

A Feature Extraction with Graph Based Clustering

Pushpa s. Ghonge¹, B. K. Patil²

Student, Computer Science & Engineering, Everest College of Engineering, Aurangabad, India¹

Asst. Prof., Computer Science & Engineering, Everest College of Engineering, Aurangabad, India²

Abstract:- this paper presents a detailed study of different graph theoretic method i.e. clustering algorithm. A cluster is collection or group of data objects that are similar to each other with the same cluster object and not similar with other Cluster object. Also it is study on different feature selection algorithm. To overcome the limitations of existing algorithm. A feature selection may be related with both the efficiency and effectiveness point of view. FAST algorithm is proposed. Features are different cluster relatively independent. Clustering based strategy has high probability of producing a subset of important and independent features. To adopt the efficiency of fast clustering feature selection algorithm. It creates efficient minimum spanning tree clustering method.

Keywords- Feature subset selection technique, feature clustering, graph-based clustering

I INTRODUCTION

The main aim of feature selection algorithm is that selecting the most related features with respective target class, this technique is a way for reduces dimensionality, remove irrelevant information, increases accuracy and improve the result. We use the Prim's algorithm for creating minimum spanning tree. The simple algorithm is used to make possible subset of features and finding the most effective one which decreases the error rate. The proposed system is used which works in two steps.

1. Features are divided into clusters by using graph based clustering method.
2. the most representative feature that is strongly related to target class is selected from every cluster to produce final subset of features.

The process of feature selection is choosing a subset of relevant features is used in system development. The central idea of using a feature selection method is that the data contains redundant or irrelevant features. The redundant features are those which provide no more information than correctly selected features .and the irrelevant features provide no useful information in any context .this feature selection process is a subset of the more field of feature extraction. Feature extraction creates new features from the already existing features, so feature selection returns a subset of the features. The feature selection used in domains where there are many features and data points. Feature selection is a way for reducing dimensionality, removing redundant data, increasing learning accurateness, and recovering improve result.

In practice we have to improve the quality of data and reduce the time i.e. it is totally related with efficiency and effectiveness of data.

II LITERATURE SURVEY

Today different types of technology are growing fast so in this clustering is also one of the important tasks for feature subset selection. Feature selection algorithm main aim is that selecting a subset of features by removing irrelevant information. It is the way of choosing a subset of original features related to target class. Irrelevant features do not provide accuracy and redundant features are that same data present in another features . Different feature selection algorithm present, most of them are useful at removing irrelevant features but there is no effective to carry redundant features. But some algorithm can remove irrelevant feature at that time it take care of redundant features [1]. Fast clustering based feature selection algorithm come in second group. The most feature selection algorithm is relief which weight every features according to its ability to discriminate instances under different criteria based on distance based target function.

Whatever Relief is not useful at removing duplicate features is two predictive but more correlated features are likely both to be highly weighted [4].Relief F [5] extends Relief, this algorithm works with the removing redundant information and eliminating irrelevant data and also deals with the multiclass problem ,but still cannot identify redundant features. The redundant features affects the accuracy of learning algorithm. So it is needy to eliminate it.CFS[6],FCBF[7]are the example that used for the redundant features.CFS[6]is represented by hypothesis that a good feature subset is one that enables relevancy of feature as well as redundancy among relevant features . Some different from above algorithm, FAST algorithm uses minimum spanning tree based method to cluster features.

Generally feature selection can be presented as the process of identifying and eliminating as irrelevant and redundant features as well as possible. The first irrelevant features that do not enable predictive accuracies And secondly redundant features that do not getting a better predictor for that they mostly provide information which is already situated in another feature.

