International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 05 | May -2017 www.irjet.net p-ISSN: 2395-0072

Fast Clustering Feature Selection Algorithm For high Dimensional Data

Kiran Mhaske, Mahesh Lagad,Tushar Mhaske ,Javed Mulani,

Guided by- Prof.Satish yedge , Dept. of computer Engineering, kj college pune, Maharashtra, India

---

Abstract *A feature choice rule could also be evaluated from each the potency and effectiveness points of read .Feature choice involves distinctive a set of the foremost helpful options that produces compatible results because the original entire set of options. Whereas the potency considerations the time needed to search out a set of options, the effectiveness is said to the standard of the set of options. Based on these criteria, a quick clustering-based feature choice rule (FAST) is projected and through an experiment evaluated during this paper. The quick rule works in two steps. Within the start, options are divided into clusters by mistreatment graph-theoretic bunch ways. Within the second step, the foremost representative feature that's powerfully associated with target categories is chosen from every cluster to make a set of options. To make sure the potency of quick, we tend to adopt the economical minimum-spanning tree (MST) bunch techniques. Options in several clusters ar comparatively freelance, the clustering-based strategy of quick includes a high likelihood of manufacturing a set of helpful and freelance options. The results, on thirty five publicly accessible real-world high-dimensional image, microarray, and text knowledge, demonstrate that the quick not solely produces smaller subsets of options however conjointly improves the performances of the four kinds of classifier.The potency associate degreed effectiveness of the quick rule are evaluated through an empirical study.*

Key Words: Filter method, Feature subset selection, graph-based clustering, feature clustering.

1. INTRODUCTION

Data mining is a process of analyzing data and summarizes it into useful information. In order to achieve successful data mining, feature selection is an essential component. In machine learning feature selection is also known as variable selection or attributes selection. Feature selection is an important and frequently used technique in data mining for dimension reduction.The main idea of feature selection is to choose a subset of features by eliminating irrelevant or no predictive information. It is a process of selecting a subset of original features according to specific criteria.It employ for removing irrelevant, redundant information from the data to speeding up a data mining algorithm, improving learning accuracy, and leading to best model comprehensibility. In cluster analysis, graph theoretic approach is used in many applications. In general graph-theoretic clustering a complete graph is formed by connecting each instance with all its neighbours.

1.Remove the inconsistent edges to form connected components and call them clusters.
2.Construct the MST for the set of n patterns given.
3.Identify inconsistent edges in MST.

A feature subset selection algorithm (FAST) is used to test high dimensional available image, microarray, and text data sets. .In the FAST algorithm, features are divided into clusters by using graph-theoretic clustering methods and then, the most representative feature that is strongly related to target classes is selected. Feature subset selection methods can be divided into four major categories: Wrapper,Embedded, Hybrid and Filter. The embedded methods has a feature selections as a part of the initial process and are usually specific to given learning algorithms, and thus possibly more efficient that the other three

2.LITRATURE SURVEY

2.1 Fast Binary Feature Selection With Conditional Mutual Information

Author-Franc¸ois Fleuret
This paper a very fast feature selection technique based on conditional mutual information. We show that this feature selection method outperforms other classical algorithms, and that a naive Bayesian classifier built with features selected that way achieves error rates similar to those of state-of-the-art methods such as boosting or SVMs.By picking features which maximize their mutual information with the class to predict conditional to any feature already picked, it ensures the selection of features which are both individually informative and two-by-two weakly dependant.

2.2 Exploitation Mutual data for choosing options in supervised Neural internet Learning

Author-R. Battiti
This paper investigates the appliance of the mutual data criterion to guage a group of candidate options and to pick out an informative set to be used as computer file for a neural network classifier. as a result of the mutual data measures capricious dependencies between random variables, it's appropriate for assessing the "information content" of options in complicated classification tasks, wherever ways bases on linear relations (like the correlation) square measure vulnerable to mistakes. An

algorithmic rule is projected that's supported a "greedy" choice of the options which takes each the mutual data with regard to the output category and with regard to the already-selected options under consideration. Finally the results of a series of experiments square measure mentioned.The very fact that the mutual data is freelance of the coordinates chosen permits a sturdy estimation. However, the utilization of the mutual data for tasks characterised by high input spatial property needs appropriate approximations thanks to the preventive demands on computation and samples.

