Based on Image Search a Unique Feature Subset Selection Process For High Proportional Clustering

Alekhya Juttiga[1], B Nandan Kumar[2]

[1]M. Tech Student
[2]Assistant Professor
[1,2]Dept. of Computer Science & Engineering, D.N.R College of Engineering & Technology, Bhimavaram, Andhra Pradesh

---***---

Abstract - *Article selection encompasses identifying a subset of the most valuable features that produces harmonious consequences as the unique entire set of features. Based on these criteria, a fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated in this paper.. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the excellence of the subset of features. The FAST algorithm works in two steps. A feature selection algorithm may be weighed from both the efficiency and efficiency points of view In the first step, features are allocated into clusters by using graph-hypothetical clustering methods. In the second step, the most expressive feature that is strongly related to aim classes is designated from each cluster to form a subset of features. To ensure the productivity of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method. Extensive experiments are carried out to compare FAST and several explanatory article selection algorithms, namely, FCBF, ReliefF, CFS, Contain, and FOCUS-SF, with respect to four types of well-known classifiers, namely, the viewpoint based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER before and after feature selection. The efficiency and effectiveness of the FAST algorithm are evaluated through an realistic study. FSS algorithm works under two different steps that are graph hypothetical clustering methods and representative feature cluster selection. Subsets of features but also improves the presentations of the four types of classifiers. FAST Algorithm can be used for identifying and removing the irrelevant data. The results, on 35 publicly obtainable real-world high-dimensional image, microarray, and text data, demonstrate that the FAST not only produces*

Key Words: Fast algorithm, Image Searching, Minimum Spanning Tree, High Dimensional Image, Graph Theoretic Cluster.

1. INTRODUCTION

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. 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. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. 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. The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce the 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. The wrapper methods are computationally expensive and tend to over fit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, we will focus on the filter method in this paper. 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. Pereira et al. Baker et al. and Dhillon etal. employed the distributional clustering of words to reduce the dimensionality of text data. In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Their results have, sometimes, the best agreement with human performance. The general graph-theoretic clustering is simple: Compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter than its neighbors. The result is a forest and each tree in the forest represents a cluster. In our study, we apply graph theoretic clustering methods to features. In particular, we adopt the minimum spanning tree (MST) based clustering algorithms, because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used.

1.1 Purpose

- 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.
- The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce the search space that will be considered by the subsequent wrapper.
- 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.
- Adopting the minimum spanning tree (MST) based clustering algorithms, because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice.

2. LITERATURE SURVEY

2.1 Feature Selection System Architecture

A simple three step feature selection approach is explained below. The goal of this architecture is to reduce a large set of features (on the order of thousands) to a small subset of features (on the order of tens), without significantly reducing the system's ability. The basic three steps of this system are:

- In first step the irrelevant features are removed.
- After that the redundant features are removed.
- And finally a feature selection algorithm is applied to the remaining features.

Which assigns relevance values to features by treating training samples as points in feature space. For each sample, it finds the nearest "hit" (another sample of the same class) and "miss" (a sample of a different class), and adjusts the relevance value of each feature according to the square of the feature difference between the sample and the hit and miss. There are several modifications to Relief to generalize it for continuous features and to make it more robust in the presence of noise. This system adopts Kononenko‟s modifications, and modify Relief again to remove a bias against non- monotonic features, as described in [13]. Within this feature selection system, Relief is used as a relevance filter. Therefore it threshold the relevance values, to divide the feature set into relevant and irrelevant features. This can be done either by thresholding the relevance value directly, or by selecting the highest n values and discarding the remaining features. In either case, relief does not detect redundancy, so the remaining feature set still contains redundant features. The second step is a redundancy filter that uses the K-means algorithm [14] to cluster features according to how well they correlate to each other. When feature clusters are not sufficiently similar, in which case the cluster is split to make sure that potentially useful features are not removed from the feature set. The third and final filter is a combinatorial feature selection algorithm.

