Improved Fast Clustering-Based Feature For Similarity Function Dice coefficient Algorithm for High-Dimensional Data

Ms Ashwini Patil
Computer Department
S.S.G.M.C.E
Shegaon,India

Prof Priti V Kale
Computer Department
S.S.G.M.C.E
Shegaon ,India

Abstract

Feature selection process is a identifying a subset of potential features that can be used to produce results similar to the set of original feature The feature selection should be characterized and viewed from the efficiency and effectiveness . Efficiency is a quality matter of time needed to find a subset of features where as effectiveness deals with the quality of features. These two factors are studied with the help of the fast clustering-based feature selection algorithm . In this paper, we have presented a novel clustering-based feature subset of selection algorithm for high dimensional data. The algorithm involves following steps Removing irrelevant features, Constructing a minimum spanning tree from relative ones, and partitioning and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treat as a single feature and thus dimensionality and dynamically reduced. In Proposed System will be Implementation of FAST algorithm Using proposed Dice Coefficient to Measure to remove irrelevant and redundant features

Keywords FAST,FCBF, RELIEF ,MST etc.

I. INTRODUCTION

With the aim of choosing a subset of good feature with respect to the target concepts, feature subset selection is an effective system for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than existing feature selection algorithms. Distributional clustering of words to reduce the dimensionality of feature text data. In cluster analysis, graph-theoretic methods have been well studied and used in various applications. The result of a forest and each tree in the forest represents a cluster. In our proposed study, by using graph-theoretic clustering methods to features. Spanning tree based clustering algorithms. Based on the MST method, we propose a Fast clustering Based feature selection Algorithm . Features in different clusters are relatively independent; the clustering based strategy of FAST has a high probability method producing a subset of useful and independent features.

II. LITERATUER SURVE

Useful Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as much as possible. This is because: irrelevant features do not contribute to the predictive accuracy, and redundant features do not getting a better predictor for they provide mostly information which is already present in other features. Many feature subset selection algorithms can effectively eliminate irrelevant features but fail to handle redundant features some of others can eliminate the irrelevant while taking care of the redundant features

Relief method is in effective at removing redundant features as two predictive but highly correlated features are likely both are highly weighted. Relief, enabling this method to work with noisy and incomplete data sets and to deal with multi class[4] FCBF method is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis[5].

CMIM method us iteratively picks features which maximize their mutual an information with the class to predict, conditionally to the response of any feature already picked[6].

III. RELATED WORK

EXISTING SYSTEM.

Traditional learning algorithms decision trees or artificial neural networks are best examples these uses embedded approaches for searching features. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the good feature of the selected subsets, the accuracy of the learning algorithms is usually high.Selected features is limited and the computing complexity is very large. The filter methods are independent of learning algorithms, with good generality. Computational complexity is very low, but the accuracy of the learning algorithms is not guaranted. Hybrid methods are used a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequence wrapper. They mainly focus on are combining filter and wrapper methods to achieve the best possible performance learning algorithm with similar time complexity of the filter methods. Relief-weights are assigned to instances Ineffective for removing redundant features Relief-F-Can work with noisy data but still cant remove redundant features CFS(Correlation Based Feature Selection) FCBF(Fast Correlation Based Filter Solution) and CMIM(Conditional Mutual Information Maximization).

Disadvantages of existing system:

1) The generality of the selected features is limited and the computational complexity is large.
2) Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.
3) 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.

GOAL

The aim of the project is to implement a Feature selection technique which reduces size of the dataset to be tested and improves quality along with efficiency.

Our objective is that to develop an algorithm which can efficiently find out irrelevant and redundant features that than of previous algorithms. We have proposed FAST algorithm using dice coefficient measure which can deliver effective results.

PROBLEM DEFINITION

Selection of a subset of good features with respect to the target concepts feature subset selection is an effective way for the reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result.For efficiency issue the time is required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on some criteria, a fast clustering based feature selection algorithm is proposed and experimentally . The FAST algorithm works in two steps. In this first step features are divided into clusters by using a graph theoretic approach and clustering methods. In the second step the most representative feature that is strongly related to target classes and selected feature from each cluster to form a subset of features..

PROPOSED SYSTEM

There are many existing feature selection techniques which are aimed at reducing unnecessary features to reduce dataset. But some of them are failed at removing redundant features after removing irrelevant features. Proposed system focuses on removing both the irrelevant and redundant features. The features are first divided into various clusters and features from each clusters are selected and which are more feasible .In this paper propose a Fast clustering based feature Selection algorithm .. Proposed system will be Implementation of FAST algorithm Using Dice

Coefficient Measure to remove irrelevant and redundant features

Advantages

Good feature subsets contain features highly correlated with the class, no correlated with each other. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset

PROPOSED METHODOLOGY

Modules Information

1. Graphical User Interface

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

2.Minimum Spanning Tree

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

3. Selection of Features

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

4.Contribution

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

FLOWCHART

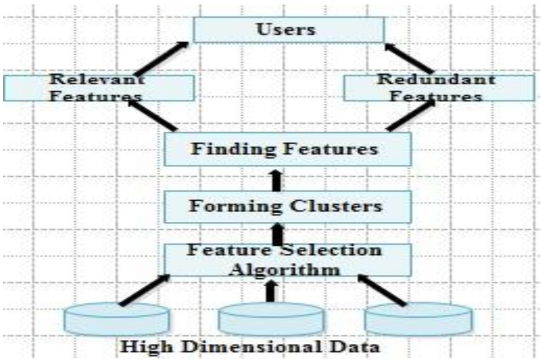

Fig 1 Flowchart FAST

PROPOSED ARCHITECTURE

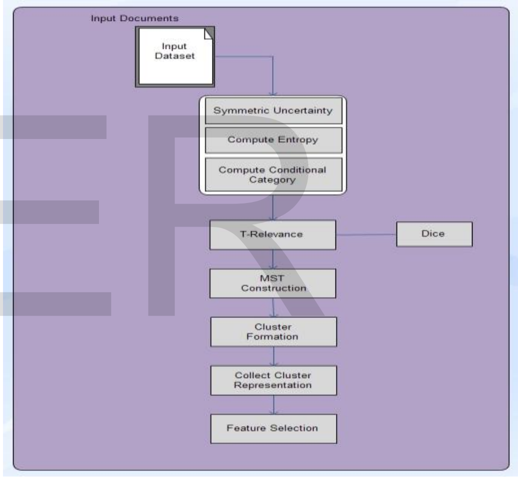

Fig 2 Proposed Architecture

The proposed system consist of Fast algorithm which removes irrelevant features and redundant features. The feature selection algorithm helps to identifying the relevant datasets. The irrelevant features will not have any relation with the logical datasets. The redundant features can be separated from cluster and can be easily eliminated. The FAST has a better performance with high dimensional data. The Subset selection algorithm (fig 2) invokes searching of the relevant datasets and cluster these datasets and eliminating the relevant features. The qualities of datasets are also efficient which satisfies the users search requirements. The Subsets are the sets containing another set that is important information which user needs at during the data retrieval. Feature selection techniques provide three main benefits when constructing predictive models such as improved model interoperability, shorter training times,

- Enhanced generalization by reducing the over fitting

Feature selection is also useful as part of the data analysis process, as it shows which features are important for prediction on the basis of how these features are related. Clustering is mainly used in grouping the datasets which are similar to the users search related query .

PROPOSED FRAMEWORK

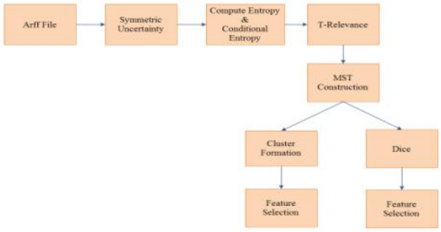

Fig 3 Proposed Framework

PROPOSED ALGORITHM

Similarity Function -Dice coefficient

Dice's Coefficient, is a term based similarity measure. It is a based similarity measure such as (0-1) where by the similarity measure is defined as twice much the number of terms common and compared entity's divided by the total number of terms and both tested entities. The Coefficient result of 1 indicates identical vectors as where a 0 equals orthogonal vectors.**Formula:**

It is calculated as follows.

$$Dice\ Coefficient = \frac{2 * C}{L1 + L2}$$

Where C is the number of character bigrams found in both strings S and T, L1 is the number of unique bigrams in string S and L2 is the number of unique bigrams in string T.

Algorithm:

The algorithm contains the following steps.

1. Split the S and T into two sets of 2-gram array S, array T.
2. Remove the duplicate 2-gram in array S, array T and get the number L1, L2 of
2-gram in array S and array T.
3. Combine two sets of 2-gram array S and array T to a new set of 2-gram Array Total, remove the duplicate 2-gram in array Total, and get the number L of 2-gram in array Total.
4. The variable C is calculated as follows.

$$C = (L1 + L2) - L$$

5. The dice coefficient is calculated as follows.

$$Dice\ Coefficient = \frac{2 * C}{L1 + L2}$$

FAST Algorithm

FAST is Tree-Based Algorithm and Advanced Chameleon is Graph-Based Algorithm. Features in different clusters are very relatively independent the clustering-based strategy of has a high probability of producing a subset of useful and independent features. To ensure that the efficiency of FAST, we adopt the efficient minimum spanning tree clustering method, for Chameleon we adopt the K means Nearest neighbor graph clustering method. Feature subset selection algorithms, most of them can effectively eliminate to the irrelevant features but the fail to handle redundant features. There are also algorithms that can be eliminate the irrelevant features also taking care of the redundant features.

