are exhibited by the feature selection techniques: Interpreting the implicit meaning using the improvised prototype. Smaller duration of training and increased generalization by overfitting are other benefits. Feature selection detects crucial features for forecasting and aids in the data analysis process. When some of the good features are selected as per the target concepts, feature selection can effectively decrease dimensionality, eliminate redundant data, enhance the precision of learning and improvise the understand ability. There are many irrelevant features as well as repeated ones contained in data of high dimensionality. While the repeated features do not provide any more useful information that the chosen ones, the irrelevant features do not provide, under any situation any useful information [1].

For training the opinion mining system, there are several techniques used by the researchers. Obtaining an effective feature set in most machine language applications is the crucial step in sentiment classification. There are many applications in which there is a huge use of sentiment analysis. There are many viable tasks it can perform including detecting the attitude or the opinion of the customers regarding the product or the services. The product feedback and the service reviews will help make reasonable decisions .For instance, a new person visiting a city can be aided by the restaurant reviews for locating a good place. In the same manner, a movie can be decided as worthy of watching or not by referring to the movie reviews [2].

Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/XRGDH

II.RELATED WORKS

For high dimensional data, an innovative feature subset selection algorithm on the basis of clustering for has been suggested by Kumari & Naidu [7]. The features of this algorithm are: (i) eliminating redundant features (ii) Building minimum spanning tree from comparative ones (iii) After portioning the MST, the representative features are selected. There are features comprised in a cluster. Every cluster is taken to be a single feature and there is a reduction in dimensionality. Five of the popular feature selection schemes exist with which the performance of the suggested algorithm can be compared. These include FCBF, ReliefF, CFS, Consist, and the FOCUS-SF on thirty-five openly accessible image microarray as well as text data from 4 various sections of the chosen features. For RIPPER, C4.5 and Naïve Bayes, the suggested algorithm has the optimal selected features, runtime and accuracy of classification.

Different techniques for decreasing the dimensionality in high dimensional microarray cancer data have been presented by Hira & Gillies [8]. There is an increase in the quantity of data which is to be analyzed. This has made it essential to allow decrease in dimensions for obtaining pertinent outcomes. There are various feature selection as well as feature extraction schemes that have been described and compared. Along with them, their benefits and disadvantages are also contrasted. Additionally, this work has also presented various schemes for including the prior knowledge from different biological sources which is a technique of enhancing the precision and decreasing the computational complexity of the schemes that are already present.

A feature selection scheme has been designed by Kaveri& Asha [9]. This scheme makes use of the accurate relevance schemes. This work uses relevance measures “Symmetric Uncertainty (SU)” for effectively selecting the relevant attributes. The chosen attributes are then partitioned into clusters on the basis of “graph theoretic” clustering scheme by making use of a relevance measure referred to as “Conditional Mutual Information (CMI)”. To choose the attributes corresponding well to target class and also the one which represents the cluster in the best possible way, the relevance measure “symmetry uncertainty” has been selected. This leads to the accurate and independent subset of features. This technique helps to produce smaller as well as more precise subset of features improving the act of machine learning functions like the Naive Bayes classifier.

Pullela et al., [3] suggested clustering scheme having two stages for feature selection. The features are clustered in the first stage and in the second, for decision making and further use in data mining, the most representative features are extracted. This data of high dimensions is used as input and for formulation of clusters, clustering has been performed. There are many features in every cluster. The selected subset of features is the best features that are representative of features. The input used is the high dimensional data. For the formation of clusters, clustering process is employed. There are several features in each cluster and the subset of features selected contains the best features that are representative of the features chosen. This has several utilities that can be used in real world applications. These chosen features help find newer avenues scope for further data mining purposes. The experimental outcomes have been promising. A significant direction for future work is the construction of a prototype comprising generalized features that can help adapt to different high dimensional datasets for various domains.

Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/XRGDH

A feature subset selection algorithm for huge dimensional data which has been based on clustering has been suggested by Elankavi et al., [4]. These are the features of the algorithm- 1) eliminating irrelevant features. 2) Using relative trees for construction of minimum spanning tree and 3) dividing the MST and choosing delegate features. There are features contained in the selected algorithm. After treating each cluster as a single feature, there is a drastic decreased of dimensionality. The best proportion of the selected features has been obtained from the suggested algorithm, including the optimal runtime, the optimal classification precision for Naïve Bayes, C4.5, RIPPER including the second best classification precision for IB1.

