IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014
ISSN: 2320 – 8791 (Impact Factor: 1.479)
www.ijreat.org

Cluster Based Speed and Effective Feature Extraction for Efficient Search Engine

Manjuparkavi A[1], Arokiamuthu M[2]

[1]PG Scholar, Computer Science, Dr. Pauls Engineering College, Villupuram, India

[2]Assistant Professor, Computer Science, Dr. Pauls Engineering College, Villupuram, India

Abstract

Giving different Keywords to the Search engine till getting the best results is tedious process. Because we get redundant and irrelevant data. User can sends request and get response to the request. Many things have been occurred between clients and server in the process of searching. Not only user most of them don't know about the internal process of searching records from a large database. Let's see how an internal process of searching taken place. Feature selection identifies a subset of the most useful features that produces compatible results as the original entire set of features. Feature selection concerns both efficiency and effectiveness. Many problems have been occurs during text classification that are handled by FAST Algorithm. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of most useful and independent features.

Index Terms *- Feature subset selection, filter method, feature clustering, graph-based clustering.*

I. Introduction

Feature subset selection technique is used to choose a subset of good features mostly related to target concepts. Also Feature subset selection reduces dimensionality, removes irrelevant data, increases learning accuracy, and improving result comprehensibility "Ref [1]". While using a feature selection technique, data may contain many redundant or irrelevant features. Redundant features won't provide information related to the current selected features and irrelevant features provide information that is not related to the target concept. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the feature. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories "Ref [2]". The wrapper methods use predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets. The accuracy of the learning algorithms is usually high. The generality of the selected features is limited and the computational complexity is large. The wrapper methods are computationally expensive and tend to over fit on small training sets "Ref [3]". Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. The hybrid methods are a combination of filter and wrapper methods "Ref [4]", by using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance. The filter methods are independent of learning algorithms, with good generality. Computational complexity is low, but the accuracy of the learning algorithms is not guaranteed "Ref [5]". The filter methods are usually a good choice when the number of features is very large. Thus, we will focus on the filter method in this paper.

II. Existing system

In existing system, Searching is a very tedious Process because, we all be giving the different Keywords to the Search engine until we land up with the Best Results. Data contain many redundant and irrelevant features. Redundant features won't provide information related to the current selected features and irrelevant features provide information that is not related to the target concept.

Disadvantages of Existing System

There is no Clustering Approach is achieved in the Existing. There is no High Time Consume Process and effective search mechanism.

III. Related works

Feature subset selection identifies and removes irrelevant and redundant features as much as possible. Irrelevant feature affects the predictive accuracy "Ref [6]", where as Redundant feature won't redound to get better predictor. It provides information that is already present in other feature. Many subset selection algorithms are there, some algorithms can effectively eliminate irrelevant features but fail to remove redundant features "Ref [7]". While some algorithms eliminate irrelevant features while taking care of redundant features. Our proposed FAST algorithm falls into second group. Normally, feature subset selection search mostly relevant features. One of the best examples is Relief "Ref [8]", but it is ineffective to remove redundant features. Relief-F "Ref [9]", extends of a Relief, it works with noisy, incomplete data sets and also with multiclass problems, but it also fail to remove redundant features. Redundant features also affect speed and accuracy of learning algorithms, it must be eliminated "Ref [10]". CFS, FCBF, CMIM is some examples.CFS "Ref [11]", is a good feature subset, it contains features mostly correlated with target, yet uncorrelated with each other. FCBF "Ref [12]", is fast filter method, it identifies both redundant and relevant features without pair wise correlation analysis. CMIM "Ref [13]", select feature that maximizes their mutual information with response to any feature that have been already picked. Our proposed FAST algorithm implies clustering-based method to choose features.

IV. The Proposed Scheme

Feature subset selection is the process of identifying and removing as many irrelevant and redundant features as possible. Because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature. Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. Our proposed FAST algorithm falls into the second group. Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief. Relief is ineffective to remove redundant features. Relief-F extends of Relief, this method work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features. Feature selection identifies a subset of the most useful features that produces results as original entire set of features. Novel algorithm can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. The user obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset. In our proposed FAST algorithm, it involves (i) the construction of the minimum spanning tree (MST) from a weighted complete graph; (ii) the partitioning of the MST into a forest with each tree representing a cluster; and (iii) the selection of representative features from the clusters.

Advantages

1)Feature subset selection contain features highly correlated with the target class only, we can avoid irrelevant, redundant data.
2) FAST algorithm efficiently deals with both irrelevant and redundant features, and obtains a good feature subset.