FAST Algorithm step

```
inputs: D($F1$, $F2$, ..., $Fm$, $C$) - the given data set
$\theta$- the T-Relevance threshold.
output: S - selected feature subset .
//==== Part 1 : Irrelevant Feature Removal ====
for i = 1 to m do
    T-Relevance = SU ($Fi$, $C$)
    if T-Relevance >$\theta$then
        S = S ∪ {$Fi$};
//==== Part 2: Minimum Spanning Tree Construction ====
G = NULL; //G is a complete graph
for each pair of features {$F'i$, $F'j$} ⊂ S do
    F-Correlation = SU ($F'$,'$j$)
    $AddF'iand/orF'jtoGwit$ F-Correlation $asteweig\ toft\ ecorrespondingedge$;
min Span Tree = KRUSKALS(G); //Using KRUSKALS Algorithm to generate the minimum spanning tree
//==== Part 3: Tree Partition and Representative Feature Selection ====
Forest = min Span Tree
for each edge ∈Forest do
    if SU($F'i$, $F'j$) <SU($F'i$, $C$) ∧SU($F'i$, $F'j$) <SU($F'j$,$C$) then
        Forest = Forest − $Eij$
S = $\phi$
for each tree ∈Forest do
    $FjR$= argmax$F'k$∈$Ti$SU($F'k$,$C$)
    S = S ∪ {$FjR$};
return S
```

MATHEMATICAL MODEL

It is calculated as follows.

$$Dice\ Coefficient = \frac{2 * C}{L1 + L2}$$

Where C is the number of character bigrams found in both strings S and T, L1 is the number of unique bigrams in string S and L2 is the number of unique bigrams in string T.
The variable C is calculated as follows.

$$C = (L1 + L2) - L$$

HARDWARE REQUIREMENT:

Processor	-	PIV
RAM	-	2GB
Hard Disk	-	40GB

SOFTWARE REQUIREMENT:

Operating System	: Windows
Programming Language	: JAVA
Java Version	: JDK

OUTPUT

Accuracy of FAST and DICE				
Algo	Naïve Bayes	C4.5	IB1	RIPPER
FAST	70	69	60	65
DICE	72	71	65	68

Table 1 Shows Accuracy of FAST & DICE

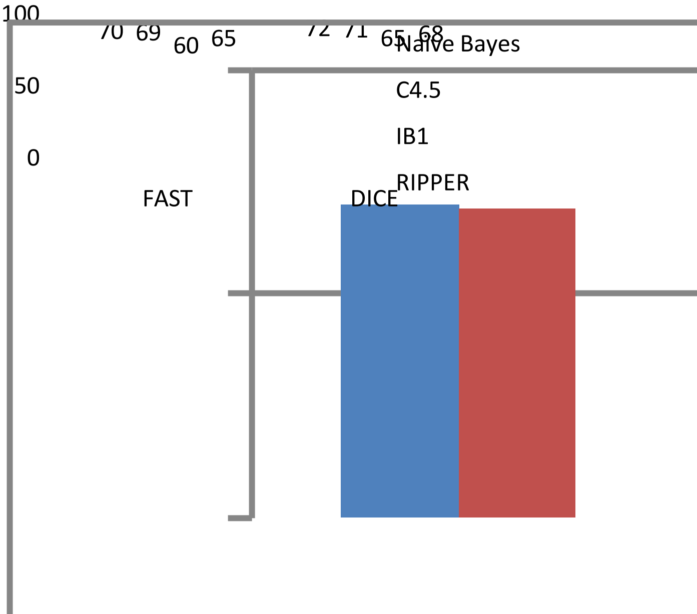

Fig 5 Performance Comparison between FAST & DICE

CONCLUSION

The performance of the proposed algorithm compare the five well-known feature selection algorithms FCBF, ReliefF, CFS, Consist, and FOCUS-SF on the image, microarray, and text data. proposed algorithm obtained the best proportion of selected features at the best runtime, and the best classification accuracy for few algorithm such as Naive Bayes, C4.5, and RIPPER, and another second best classification accuracy for IB1FAST is the best algorithm amongst available algorithm for all kind of data including text, image, microarray Its efficiency can be increased by using different similarity measures like dice coefficient

FUTUER ENHANCEMENT

For the future work plan to explore different types of correlation measures, and study some formal properties of feature space.

REFERENCES

[1]. QinBao , “A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data - Song in "IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING” VOL:25 NO:1 YEAR 2013.

[2]. 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.

[3]. Koller D. and Sahami M., “Toward optimal feature selection”, In Proceedings of International Conference on Machine Learning, pp 284-292, 1996.

[4]. Kononenko I., Estimating Attributes “ Analysis and Extensions of RELIEF”, In Proceedings of the 1994 European Conference on Machine Learning, pp 171-182, 1994.

[5]. 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.

[6]. Fleuret F., “ Fast binary feature selection with conditional mutual Information”, Journal of Machine Learning Research, 5, pp 1531-1555, 2004.

[7]. 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.

[8]. 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

[9]. 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.