INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

IMPLEMENT EFFICIENT AND EFFECTIVE FAST CLUSTERING-BASED FEATURE SELECTION ALGORITHM FOR HIGH-DIMENSIONAL DATA

MS ASHWINI PATIL, PROF PRITI V KALE

Computer Department, S.S.G.M.C.E, Shegaon, India.

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract: Feature selection is widely used in preparing high-dimensional data for effective data mining. Getting fast popularity in the social media dataset presents new challenges for feature selection. Social media data consists of traditional high-dimensional, attribute-value data such as posts, tweets, comments, and images, and linked data that describes the relationships between social media users as well as who post the posts, etc. The nature of social media also determines that its data is massive, noisy, and incomplete, which exacerbates the al-ready challenging problem of feature selection. a fast clustering-based feature selection algorithm, FAST, is proposed and experimentally evaluated in this paper. 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 a subset of features. In this paper, we illustrate the various differences between attribute-value dataset and social media information, that investigate if linked data can exploited in a proposed feature selection frame-work by giving advantage of social science theories, Hence it extensively evaluated the effects of user-user and user-post relationships manifested in linked data on feature selection, and discuss some research issues for proposed work.

Keywords: FAST, FCBF, RELIEF, MST etc.

Corresponding Author: MS ASHWINI PATIL

Access Online On:

www.ijpret.com

How to Cite This Article:

Ashwini Patil, IJPRET, 2016; Volume 4 (9): 43-57

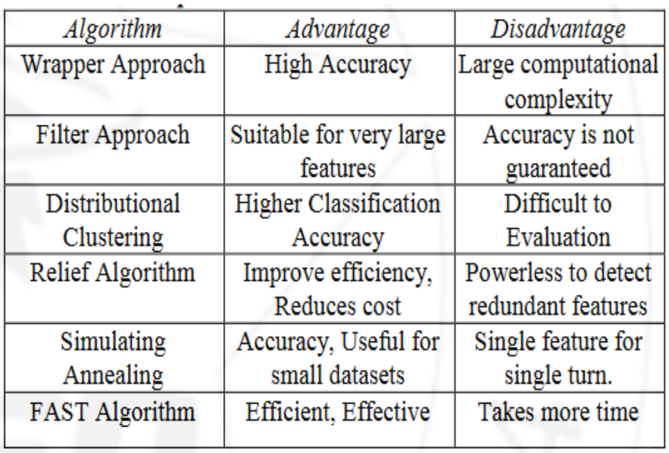
PAPER-QR CODE

INTRODUCTION

The feature selection algorithm may be seen as the combination of a search technique and with an evaluation measure which scores the different feature subsets. The simplest algorithm is that algorithm to test each possible subset of features finding and minimizes the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods [1]. The criterion function guiding the search for the best features is usually some kind of reparability measure between classes. It can be either classifier independent (i.e., filter approach) or classifier specific (i.e., wrapper approach or embedded method). Wrapper methods use as a predictive model for score feature subsets selection. The wrapper methods train by a new model for each subset, they are very computationally intensive, but this method provides the best performing feature subset for that particular type of model [6]. Filter approaches methods are use a proxy measure instead of the error rate to score a feature subset. The Filter methods are usually less computationally intensive than wrappers, but they produce a feature set which is not tuned to a specific type of predictive model. Many filters methods provide a feature ranking rather than an explicit of best feature subset selection, and the cut - off point in the ranking is chosen by cross-validation [7].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.

- **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



Algorithm	*Advantage*	*Disadvantage*
Wrapper Approach	High Accuracy	Large computational complexity
Filter Approach	Suitable for very large features	Accuracy is not guaranteed
Distributional Clustering	Higher Classification Accuracy	Difficult to Evaluation
Relief Algorithm	Improve efficiency, Reduces cost	Powerless to detect redundant features
Simulating Annealing	Accuracy, Useful for small datasets	Single feature for single turn.
FAST Algorithm	Efficient, Effective	Takes more time

Fig 1 Comparison of different algorithm

FETURE SUBSET SELECTION ALGORITHM

Feature subset selection is a long existing technique to deal with problems brought by too many features [1]. A feature subset selection method is usually made up of two parts: a feature subset generator and an evaluator. The two parts work together to find the feature subset which meets evaluation criteria best. Feature subset generator can also be seen as a search engine, which can be divided into three categories: exhaustive search engine, heuristic search engine and nondeterministic search engine [1]. Throughout the ([5, 6, 7, and 9]), feature subset selection approaches are categorized into three main groups: filter methods, wrapper methods and embedded approaches. Filter methods rely on general characteristics of the training data to estimate and select subsets of features without involving a learning algorithm. Contrary to that, wrapper approaches use a classification algorithm as a black box to assess the prediction accuracy of various subsets. The last group, embedded approaches, performs the feature selection process as an integral part of the machine learning algorithm. In the following, these three techniques are described in detail. The overview of these three approaches is given in Figure 1