V. System Model

The proposed system model for feature subset selection algorithm is explained. Irrelevant features and redundant features severely attack the accuracy of learning machines. Feature subset selection algorithm will identify and remove irrelevant and redundant feature as much as possible. Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other. An example for getting data without redundant and irrelevant data is illustrated in figure1 and modules are briefly explained below.

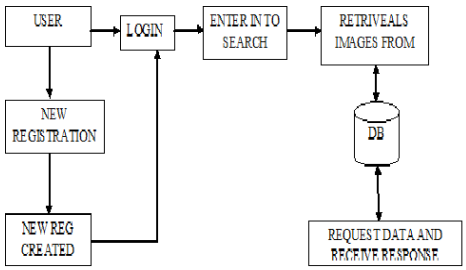

Fig. 1 System model

VI. Modules

The proposed system consists of five modules

A. User Login

Users are having authentication and security to access detail which is presented in the ontology system. Before accessing or searching details user should have the account, otherwise they should register first.

B. Distributed Clustering

Clustering is a combination of various features including text subsets. Distributed clustering is used to cluster words into groups based on their contribution in particular grammatical relations with other words. Here the distributed clustering focuses on the cluster with various text subsets. In this module the system can manage the cluster with various classifications of data.

C. Subset Selection Algorithm

Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Feature subset selection should be able to identify and remove as much as irrelevant and redundant features. We develop a novel algorithm that can efficiently deal with both irrelevant and redundant features, to obtain good feature subset.

D. Association Rule Mining

Association rule mining is the best method for discovering interesting relations between variables in large databases or data warehouse. It is intended to identify strong rules discovered in databases using different measures.

E. Text Representation

With the help of association rule mining the cluster is assembled with proper subset and correct header representations, in this stage the system can easily find out the text representation with maximum threshold value.

VII. Results And Analysis

In this section we present the experimental results in terms of the proportion of selected features, the time to obtain the feature subset, the classification accuracy, and the Win/Draw/Loss record. For the purpose of exploring the statistical significance of the results, we performed a Friedman and the Nemenyi test results are reported as well

A. Proportion of selected features

Proportion of selected features						
Data set	FAST	FCBF	CFS	Relief-F	Consist	FOCUS-SF
Chess	6.22	21.62	10.81	62.16	81.08	18.92
Horse	0.86	3.88	5.60	6.03	0.86	0.86
Rose	2.50	4.64	9.29	50.00	8.93	8.93
Lily	0.30	0.75	1.35	39.13	0.30	0.30
Dog	0.52	1.61	1.07	30.49	0.10	0.10
Hills	1.66	6.13	3.85	96.87	0.15	0.15
Nature	2.06	7.95	4.20	98.24	0.12	0.12
Image	3.59	10.04	6.68	79.85	3.48	3.48
Microarry	0.71	2.34	2.50	52.92	0.91	0.91
Text	2.05	3.25	2.64	10.87	11.46	2.53
Avg	1.82	4.27	3.42	42.54	5.44	2.06
Win/Draw/Loss	33/0/2	31/0/4 s	29/1/5	20/2/13	19/2/14	

Table 1 Proportion of selected features

Here we are comparing the proportion of selected features with other five algorithms. FAST on average obtains the best proportion of selected features. The Win/Draw/Loss records show FAST wins other algorithms as well. For image data, the five algorithms are not very suitable to choose features for image data compared with microarray and text data. FAST ranks three. For microarray data, six algorithms work well with microarray data. FAST ranks first. For text data, FAST ranks first and the second best algorithm is FOCUS-SF. The Friedman test "Ref [14]", can be used to compare k algorithms over N data sets by ranking each algorithm on each data set separately. The algorithm obtained the best performance gets the rank of 1, the second best ranks 2, and so on

B. Run Time

Here we are comparing the runtime of selected features with other five algorithms. Individual evaluation based feature selection algorithms of FAST, FCBF and Relief -F are much faster than the subset evaluation based algorithms of CFS, Consist and FOCUS-SF. The Win/Draw/Loss records show that FAST outperforms other algorithms as well. For image data, FAST obtains the rank first. FAST is more efficient than others when choosing features for image data. For microarray data, FAST ranks two. For text data, FAST ranks first. This indicates that FAST is more efficient than others when choosing features for text data as well.

Runtime of selected feature (in ms)						
Data set	FAST	FCBF	CFS	Relief F	Consist	FOCUS-SF
Chess	105	60	352	2660	1999	653
Horse	783	312	905	2099	2439	1098
Rose	110	115	821	3684	3492	2940
Lily	166	148	1224	744	1624	960
Dog	160	248	1038	1162	2476	1310
Hills	626	1618	9304	4334	5102	2556
Nature	635	2168	1097	7001	4666	2348
Image	1520	4090	9315	8213	6014	2793
Microarry	146	1169	1152	8059	9590	5385
Text	6989	8808	1372	1075	3815	3273
Average	3573	4671	2456	4582	14993	1269
Win/Draw/Loss - 22/0/13		33/0/2		29/0/6	35/0/0	34/0/1

Table 2 Runtime of selected features

VIII. Classification accuracy

Classification accuracy of six feature selection algorithms is compared by Naive Bayes, C4.5, IB1 and RIPPER algorithms against each other with the Nemenyi test.

