A Modified Approach for Feature Subset Selection using Multithreaded Boruvka's Algorithm

Vaidehi Bhavsar[1] Ompriya Kale[2]
[1]M.E Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]L. J. Institute of Technology, Ahmedabad, Gujarat, India

Abstract— Data mining is a multidisciplinary effort to extract a small chunk of knowledge from data. Feature Subset selection is an effective technique in dealing with dimensionality reduction. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. Based on these criteria, a fast clustering-based feature selection algorithm, FAST, is given. It is based on the minimum spanning tree method. The algorithm is a two steps process in which, 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 the final subset of features. The clustering-based scheme of FAST has a high possibility of producing a subset of constructive and independent characteristics. In this FAST algorithm, I have used Boruvka's Algorithm. By using Boruvka's Algorithm, parallel computing can be done. The proposed system will decrease the computational time of creating MST in the FAST Algorithm. I also used Chi-Square test that is used to find correlation among attributes for removing irrelevant features. Hence, this will improve the accuracy.

Key words: Subset Selection, Boruvka's Algorithm

I. INTRODUCTION

Data mining is an evolving area in information technology because of the ready availability of large amount of data. It is the process of, extracting hidden predictive information from huge databases, and it is a powerful new technology with great potential to focus on the most important information in the data warehouses[5]. Data mining is a multidisciplinary effort to extract a small chunk of knowledge from data.

The rapid increase of large data sets within many domains poses never known challenges to data mining. Along with data sets getting larger, new types of data have also evolved, such as microarrays in proteomics and genomics, data streams on the Web and networks in system biology and social computing.

Feature subset selection also known as Attribute subset selection, is data mining enhancement technique which reduces the number of fields in the database to a highly predictive subset[5]. The main objective of feature selection is to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information. Feature selection algorithms have two basic criteria namely, quality and time requirement[9].

Feature selection techniques are frequently used in domains where there are many features and comparatively few samples (or data points)[9].

The feature subset selection methods in machine learning applications are classified into four categories[7]:

- Embedded approach
- Wrapper approach
- Filter approach
- Hybrid approach

II. EXISTING SYSTEM

In the past approach there are several algorithm which illustrates how to maintain the data into the database and how to retrieve it faster, but the problem here is no one cares about the database maintenance with ease manner and safe methodology. 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. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches.

The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

The hybrid methods are a combination of filter and wrapper methods 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 with a particular learning algorithm with similar time complexity of the filter methods.

A. *Drawbacks of Existing System*

- Performance Related Issues
- Security Issues
- Lacks speed
- The generality of the selected features is limited
- Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

Hence the focus of our new system is to enhance the throughput for any basis to eliminate the data security lacks therein and make a newer system prominent handler for handling data in an efficient manner.

III. GENERAL FEATURE SELECTION STRUCTURE

It is possible to derive a general architecture from most of the feature selection algorithms. It consists of four basic steps: subset generation, subset evaluation, stopping criterion, and result validation. The feature selection algorithms create a subset, evaluate it, and loop until an ending criterion is satisfied. Finally, the subset found is validated by the classifier algorithm on real data.

Subset Generation Subset generation is a search procedure; it generates subsets of features for evaluation. The total number of candidate subsets is 2N, where N is the number of features in the original data set, which makes exhaustive search through the feature space infeasible with even moderate N. Non-deterministic search like evolutionary search is often used to build the subsets. It is also possible to use heuristic search methods. There are two main families of these methods: forward addition (starting with an empty subset, we add features after features by local search) or backward elimination (the opposite)[4].

Subset Evaluation Each subset generated by the generation procedure needs to be evaluated by a certain evaluation criterion and compared with the previous best subset with respect to this criterion. If it is found to be better, then it replaces the previous best subset. The method is classified as a wrapper, because in this case, the classifier algorithm is wrapped in the loop. In contrast, filter methods do not rely on the classifier algorithm, but use other criteria based on correlation notions[4].

