International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 4, Issue 6, June 2015 **ISSN 2319 - 4847**

Improvised Method Of FAST Clustering Based Feature Selection Technique Algorithm For High Dimensional Data

Avinash Godase[1], Poonam Gupta[2]

[1]Post Graduate Student ,G.H.Raisoni College Of Engg & Mgmt,Wagholi,Pune

[2]H.O.D, Computer Engg, G.H.Raisoni College Of Engg & Mgmt,Wagholi,Pune

ABSTRACT

A high dimensional data is data consisting thousands of attributes or features. Nowadays for scientific and research applications high dimensional data is used.But.as there are thousands of features present in the data, We need to select the features those are non-redundant and most relevant in order to reduce the dimensionality and runtime, and also improve accuracy of the results. In this paper we have proposed FAST algorithm of feature subset selection and improved method of FAST algorithm. The efficiency and accuracy of the results is evaluated by empirical study. In this paper, we have presented 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 highly reduced. The Proposed System will be Implementation of FAST algorithm Using Dice Coefficient Measure to remove irrelevant and redundant features.

Keywords: FAST, Feature Subset Selection, Graph Based Clustering, Minimum Spanning Tree.

1. INTRODUCTION

With the goal of choosing a subset of good features with respect to the target classes, feature subset selection is an proper way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility[01][02]. Basically there are four methods for features selection i. e Wrapper , Filter , Embedded and Hybrid With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms[04]. Filter approach uses intrinsic properties of data for feature selection. This is the unsupervised feature selection approach. This approach performs the feature selection without using induction algorithms which is display in the figure. This method is used for the transformation of variable space. This transformation of variable space is required for the collation and computation of all the features before dimension reduction can be achieved [05][06]. In particular, we accept the minimum spanning tree based clustering algorithms, for the reason that they do not imagine that data points are clustered around centers or separated by means of a normal geometric curve and have been extensively used in tradition [02]. In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Select the features from the generated cluster Which remove the redundant and irrelevant attributes [03][13]. This method is use for selecting the interesting features from the clusters. Clustering is a semi-supervised learning problem, which tries to group a set of points into clusters such that points in the same cluster are more similar to each other than points in different clusters, under a particular similarity matrix[11][12]. Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because 1) irrelevant features do not contribute to the predictive 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)[07][09][10].

Using this method we can get the quality of feature attributes In this paper we have focused on using best similarity measure to calculate relevance between the features. In this paper we have compared results with few traditional feature selection algorithms like CFS,FAST..The goal of this paper includes focusing on using best algorithm i.e improved FAST for feature subset selection so that we will get effective accuracy[08].

2. LITERATURE SURVEY

In [02], Qinbao Song et al, proposed a new FAST algorithm that gain more accuracy and reduce time complexity than traditional feature selection algorithm like, FCBF, Relief, CFS, FOCUS-SF, Consist and also compare the classification accuracy with prominent classifiers. Graph theoretic clustering and MST based approach is used for ensure the efficiency of feature selection process. Classifiers plays vital roles in feature selection operation since accuracy of

selected features are measured using the progression of classifiers[06]. The following classifiers are utilized to classify the data sets [2], [3], Naïve Bayes: it works under Bayes theory and is based on probabilistic approach and yet then offers first-rate classification output. C4.5 is the successor of ID3 [4] support of decision tree induction method. Gain ratio, gini index information gain are the measures used for the process of attribute selection. Simplest algorithm is IB1 (instance based) [5]. Based on the distance vectors, it performs the classification process. RIPPER [6] is the rule based technique, it make a set of rules for the purpose of classify the data sets. Classifier is one of the evaluation parameter for measuring the accuracy of the process.

Author	Description
Kononenko	Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief, enabling this method to work with noisy and incomplete data sets and to deal with multi-class[2]
Yu L. and Liu H	FCBF is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis[01].
Fleuret F	CMIM iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked[14].
Krier C., Francois D	In this paper presented a methodology combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information[15].
Qinbao	The FAST algorithm works in two steps. 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. 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 feature[02].