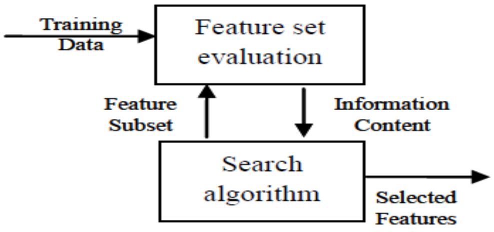


Figure 1: Basic feature (Filter) selection

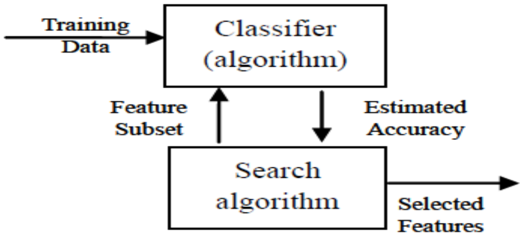


Figure 2 Basic feature (Wrapper) selection

- **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 guaranteed. 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 can't 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.

CHALLENGES

In feature extraction scenario, various algorithms have been proposed for feature selection In the existing FAST feature extraction algorithm, main focus is on both removing unnecessary as well as immaterial data that means it extracts only targeted features. Problem in existing algorithm is that when it removes irrelevant features, it considers only single feature and match up to that to the target feature. If that feature matches then it extracts that feature otherwise remove it. So in this existing method the challenge lies in the fact that if more than one

feature are joint and they suit the target feature then it can be treated as relevant. Feature interface is the new challenge for identifying the applicable feature. In existing system where feature interface is not supported that proves to be the modification criteria for this work.

PURPOSE

Feature selection, also recognized as a changeable selection, attribute collection or variable extraction, is the process of selecting a compartment of applicable features for use in model creation. The central supposition, when using a feature selection method, is that the data contains many redundant or irrelevant features. Unneeded features are those which offer no extra information than the presently

selected features, and irrelevant features supply no useful information in any context. The main and important reason of the feature extraction algorithm is that they discover out or extract only targeted features out of many features. They don't measure the irrelevant and redundant data because irrelevant and redundant data affects the competence and effectiveness of the algorithm. In existing algorithm that uses a variety of different techniques to decide relevant features or removing

irrelevant or redundant features , when it removes the irrelevant features it does not consider the interface of different features. Proposed algorithm not only removes the irrelevant features but also focus on interaction of the features.

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 ARCHITECTURE

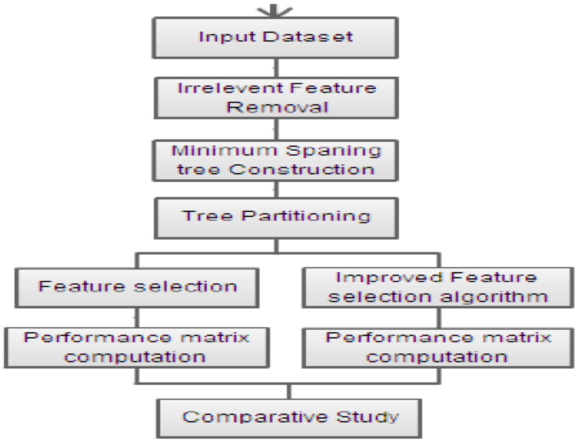

Input Dataset
Irrelevent Feature Removal
Minimum Spaning tree Construction
Tree Partitioning
Feature selection
Improved Feature selection algorithm
Performance matrix computation
Performance matrix computation
Comparative Study

Fig 3 Proposed Architecture

MODULES

User Module

In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first.

Distributed Clustering

The Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Pereira et al. or on the distribution of class labels associated with each word by Baker and McCallum . As distributional clustering of words are agglomerative in nature, and result in suboptimal word clusters and high computational cost, proposed a new information-theoretic divisive algorithm for word clustering and applied it to text classification. proposed to cluster features using a special metric of distance, and then makes use of the of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluation measure based on distance does not identify a feature subset that allows the classifiers to improve their original

l performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower.

Subset Selection Algorithm

The Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, “good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. Keeping these in mind, we develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

Time Complexity

The major amount of work for this algorithm involves the computation of SU values for TR relevance and F-Correlation, which has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity in terms of

the number of features m. Assuming features are selected as relevant ones in the first part, when k ¼ only one feature is selected.