2.3Algorithms for distinctive Relevant options

Author-H. Almuallim and T.G. Dietterich.
This paper describes completely different ways for precise and approximate implementation of the MIN-FEATURES bias, that prefers consistent hypotheses determinable over as few options as attainable. This bias is beneficial for learning domains wherever several moot options square measure gift within the coaching information. We have a tendency to 1st introduce FOCUS-2, a replacement algorithmic rule that specifically implements the MIN-FEATURES bias. This algorithmic rule is by trial and error shown to be considerably quicker than the main target algorithmic rule antecedently given in [Almuallim and Dietterich 91]. We have a tendency to then introduce the Mutual-Information-Greedy, Simple-Greedy and Weighted-Greedy Algorithms, that apply economical heuristics for approximating the MIN-Features bias

3.EXISTING SYSTEM

A Feature subset selection can be seen as the process of identifying and removing as irrelevant and redundant data features as possible. This is because 1) irrelevant features do not contribute to the accuracy and 2) redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s).With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. The irrelevant feature removal is straight- forward once the right relevance measure is selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the cluster .Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible.We achieve this through a new feature selection framework which composed of the two unwanted connected components of irrelevant feature data removal and the redundant feature elimination. The former 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.

4.PROPOSED SYSTEM

Feature set choice will be viewed because the method of distinguishing and removing as several unsuitable and redundant options as doable. This can be as a result of unsuitable options don't contribute to the prognostic accuracy and redundant options don't redound to obtaining a far better predictor for that they supply principally data that is already gift in alternative feature(s). Of the numerous feature set choice algorithms, some will effectively eliminate unsuitable options however fail to handle redundant options nonetheless a number of others will eliminate the unsuitable whereas taking care of the redundant options. Our projected quick formula falls into the second cluster. Historically, feature set choice analysis has targeted on sorting out relevant options. A widely known example is Relief that weighs every feature in line with its ability to discriminate instances below totally different targets supported distance-based criteria operate. However, Relief is ineffective at removing redundant options as two prognostic however extremely related options area unit possible each to be extremely weighted. Relief-F extends Relief, sanctionative this methodology to figure with droning and incomplete knowledge sets and to traumatize multiclass issues, however still cannot establish redundant options

5 MATHEMATICAL MODEL

INPUT:-
Let S is the Whole System Consist of

S= {I, P, O}

I = Input.

I = {U, Q, D}

U = User

U = {u1,u2….un}

Q = Query Entered by user

Q = {q1, q2, q3…qn}

D = Dataset

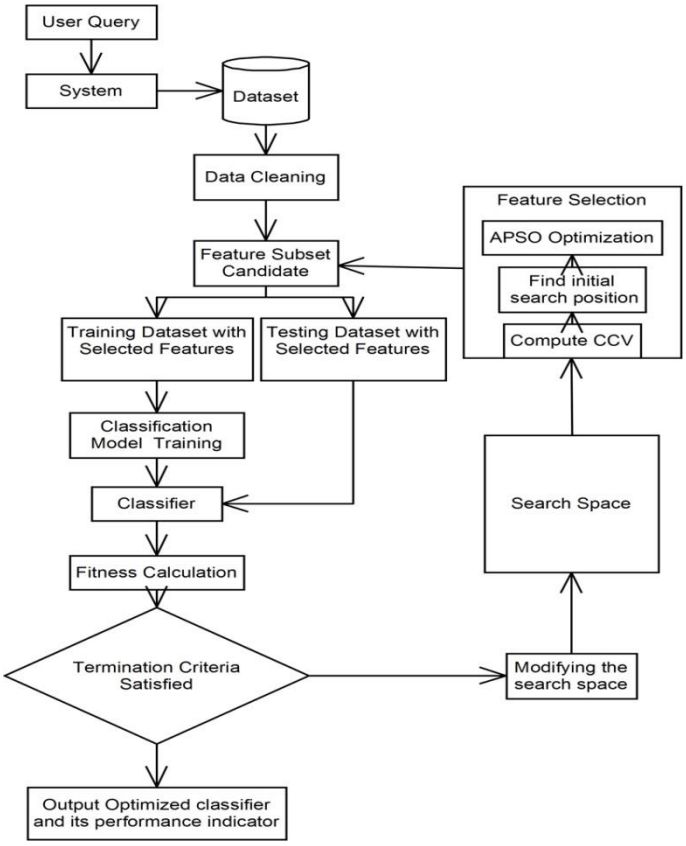
P = Process:

Step1: User will enter the query.