III. METHODOLOGY

Extracting opinion of the users regarding a product or a service from the review documents is referred to as sentiment classification. When machine learning techniques are used in sentiment classification, the feature vectors have high dimensionality problems. Hence there is a need for the feature selection scheme to remove the unrelated and loud features from the feature vector for the effective functioning of the machine learning algorithms. In this section, the MRMR feature selection, FCM clustering, MST based clustering and GSO optimized MST clustering methods are discussed.

3.1 Minimum Redundancy Maximum Relevance (MRMR) Feature Selection

The discriminate features of a class can be recognized using the MRMR feature selection. Features having the maximum relevancy/high dependency to the class and minimum redundancy / minimal dependence among the features are selected by the MRMR techniques. At times there is a lot of redundancy among the features for the pertinent features with maximum relevancy with the class. Given two features that are redundant, eliminating one of the features will not cause much difference in the discrimination of class. The correlation or the dependence among the features and the class attributes and amongst the features can be calculated using the mutual information [12]. Features that comprise enhanced mutual information/maximum relevance with class attribute are selected by the mRMR technique while the ones having high mutual information and high correlation among themselves/ minimum redundancy are eliminated.

Some of the benefits of mRMR are: estimation of both relevance and redundancy are problems of lower dimensions which have only about two variables. Thus, it is easier when compared to direct estimation of multivariate density or mutual information in huge dimensional area. This estimation is not only quicker but also more consistent. MRMR is an optimal first order approximation of I(.), maximization, and relevance-only ranking only maximizes J(.).

3.3 Minimum Spanning Tree (MST) Based Clustering

MST is employed for generating highly complex clusters and is a graph based prototype. The MST edges can be either chosen or discarded by it. A spanning tree is an acyclic sub-graph of graph G comprising all of G's vertices. The least weight spanning tree of a graph is the MST of that weighted graph. The expense to construct a minimum spanning tree in conventional MST is O (m log n). Here, n is representative of vertices and m is representative of edges. Regarded as a hierarchical clustering algorithm, MST can follow divisive clustering scheme.

Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/XRGDH

The divisive clustering scheme is tracked by MST clustering which is a hierarchical clustering algorithm. For constructing MST of point set and for deleting conflicting edges, clustering algorithm on the basis of minimum and maximum spanning tree have been examined. Their weights are expansively greater compared to the standard weight of the closest edges of the tree. The objective is maximizing the minimum inter cluster distance.

IV. RESULTS AND DISCUSSION

Hence, the MRMR, FCM based feature selection, MST based clustering and proposed MST optimized clustering methods are used. Table 1 shows the summary of results. The classification accuracy, positive predictive value for positive, neutral and negative opinion and hitrate for positive, neutral and negative opinion as shown in figures 1.

Table 1 Summary of Results

	MRMR	FCM based Feature Selection	MST based Clustering	Proposed MST optimized Clustering
Classification Accuracy	79.26	81.11	84.33	87.37
Positive predictive value for Positive Opinion	0.8049	0.8385	0.8441	0.8834
Positive predictive value for Neutral Opinion	0.7989	0.8056	0.8352	0.8606
Positive predictive value for Negative Opinion	0.7766	0.794	0.8509	0.8777
Hitrate for Positive Opinion	0.7333	0.75	0.8422	0.8422
Hitrate for Neutral Opinion	0.7944	0.8056	0.8444	0.8778
Hitrate for Negative opinion	0.85	0.8778	0.8433	0.9011

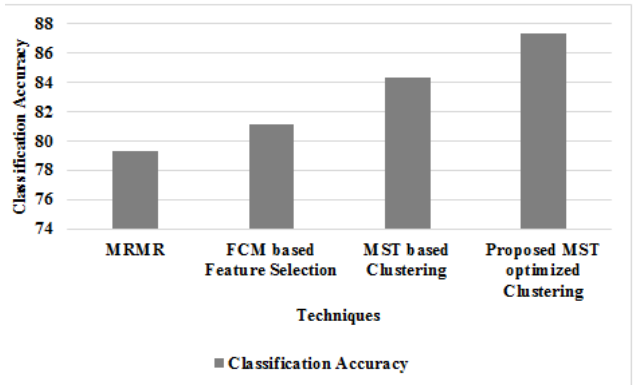

Figure 1 Classification Accuracy for Proposed MST Optimized Clustering

From the figure 1, it can be observed that the proposed MST optimized clustering has higher classification accuracy by 9.73% for MRMR, by 7.43% for FCM based feature selection and by 3.54% for MST based clustering.