A. Classification accuracy of Naive Bayes

Compared with original data, classification accuracy of Naive Bayes algorithms has been improved by FAST, CFS, and FCBF Unfortunately, Relief-F, Consist, and FOCUS-SF have decreased classification accuracy. For image data, FAST ranks third and the best accuracy is FCBF. For microarray data, FAST ranks first and the second best accuracy is CFS. For text data FAST ranks first and the second best accuracy is CFS.

B. Classification accuracy of C4.5

Compared with original data, the classification accuracy of C4.5 algorithms has been improved by FAST, FCBF and FOCUS-SF. and Unfortunately, Relief-F, Consist and CFS have decreased classification accuracy. For image data, FAST ranks two and the best accuracy is FCBF. For microarray data, FAST ranks first and the second best accuracy is CFS. For text data, FAST ranks third and the second best accuracy is FOCUS-SF.

C. Classification accuracy of IB1

Compared with original data, the classification accuracy of IB1 algorithms has been improved by FAST, FCBF and FOCUS-SF. and Unfortunately, Relief-F, Consist and CFS have decreased classification accuracy. For image data, FAST ranks four and the best accuracy is CFS. For microarray data, FAST ranks first and the second best accuracy is CFS. For text data, FAST ranks third and second best accuracy is CFS.

D. Classification accuracy of RIPPER

Compared with original data, the classification accuracy of RIPPER algorithms has been improved by FAST, FCBF and FOCUS-SF. and Unfortunately, Relief -F, Consist and CFS have decreased classification accuracy. For image data, FAST ranks first and the best accuracy is Consist. For microarray data, FAST ranks first and the second best accuracy is Consist and FOCUS-SF. For text data, FAST ranks five and the second best accuracy is CFS. For all data, FAST ranks 1 and should be the undisputed first choice, and FCBF, CFS are good alternatives.

From the above analysis we know that FAST performs very well on the microarray data. Our proposed FAST effectively filters out a mass of irrelevant features in the first step. This reduces the possibility of improperly bringing the irrelevant features into the subsequent analysis. Then, in the second step, FAST removes a large number of redundant features by choosing a single representative feature from each cluster of redundant features. As a result, only a very small number of discriminative features are selected. This coincides with the desire happens of the microarray data analysis.

IX. Conclusion

We have presented a novel clustering- based feature subset selection algorithm for high dimensional data. The algorithm removes irrelevant features, constructs minimum spanning tree from relative ones, and partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. We have compared the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, Relief, CFS, Consist, and FOCUS-SF on the 35 publicly available image, microarray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record.

References

[1] Liu H., Motoda H. and Yu L., Selective sampling approach to active feature selection, Artif. Intell., 159(1-2), pp 49-74 (2004).

[2] Guyon I. and ElisseeffA.,introduction to variable and feature selection, Journal of Machine Learning Research, 3, pp 1157-1182, 2003.

[3] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.

[4] Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.

[5] Langley P., Selection of relevant features in machine learning, In Proceedings of the AAAI Fall Symposium on Relevance, pp 1-5, 1994.

[6] John G.H., Kohavi R. and Pfleger K., Irrelevant Features and the Subset Selection Problem, In the Proceedings of the Eleventh International Conference on Machine Learning, pp 121-129, 1994.

[7] Forman G., An extensive empirical study of feature selection metrics for text classification, Journal of Machine LearningResearch, 3, pp 1289-1305, 2003.

[8] Kira K. and Rendell L.A., The feature selection problem: Traditional methods and a new algorithm, In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134, 1992.

9] Koller D. and Sahami M., Toward optimal feature selection, In Proceedings of International Conference on Machine Learning, pp 284-292, 1996.

[10] Forman G., An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 3, pp 1289-1305,2003.

[11] Hall M.A., Correlation-Based Feature Subset Selection for Machine Learning, Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.

[12] Kira K. and Rendell L.A., The feature selection problem: Traditional methods and a new algorithm, In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134, 1992

[13] Kononenko I., Estimating Attributes: Analysis and Extensions of RELIEF, In Proceedings of the 1994 European Conference on Machine Learning, pp 171-182, 1994.

[14] Friedman M., A comparison of alternative tests of significance for the problem of m ranking, Ann. Math. Statist., 11, pp 86-92, 1940.