DESIGN AND ANALYSIS OF IMAGE SEARCH BASED ON UNIQUE FEATURE SUBSET SELECTION PROCESS

ALEKHYA JUTTIGA [#1], B. NANDAN KUMAR[#2]

[#1] M.Tech Scholar, Department of Computer Science and Engineering,
DNR College of Engineering and Technology, Sri RamaPuram, Balusumudi,
Bhimavaram - 534202.

[#2] Assistant Professor, Department of Computer Science and Engineering,
DNR College of Engineering and Technology, Sri RamaPuram, Balusumudi,
Bhimavaram - 534202.

ABSTRACT

In current days feature selection occupied a lot of users attention towards it in identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature/component determination calculation might be assessed from both the productivity and adequacy perspectives. While the productivity concerns the time required to discover a subset of highlights, the viability is identified with the nature of the subset of highlights. Based on these concepts and advantages, a fast clustering-based feature selection algorithm (FAST) is proposed and tentatively assessed in this paper. The FAST calculation method will mainly work in two forms or levels. In the starting level this FAST algorithm try to highlight the groups which are almost nearly same and then are formed as one cluster and remaining are formed as different clusters by utilizing the most popular graph-theoretic clustering methods. In the next stage we try to highlight the components which are very nearer to the target classes and try to form them as a subset of main cluster. To guarantee the efficiency and performance of FAST, we embrace the we adopt the efficient minimum-spanning tree (MST) clustering method in order to increase the efficiency of FAST algorithm. Our experimental results clearly state that proposed FAST is very efficient in image search based on feature and sub features extraction.

Key Words: Minimum-Spanning Tree, Feature Selection Algorithm, Cluster, Graph-Theoretic, Component Determination.

1. Introduction

From the point of picking a subset of good highlights regarding the objective ideas, include subset determination is a viable path for lessening dimensionality, evacuating immaterial information, expanding learning exactness, and improving outcome conceivability. Many feature subset choice techniques have been proposed and read for AI applications. They can be separated into four general classes: the Embedded, Wrapper, Filter, and Hybrid methodologies. The implanted strategies fuse includes determination as a piece of the preparation procedure and are typically explicit to given learning calculations, and subsequently might be more effective than the other three classifications. Customary AI calculations like choice trees or counterfeit neural systems are instances of implanted methodologies. In any case, the all-inclusive statement of the chose highlights is constrained and the computational unpredictability is enormous.

The channel techniques are free of learning calculations, with great consensus. Their computational unpredictability is low, yet the exactness of the learning calculations isn't ensured. The half breed techniques region mix of channel and covering strategies by utilizing a channel strategy to diminish search space that will be considered by the resulting wrapper. They mostly center on joining channel and covering strategies to accomplish the most ideal presentation with a specific learning calculation with comparative time complexity of the channel techniques. The covering techniques are computationally costly and tend to over fit on small training sets. The channel techniques, in addition to their all-inclusive statement, are normally a decent decision when the number of highlights is enormous. Subsequently, we will center on the channel strategy in this paper. With regard to the channel include determination techniques, the application of bunch examination has been shown to be more viable than customary element choice algorithms. Pereira et al. Dough puncher et al. also, Dhillonetal. Utilized the distributional bunching of words to decrease the dimensionality of text data.

In cluster investigation, diagram hypothetical strategies have been well examined and utilized in numerous applications. Their results have, in some cases, the best concurrence with human performance. The overall diagram hypothetical clustering is basic: Compute a local chart ofinstances,then erase any edge in the diagram that is much longer/shorter than its neighbors. In

our investigation, we apply graph theoretic clustering strategies to highlights. In particular, we embrace the base spreading over tree (MST) based clustering calculations.

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 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

Literature survey is that the most vital step in software development process. Before developing the tool, it's necessary to work out the time factor, economy and company strength. Once this stuff is satisfied, ten next steps are to work out which OS and language used for developing the tool. This literature survey is mainly used for identifying the list of resources to construct this proposed application.

MOTIVATION

An "element" or "trait" or "variable" alludes to a part of the information. As a rule before gathering information, highlights are indicated or picked. Highlights can be discrete, constant, or ostensible.

By and large, highlights are portrayed as:

1. **Relevant:** These are highlights which have an impact on the yield and their job can not be accepted by the rest
2. **Irrelevant:** Irrelevant highlights are characterized as those highlights not having any impact on the yield, and whose qualities are created indiscriminately for every model.
3. **Redundant:** A repetition exists at whatever point an element can play the job of another (maybe the least complex approach to display excess).

FEATURE DETERMINATION

There are several issues while choosing the best feature selection algorithm for identifying the factors which are mapped from an image. So mostly we try to check the highlight determination component as the best component among all the highlights since all the highlights are not helpful in building the bunches: a few highlights might be excess or unimportant in this way not adding to the learning procedure.

