|| Volume 2||Issue 7 || 2017||ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

Efficient Feature Selection Through Graph Based Clustering

Pushpa Santosh Ghonge
ME 2nd Year, Computer Science & Engineering, Dr. Seema Quadri College of Engineering & Technology, Aurangabad, India

***Abstract*—In data mining the Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contain many redundant or irrelevant features. 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. Fast clustering based feature selection 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- subset selection, clustering, Fast Algorithm

I INTRODUCTION

We use the minimum spanning tree clustering algorithm. The simple algorithm is used to make possible subset of features and finding the one which minimizes the error rate. The proposed algorithm is used which is 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 selecting a subset of relevant features is used in model construction. 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 techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas 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 dimensionality reduction, elimination of inappropriate data, rising 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

Different feature selection algorithm present, most of them are useful at removing irrelevant features but not effective to handle 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. 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 choosing a subset of features by removing irrelevant information. it is the process of selecting 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 .

Whatever Relief is not useful at removing redundant features as two predictive but highly 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 relevant as well as redundancy among relevant features without pair wise correlation analysis.

Some different from above algorithm ,fast clustering based feature selection 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 accuracy. and secondly redundant features that do not redound to getting a better predictor for that they mostly provide information which is already situated in another feature.

III FEATURE SUBSET SELECTION

Feature subset selection must be 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.

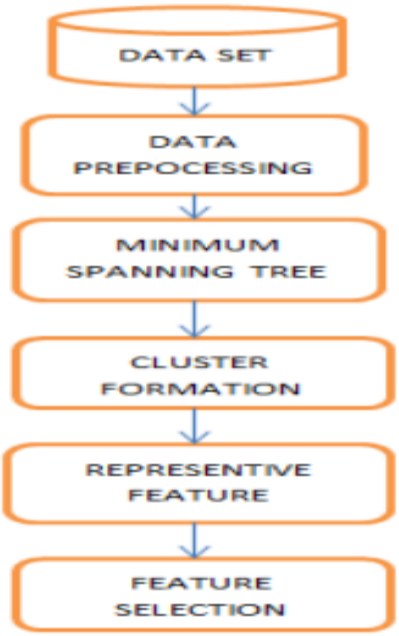


Figure 1: Feature selection process

In this feature subset algorithm the role of Symmentric 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 **F**ast clustering b**a**sed feature **S**election algori**T**hm (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 i €F be a feature, Si=F-{f i}and $Si' \subseteq Si$ .let si be the value assignment of all features in Si, f i is a value assignment of all features Fi and c be the value assignment of target class C. The Symmentric uncertainty is as follows

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

Where,

$Gain\ X(\ Y) = H\ (X) - H(X|Y)$

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

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

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

Where, H(X) is the entropy of discrete random variable X.And H(Y) is the entropy of discrete random variable Y. 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 instances, then delete 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.

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. Figure shows the proportion of selected features with all feature selection algorithm. And the graph of proportion of selected feature.

Ratio Comparison | Time Comparison

Graph

Selected Features ratio

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

Figure 2: Proportion of selected features

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

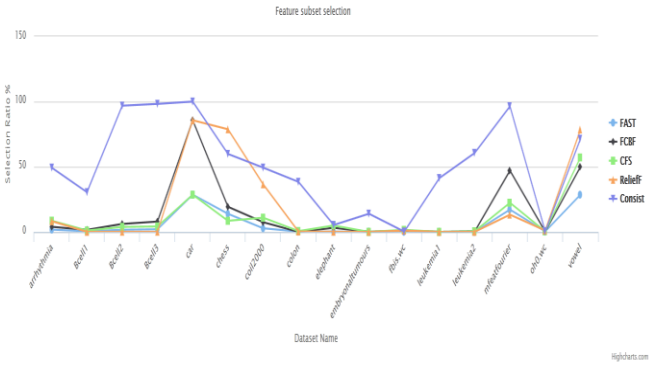


Figure 3: Graph of proportion of selected feature

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

Ratio Comparison | Time Comparison

Graph

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 4 : Time taken to select feature from dataset

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

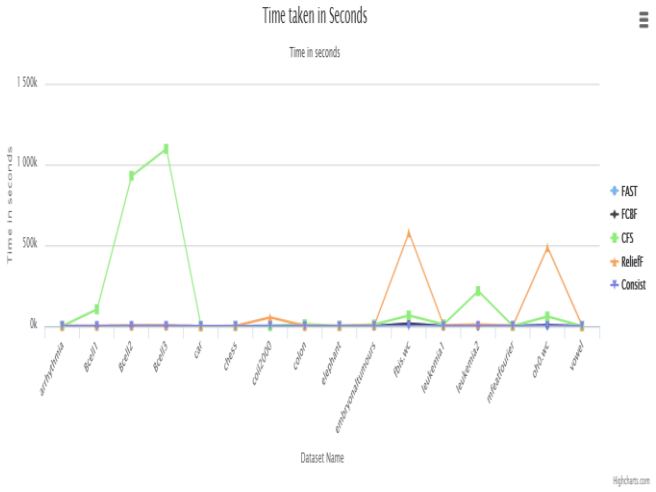
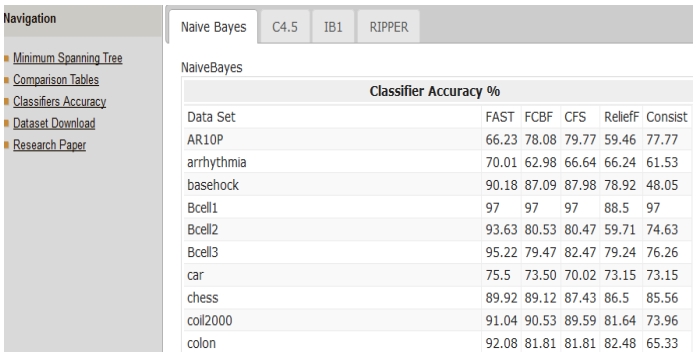


Figure 5 : Graph for time taken to select feature from dataset with respective feature selection algorithm

Accuracy of Naïve Bayesian classifier with feature selection algorithm.



Classifier Accuracy %

Data 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

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 Leaning, 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" , pp. 235-239, 1999.