III FEATURE SUBSET SELECTION

Feature selection able to identify and eliminate as much of the irrelevant and redundant feature as possible. However good

features subset contains highly correlated features with the class, yet uncorrelated with other features

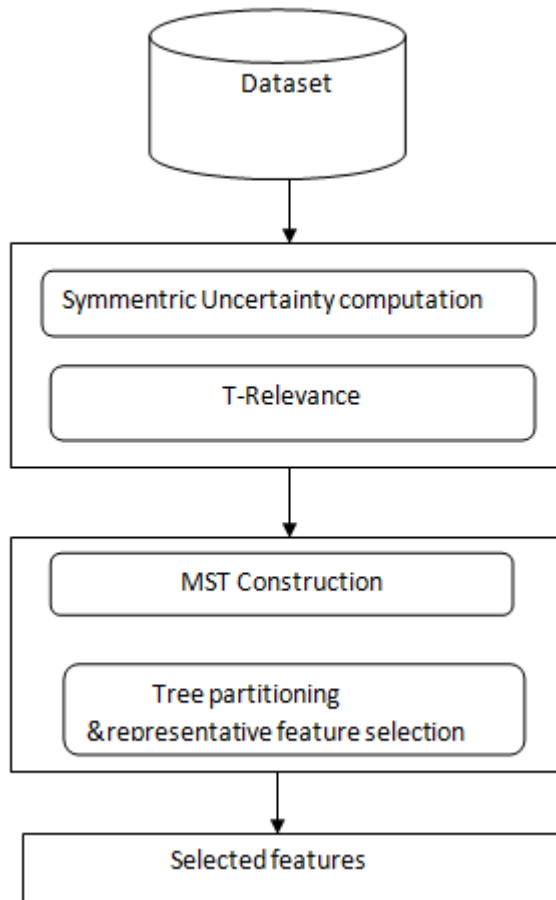


Figure 1 Feature selection process

In this feature subset algorithm the role of Symmetric uncertainty and T-relevance is irrelevant feature removal and the role of minimum spanning tree and tree partitioning with representative feature selection is redundant feature elimination. In Fast clustering based feature Selection algorithm (FAST) involves in first step it construct the minimum spanning tree from weighted complete graph, In second step the partitioning of MST into forest with each tree representing cluster. And in third step the selection of representative features from the clusters. Consider F is full set of features, $f \in F$ be a feature, $S_i = F - \{f\}$ and $S_i' \subseteq S_i$. let s_i be the value assignment of all features in S_i , f_i is a value assignment of all features F_i and c be the value assignment of target class C . The SU (symmetric uncertainty) is as follows

$$SU(X, Y) = \frac{2 \times \text{Gain}(X|Y)}{H(X) + H(Y)}$$

Where,

$$\text{Gain}(X|Y) = H(Y) - H(X|Y)$$

$$\text{Gain}(X/Y) = H(Y) - H(X|Y)$$

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x)$$

$$H(X|Y) = - \sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y)$$

Where, $H(X)$ is the entropy of discrete random variable X . And $H(Y)$ is the entropy of discrete random variable Y . $\text{Gain}(X/Y)$ is the amount by which entropy of Y decreases. It reflects the additional information about Y provided by x , so it is called as information gain.

The general graph-theoretic clustering method is introduced: Calculate a neighborhood graph of vertices, then remove any edge in the graph that is much larger/lesser than its neighbors. Result is a forest and each tree in the forest represents a cluster. We use dataset like Cancer, diabetes, car etc which is text, image and microarray data set.

There are four different classification algorithm used to increase the accuracy of classifier .they are i. Naive Bayes –which is probability based classifier (NB).ii. C4.5-It is tree based classifier. iii. IB1-instanced based lazy learning algorithm .iv. RIPPER- It is rule-based algorithm. Accuracy of all these classifier with respective different feature selection algorithm is implemented in this paper as well as total no of selected feature and time taken to select the features .these two things are related with Efficiency and effectiveness respectively