2.2 Characteristics of Feature Selection Algorithms

Feature selection algorithms (with a few notable exceptions) perform a search through the space of feature subsets, and, as a consequence, must address four basic issues affecting the nature of the search [21]:

1. Starting point. Selecting a point in the feature subset space from which to begin the search can affect the direction of the search. One option is to begin with no features and successively add attributes. In this case, the search is said to proceed forward through the search space. Conversely, the search can begin with all features and successively remove them. In this case, the search proceeds backward through the search space. Another alternative is to begin somewhere in the middle and move outwards from this point.
2. Search organization. An exhaustive search of the feature subspace is prohibitive for all but a small initial number of features. With N initial features there exist 2N possible subsets. Heuristic search strategies are more feasible than exhaustive ones and can give good results, although they do not guarantee finding the optimal subset.
3. Evaluation strategy. How feature subsets are evaluated is the single biggest differentiating factor among feature selection algorithms for machine learning. One paradigm, dubbed the filter [23, 24] operates independent of any learning algorithm, undesirable features are filtered out of the data before learning begins. These algorithms use heuristics based on general characteristics of the data to evaluate the merit of feature subsets. Another school of thought argues that the bios of a particular induction algorithm should be taken into account when selecting features. This method, called the wrapper, uses an induction algorithm along with a statistical re-sampling technique such as cross-validation to estimate the final accuracy of feature subsets. Figure No. 4.2 and 4.3 illustrates the filter and wrapper approaches to feature selection.
4. Stopping criterion. A feature selector must decide when to stop searching through the space of feature subsets. Depending on the evaluation strategy, a feature selector might stop adding or removing features when none of the alternatives improves upon the merit of a current feature subset. Alternatively, the algorithm might continue to revise the feature subset as long as the merit does not degrade. A further option could be to continue generating feature subsets until reaching the opposite end of the search space and then select the best.

2.3 Feature Selection Benefits

Feature selection method consist of potential benefits are
a. A reduction in the amount of training data needed to achieve learning.
b. The generation of learning models with improved predictive accuracy.
c. Learned knowledge more compact, simpler and easier to understand.
d. The Reduced execution time required for learning.

2.4 Irrelevant Features

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. We achieve this through a new feature selection framework which composed of the two connected components of irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset. The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves 1) the construction of the minimum spanning tree of a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features of the clusters.

2.5 Minimum Spanning Tree

Minimum spanning tree MST is a graph based model in producing the clusters from high computational complexity, it selects or rejects the edges in MST. Spanning tree with their weight less than or equal to the weight of every other spanning tree. Clustering by Minimal Spanning Tree can be view as a hierarchical clustering algorithm which track the divisive clustering approach. Clustering algorithm based on minimum and maximum spanning tree were generally studied to construct MST of point set and delete conflicting edges. Whose weights are expansively larger than the standard weight of the close proximity edges in the tree. The goal to maximize the minimum inters cluster distance. MST based image segmentation is based on select the edges from the graph, where each pixel correspond to a node in the graph. Weights on every edge calculate the dissimilarity between pixels. The segmentation algorithm define the restrictions between regions by comparing two quantities Intensity difference across the boundary and Intensity difference between neighboring pixels with all region. This is useful knowing that the intensity differences across the boundary are important if they are huge comparative to the concentration distinction inside the at least on of the regions.

3. SYSTEM ANALYSIS

3.1 Existing System

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 resourceful than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the extrapolative 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 difficulty is low, but the accuracy of the learning algorithms is not guaranteed. 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. 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.

Disadvantages

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.

3.2 Proposed System

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. 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 as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

4. SYSTEM DESIGN

4.1 Input Design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- ➢ What data should be given as input?
- ➢ How the data should be arranged or coded?
- ➢ The dialog to guide the operating personnel in providing input.
- ➢ Methods for preparing input validations and steps to follow when error occur.