The primary point of highlight determination is to decide an insignificant component subset from a difficult space while holding an appropriately high exactness in speaking to the first highlights. In numerous genuine issues Feature choice is an absolute necessity because of the wealth of boisterous, unessential or deluding highlights. For example, by evacuating these variables, gaining from information strategies can profit.

To be totally certain about the trait determination, we would in a perfect world need to test all the lists of characteristic subsets, which is infeasible much of the time as it will bring about 2n subsets of n properties.

3. EXISTING SYSTEM AND ITS LIMITATIONS

In this section we will mainly discuss about existing components and the limitations that are present in the image based search. Now let us discuss about this proposed model in detail as follows:

The existing methods consolidate highlight choice as a piece of the preparation procedure and are generally explicit to given learning calculations, and in this way might be more proficient than the other three classes. All the primitive ML based algorithms are used as embedded approaches for identifying the features and some of the best algorithms which come under that category are decision trees or artificial neural networks and a lot more. In general the primitive methods have the feature identification facility with very less accuracy and there is a lot of time complexity for those methods to extract the features from that input data. All the primitive methods has the computational unpredictability is low, yet the exactness of the learning

calculations isn't ensured. The crossover techniques are a mix of channel and covering strategies by utilizing a channel strategy to decrease search space that will be considered by the resulting covering. They for the most part center around joining channel and covering strategies to accomplish the most ideal exhibition with a specific learning calculation with comparative time multifaceted nature of the channel techniques. The following are the limitations of the existing methods:

1. The generality of the primitive algorithm to identify the features is limited and the computational complexity is very more.
2. If we assume the computational complexity as low in other case, parallel the accuracy of the algorithm is not up to the mark or we can’t assume the guarantee results.

4. PROPOSED SYSTEM AND ITS ADVANTAGES

In this section we will mainly discuss about proposed components and the advantages that are present in the image based search. Now let us discuss about this proposed model in detail as follows:

MOTIVATION

Highlight subset determination can be seen as the way toward distinguishing and expelling whatever number superfluous and excess highlights as could be expected under the circumstances. This is on the grounds that unimportant highlights don't add to the prescient precision and repetitive highlights don't redound to improving indicator for that they give for the most part data which is already present in different feature(s). Of the many component subset determination calculations, some can adequately take out unimportant highlights yet neglect to deal with excess highlights yet some of others can take out the superfluous while dealing with the repetitive highlights. Our proposed FAST calculation falls into the subsequent gathering. Customarily, include subset choice examination has concentrated on scanning for pertinent highlights. A notable model is Relief which gauges each element as per its capacity to separate occasions under various targets dependent on separation based standards work. Be that as it may, Relief is insufficient at evacuating repetitive highlights as two prescient however profoundly

associated highlights are likely both to be exceptionally weighted. Alleviation F expands Relief, empowering this technique to work with loud and deficient informational indexes and to manage multiclass issues, yet at the same time can't distinguish excess highlights.

ADVANTAGES OF THE PROPOSED SYSTEM

1. Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with each other.
2. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

5. IMPLEMENTATION PHASE

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. The front end of the application takes JSP,HTML & CSS programming and as a Back-End Data base we took My SQL data base. The application is divided mainly into following 2 modules. They are as follows:

1. USER MODULE

In this module, initially the user need to register into the application with all his personal details and then he try to get authorized by the admin for getting login into the system. Once the user is authorized he can able to login into his account by substituting his login credentails.After login the user can perform following operations like:

1. User can able to edit his profile
2. He can able to see his Profile or Account Details
3. He can able to search the data like Image ,News, Sports
4. He can able to view the list of images which are matched with the search keyword

2. ADMIN MODULE

In this module, admin can able to login into his account with his valid id and password. Once the admin gets login the admin can able to add images into the database, this can also add some useful information related to sports, news and other job related useful information into the database. While adding the data into the database the admin try to select some search keywords

and add those keywords along with the information. These search keywords act as features while extracting the data from the user module.