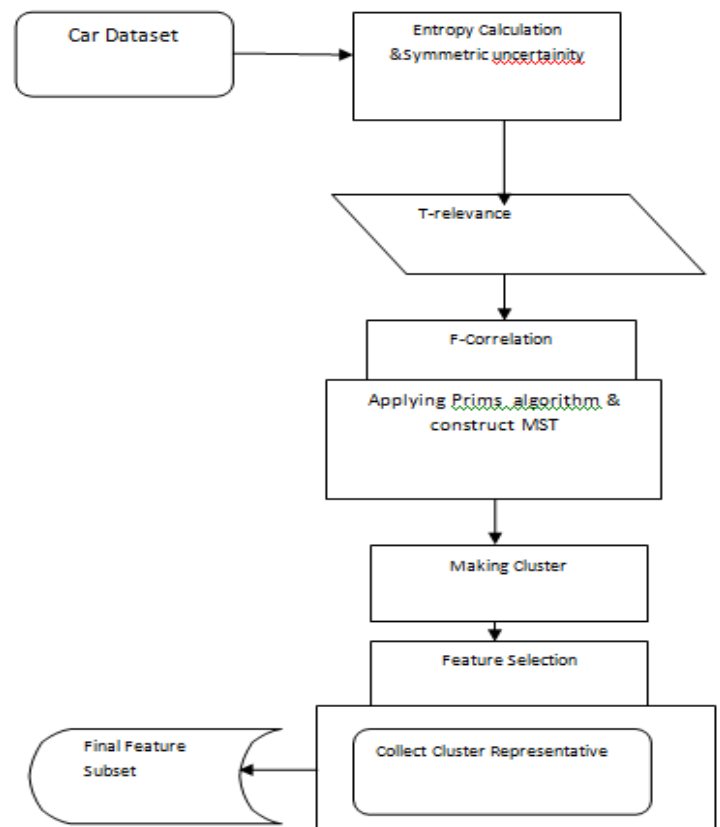
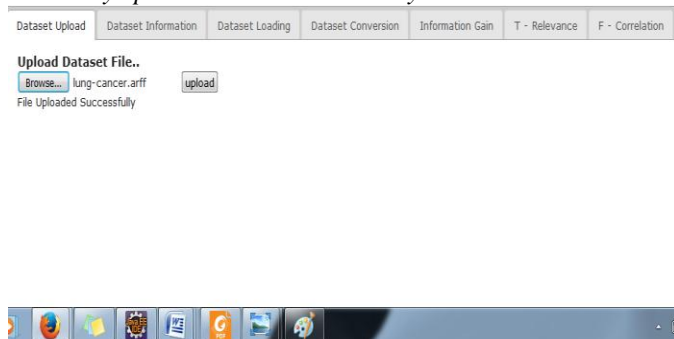


Figure 2 System Architecture

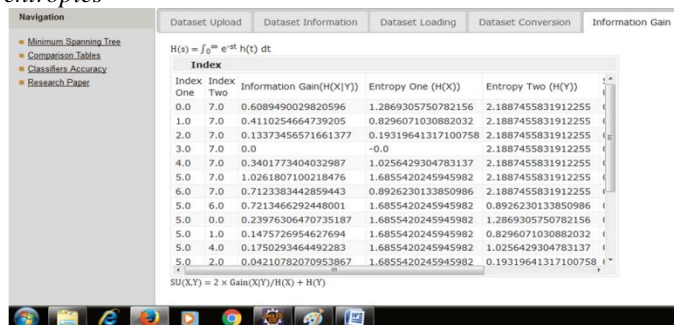
IV RESULT ANALYSIS

The Proportion of selected features is the ratio of no. of features selected by feature selection algorithm to original no. of features to the dataset. According to different dataset like text, image, microarray is included and implemented in this paper.