4.1.1 Objectives

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

4.2 OUTPUT DESIGN

Confirm an action. A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- ❖ Convey information about past activities, current status or projections of the
- ❖ Future.
- ❖ Signal important events, opportunities, problems, or warnings.
- ❖ Trigger an action.

4.3 Dataflow Diagram

4.3.1 User Data Flow Diagram

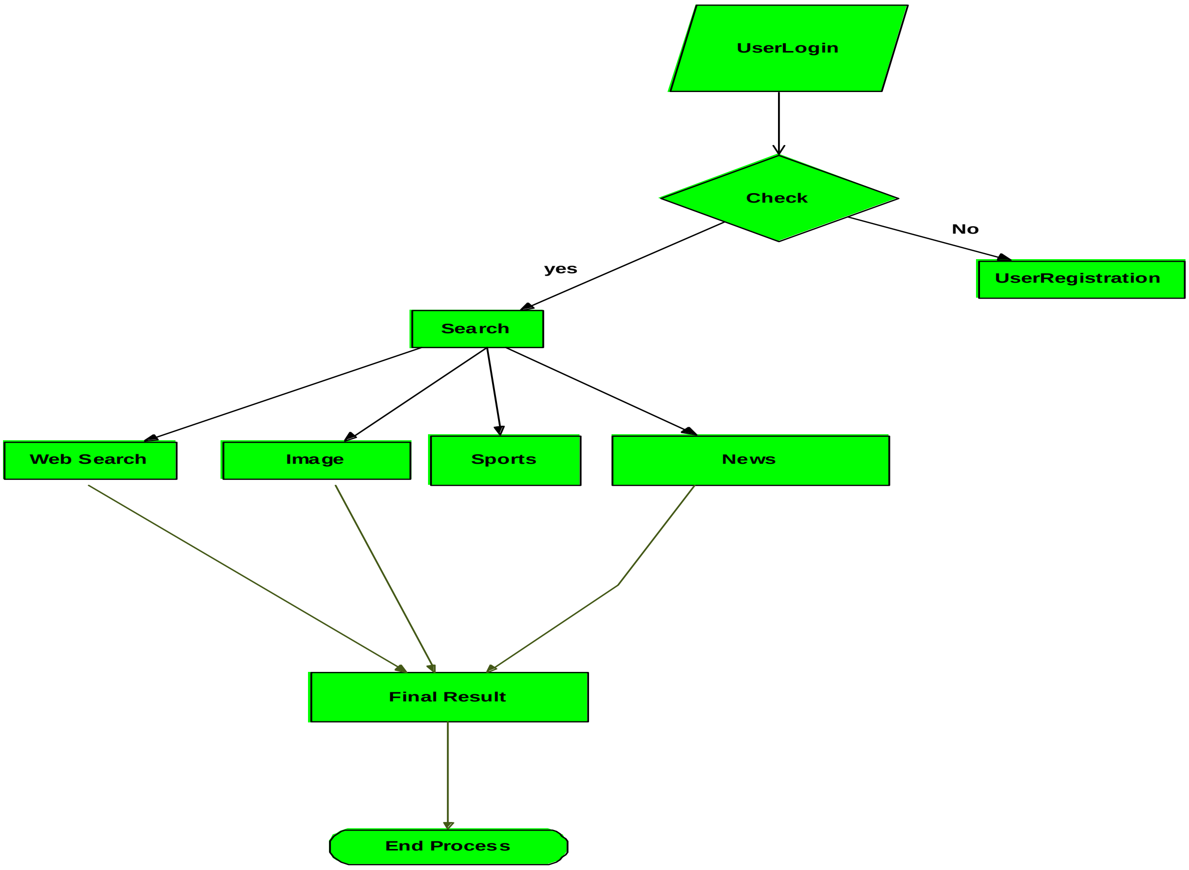


Figure 4.4.1 : User Data Flow Diagram

4.3.2 Admin Dataflow Diagram:

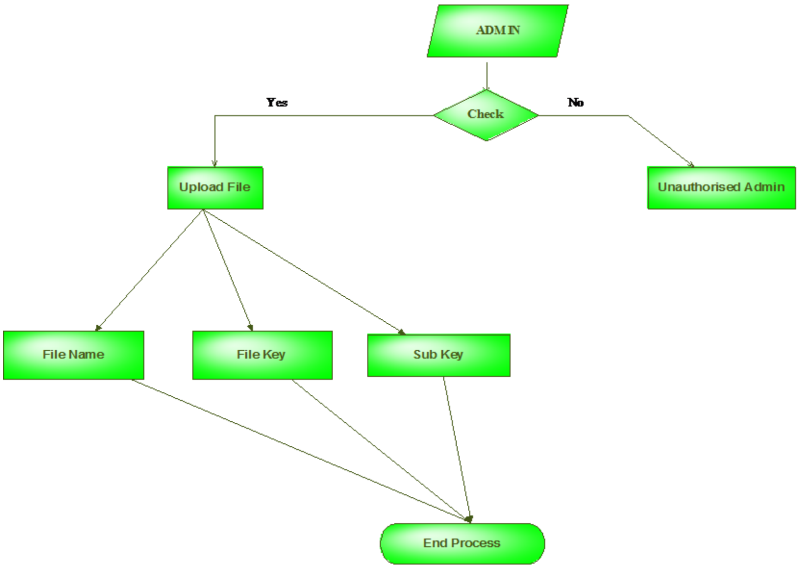