Flow chart:

The following Diagram shows the flow chart for implementing the clustering based feature selection algorithm. Feature Selection Algorithm Forming Clusters Finding Features

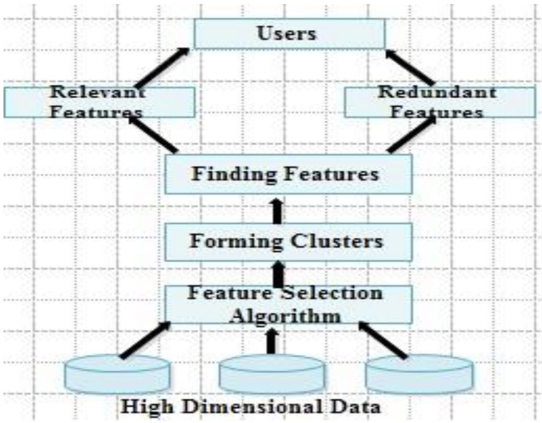


Fig 4 Flowchart

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 and clustering algorithm for clustering and selecting most relevant features from cluster.

MATHEMATICAL MODEL

John et al. [8] presented a definition of relevant features.

Suppose F to be the full set of features, Fi # F be a feature, Si

= F −$\{Fi\}$ and $S'i \subseteq Si$.

Relevant feature:

Fi is relevant to the target concept C if and only if there exists

some $s'\ i$, fi and c, such that, for probability $p(S'\ i$ = $s'\ i$,$F\ i$ =

fi) > 0, $p(C$ = c | $S'\ i$ = $s'\ i$,$F\ i$ = fi) ⁄= $p(C$ = c | $S'\ i$ = $s'\ i$).

Markov blanket:

Given a feature $Fi \in F$, let $Mi \subset F$ (Fi ⁄∈ Mi), Mi is said to be

a Markov blanket for Fi if and only if $p(F$−Mi −$\{Fi\}$,C| Fi,M

i)=$p(F$ −Mi −$\{Fi\}$,C| Mi)**.**

Redundant feature: Let S be a set of features, a feature in S is redundant if and only if it has a Markov Blanket. The symmetric uncertainty

(SU) [12] is derived from the mutual information by

normalizing it to the entropies of feature values or feature values and target classes.

The symmetric uncertainty is defined as follows

$SU(X,Y\ )$=2×$Gain(X|Y\ )\ H(X)$+$H(Y\ )$

H(X)=−Σ $x \in X\ p(x)$log2 $p(x)$.

$Gain(X|Y)$=$H(X)$−$H(X|Y)$

=$H(Y)$−$H(Y|X)$.

$H(X|Y\ )$=−Σ $y \in Y\ p(y)$ Σ $x \in X\ p(x|y)$log2 $p(x|y)$.

T-Relevance:

The relevance between the feature $Fi \in F$ and the target concept C is referred to as the T-Relevance of Fi and C, and denoted by $SU(Fi,C)$.

F-Correlation:

The correlation between any pair of features Fi and Fj ($Fi,F\ j$

$\in F \wedge i \not= j$) is called the F-Correlation of Fi and Fj, and

denoted by $SU(Fi,F\ j)$**.**

F-Redundancy:

Let S = {$F1$,$F2$,...,Fi,...,$F\ k$<|F|} be a cluster of features. if

$\exists Fj \in S,SU\ (Fj,C) \geq SU(Fi,C) \wedge SU(Fi,F\ j) > SU\ (Fi,C)$ is

always corrected for each $Fi \in S$.

R-Feature:

A feature $Fi \in S$ = {$F1$,$F2$,...,Fk} (k<|F|) is a representative

feature of the cluster S (i.e. Fi is a R-Feature) if and only if,

Fi = arg maxFj∈S $SU(Fj,C)$**.**

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

θ- 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
```

DATASET

The data has to be pre-processed for removing missing values, noise and outliers. Then the given dataset must be converted into the arff format. From the arff format, only the attributes and the values are extracted and stored into the database. By considering the last column of the dataset as the class attribute and select the distinct class labels from that and classify the entire dataset with respect to class labels.

CONCLUSION

An Efficient FAST clustering-based feature subset selection algorithm for high dimensional data improves the efficiency of the time required to find a subset of features. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) 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 and improved the classification accuracy.

FUTUER ENHANCEMENT

In future scope, different correlation Measures along with fuzzy logic can be included in the present algorithm to improve performance of a system. We can enhance this work by extending

The symmetric uncertainty for extracting

The feature subset selection.

REFERENCES

1. Qin Bao, “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.

10. Guyon I. and Elisseeff A., “An introduction to variable and feature selection,” The Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

11. 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.

12. Peng H. C., Long F. H., and Ding C., “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, 2005.

13. Banks A., Vincent J., and Anyakoha C., “A review of particle swarm optimization. Part i: background and development,” Natural Computing, vol. 6, no. 4, pp. 467-484, 2007.

14. Azevedo G. L. F. B. G., Cavalcanti G. D. C., and Filho E. C. B. C., "An approach to feature selection for keystroke dynamics systems based on pso and feature weighting," in Proc. IEEE Congress on Evolutionary Computation (CEC'97), pp. 3577-3584, 2007.