Survey On Effective Feature Selection Methods And Algorithms For High Dimensional Data

S.I.S.Jaffarvalli
M.Tech student, Department of CSE
AITS (autonomous)
Rajampet, India
ac.jafar2@gmail.com

G.Surya Narayana
Assistant professor, Department of CSE
AITS (autonomous)
Rajampet, India
surya.aits@gmail.com

Abstract— In terms of high dimensional data a cluster is defined is a connected region of multi dimensional space containing a relative high density of points separated from other such regions contains a relative low density of points. Cluster analysis is an unsupervised learning. The main aim of this paper is to remove irrelevant and redundant features and increase the level of accuracy. In the proposed Feature selection and extraction is the special form of dimensionality decline where feature selection is the subfield of feature extraction mining. The efficiency concerns related to time for finding subset of features and the effectiveness is related to the quality of the subset selections. To ensure the quality and efficiency of FAST, the efficient minimum spanning tree algorithm called MST clustering method is implemented. The proposed extensive experiments are carried out to compare FAST algorithm and several representative feature subset selection algorithms, namely, ReliefF, FCBF, CFS, FOCUS and Consist with respect to various types of well-known classifiers.

Keywords- ***High dimensional data ,Cluster,FCBF, Irrelevant and redundant features.***

I. Introduction

Data mining provides various analysis functionalities, algorithms for ascertaining the interesting knowledge from huge amounts of data warehouses or information repositories. Data mining tasks and activities are specified by its functionalities that mining tasks are classified into two forms:

1. Descriptive mining tasks are a represent to provide general properties of the data.
2. Predictive mining tasks are various implications on the current data order to craft Data prediction.

Feature subset selection is the mode of identifying the good number of features that fabricate well suited and the outcome as the unique entire set of features. Feature data Extraction is the special form of dimensionality reducing where feature selection is the subfield of feature extraction. Feature subset selection algorithms essentially have two basic constraints named as, quality and time requirement. Feature subset selection involves identifying a subset of the most useful features that produces compatible output results as the original entire set of features.

II. Related work

A. *Survey views of Authors*

According R. Agrawal and R. Srikant Fast algorithms for mining association rules in large databases provides information of mining association rules in large databases. According to H. Grosskreutz, B. Lemmen, and S.

R¨uping. Secure distributed subgroup discovery in horizontally partitioned data provide a new technique for secured distributed sub groups in horizontally portioned data. According to A.V. Evfimievski, R. Srikant R.Agrawal, and Gehrke. Privacy preserving mining of association rules , it privacy preserving mining of association rules. According to D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. A fast distributed algorithm for mining association rules, it fast distributed algorithm. According to Brickell and V. Shmatikov. Privacy-preserving graph algorithms in the semi-honest model it privacy preserving graph algorithms.

B. System modules

- Removal of Irrelevant features:

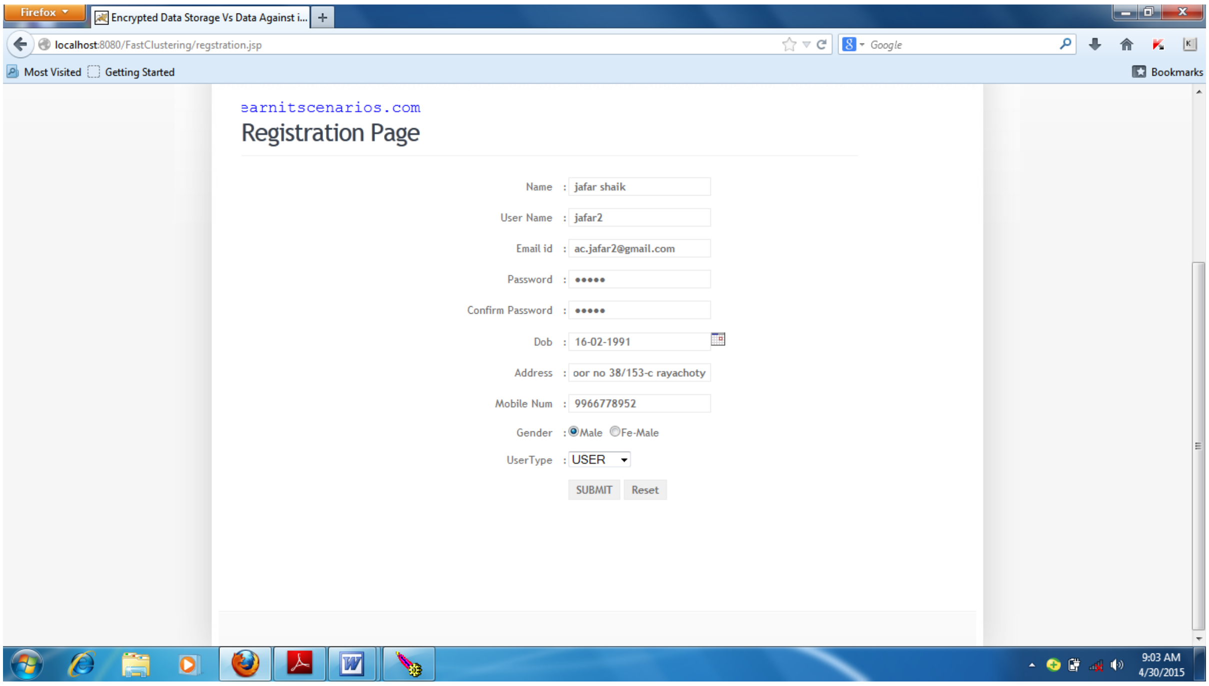
An effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed for machine learning applications. if we take a Dataset 'D' with m features F={F1,F2,..,Fn} and class C, automatically features are available with target relevant feature. The generality of the selected features is limited and the computational complexity is large. 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.

- MST construction:

To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. Extensive experiments are carried out to compare FAST and several representative feature selection algorithms. We construct a Minimal spanning tree with weights. 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.

III. RESULTS

An old user's can directly logged in into the clustering system by using there Id's. If any new users want to enter into the clustering folders then he/she must be signup into the system at first then they are able to get clustering the files by using there user ID.


Firefox
Encrypted Data Storage Vs Data Against i...
localhost:8080/FastClustering/regstration.jsp
Google
Most Visited
Getting Started
Bookmarks
9:03 AM
4/30/2015

earnitscenarios.com

Registration Page

Name	:	jafar shaik
User Name	:	jafar2
Email id	:	ac.jafar2@gmail.com
Password	:	●●●●●
Confirm Password	:	●●●●●
Dob	:	16-02-1991
Address	:	oor no 38/153-c rayachoty
Mobile Num	:	9966778952
Gender	:	● Male ○ Fe-Male
UserType	:	USER

SUBMIT Reset

Figure 1. user registation

the details of the clustering items like web searching, image, News, Sports. The user will be selected by using their requirements. This used for searching the items. Feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features.Idea behind feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features.