Figure 4.4.2: Admin Dataflow Diagram

4.4 UML DIAGRAMS

Use Case Diagram: A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

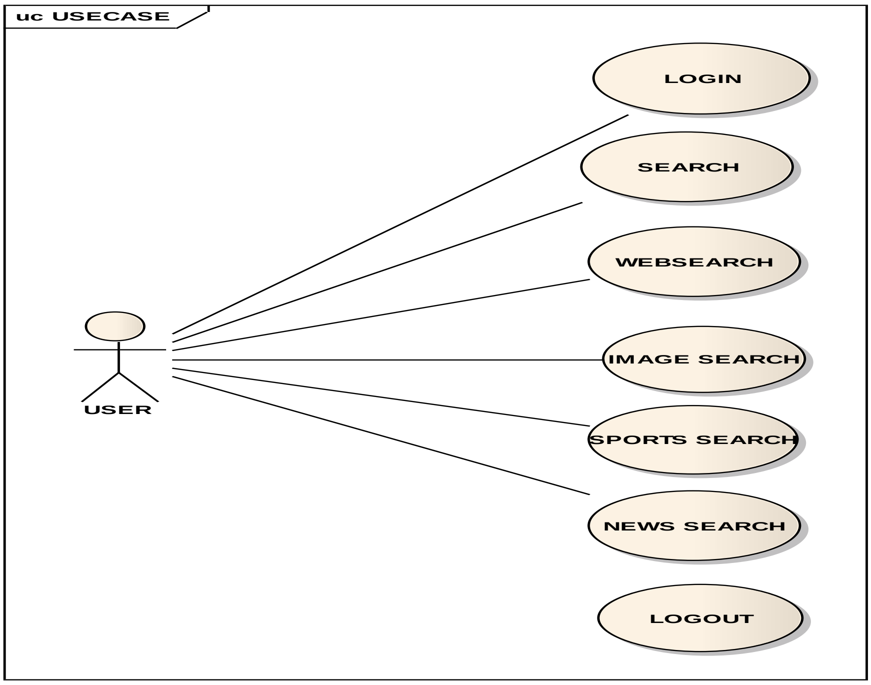


Figure 4.5.1.: Use case Diagram

Admin User case Diagram

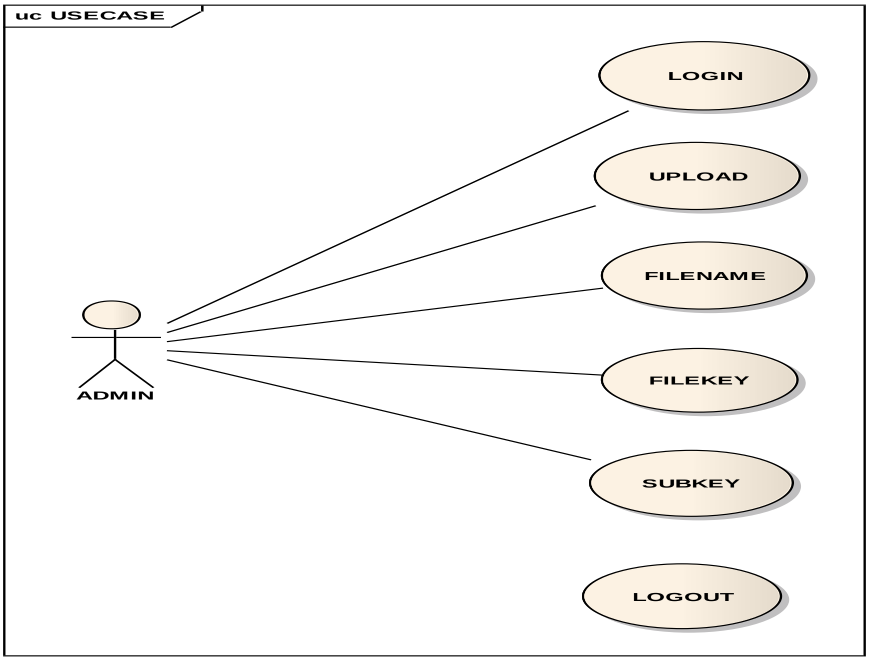


Figure 4.5.2: Admin Use case Diagram

5. TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

5.1 Types of Testing

* Unit Testing
* Integration Testing
* System Testing
* White Box Testing
* Black Box Testing
* Acceptance Testing

5.2 Testing Strategies

