Efficient Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data

P.Madhavi Latha
Assistant Professor, Department of Information Technology,
VR Siddhartha Engineering College, Vijayawada,
Andhra Pradesh, India,

Madhavi katamaneni
Assistant Professor, Department of Information Technology,
VR Siddhartha Engineering College, Vijayawada,
Andhra Pradesh, India,

Geetha.G
Assistant Professor, Department of Information Technology,
VR Siddhartha Engineering College, Vijayawada,
Andhra Pradesh, India,

Abstract-Feature subset selection is a 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. Many feature subset selection methods have been proposed and studied for machine learning applications. In the existing system they can be divided into three broad categories: the Embedded, Wrapper, 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. 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. . The hybrid methods are a combination of filter and wrapper methods by using a embedded method to reduce search space that will be considered by the subsequent wrapper. In the proposed algorithm not only reduces the number of features, but also improves the performances, It efficiently and effectively deals with both irrelevant and redundant features and Partitioning of the Minimal Spanning Tree(MST) into a forest with each tree representing a cluster.

Keywords: feature subset selection, filter method, feature clustering, graph-based clustering

I. INTRODUCTION

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. 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 [1]. 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 [5]. 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.

The wrapper methods are computationally expensive and tend to over fit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, we will focus on the filter method.

II.PROPOSED SYSTEM

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is 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(s) [16].

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 [3]. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function.

However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

A. Functional Diagram

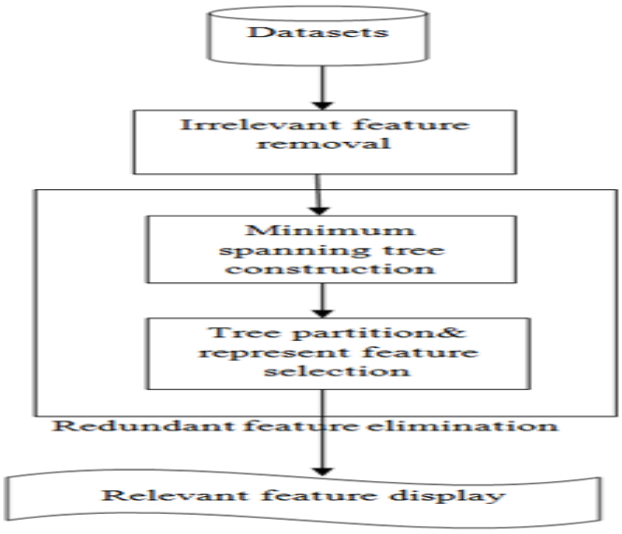

Fig 1: Functional Diagram proposed feature subset selection algorithm

B. Advantages of Proposed system:

- ✓ The proposed algorithm not only reduces the number of features, but also improves the performances
- ✓ It efficiently and effectively deals with both irrelevant and redundant features.

- ✓ Partitioning of the Minimal Spanning Tree (MST) into a forest with each tree representing a cluster.

C. Fast Algorithm

Step1: identifying the datasets

step2: irrelevant feature removal

Step3: for i=m

Step4:T relevance=SU(Fi,C)

Step5: if T-relevance>ø

Step6: S=SU{Fi}

Step7: for each pair of features {Fi ,Fj }then s do

Step8:F-corerealtion=SU {Fi ,Fj }

Step9: adding the all features as a weight of corresponding edge

Step10: minimum value of G huge spanning tree

Step11:for each edge Ei € forest edge

Step12:if SU {F'i ,F'j }< S U (Fi ,C)^SU {F'i ,F'j }<SU (Fj, c) then Forest=forest-edge

Step13:for each tree €Forest

Step14:F=max SU (Fj, c)

Step15:S=SU (Fj, c)

Step16: return S;

III.EXPERIMENTAL RESULTS

3.1. Input Datasets:

A1 duration

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	duration	protocol	service	flag	src_byte	dst_bytes	land	wrong frag	urgent	not	num_faile	logged_in	num_com	root_shell	su_attemp
2	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0
3	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0
4	2	tcp	ftp_data	SF	12983	0	0	0	0	0	0	0	0	0	0
5	0	icmp	eco_i	SF	20	0	0	0	0	0	0	0	0	0	0
6	1	tcp	telnet	RSTO	0	15	0	0	0	0	0	0	0	0	0
7	0	tcp	http	SF	267	14515	0	0	0	0	0	1	0	0	0
8	0	tcp	smtp	SF	1022	387	0	0	0	0	0	1	0	0	0
9	0	tcp	telnet	SF	129	174	0	0	0	0	1	0	0	0	0
10	0	tcp	http	SF	327	467	0	0	0	0	0	1	0	0	0
11	0	tcp	ftp	SF	26	157	0	0	0	0	1	0	0	0	0
12	0	tcp	telnet	SF	0	0	0	0	0	0	0	0	0	0	0
13	0	tcp	smtp	SF	616	330	0	0	0	0	0	1	0	0	0
14	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0
15	0	tcp	telnet	S0	0	0	0	0	0	0	0	0	0	0	0
16	37	tcp	telnet	SF	773	364200	0	0	0	0	0	1	0	0	0
17	0	tcp	http	SF	350	3610	0	0	0	0	0	1	0	0	0
18	0	tcp	http	SF	213	659	0	0	0	0	0	1	0	0	0
19	0	tcp	http	SF	246	2090	0	0	0	0	0	1	0	0	0
20	0	udp	private	SF	45	44	0	0	0	0	0	0	0	0	0
21	0	tcp	private	REJ	0	0	0	0	0	0	0	0	0	0	0
22	0	tcp	ldap	REJ	0	0	0	0	0	0	0	0	0	0	0
23	0	tcp	pop_3	S0	0	0	0	0	0	0	0	0	0	0	0
24	0	tcp	http	SF	196	1823	0	0	0	0	0	1	0	0	0
25	0	tcp	http	SF	277	1816	0	0	0	0	0	1	0	0	0
26	0	tcp	courier	REJ	0	0	0	0	0	0	0	0	0	0	0