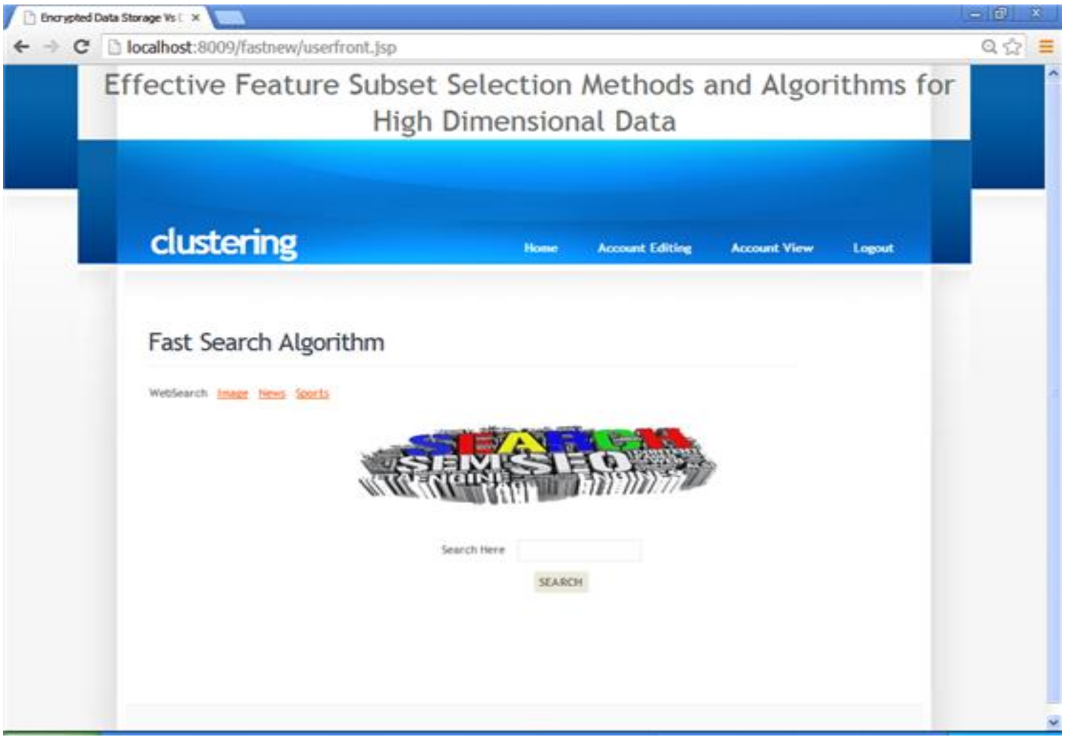

Encrypted Data Storage Vs
localhost:8009/fastnew/userfront.jsp
Effective Feature Subset Selection Methods and Algorithms for High Dimensional Data
clustering
Home
Account Editing
Account View
Logout
Fast Search Algorithm
WebSearch
Image
News
Sports
SEARCH
SEM SEO
Search Here
SEARCH