Field testing will be performed manually and functional tests will be written in detail.

5.2.1 Test Objectives

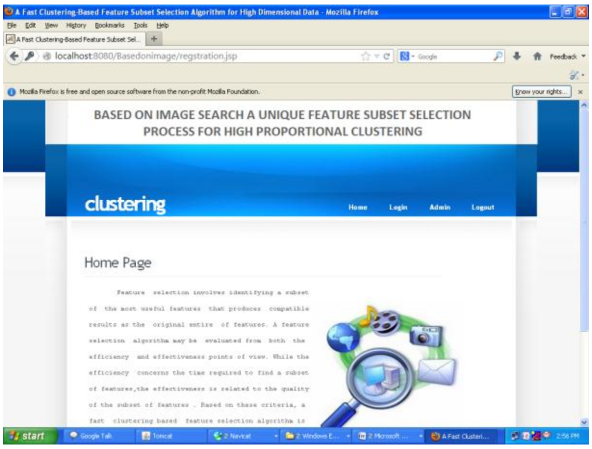
- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

5.2.2 Features To Be Tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

6. SCREENSHOTS

HOME PAGE


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

localhost:8080/Basedonimage/regstration.jsp

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

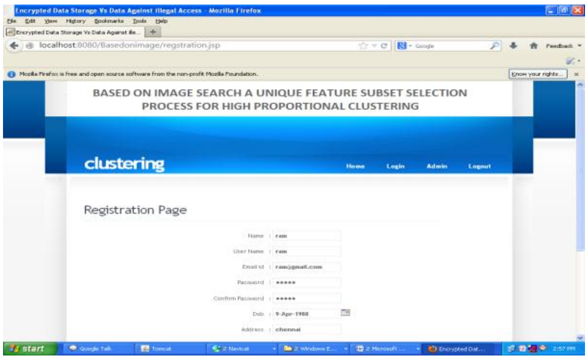
BASED ON IMAGE SEARCH A UNIQUE FEATURE SUBSET SELECTION PROCESS FOR HIGH PROPORTIONAL CLUSTERING