1. Initially upload the dataset on the system.

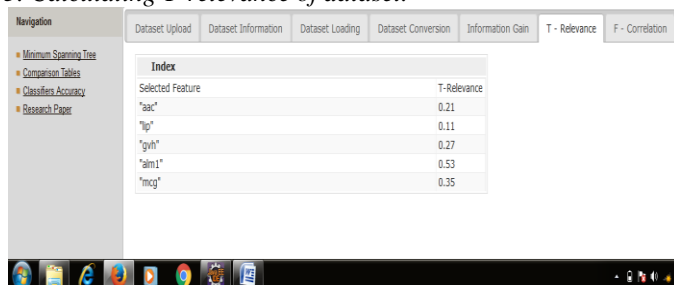


2. Reading of symmetric uncertainty, Information Gain and their entropies



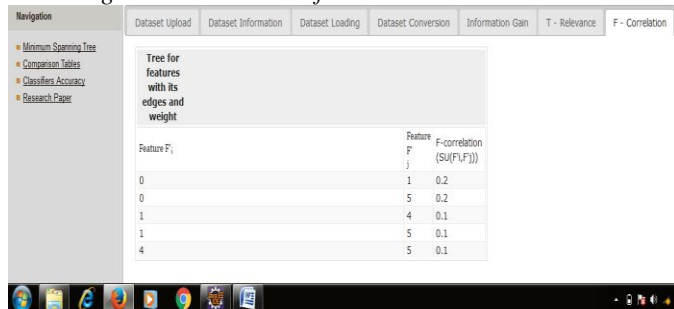
Index	Information Gain (H(X Y))	Entropy One (H(X))	Entropy Two (H(Y))
0.0	0.6089490029820596	1.2869305750782156	2.1887455831912255
1.0	0.4110254664739205	0.8296071030882032	2.1887455831912255
2.0	0.13373456571661377	0.19319641317100758	2.1887455831912255
3.0	0.0	-0.0	2.1887455831912255
4.0	0.3401773404032987	1.0256429304783137	2.1887455831912255
5.0	0.701261807100218476	1.6855420245945982	2.1887455831912255
6.0	0.7123383442859443	0.8926230133850986	2.1887455831912255
5.0	0.7213466292448001	1.6855420245945982	0.8926230133850986
5.0	0.0	0.23976306470735187	1.6855420245945982
5.0	0.1475726954627694	1.6855420245945982	0.8296071030882032
5.0	0.4.0	0.1750293464492283	1.6855420245945982
5.0	0.04210782070953867	1.6855420245945982	0.19319641317100758

3. Calculating T-relevance of dataset.



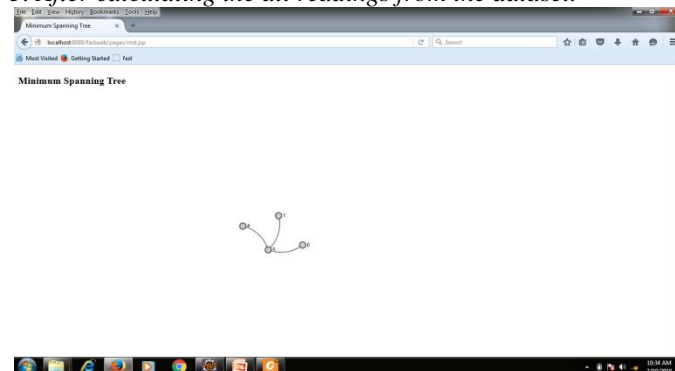
Selected Feature	T-Relevance
"aac"	0.21
"tip"	0.11
"gph"	0.27
"alm1"	0.53
"mcg"	0.35

4. Finding the Correlations of dataset.

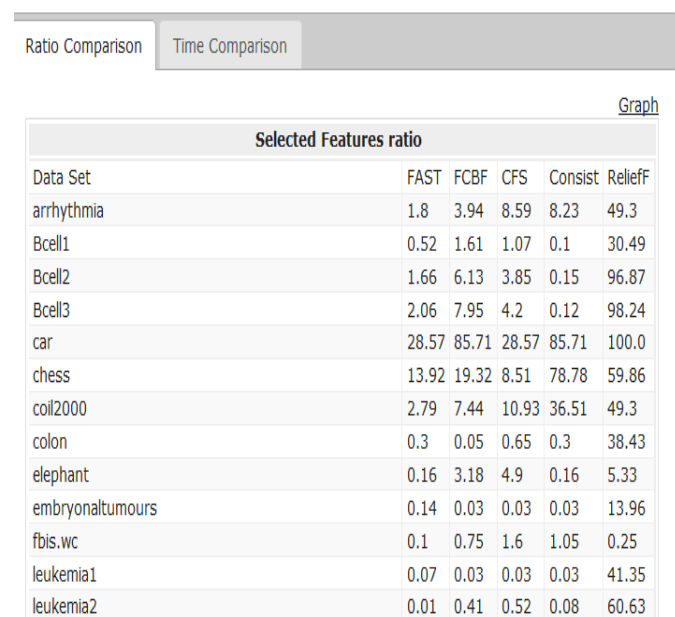


Feature F1	Feature F2	F-correlation (SU(F1,F2))
0	1	0.2
0	5	0.2
1	4	0.1
1	5	0.1
4	5	0.1

