Analysis of New Clustering Algorithm on High Dimensional Data based on Feature Sub Set Selection

Ponduru Praveen Kumar [#1]

M.Tech Scholar [#1],

Department of Computer Science & Engineering,

Vignan Institute of Information Technology,
Visakhapatnam, AP, India.

.

Abstract

Clustering high-dimensional data is the cluster analysis of data with anywhere from a few dozen to many thousands of dimensions. While doing clustering feature selection mainly involves in identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency of this feature selection algorithm mainly tells the time required to find a subset of features, the effectiveness is related to the quality of the subset of extracted features. Based on these criteria, a fast clustering-based feature selection algorithm, FAST, is proposed and experimentally evaluated in this paper. The FAST algorithm which we proposed in this paper 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. 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 features. In order to ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree clustering method. We have conducted several experiments to compare FAST and several representative feature selection algorithms that are already available in order to extract feature selection. Finally our experimental results tells that FAST not only produces smaller subsets of features but also improves the performances of the all the existing classifiers.

Keywords

Clustering, Graph-Based Clustering, Filter Method, Feature Subset Selection, Feature Clustering

1. Introduction

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 central 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 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. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points). The archetypal case is the use of feature selection in analysing DNA microarrays, where there are many thousands of features, and a few tens to hundreds of samples. Feature selection techniques provide three main benefits when constructing predictive models:

1. Improved Model Interpretability,
2. Shorter Training Times,
3. Enhanced Generalization By Reducing Over Fitting.

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. With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility [1], [2].

Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories [3]. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches[4]. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed [5], [6], [7]. The hybrid methods are a combination of filter and wrapper methods [8], [9], [10] by using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods.

2. Background Knowledge

In this section we will describe the assumptions and background knowledge that is used for developing the new feature selection process.

Feature Selection (or) Extraction

A "**feature**" or "**attribute**" or "**variable**” refers to an aspect of the data. Usually before collecting data, features are specified or chosen. Features can be discrete, continuous, or nominal. Generally, features are characterized as: 3 ways.

a) Relevant
b) Irrelevant
c) Redundant

a) Relevant

These are features which have an influence on the output and their role cannot be assumed by the rest. This type of feature extraction or selection is known as relevant feature selection mechanism.

b) Irrelevant

Irrelevant features are defined as those features not having any influence on the output, and whose values are generated at random for each example.

c) Redundant

Redundancy exists whenever a feature can take the role of another (perhaps the simplest way to model redundancy).

Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features [11],

[12], [13], [14] yet some of others can eliminate the irrelevant while taking care of the redundant features [15]. Our proposed FAST algorithm falls into the second group.

Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is ineffective at removing redundant features are likely both to be highly weighted [16]. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multi-class problems, but still cannot identify redundant features.

3. Proposed Algorithm

The following is the main feature subset selection algorithm that we have proposed in this paper for extracting a best feature sub set selection from a set of multi-dimensional data.

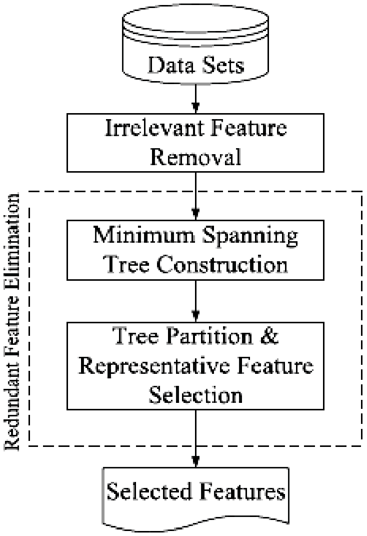

Data Sets
Irrelevant Feature Removal
Redundant Feature Elimination
Minimum Spanning Tree Construction
Tree Partition & Representative Feature Selection
Selected Features

Figure. 1: Framework of the proposed feature subset selection algorithm

3.1 Constructing an Feature Subset Selection Algorithm

Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines [17]. 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.”, as shown in figure 1.

3.2 Algorithm Design

The proposed FAST algorithm logically consists of three steps:

(i) Removing irrelevant features,
(ii)Constructing a MST from relative ones, and
(iii) Partitioning the MST and selecting representative features.