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

Screen 6.1: Home Page

REGISTRATION PAGE


Encrypted Data Storage Vs Data Against Illegal Access - Mozilla Firefox

localhost:8080/Basedonimage/regstration.jsp

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

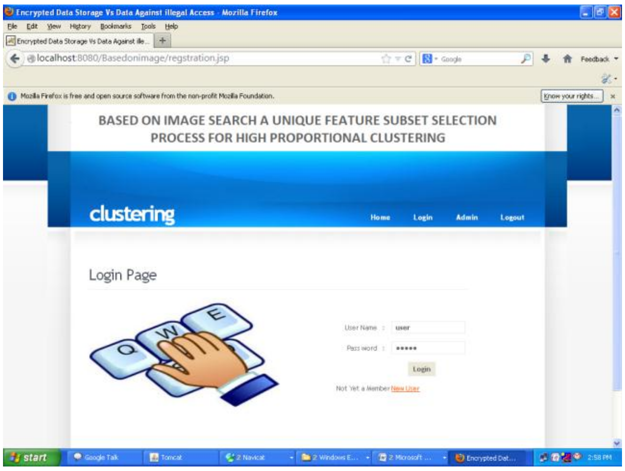
BASED ON IMAGE SEARCH A UNIQUE FEATURE SUBSET SELECTION PROCESS FOR HIGH PROPORTIONAL CLUSTERING

Registration Page

Name : ram
User Name : ram
Email id : ram@gmail.com
Password : *****
Confirm Password : *****
Dob : 9-Apr-1988
Address : chennai

Screen6.2: Screen for Register Page

LOGIN PAGE:


Encrypted Data Storage Vs Data Against Illegal Access - Mozilla Firefox

localhost:8080/Basedonimage/regstration.jsp

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

BASED ON IMAGE SEARCH A UNIQUE FEATURE SUBSET SELECTION PROCESS FOR HIGH PROPORTIONAL CLUSTERING

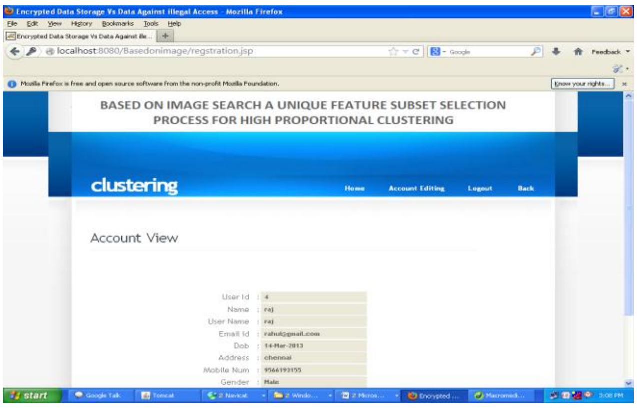
Login Page

User Name : user
Pass word : *****

Not Yet a Member New User

Screen6.3: Screen for Login Page

ACCOUNT VIEW:


Encrypted Data Storage Vs Data Against Illegal Access - Mozilla Firefox

localhost:8080/Basedonimage/regstration.jsp

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

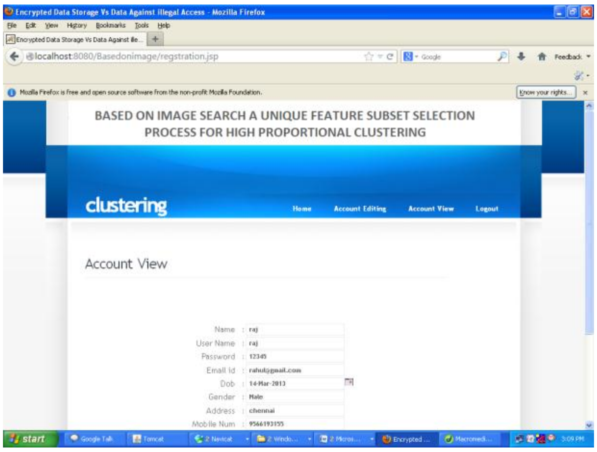
BASED ON IMAGE SEARCH A UNIQUE FEATURE SUBSET SELECTION PROCESS FOR HIGH PROPORTIONAL CLUSTERING