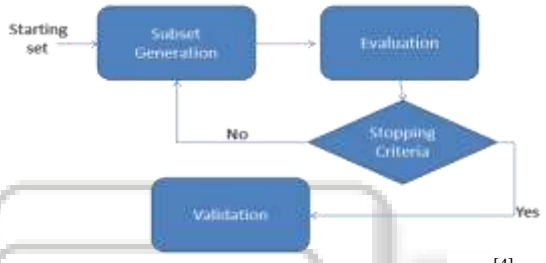

Fig. 1: General Feature selection structure [4]

A. *Stopping Criteria:*

Without a suitable criteria of stopping, the feature selection process may run exhaustively before it stops. A feature selection process may stop under one of the following reasonable criteria[4]:

1) a predefined number of features are selected
2) a predefined number of iterations are reached
3) in case addition (or deletion) of a feature fails to produce a better subset
4) obtained an optimal subset according to the evaluation criterion

B. *Validation*

The selected best feature subset needs to be validated by carrying out different tests on both the selected subset and the original set and comparing the results using artificial data sets and/or real-world data sets[4].

IV. LITERATURE REVIEW AND RELATED WORK

Literature survey is the most important step in software development process. The major part of the project development sector considers and fully survey all the required needs for developing the project. Before developing the tools and the associated designing it is necessary to determine and survey the resource requirement, man power, company strength, time factor and economy.

A. A Fast Clustering-Based Feature Subset Selection Algorithm[1]

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. The efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, FAST is proposed, a fast clustering based feature selection algorithm.

The FAST algorithm works in two steps[1]:

1) Features are divided into clusters by using graph-theoretic clustering methods
2) 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 useful and independent features. To ensure the efficiency of FAST, we have adopted the efficient minimum-spanning tree clustering method.

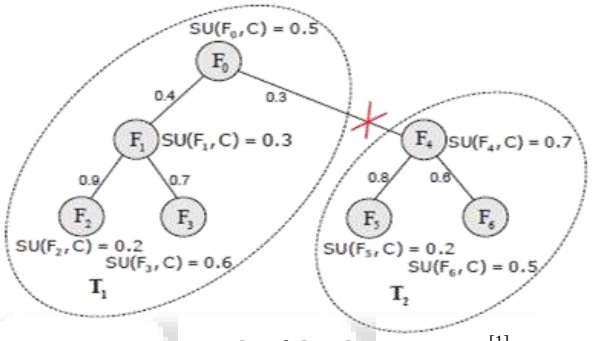

Fig. 2: Example of the clustering step[1]

Based on the MST method, a **F**ast clustering-b**A**sed feature **S**election algori**T**hm (FAST) is proposed.

Algorithm Steps:

1) removing irrelevant features
2) constructing a MST from relative ones
3) Partitioning the MST and selecting representative features.

V. PROPOSED SYSTEM

A. *Proposed Model:*

By using Boruvka's Algorithm, parallel computing can be done. Hence in the proposed model I have used Boruvka's Algorithm instead of Prim's Algorithm to create Minimum Spanning Tree in FAST Algorithm.

The proposed system will decrease the computational time of creating MST in the FAST Algorithm.

I also used Chi-Square test that is used to find correlation among attributes for removing irrelevant features. Hence, this will improve the accuracy.

B. *Steps of Proposed Model:*

1) Irrelevant Feature removal (Using Chi-Square Test)
2) Minimum Spanning Tree Construction (Using Boruvka's Algorithm)

– For Graph G(weighted graph)
– Component = Vertices of Graph (Initially every vertices are component)
– For each Component in Components Edge E = Select random node and connect with Minimal Edge

– Tree = Add (E) (Add edges to Spanning Tree)
– Return Tree

3) Tree partition and representative Features Selection

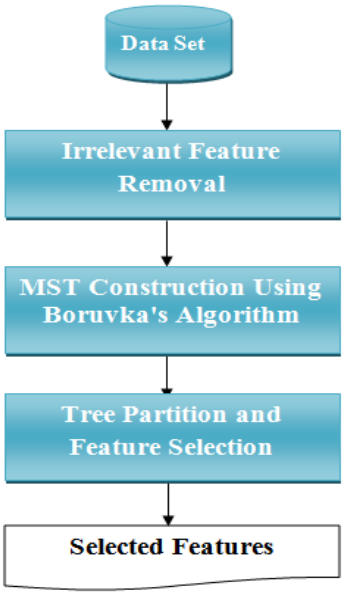

Fig. 3: Flow chart of proposed model

C. Comparison:

Data set	Approach	Accuracy	Time(ms)
Iris	Existing System	41.33	321
	Proposed multi-threaded Boruvka's Algorithm	52.66	120