Figure 2. user searching images

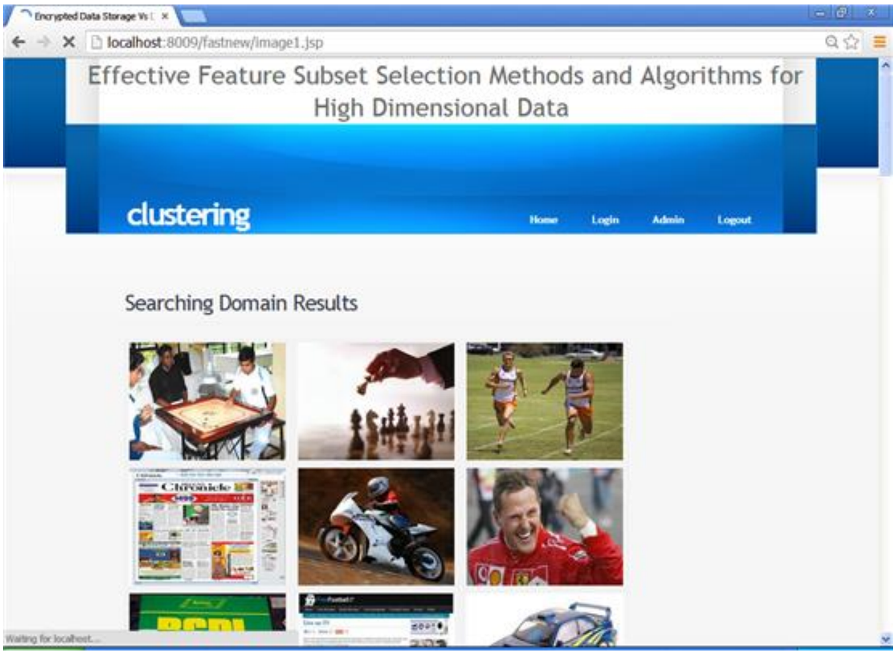
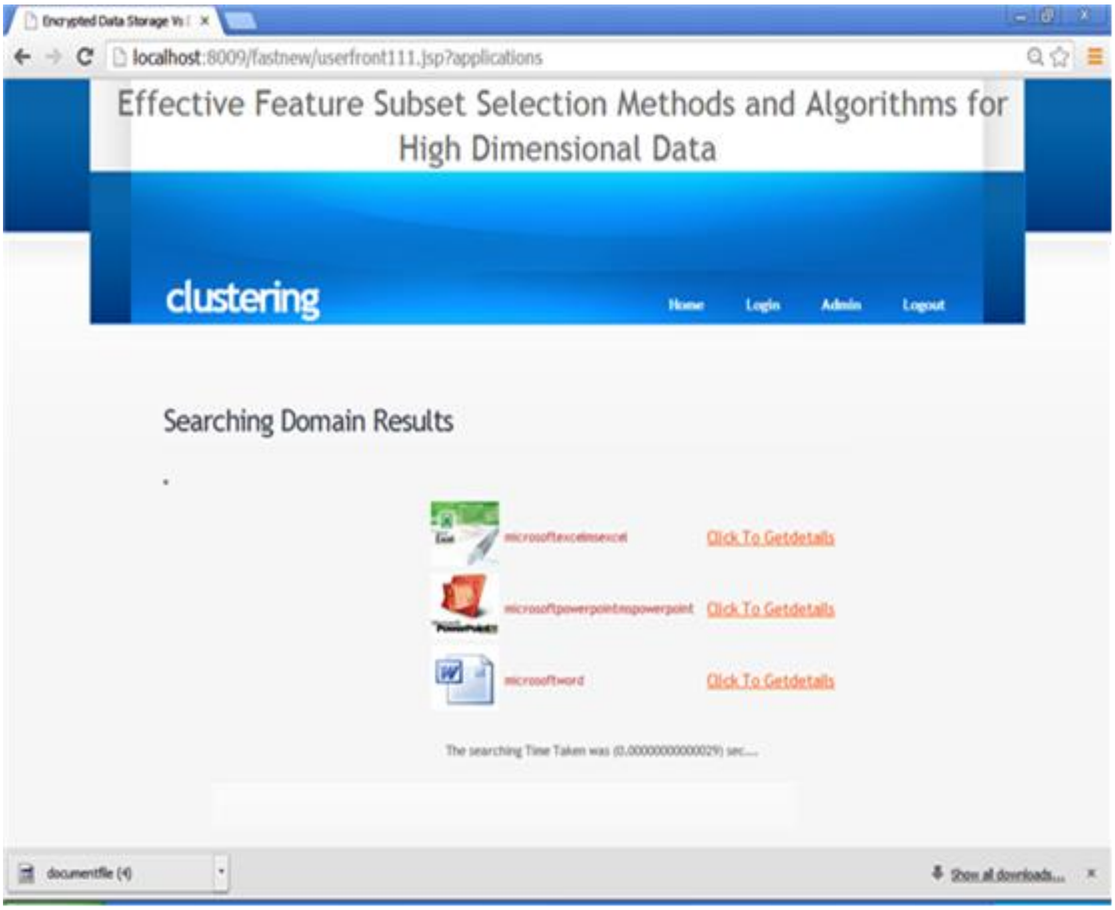

Encrypted Data Storage Vs
localhost:8009/fastnew/image1.jsp
Effective Feature Subset Selection Methods and Algorithms for High Dimensional Data
clustering
Home
Login
Admin
Logout
Searching Domain Results
Waiting for localhost...

Figure 3. results of searched information



Effective Feature Subset Selection Methods and Algorithms for High Dimensional Data

Searching Domain Results

microsoftexcelmsexcel

microsoftpowerpointmspowerpoint

microsoftword

The searching Time Taken was (0.00000000000029) sec....

Figure 4. Minimzed time result

IV. CONCLUSION

The main aim of activity of fast subset is to determine a minimal activity subset from a problem domain while keeping a suitably high correctness in representing the unique features. In present real world problems fast subset is a must due to the abundance of too many noisy, irrelevant or misleading subset features. The 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 IB1.when compared with other algorithms fast algorithm is capable of searching the information with a few nanoseconds whaich we have proved in the above results. Feature Selection is an important research direction of rough set application. In feature we implement this algorithm for high dimensional data and variable data.

REFERENCES

[1] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.

[2] 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.

[3] 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

[4] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96- 103, 1998.

[5] *Jaffarvalli S I S, Suryanarayana G* ,Eliminating Two-Shake Problem in Cluster Analysis using Jafar Surya Algorithm, *in proceedings of* IEEE ICECCT 2015 International conference, pp 617-620 , 2015.