Account View

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

Screen 6.4: Screen for Account View

ACCOUNTING EDITING:

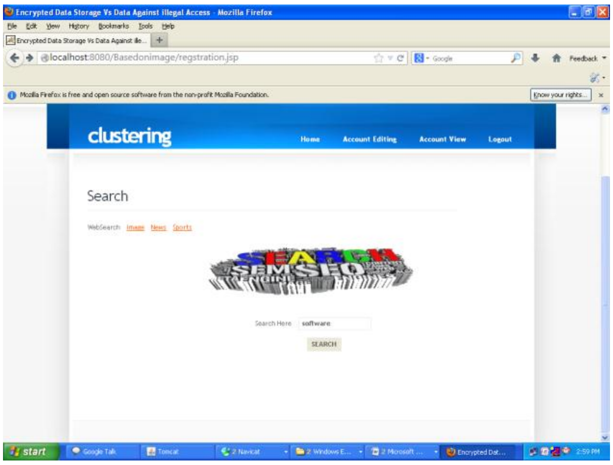

Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
localhost:8080/Basedonimage/regstration.jsp
BASED ON IMAGE SEARCH A UNIQUE FEATURE SUBSET SELECTION PROCESS FOR HIGH PROPORTIONAL CLUSTERING
clustering
Home
Account Editing
Account View
Logout

Account View

Name : raj
User Name : raj
Password : 12345
Email id : rahul@gmail.com
Dob : 14-Mar-2013
Gender : Male
Address : chennai
Mobile Num : 9566193155

Screen 6.5: Screen for Account Editing

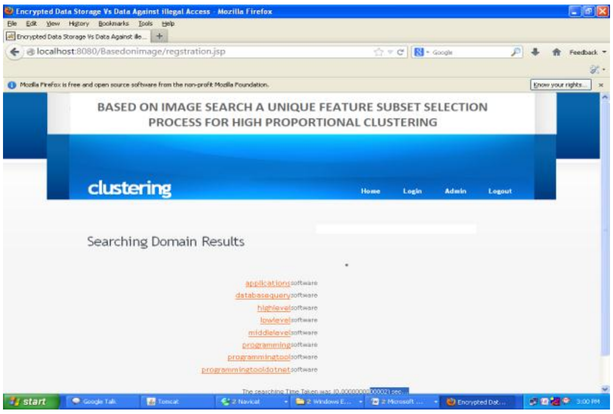
ACCOUNT SEARCHING:


Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
localhost:8080/Basedonimage/regstration.jsp
clustering
Home
Account Editing
Account View
Logout
WebSearch
Image
News
Sports
SEARCH

Search

Search Here : software

Screen 6.6: Screen for Account Searching 1


Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
localhost:8080/Basedonimage/regstration.jsp
BASED ON IMAGE SEARCH A UNIQUE FEATURE SUBSET SELECTION PROCESS FOR HIGH PROPORTIONAL CLUSTERING
clustering
Home
Login
Admin
Logout

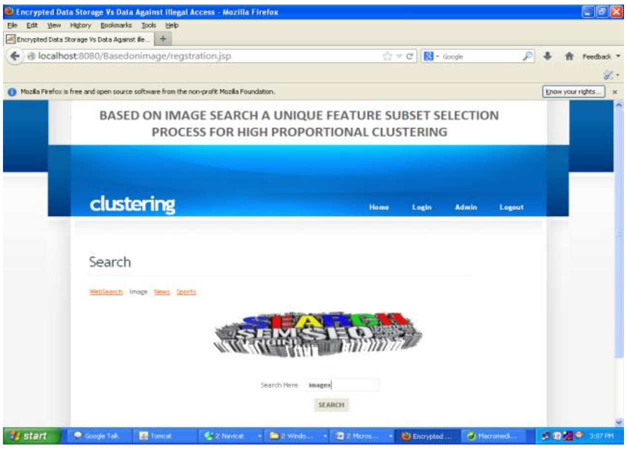
Searching Domain Results

applicationsoftware
databasequerysoftware
highlevelsoftware
lowlevelsoftware
middlelevelsoftware
programmingsoftware
programmingtoolsoftware
programmingtooldotnetsoftware