3.PROPOSED SYSTEM

In proposed sytem i.e improved FAST ,we use relevance between features and the relevance between feature and the target concept.We have used dice coefficient for calculating the relevance between the features.We extract the best representative features from cluster using relevance between the features and the relevance between feature and target class. There exist many feature selection techniques which are aimed at reducing unnecessary features to reduce dataset. But some of them failed at removing redundant features after removing irrelevant features. Proposed system focuses on removing both irrelevant and redundant features. The features are first divided into clusters and features from each clusters are selected which are more representative to target class. System provides MST (Minimum Spanning Tree) method, using which we propose a Fast clustering based feature Selection algorithm (FAST).. Proposed System will be Implementation of FAST algorithm Using Dice Coefficient Measure to remove irrelevant and redundant features.

Feature selection is also useful as part of the data analysis process, as it shows which features are important for prediction, and how these features are related. Clustering is mainly used in grouping the datasets which are similar to the users search. The datasets which are irrelevant can be easily eliminated and redundant features inside the datasets are also removed. The clustering finally produces the selected datasets. The clustering uses MST for selecting the related datasets and finally the relevant datasets.

A minimum spanning tree (MST) or minimum weight spanning tree is then a spanning tree with weight less than or equal to the weight of every other spanning tree. More generally, any undirected graph (not necessarily connected) has a minimum spanning forest, which is a union of minimum spanning trees for its connected components. It is the cluster analysis of data with anywhere from a few dozen to many thousands of dimensions. Such high-dimensional data spaces are often encountered in areas such as medicine, where DNA microarray technology can produce a large number of measurements at once, and the clustering of text documents, where, if a word frequency vector is used, the number of dimensions equals the size of the dictionary.

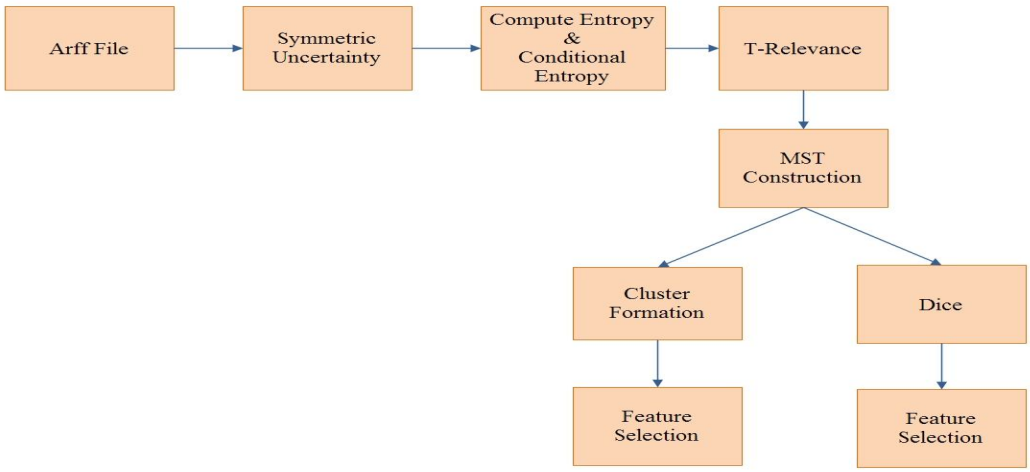

Fig.1 Proposed Framework

3.1 Proposed Algorithm

Input: D (F1, F2…, C) - the given data set
Θ - the T-Relevance threshold
Output: S – selected feature subset.
Start
//------Irrelevant Feature Removal---------//
1. for i=1 to m do
2. T-Relevance = SU (Fi, C)
3. if T-Relevance > Θ then
4. S = S U {Fi};
5. Feature f = Dice (Fi, Fj) // use dice to remove irrelevant feature.
//-------Minimum Spanning Tree Construction--------//
6. G = Null; //G is a complete Graph
7. for each pair of features {Fi, Fj} C S do
8. F- Correlation = SU (Fi,Fj)
9. Add Fi and Fj to G with F-Correlation as the weight of
The corresponding edge.
10. minSpanTree = Prim(G); //Using Prims Algorithm to generate the Minimum spanning tree
//-------Tree partition and Representative Feature Selection-------//
11. Forest = minSpanTree
12. for each edge Eij € Forest do
13. if SU(Fi , Fj) < SU (Fi , C) □ SU (Fi , Fj) < SU (Fj , C) then
14. Forest = Forest - Eij
15. S = φ
16. for each tree Ti € Forest do
17. Fr = argmax Fk € Ti SU (Fk, C)
18. S = S U {Fr}
19. return S
End