V. CONCLUSION

In the classification problems that involve high dimensional data, there are many features. However, not all of them are suitable for classification. Better classification performance can be achieved using filter selection which involves selecting small number of related features. The presence of unrelated, noisy and redundant features may degrade the performance. An important objective of feature selection is increasing the classification performance and minimizing the number of features. The input space can be segregated into decision regions using fuzzy entropy which can select appropriate features with good partitioning for classification. This work suggests the GSO based feature selection optimized MST clustering algorithm. Clusters with irregular edges can be detected using the MST clustering algorithms. GSO is motivated by the searching behaviour and the group living theory of the animals. PS model forms the basis of the algorithm in which it is assumed that the group members 'finding' are producers or the ones that are 'joining' are scroungers. The algorithm is developed using the animal scanning scheme like vision. Also, for evading the local minima and for performing random walks, rangers can be employed Results show that the proposed MST optimized clustering has higher classification accuracy by 9.73% for MRMR, by 7.43% for FCM based feature selection and by 3.54% for MST based clustering.

Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/XRGDH

REFERENCES

1. Khedkar, S. A., Bainwad, A. M., &Chitnis, P. O. (2014). A survey on clustered feature selection algorithms for high dimensional data. Int J ComputSciInfTechnol (IJCSIT), 5(3), 3274-3280.
2. Manek, A. S., Shenoy, P. D., Mohan, M. C., &Venugopal, K. R. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World wide web, 20(2), 135-154.
3. Pullela, V., Kumar, V. S., & Yadav, C. R. (2014). A Framework for Mining High Dimensional Data for Feature Subset Selection. International Journal of Computer Science and Mobile Computing, 3 (12), 50-55.
4. Elankavi, R., Kalaiprasath, R., & Udayakumar, D. R. (2017). A fast clustering algorithm for high-dimensional data. International Journal Of Civil Engineering And Technology (Ijciet), 8(5), 1220-1227.
5. Song, Q., Ni, J., & Wang, G. (2013). A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE transactions on knowledge and data engineering, 25(1), 1-14.
6. Torkestani, J. A., & Meybodi, M. R. (2012). A learning automata-based heuristic algorithm for solving the minimum spanning tree problem in stochastic graphs. The Journal of Supercomputing, 59(2), 1035-1054.
7. Kumari, B. S., & Naidu, M. D. (2014). Feature Subset Selection Algorithm for Elevated Dimensional Data By using Fast Cluster. International Journal Of Engineering And Computer Science, 3(07).
8. Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015.
9. Kaveri, B. V., & Asha, T. (2015). An Ameliorated Methodology for Feature Subset Selection on High Dimensional Data using Precise Relevance Measures. International Journal of Computer Applications, 127 (7).
10. Singh, A. (2009). An artificial bee colony algorithm for the leaf-constrained minimum spanning tree problem. Applied Soft Computing, 9(2), 625-631.
11. Gu, S., Cheng, R., & Jin, Y. (2018). Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Computing, 22(3), 811-822.
12. Elawady, R. M., Barakat, S., & Elrashidy, N. M. (2014). Different feature selection for sentiment classification. International Journal of Information Science and Intelligent System, 3(1), 137-150.
13. Manjula, S. P., &Jagadeesan, J. (2014). Fuzzy C Means Clustering Algorithm for High Dimensional Data Using Feature Subset Selection Technique. IOSR Journal of Computer Engineering (IOSR-JCE), 16 (2), 64-69.
14. Jeyarani, D. S., &Pethalakshmi, A. (2016). Optimized Feature Selection Algorithm for High Dimensional Data. Indian Journal of Science and Technology, 9(31).
15. Nagendrudu, S., & Reddy, V. R. (2015). Enhanced Clustering of High Dimensional Data Using Fast Cluster Based Feature Selection. International Journal of Science, Engineering and Computer Technology, 5(5), 113.
16. Alapati, Y. K., Sindhu, K., & Suneel, S. (2015). Relevant Feature Selection from High-Dimensional Data Using MST Based Clustering. International Journal of Emerging Trends in Science and Technology, 2(03).