5. After calculating the all readings from the dataset.



6. Figure shows the proportion of selected features with all feature selection algorithm, and the graph of proportion of selected feature.



Data Set	FAST	FCBF	CFS	Consist	ReliefF
arrhythmia	1.8	3.94	8.59	8.23	49.3
Bcell1	0.52	1.61	1.07	0.1	30.49
Bcell2	1.66	6.13	3.85	0.15	96.87
Bcell3	2.06	7.95	4.2	0.12	98.24
car	28.57	85.71	28.57	85.71	100.0
chess	13.92	19.32	8.51	78.78	59.86
coil2000	2.79	7.44	10.93	36.51	49.3
colon	0.3	0.05	0.65	0.3	38.43
elephant	0.16	3.18	4.9	0.16	5.33
embryonaltumours	0.14	0.03	0.03	0.03	13.96
fbis.wc	0.1	0.75	1.6	1.05	0.25
leukemia1	0.07	0.03	0.03	0.03	41.35
leukemia2	0.01	0.41	0.52	0.08	60.63

7. The selected feature graph shows the selected feature percentage with dataset.

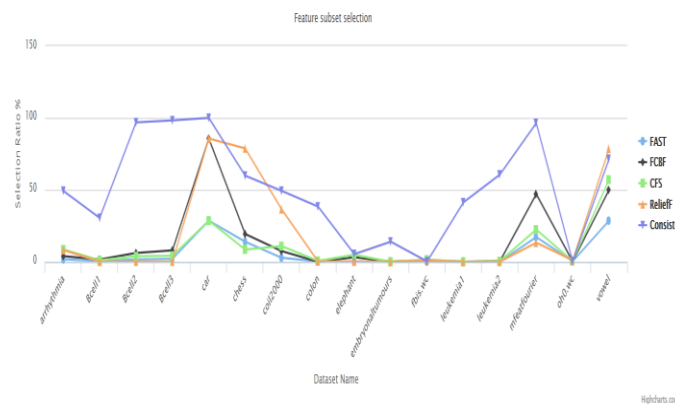


Figure 7: Graph of proportion of selected feature

8. This figure shows time taken to select feature from the dataset.

Time Comparison					
Time taken in Milliseconds					
Data Set	FAST	FCBF	CFS	Consist	Relieff
arrhythmia	125.0	130.0	836.0	3507.0	3699.0
Bcell1	175.0	263.0	103886.0	2491.0	1177.0
Bcell2	641.0	1633.0	930480.0	5117.0	4349.0
Bcell3	650.0	2183.0	1097137.0	4681.0	7016.0
car	20.0	16.0	0.0	764.0	141.0
chess	120.0	75.0	367.0	2014.0	12675.0
coil2000	881.0	890.0	1498.0	53865.0	304177.0
colon	181.0	163.0	12264.0	1639.0	759.0
elephant	798.0	327.0	920.0	2454.0	21006.0
embryonaltumours	769.0	329.0	10169.0	5060.0	1696.0
fbis.wc	14776.0	16222.0	66073.0	579391.0	79542.0
leukemia1	464.0	293.0	10915.0	5723.0	805.0
leukemia2	1156.0	471.0	216903.0	10422.0	909.0
mfeatfourier	1487.0	731.0	953.0	3242.0	13933.0