3.2 Proposed Methodology

Modules Information

1. Module (GUI Design and calculate Symmetric Uncertainty)

First module consists of development of application GUI in Java. Also includes the development of user registration and login parts. Again this module contains calculation of Symmetric Uncertainty to find the relevance of particular feature with target class

2. Module (Minimum Spanning Tree Construction)

In this module the construction of the minimum spanning tree from a weighted complete graph and then partitioning of the MST into a forest with each tree representing a cluster.

3. Module (Selection of Features)

In this module we do selection of most relevant features from the clusters which give us the reduced training dataset containing relevant an useful features only which improves efficiency.

4. Module (Dice)

In this module we will use similarity function dice coefficient clustering algorithm for clustering and selecting most relevant features from cluster.

3.3 Flow Chart Of The System

The following Diagram shows the flow chart for implementing the clustering based feature selection algorithm.

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The main assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques provide three main benefits when constructing predictive models:

• Improved model interpretability,
• Shorter training times,
• Enhanced generalization by reducing over fitting.

Feature selection is also useful as part of the data analysis process, as shows which features are important for prediction, and how these features are related.

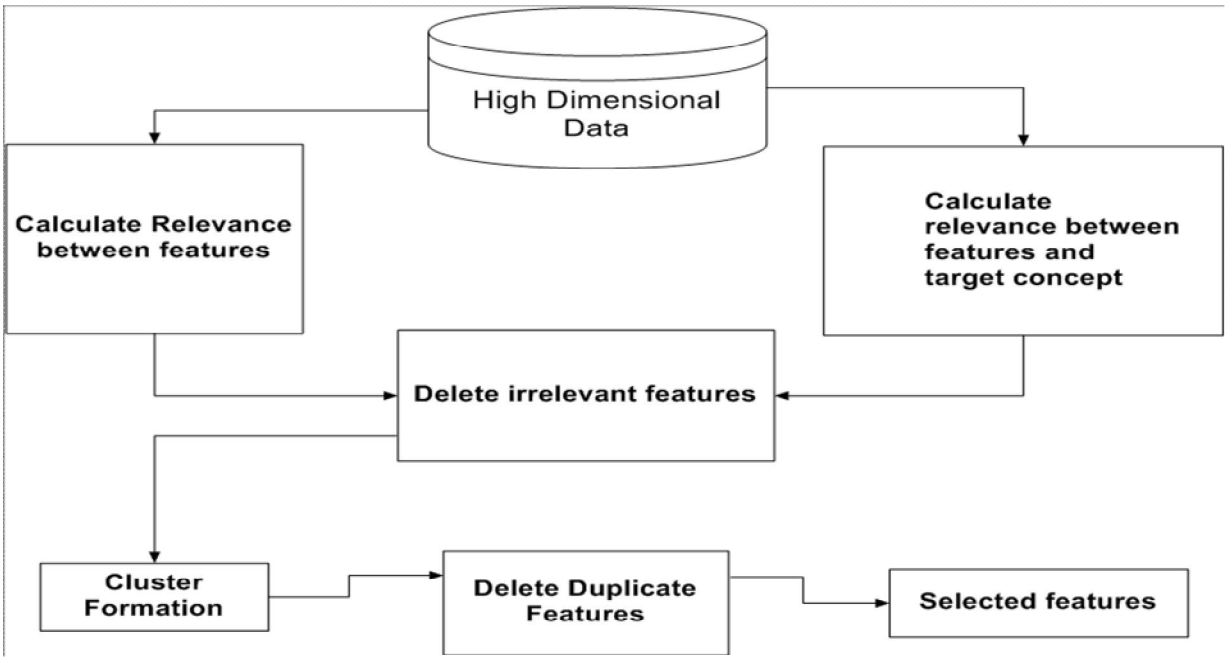

Fig.2 Flow Chart Of The System

4.RESULTS AND ANALYSIS

In this section we present the experimental results.The parameters used are

1. Proportion of the selected features,
2.The time to select the features
3.Classification accuracy.

We have performed experiments on three different algorithms i.e CFS,FAST, and Improved FAST.Proposed algorithm shows best results for these 3 cosidered parameters.