For a data set D with m features F = {F1, 2,...,Fm} and class C, we compute the T-Relevance $SU(Fi,C)$ value for each feature Fi ($1 \leq i \leq m$) in the first step. The features whose (Fi,) values are greater than a predefined threshold θ comprise the target-relevant feature subset F' = {F' 1,′ 2,...,F' k} ($k \leq m$). In the second step, we first calculate the F-Correlation (F' i,′ j) value for each pair of features F' i and F' j (F' i,F' $j \in F' \wedge i \neq j$). Then, viewing features F' i and F' j as vertices and $SU(F'$ i,F' $j)$($i \neq j$) as the weight of the edge between vertices F' i and F' j, a weighted complete graph G =(V,E) is constructed where V = {F' i | F' $i \in F$ ′ $\wedge$ $i \in$ [1,k]} and E = {(F' i,F' j) | (F' i,F' $j \in F' \wedge i,j \in$[1 ,k] $\wedge$ $i \neq j$}. As symmetric uncertainty is symmetric further the F-Correlation (F' i,′ j) is symmetric as well, thus G is an undirected graph. The complete graph G reflects the correlations among all the target-relevant features. Unfortunately, graph G has k vertices and (k−1)/2 edges. For high dimensional data, it is heavily dense and the edges with different weights are strongly interweaved. Moreover, the decomposition of complete graph is NP-hard. Thus for graph G, we build a MST, which connects all vertices such that the sum of the weights of the edges is the minimum, using the well-known Prim algorithm [54]. The weight of edge $(F'\ i,'\ j)$ is F-Correlation $SU(F'\ i,F'\ j)$. After building the MST, in the third step, we first remove the edges $E = \{(F'\ i,F'\ j) \mid (F'\ i,F'\ j \in F' \wedge i,j \in [1\ ,k] \wedge i \neq j\}$, whose weights are smaller than both of the T-Relevance $SU(F'\ i,C)$ and $SU(F'\ j,C)$, from the MST. Each deletion results in two disconnected trees $T1$ and $T2$. Assuming the set of vertices in any one of the final trees to be $V\ (T)$, we have the property that for each pair of vertices $(F'\ i,'\ j \in V\ (T))$, $SU(F'\ i,F'\ j) \geq SU(F'\ i,C) \vee SU\ (F'\ i,F'\ j) \geq SU(F'\ j,C)$ always holds. From Definition 6 we know that this property guarantees the features in $V\ (T)$ are redundant.

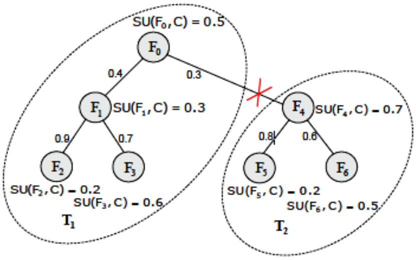
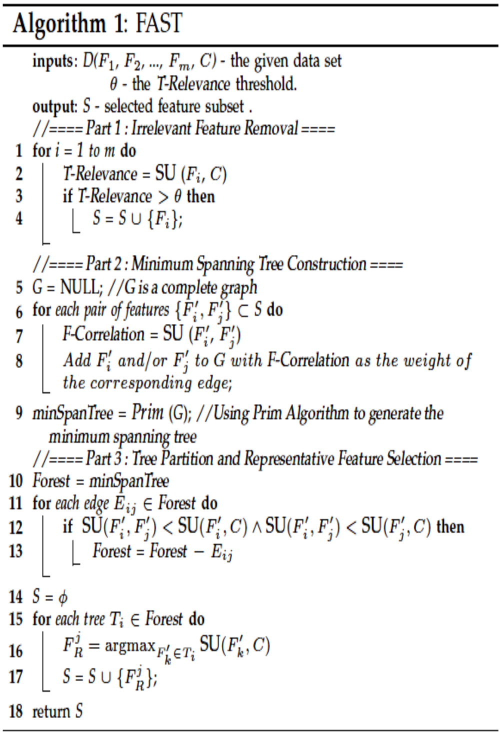


Figure. 2: Example of the clustering step

This can be illustrated by an example. Suppose the MST shown in Figure.2 is generated from a complete graph G. In order to cluster the features, we first traverse all the six edges, and then decide to remove the edge $(F0,\ F4)$ because its weight $SU(F0,\ F4) = 0.3$ is smaller than both $SU(F0,\ C) = 0.5$ and $SU(F4,\ C) = 0.7$. This makes the MST is clustered into two clusters denoted as $V\,(T1)$ and $V\,(T2)$. Each cluster is a MST as well. Take $V\,(T1)$ as an example.