Step2: After entering query the following operations will be performed.

Step3: Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features.

6. System Architecture



7. Motivation

A feature selection algorithm may be evaluated from both the effciency and effectiveness points of view. While the e ciency concerns the time required to and a subset of features, the effectiveness is related to the quality of the subset of features.

8. Objectives

The main objective of this system is to remove irrelevant data, to reduce redundancy of data. Features in different clusters are not dependent, the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. In this system, features are divided into clusters by using graph-theoretic clustering methods. In the next step, the most simillar feature that is mainly related to particular classes is selected from each cluster to form a subset of features.The efficiency and effictiveness of the FAST algorithm are evaluated through an empirical study. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method.

9.SCOPE OF PROJECT

We conjointly found that quick obtains the rank of one for microarray knowledge, the rank of two for text knowledge, and therefore the rank of three for image knowledge in terms of classification accuracy of the four different types of classifiers, and CFS could be a smart various. At a similar time, FCBF could be a smart various for image and text data. Moreover, Consist, and FOCUS-SF square measure alternatives for text knowledge. For the future work, we tend to attempt to explore differing kinds of correlation measures, and study some formal properties of feature area.

10.CONCLUSION

Fast cluster based subset selection algorithm involves three important steps: 1.Elimination of Redundant features using minimum spanning tree 2.Removal of irrelevant features Elimination of Redundant features using minimum spanning tree. 3. Partitioning the MST and collect the selected features. Each cluster consists of redundant features and which is treated as single feature, so that dimensionality is reduced.

11. ACKNOWLEDGEMENT

We might want to thank the analysts and also distributers for making their assets accessible. We additionally appreciative to commentator for their significant recommendations furthermore thank the school powers for giving the obliged base and backing.

12.REFERENCES

[1] H. Almuallim and T.G. Dietterich, Algorithms for IdentifyingRelevant Features, Proc. Ninth Canadian Conf. Articial Intelligence, pp.
38-45, 1992.
[2] J. Biesiada and W. Duch, Features Election for High-Dimensional data a Pearson Redundancy Based Filter, Advances in Soft Computing, vol. 45, pp. 242-249, 2008.
[3] H. Almuallim and T.G. Dietterich, Learning Boolean Concepts in the Presence of Many Irrelevant Features, Articial Intelligence, vol. 69, nos. 1/2, pp. 279-305, 2001.
[4] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.
[5] C. Cardie, "Using Decision Trees to Improve Case- Based Learning," Proc. 10th Int'l Conf. Machine Learning,pp. 25-32, 1993.
[6] M. Dash and H. Liu, "Feature subset Selection for Classification," Intelligent Analysis Data, vol. 1, no. 3, pp. 131-156, 1998.
[7] W. Cohen, "Fast Effective Rule Induction," Proc. 12th Int'l Conf. Machine Learning (ICML '95), pp. 115-123,1995.
[8] M. Dash, H. Liu, and H. Motoda, "Consistency Based Feature Selection," Proc. Fourth Pacific Asia Conf.

Knowledge Discovery and Data Mining, pp. 98-109, 2000.
[9] P. Chanda, Y. Cho, A. Zhang, and M. Ramanathan, Mining of Attribute Interactions Using Information Theoretic Metrics, Proc. Intl Conf. Data Mining Workshops, pp. 350-355, 2009.
[10] D.A. Bell and H. Wang, "A Formalism for Relevance and Its Application in Feature Subset Selection," Machine Learning, vol. 41, no. 2, pp. 175-195, 2000.

13.RESULTS-