1) Proportion Of The Selected Features:

Table 1: Proportion Of The Selected Features

Dataset	CFS	FAST	Improved FAST
Fbis.wc.arff	50	29	19
New3s.wc.arff	40	30	25
Oh10.wc.arff	28	08	06
Oh0.wc.arff	41	11	10

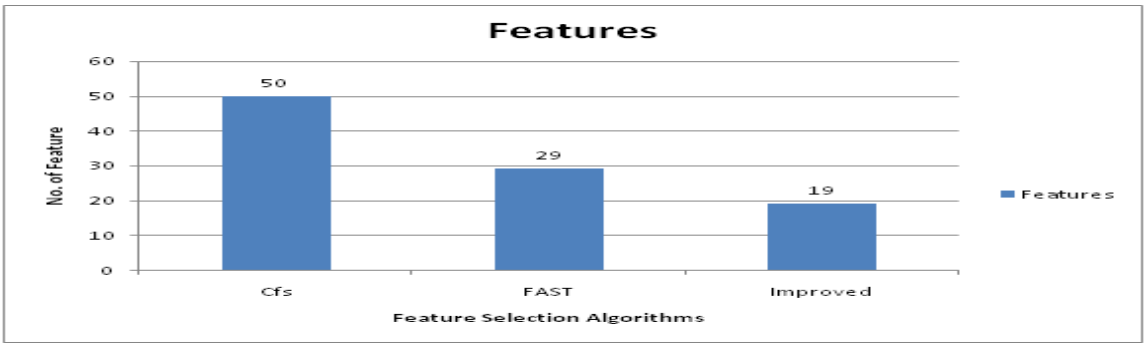

Fig3: Implementation Results Of Proportion Of The Selected Features

2) Time Required To Select Features

Table 2: Time Required To Select The Features

Dataset	CFS	FAST	Improved FAST
Fbis.wc.arff	17382ms	4416ms	3775ms
New3s.wc.arff	13236ms	3456ms	2800ms
Oh10.wc.arff	40486ms	2418ms	1846ms
Oh0.wc.arff	44486ms	2918ms	2246ms

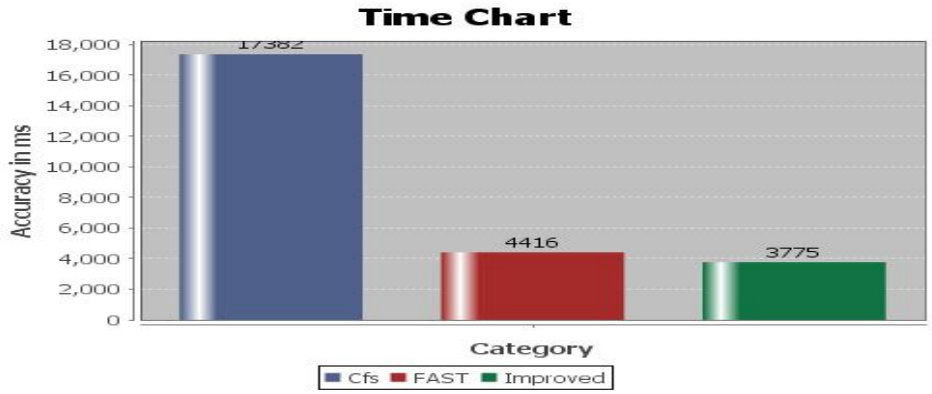

Fig 4: Implementation Results Of Time Required To Select The Features

3) Classification Accuracy:

Table 3: Classification Accuracy Of The Selected Features

Dataset	CFS	FAST	Improved FAST
Fbis.wc.arff	99.14	99.75	99.81
New3s.wc.arff	92.34	94.45	99.10
Oh10.wc.arff	94.71	96.65	98.69
Oh0.wc.arff	98.71	99.65	99.69