From Fig.2 we know that $SU(F0,\ F1) > SU(F1,\ C)$, $SU(F1,\ F2) > SU(F1,\ C) \wedge SU(F1,\ F2) > SU(F2,\ C)$, $SU(F1,\ F3) > SU(F1,\ C) \wedge SU(F1,\ F3) > SU(F3,\ C)$. We also observed that there is no edge exists between $F0$ and $F2$, $F0$ and $F3$, and $F2$ and $F3$. Considering that $T1$ is a MST, so the $SU(F0,\ F2)$ is greater than $SU(F0,\ F1)$ and $SU(F1,\ F2)$, $SU(F0,\ F3)$ is greater than $SU(F0,\ F1)$ and $SU(F1,\ F3)$, and $SU(F2,\ F3)$ is greater than $SU(F1,\ F2)$ and $SU(F2,\ F3)$. Thus, $SU(F0,\ F2) > SU(F0,\ C) \wedge SU(F0,\ F2) > SU(F2,\ C)$, $SU(F0,\ F3) > SU(F0,\ C) \wedge SU(F0,\ F3) > SU(F3,\ C)$, and $SU(F2,\ F3) > SU(F2,\ C) \wedge SU(F2,\ F3) > SU(F3,\ C)$ also hold. As the mutual information between any pair $(Fi,\ Fj)(i,\ j = 0,\ 1,\ 2,\ 3 \wedge i \neq j)$ of $F0$, $F1$, $F2$, and $F3$ is greater than the mutual information between class C and Fi or Fj, features $F0$, $F1$, $F2$, and $F3$ are redundant.

After removing all the unnecessary edges, a forest $Forest$ is obtained. Each tree $Tj \in Forest$ represents a cluster that is denoted as $V\,(Tj)$, which is the vertex set of Tj as well. As illustrated above, the features in each cluster are redundant, so for each cluster $V\ (Tj)$ we choose a representative feature Fj_R Whose *T-Relevance* $SU(F^j{}_R,\ C)$ is the greatest. All $F^j{}_R\ (j = 1...|Forest|)$ comprise the final feature subset $\cup F^j{}_R$. The details of the FAST algorithm is shown in Algorithm 1.



Algorithm 1: FAST

```
inputs: $D(F_1, F_2, ..., F_m, C)$ - 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 $(F_i, C)$
    if T-Relevance $> \theta$ then
        $S = S \cup \{F_i\}$;
//==== Part 2 : Minimum Spanning Tree Construction ====
G = NULL; //G is a complete graph
for each pair of features $\{F'_i, F'_j\} \subset S$ do
    F-Correlation = SU $(F'_i, F'_j)$
    Add $F'_i$ and/or $F'_j$ to G with F-Correlation as the weight of
    the corresponding edge;
minSpanTree = Prim (G); //Using Prim Algorithm to generate the
    minimum spanning tree
//==== Part 3 : Tree Partition and Representative Feature Selection ====
Forest = minSpanTree
for each edge $E_{ij} \in$ Forest do
    if $\mathrm{SU}(F'_i, F'_j) < \mathrm{SU}(F'_i, C) \wedge \mathrm{SU}(F'_i, F'_j) < \mathrm{SU}(F'_j, C)$ then
        Forest = Forest $- E_{ij}$
$S = \phi$
for each tree $T_i \in$ Forest do
    $F^j_R = \mathrm{argmax}_{F'_k \in T_i} \mathrm{SU}(F'_k, C)$
    $S = S \cup \{F^j_R\}$;
return $S$
```

4. Implementation Modules

Implementation is the stage where the theoretical design is automatically converted into practically by dividing this into various modules. We have implemented the current application in Java Programming language with JEE as the main interface for developing the proposed application with Front End as HTML, JSP Pages and Back end as MY SQL data base for storing and retrieving the records. Our proposed application is divided into following 4 modules. They are as follows:

- **a)** User Module
- **b)** Distributed Clustering Module
- **c)** Subset Selection Module
- **d)** Time Complexity Module

a) 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.

b) Distributed Clustering Module

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

c) Subset Selection Module

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.