The searching Time Taken was :0.000000000000021 sec

Screen 6.7: Screen for Account Searching 2

SEARCHING IMAGE:

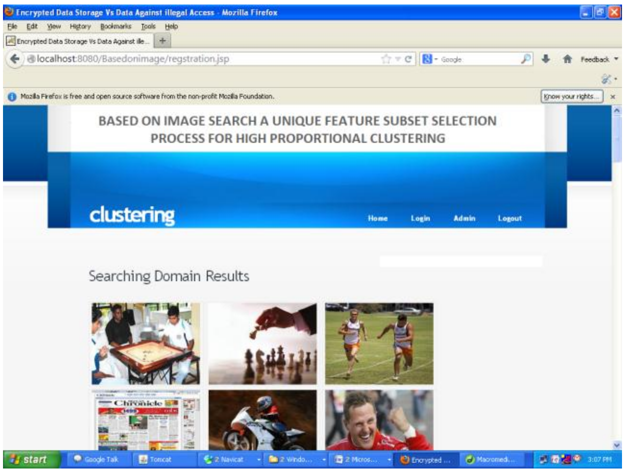

Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
localhost:8080/Basedonimage/regstration.jsp
BASED ON IMAGE SEARCH A UNIQUE FEATURE SUBSET SELECTION PROCESS FOR HIGH PROPORTIONAL CLUSTERING
clustering
Home
Login
Admin
Logout
WebSearch
Image
News
Sports
SEARCH

Search

Search Here : images

Screen 6.8: Screen for Searching Image

IMAGE RESULT:


Encrypted Data Storage Vs Data Against illegal Access - Mozilla Firefox
localhost:8080/Basedonimage/regstration.jsp
BASED ON IMAGE SEARCH A UNIQUE FEATURE SUBSET SELECTION PROCESS FOR HIGH PROPORTIONAL CLUSTERING
clustering
Home
Login
Admin
Logout

Searching Domain Results

Screen 6.9 : Screen for Image Result

7. CONCLUSIONS

In this project, 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 have compared the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, Relief F, CFS, Consist, and FOCUS-SF on the 35 publicly available image, micro array, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record. Generally, the proposed algorithm obtained the best proportion of selected features, the best run time, and the best classification accuracy for Naive Bayes,C4.5, and RIPPER, and the second best classification accuracy for IB1. The Win/Draw/Loss records confirmed the conclusions. We also found that FAST obtains the rank of 1 form micro array 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, and CFS is a good alternative. At the same time, FCBF is a good alternative for image and text data. Moreover, Consist and FOCUSS Fare alternatives for text data .For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space.

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] John G.H., Kohavi R. and Pfleger K., Irrelevant Features and the Subset Selection Problem, In the Proceedings of the Eleventh International Conference on Machine Learning, pp 121-129, 1994.

[3] Hall M.A. and Smith L.A., Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper, In Proceedings of the Twelfth international Florida Artificial intelligence Research Society Conference, pp 235-239, 1999.

[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] Yu L. and Liu H., Efficiently handling feature redundancy in highdimensional data, in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03). ACM, New York, NY, USA, pp 685-690, 2003.

[6] Yu L. and Liu H., Redundancy based feature selection for microarray data,In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 737-742, 2004.

[7] 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.

[8] 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.

[9] Biesiada J. and Duch W., Features election for high-dimensionaldatała Pearson redundancy based filter, AdvancesinSoftComputing, 45, pp 242C249, 2008.