Figure 8: Time taken to select feature from dataset

9. Below graph shows time taken in ms with dataset and selected feature algorithm.

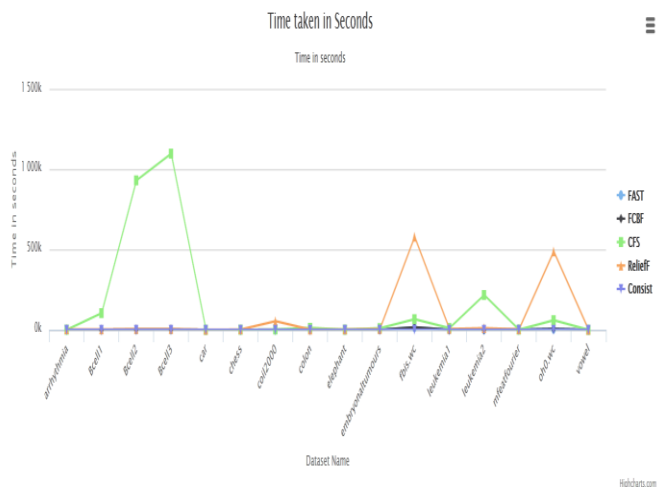
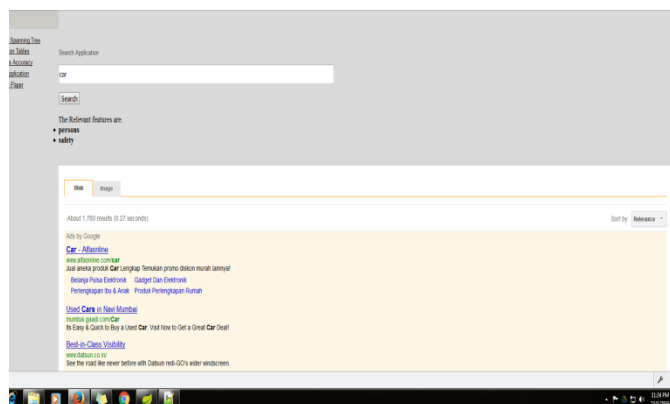


Figure 9: Graph for time taken to select feature from dataset with respective feature selection algorithm

10. Accuracy of Naïve Bayesian classifier with feature selection algorithm.

Naive Bayes					
Classifier Accuracy %					
Delta Set	FAST	FCBF	CFS	Relieff	Consist
AR10P	66.23	78.08	79.77	59.46	77.77
arrhythmia	70.01	62.98	66.64	66.24	61.53
basehock	90.18	87.09	87.98	78.92	48.05
Bcell1	97	97	97	88.5	97
Bcell2	93.63	80.53	80.47	59.71	74.63
Bcell3	95.22	79.47	82.47	79.24	76.26
car	75.5	73.50	70.02	73.15	73.15
chess	89.92	89.12	87.43	86.5	85.56
coil2000	91.04	90.53	89.59	81.64	73.96
colon	92.08	81.81	81.81	82.48	65.33

11. Implementation of feature selection i.e. search application. User can enter the dataset it gives the result of selected features.



V CONCLUSION

We implement the feature subset selection using graph based clustering to evaluate the performance, accuracy and capability of features from huge amount of data for that FAST algorithm to reduce memory usage. Fast Clustering based feature selection algorithm can be compared with existing feature algorithm. FAST get the first rank for Text data and second rank for image data as well as Microarray dataset. the response of FAST algorithm i.e. feature selection which is search algorithm.

REFERENCES

- [1] Yu L. and Liu H., "Efficient feature selection via analysis of relevance and redundancy," The Journal of Machine Learning Research, vol. 25, pp. 1205-1224, 2004
- [2] L. Yu and H. Liu, "Feature Selection for High Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Learning, vol. 20, no. 2, pp. 856-863, 2003.
- [3] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.



- [4] 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.
- [5] Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure based on relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109, 2004.
- [6] Hall M.A. and Smith L.A., "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper", Proceedings of the Twelfth international Florida Artificial intelligence Research Society Conference, pp. 235-239, 1999
- [7] Yu L. and Liu H., "Redundancy based feature selection for microarray data", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 737-742, 2004