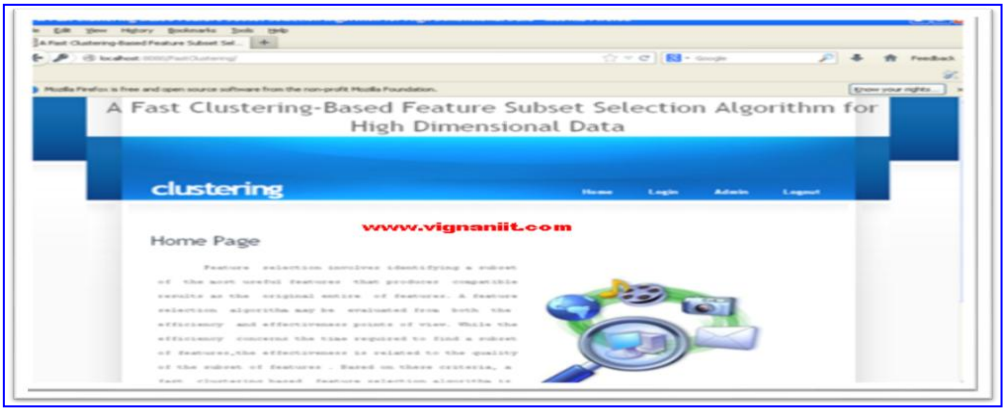
d) Time Complexity Module

The major amount of work for Algorithm 1 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 = only one feature is selected.

5. Experimental Results

In this paper, we have proposed a new Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data for reducing the processing time as well as the search time. For this we have shown the application in a web interface with JEE 6.0 edition .In this JEE we are using front end as Java Server pages (JSP) and HTML pages. As we are deploying the application in web interface, we are using tomcat server for deploying the application .Hence we use tomcat 7.0 as the deployment web server. This application is mainly implemented in a simulation manner where we are not directly connecting with Google server for retrieving the data from the Google server as that was already implemented in real time, but we are using My Sql data base for storing the data temporarily on to our system and then retrieve the same data whenever needed in a tree fashion.

5.1 Main Page



A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data

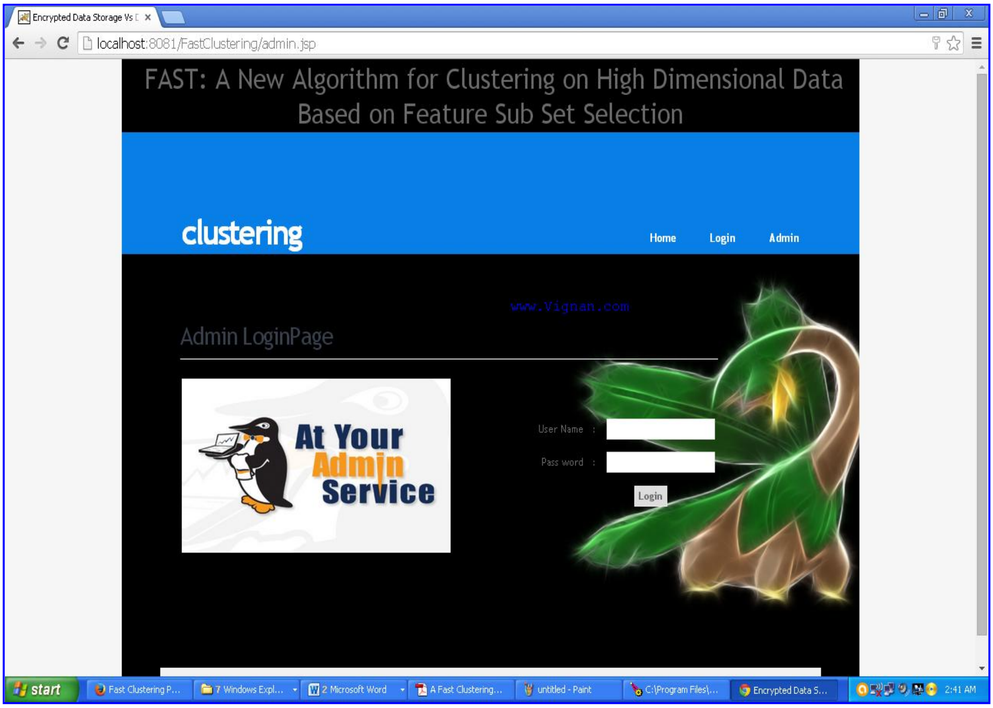
www.vignaniit.com

Home Page

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features,the effectiveness is related to the quality of the subset of features . Based on these criteria, a fast clustering based feature selection algorithm is