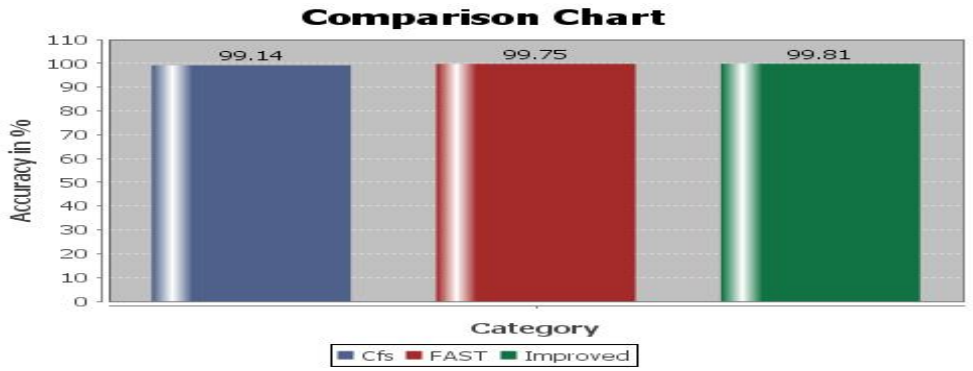

Fig 5: Implementation Results Of Accuracy Of The Algorithms To Select The Features

5.CONCLUSION

An improved clustering based feature subset selection algorithm for high dimensional data. The algorithm involves (i) deleting irrelevant features, (ii) devloping a minimum spanning tree from relative ones, and (iii) dividing 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 highly reduced. The performance of the proposed algorithm with feature selection algorithms like CFS, FAST on the different datasets is analysed. Proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy for Naive Bayes, C4.5, and RIPPER, and the second best classification accuracy for IB1FAST is the best algorithm amongst available algorithm for all kind of datasets Its efficiency can be increased by using different similarity measures like dice coefficient.

REFERENCES

[1] Mr.Avinash Godase, Mrs. Poonam Gupta, "A survey on Clustering Based Feature Selection Technique Algorithm For High Dimensional Data",International journal of emerging trends & technology in computer science,Volume 4, Issue 1, January-February 2015, ISSN 2278-6856.
[2] QinBao , "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data -in "ieee transactions on knowledge and data engineering" vol:25 no:1 year 2013.
[3] Kira K. and Rendell L.A., "The feature selection problem: Traditional methods and a new algorithm", In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134, 1992.
[4] Yu L. and Liu H., "Feature selection for high-dimensional data: a fast correlation-based filter) solution", in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.
[5] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., "On Feature Selection through Clustering", In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.
[6] Yu L. and Liu H.,Efficient feature selection via analysis of relevance and redundancy,journal of machine learning research,10(5),pp 1205-1224,2004
[7] Van Dijk G. and Van Hulle M.M., "Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis", International Conference on Artificial Neural Networks, 2006
[8] Krier C., Francois D., Rossi F. and Verleysen "M., Feature clustering and mutual information for the selection of variables in spectral data, In Proc European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning", pp 157-162, 2007.
[9] Zheng Zhao and Huan Liu in "Searching for Interacting Features", ijcai07
[10] P. Soucy, G.W. Mineau, A simple feature selection method for text classification, in: Proceedings of IJCAI-01, Seattle, WA, 2001, pp. 897–903
[11] Kohavi R. and John G.H., Wrappers for feature subset selection, Artif.Intell., 97(1-2), pp 273-324, 1997.
[12] Chanda P., Cho Y., Zhang A. and Ramanathan M., Mining of Attribute Interactions Using Information Theoretic Metrics, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355, 2009.
[13] Forman G., An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 3, pp 1289-1305,2003.
[14] Fleuret F., Fast binary feature selection with conditional mutual Information, Journal of Machine Learning Research, 5, pp 1531- 1555, 2004.
[15] C.Krier, D.Francois, F. Rossi, and M. Verleysen, "Feature Clustering and Mutual Information for the Selection of Variables in Spectral Data," Proc. European Symp. Artificial Neural Networks Advances in Computational Intelligence and Learning, pp. 157-162, 2007.