Sheet1 Sheet2 Sheet3

Ready 100%

Fig 2.Excel Datasets

Results:

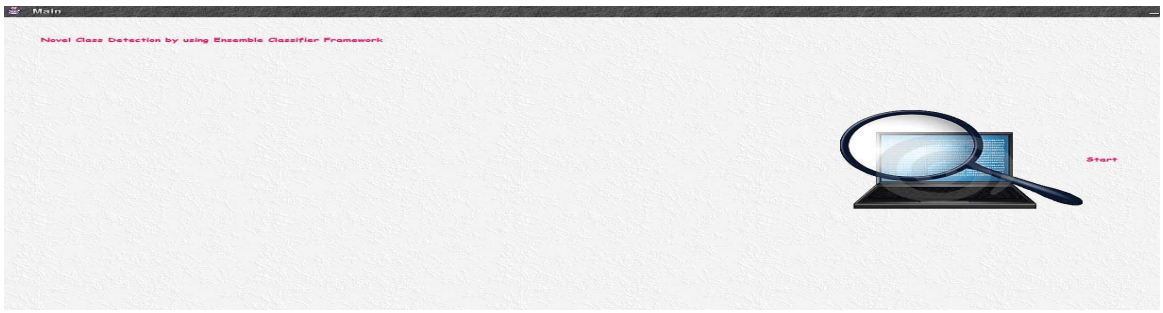

Fig 3.home page

Initial load the project in net beans. After dumping the project in net beans then right on the fast clustering icon then we select the run option.home page will be displayed.

To Load TheDatasests

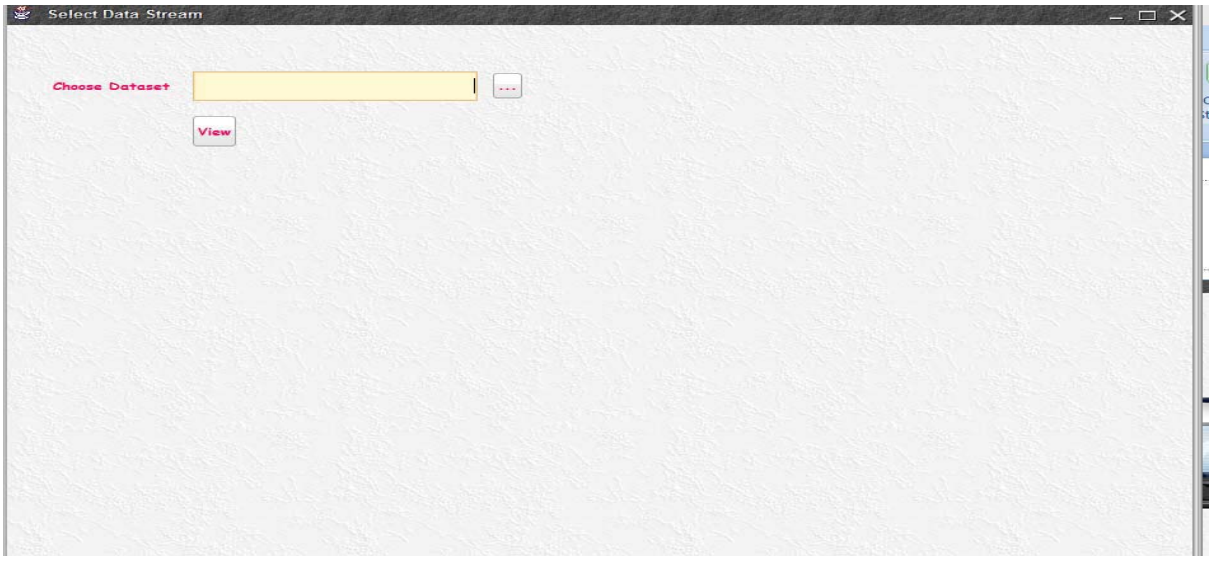

Fig 4 Load the Datasets

Identifying and loading the excel datasets in our computer. Then press the view button. After pressing the view button the following window will be displayed. Select labeled button.

Fig 5. Total Datasets

Fig 6. Constructing the MST

The user required information will be selected in the list. i.e protocols, service, source data etc.

Fig 7. Selection of the Subsets

After selecting user required fields the above window will be displayed on user needs.

IV.CONCLUSION

A novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, and (iii) 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.

The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. Extensive experiments are carried out to compare FAST and several representative feature selection algorithms, namely, FCBF, Relief, CFS, Consist, and FOCUS-SF, with respect to four types of well-known classifiers, namely, the probability-based Naive Bayes, the tree-based, the instance-based IB1, and the rule-based RIPPER before and after feature selection.

REFERENCES

[1] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.
[2] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279-305, 1994.
[3] Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure basedon relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109, 2004.
[4] Baker L.D. and McCallum A.K., Distributional clustering of words for textclassification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96- 103, 1998.
[5] Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537- 550, 1994.
[6] Bell D.A. and Wang, H., A formalism for relevance and its application infeature subset selection, Machine Learning, 41(2), pp 175-195, 2000.
[7] Biesiada J. and Duch W., Features election for high-dimensionaldatała Pearsonredundancy based filter, AdvancesinSoftComputing, 45, pp 242C249, 2008.
[8] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Selection