A Fast Clustering-Based Feature Subset Sel...
localhost:8080/FastClustering/
Mozilla Firefox is free and open source software from the non-profit Mozilla Foundation.
Know your rights...
clustering
Home
Login
Admin
Logout

5.2 Admin Login Page



FAST: A New Algorithm for Clustering on High Dimensional Data Based on Feature Sub Set Selection

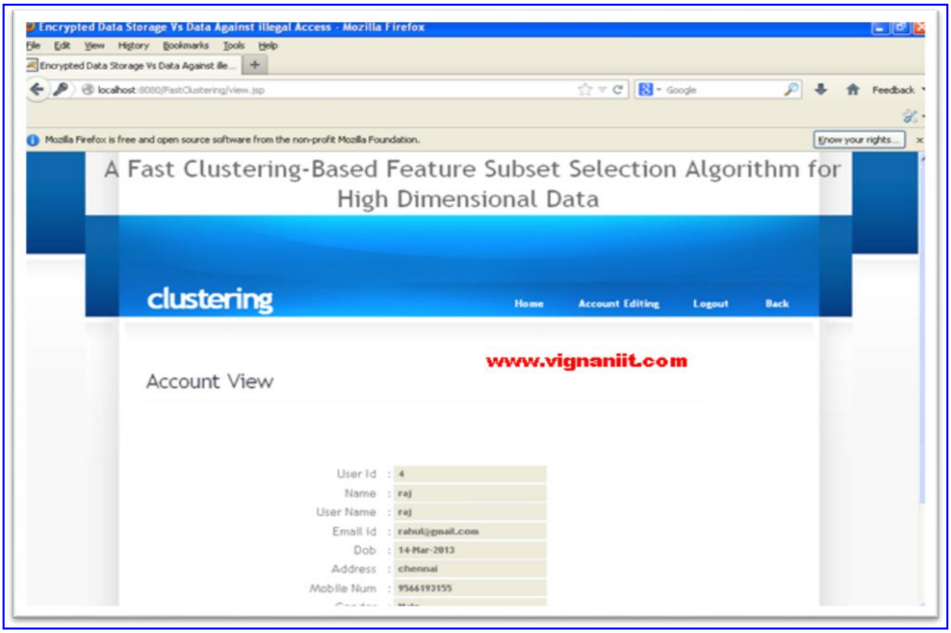
www.Vignan.com

Admin LoginPage

User Name :

Pass word :