SCREEN FOR ACCOUNT SEARCHING

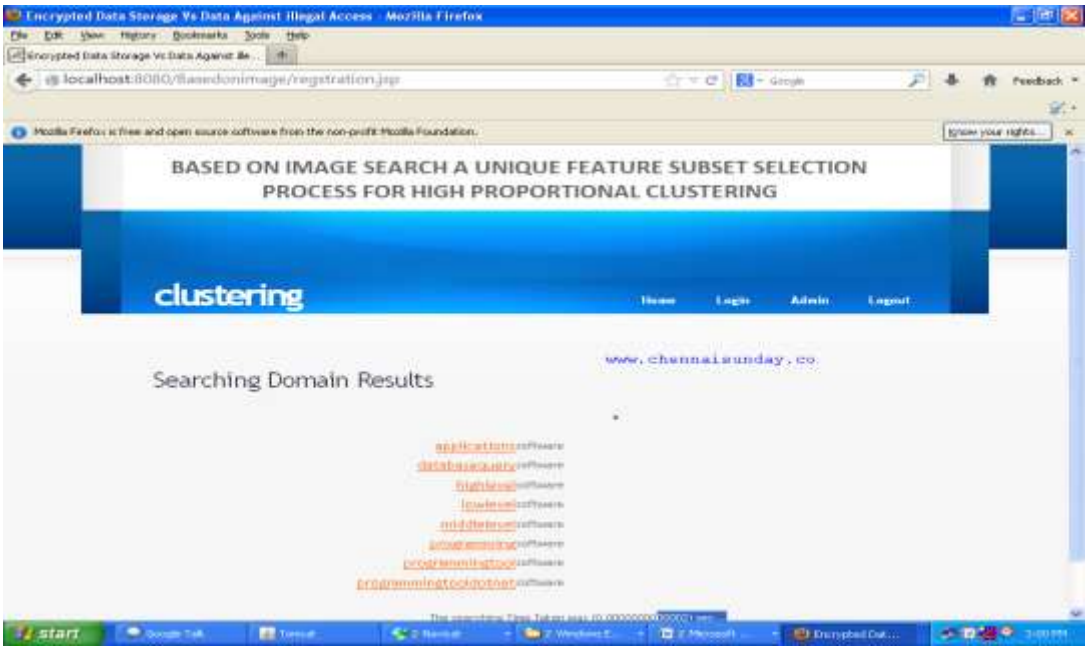

Figure. Represent the User can able to Search his Account

SCREEN FOR USER SEARCH FOR IMAGE

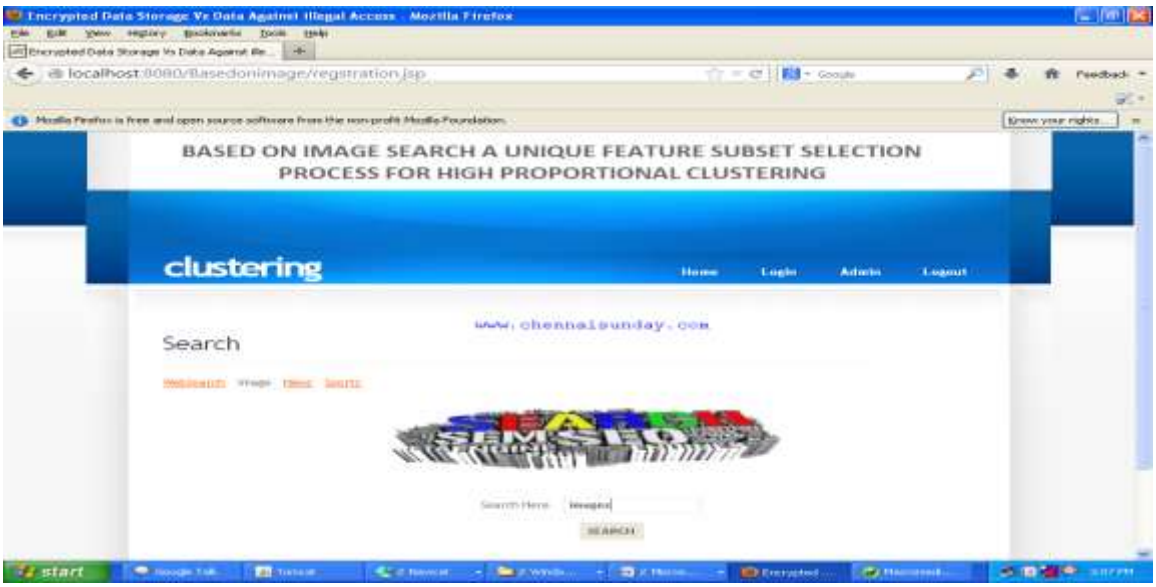

Figure. Represent the User can able to Search Images

SCREEN FOR USER SEARCHED IMAGES

Figure. Represent the User can able to retrieve the search images for his input keyword

6. CONCLUSION

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, 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, CFS, Consist, and FOCUS-SF on the 35 publicly available image, microarray, 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.

7. 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, AdvancesingSoftComputing, 45, pp 242C249, 2008.

[10] Dash M., Liu H. and Motoda H., Consistency based feature Selection, In Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109, 2000.

8. ABOUT THE AUTHORS

ALEKHYA JUTTIGA is currently pursuing her M.Tech in the Computer Science and Engineering at DNR College of Engineering and Technology, Sri RamaPuram, Balusumudi, Bhimavaram - 534202.Her area of interest includes Data Mining and Image Processing.

B.NANDANA KUMAR is currently working as an Assistant Professor in the Department of Computer Science and Engineering at DNR College of Engineering and Technology, Sri RamaPuram, Balusumudi, Bhimavaram - 534202.He has more than 10 years of teaching experience in various engineering colleges. His research interest includes Cloud Computing.