Table 1: Comparison of existing and proposed work of Iris Dataset

Data set	Approach	Accuracy	Time(ms)
Chess	Existing System	80.76	2166
	Proposed multi-threaded Boruvka's Algorithm	84.60	1908

Table 2: Comparison of existing and proposed work of Chess Dataset

Data set	Approach	Accuracy	Time(ms)
mfeat-Fourier	Existing System	80.15	2390
	Proposed multi-threaded Boruvka's Algorithm	84.1	2338

Table 3: Comparison of existing and proposed work of mfeat-Fourier Dataset

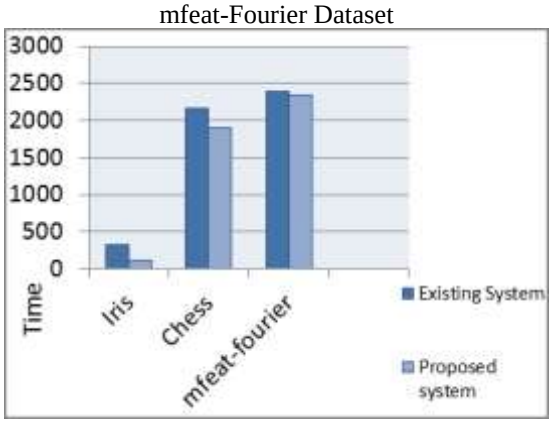

Fig. 4: Graphical representation of Time comparison with different dataset

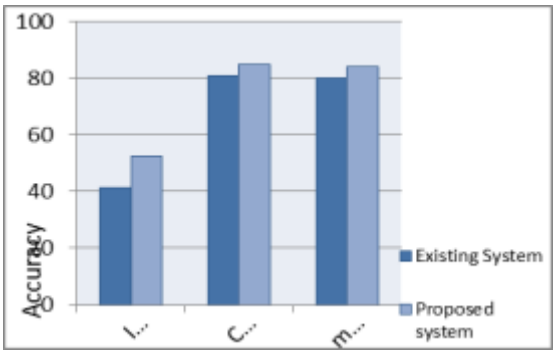

Fig. 5: Graphical representation of Accuracy comparison with different dataset

VI. Conclusion

Through Feature Subset selection we may know about how an internal process of searching works. Feature subset selection study has targeted on searching for relevant features. And according to the proposed model, I have used Boruvka's Algorithm to construct Minimum Spanning Tree. Hence the system will decrease the computational time of creating MST in the FAST Algorithm. I also used Chi-Square test that is used to find correlation among attributes for removing irrelevant features.

References

[1] 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
[2] K.Suman, S.Thirumagal "Feature Subset Selection with Fast Algorithm Implementation", International Journal of Computer Trends and Technology (IJCTT) – volume 6 number 1 – Dec 2013
[3] P.Abinaya, Dr.J.Sutha "Effective Feature Selection For High Dimensional Data using Fast Algorithm", International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), Vol. 2 Issue Special 1 Jan-March 2014
[4] Jasmina NOVAKOVIĆ, Perica STRBAC, Dusan BULATOVIĆ "Toward Optimal Feature Selection Using Ranking Methods And Classification Algorithms", Yugoslav Journal of Operations Research, March 2011
[5] Sujatha Kamepalli, Radha Mothukuri, " Implementation of Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data ", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 3, May – June 2014
[6] Jiawei Han, "Data Mining: Concepts and Techniques", Data Mining, p 5.
[7] A.GowriDurga, A.Gowri Priya, " Feature Subset Selection Algorithm for High Dimensional Data using Fast Clustering Method ", IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014
[8] S.Swetha, A.Harpika, "A Novel Feature Subset Algorithm For High Dimensional Data ", IJRRECS/October 2013/Volume-1/Issue-6/1295-1300

[9] K.Revathi, T.Kalai Selvi, " Survey: Effective Feature Subset Selection Methods and Algorithms for High Dimensional Data ", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 12, December 2013
[10] Igor Podsechin, "Comparing minimum spanning tree algorithms
[11] Anal Acharya, Devadatta Sinha, "Application of Feature Selection Methods in Educational Data Mining", International Journal of Computer Applications Volume 103 – No.2, October 2014. ISSN: 0975-8887