5.3 User Registration Page

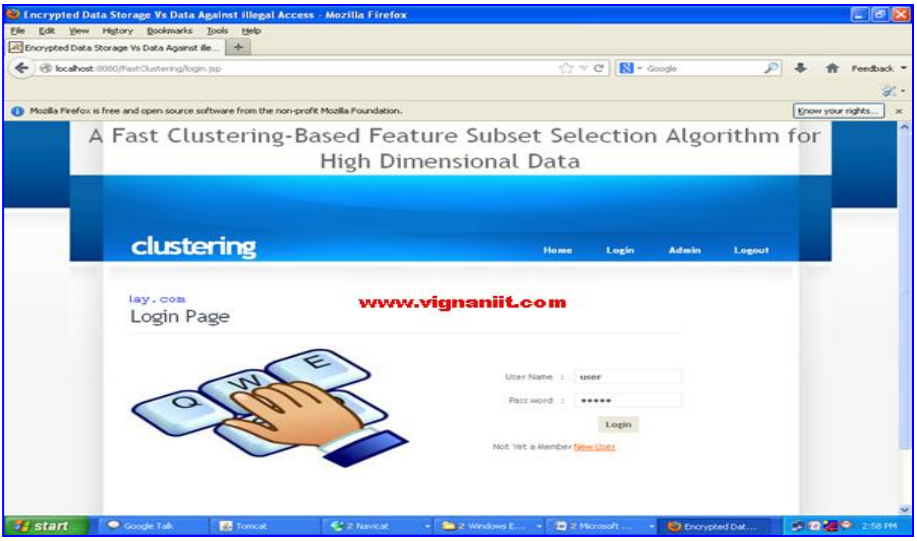

Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Encrypted Data Storage Vs Data Against ille...
localhost:8080/FastClustering/view.jsp
Google
Feedback
Mozilla Firefox is free and open source software from the non-profit Mozilla Foundation.
Know your rights...
clustering
Home
Account Editing
Logout
Back
www.vignaniit.com

A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data

Account View

User Id : 4
Name : raj
User Name : raj
Email Id : rahul@gmail.com
Dob : 14-Mar-2013
Address : chennai
Mobile Num : 9566193155

5.4 User Login Page


Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Encrypted Data Storage Vs Data Against ille...
localhost:8080/FastClustering/login.jsp
Google
Feedback
Mozilla Firefox is free and open source software from the non-profit Mozilla Foundation.
Know your rights...
clustering
Home
Login
Admin
Logout
lay.com
www.vignaniit.com
Q
W
E
start
Google Talk
Tomcat
2 Navicat
2 Windows E...
2 Microsoft ...
Encrypted Dat...
2:58 PM

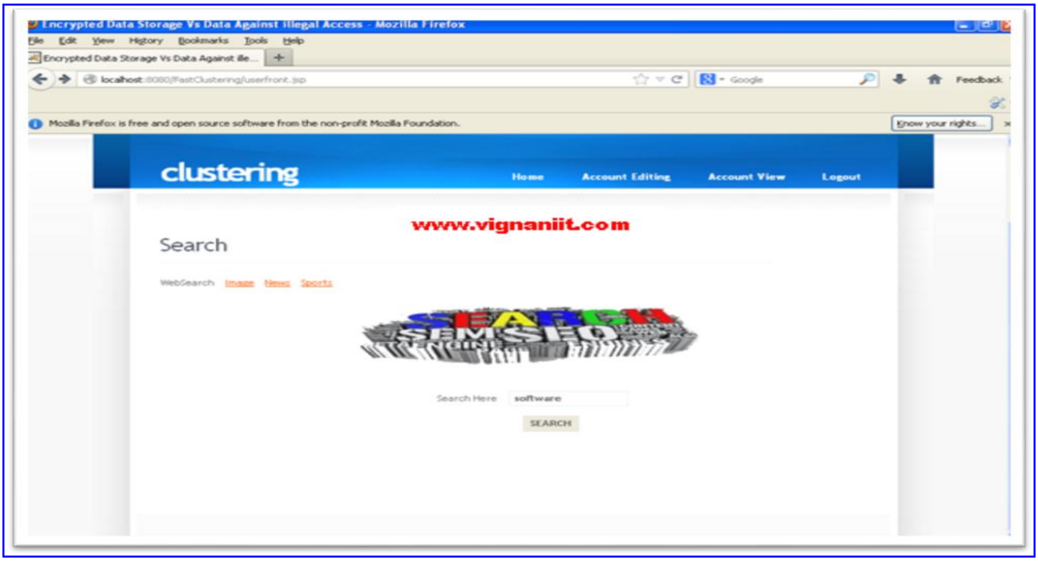
A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data

Login Page

User Name : user
Pass word : •••••
Login
Not Yet a Member New User

5.5 Search Main Page

In the below window, as our proposed application is using FAST algorithm ,we extended that proposed algorithm for analyzing with different types of search like Web Search,Image,News,Sports.


Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Encrypted Data Storage Vs Data Against ille...
localhost:8080/FastClustering/userfront.jsp
Google
Feedback
Know your rights...
clustering
Home
Account Editing
Account View
Logout
www.vignaniit.com
WebSearch
Image
News
Sports
SEARCH
SEARCH

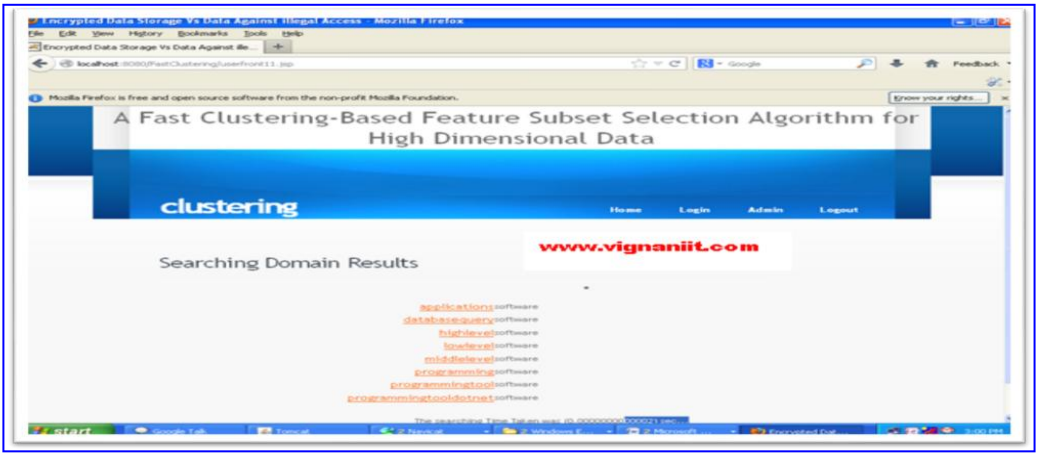
Mozilla Firefox is free and open source software from the non-profit Mozilla Foundation.

Search

Search Here software

5.6 Domain Search Main Page

In the below window, we got a result from the domain search if we give the keyword as software. We will get different types of software related links which all come under category of software.


Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Encrypted Data Storage Vs Data Against ille...
localhost:8080/FastClustering/userfront11.jsp
Google
Feedback
Know your rights...
clustering
Home
Login
Admin
Logout
www.vignaniit.com
start
Google Talk
Tomcat
2 Navicat
2 Windows E...
2 Microsoft ...
Encrypted Dat...
3:00 PM

Mozilla Firefox is free and open source software from the non-profit Mozilla Foundation.

A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data

Searching Domain Results

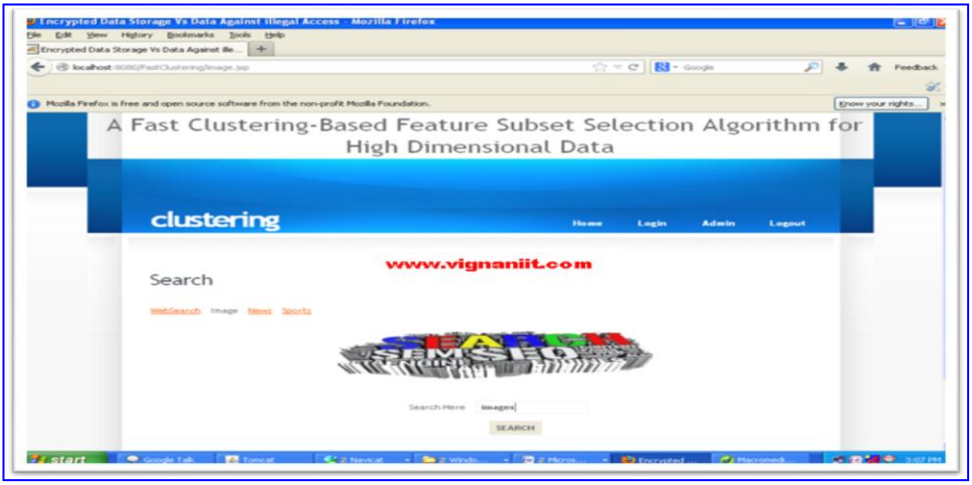
applicationsoftware
databasequerysoftware
highlevelsoftware
lowlevelsoftware
middlelevelsoftware
programmingsoftware
programmingtoolsoftware
programmingtooldotnetsoftware

The searching Time Taken was (0.000000000000021 sec...

5.7

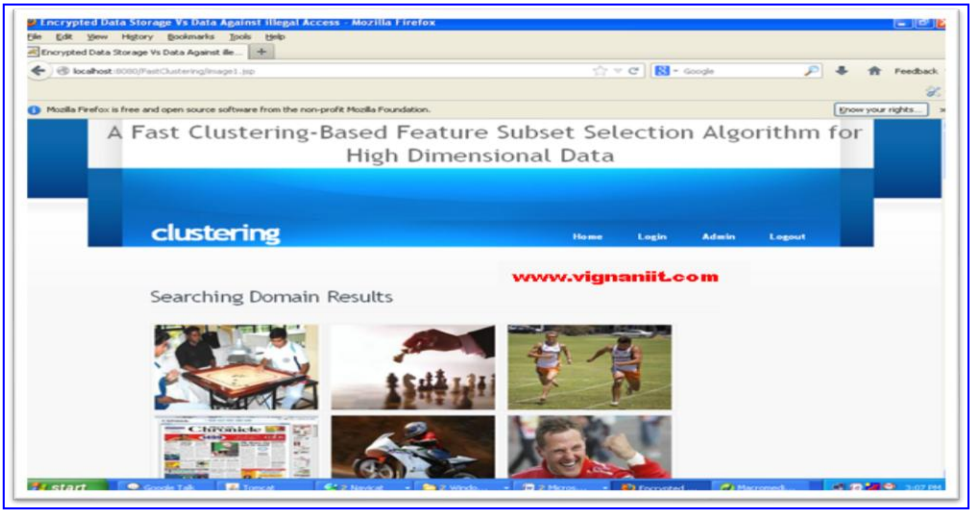
5.8 Search Based on Image

In the below window, we got a result from the image search if we give the keyword as image. We will get different types of images which come under category of images.


Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
File Edit View History Bookmarks Tools Help
localhost:8080/FastClustering/image.jsp
Mozilla Firefox is free and open source software from the non-profit Mozilla Foundation.
Know your rights...
A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data
clustering
Home
Login
Admin
Logout
www.vignaniit.com
Search
WebSearch
Image
News
Sports
Search Here
images
SEARCH
start

5.9 Search Result Based on Image

In the below window we will find the images what we got after search from the keyword like image.


Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
File Edit View History Bookmarks Tools Help
localhost:8080/FastClustering/image1.jsp
Mozilla Firefox is free and open source software from the non-profit Mozilla Foundation.
Know your rights...
A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data
clustering
Home
Login
Admin
Logout
www.vignaniit.com
Searching Domain Results
start

6. Conclusion

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 drastically reduced. We also found that FAST obtains the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data in terms of classification accuracy of the four different types of classifiers.

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

7. References

[1] Liu H., Motoda H. and Yu L., Selective sampling approach to active feature selection, Artif. Intell., 159(1-2), pp 49-74 (2004).

[2] Molina L.C., Belanche L. and Nebot A., Feature selection algorithms: A survey and experimental evaluation, in Proc. IEEE Int. Conf. Data Mining, pp 306-313, 2002.

[3] Guyon I. and Elisseeff A., An introduction to variable and feature selection, Journal of Machine Learning Research, 3, pp 1157-1182, 2003.

[4] Mitchell T.M., Generalization as Search, Artificial Intelligence, 18(2), pp 203-226, 1982.

[5] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.

[6] Souza J., Feature selection with a general hybrid algorithm, Ph.D, University of Ottawa, Ottawa, Ontario, Canada, 2004.

[7] Langley P., Selection of relevant features in machine learning, In Proceedings of the AAAI Fall Symposium on Relevance, pp 1-5, 1994.

[8] Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.

[9] Xing E., Jordan M. and Karp R., Feature selection for high-dimensional genomic microarray data, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 601-608, 2001.

[10] Yu J., Abidi S.S.R. and Artes P.H., A hybrid feature selection strategy for image defining features: towards interpretation of optic nerve images, In Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 8, pp 5127-5132, 2005.

[11] Forman G., An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 3, pp 1289-1305, 2003.

[12] Hall M.A., Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, In Proceedings of 17th International Conference on Machine Learning, pp 359-366, 2000.

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

[14] 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.

[15] 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.

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

[17] Kohavi R. and John G.H., Wrappers for feature subset selection, Artif. Intell., 97(1-2), pp 273-324, 1997.

8. About the Authors

Ponduru Praveen Kumar is currently pursuing his 2 Years M.Tech (CSE) in Department of Computer Science and Engineering at Vignan Institute of Information Technology (VIIT),Visakhapatnam. His area of interests includes Data Mining.