Journal of Cardiovascular Disease Research

ISSN: 0975-3583,0976-2833 VOL08, ISSUE 03, 2017

# ENHANCING MODEL ACCURACY USING FEATURE SELECTION TECHNIQUE

**M. Umamaheswari,**

Assistant Professor in CSE,

KSR College of Engineering,Tiruchengode-637 215.

umadeena@gmail.com

**A. Viswanathan,**

Associate Professor in CSE,

KSR College of Engineering,Tiruchengode-637 215.

professorvichu@gmail.com

**R. Mohanasundaram**

Assistant Professor (Senior)

School of Computer Science and Engineering,

VIT University, Vellore – 632014,

mohanasundaramr@vit.ac.in

**M. Sathya,**

Assistant Professor in CSE

K.S.R College of Engineering, Tiruchengode 637215

sathimanogaran@gmail.com

## ABSTRACT

For feature selection, it is necessary to choose a subset of the most important characteristics that produces the same outcomes as the entire set of data. It is possible to evaluate the effectiveness and efficiency of a feature selection approach. Efficiency is the amount of time it takes to find a subset of characteristics, whereas effectiveness is the caliber of the subset. Based on these variables, this study suggests and evaluates a feature selection method called clustering. There are two components to the Clustering Based Feature Selection technique.

ISSN: 0975-3583,0976-2833 VOL08, ISSUE 03, 2017

Initially, graph-theoretic clustering methods are employed to divide features into groups. The Clustering Based Feature Selection clustering-based approach is anticipated to produce a subset of useful and independent features since the attributes in separate clusters are often independent.

## 1. INTRODUCTION

The act of analyzing data from many angles and synthesizing it into useful information that may be used to increase income, decrease costs, or do both is known as data mining (also known as data or knowledge discovery). One of the various analytical methods accessible for data analysis is data mining software. Users may look at the data from a number of angles, classify it, and explain the correlations they find. In large relational databases, data mining is a method for identifying patterns or connections between a variety of variables. The gap left by the introduction of various transaction and analytical systems by large-scale information technology is filled by data mining. Data mining software investigates connections and patterns in transaction data that has been saved based on broad user searches.

Analytical software includes statistical, machine learning, and neural networks, for instance. Most people are seeking one of the following four types of relationships: Using stored data, data is organized into designated groupings. For instance, a restaurant chain may look at customer purchase information to understand when people arrive and what they frequently order. By marketing daily offers, this information might be leveraged to increase visitor numbers. Clusters: Based on logical correlations or consumer preferences, data is sorted into clusters. For instance, market classifications or customer affinities can be discovered using data mining. Data mining may be used to find relationships. The beer-diaper scenario serves as an illustration of associative mining. Data is mined to predict sequential patterns of behavior and trends. An outdoor equipment retailer, for instance, may predict that a consumer would purchase based on prior sales of sleeping bags and hiking boots.

**Classes:** Using stored information, data is organized into designated groups. For instance, a restaurant chain may utilize data on consumer purchases to determine when customers arrive and what they often buy. This information may be utilized to increase visitors by presenting daily deals.

**Logic correlations or customer preferences** are used to group information into groups. Data may be mined, for instance, to identify market niches or consumer preferences.

**Associations:** Relationships can be found via mining data. The beer-diaper example demonstrates associative mining.

**Sequential patterns:** Behaviour patterns and trends are predicted via data mining. An outdoor equipment retailer may foresee the potential purchase of a backpack, for instance, based on a customer's purchases of sleeping bags and hiking boots.

ISSN: 0975-3583,0976-2833 VOL08, ISSUE 03, 2017

For data scientists, feature selection is a priceless skill. The effectiveness of the machine learning algorithm depends on being able to choose key characteristics effectively. An algorithm can get contaminated by irrelevant, redundant, and noisy features, which will have a poor effect on learning efficiency, accuracy, and computing cost. As the size and complexity of the typical dataset continue to increase rapidly, feature selection becomes more and more crucial.

To put it another way, feature selection enables developers to use just the most pertinent and helpful data in machine learning training sets, thereby lowering expenses and data volume.

One illustration is the notion of sizing up a complicated shape. As the software scales, more data points are identified, and the system becomes considerably more complicated. A machine learning system does not typically use data sets with complicated shapes. These systems could make use of data sets with vastly differing amounts of variation between certain variables. Engineers can use feature selection, for instance, to only investigate the factors that would produce the most focused findings when categorizing species.

Machine learning systems are guided toward a target by engineers through feature selection, a discriminating process. Feature selection can be helpful in optimizing components of what experts refer to as the "bias variance trade-off" in machine learning, in addition to the notion of eliminating complexity from systems at scale.

The benefits of feature selection for bias and variance analysis are more intricately explained. An example of how feature selection helps projects is provided by a Cornell University study on feature selection, bias variance, and bagging.

## 2. RELATED WORKS

This paper demonstrates how to apply the MINFEATURES bias, which favors consistent hypotheses that can be specified with the fewest number of features possible. This bias is advantageous for learning domains when the training data has a high concentration of irrelevant features. First, FOCUS-2, a cutting-edge technique for precisely applying the MINFEATURES bias, is demonstrated. It has been shown that this method is substantially faster than the FOCUS algorithm. The Mutual-Information-Greedy, Simple-Greedy, and Weighted-Greedy algorithms are next shown. These algorithms employ effective heuristics to approximate the MIN-FEATURES bias. It has been shown that this method is substantially faster than the FOCUS algorithm. The Mutual-Information-Greedy, Simple-Greedy, and Weighted-Greedy algorithms are next shown. These algorithms employ effective heuristics to approximate the MIN-FEATURES bias.These algorithms use heuristics that trade off optimality for computer effectiveness, or greedy heuristics. Experiments demonstrate that using these techniques to preprocess the training data significantly improves ID3's learning performance by excluding unnecessary characteristics from its consideration.The article

ISSN: 0975-3583,0976-2833 VOL08, ISSUE 03, 2017

"examines the process by which feature selection increases the accuracy of supervised learning," according to the authors.

The report also claims:

The most accurate feature set corresponds to the optimal bias-variance tradeoff point for the learning algorithm, according to an empirical bias/variance analysis as feature selection advances.

The authors refer to feature selection as "a variance reduction strategy" when addressing the usage of strong or weak relevance; this makes sense when you consider that variance is simply the degree of variation in a particular variable. The data point or array may be practically meaningless if there is no variation. As a result, feature selection is an essential component of machine learning design.

If the variation is really high, it can turn into what engineers might refer to as "noise" or random, irrelevant outputs that are challenging for the machine learning system to handle.

## 3. METHODOLOGY

Feature subset selection is the process of finding and eliminating as many undesirable and redundant traits as is practical. This is because redundant characteristics do not provide a stronger predictor since they provide information that is already present in other features, and irrelevant qualities do not increase predictive accuracy (s). While disregarding superfluous features, certain feature subset selection methods may efficiently reduce redundant features, whilst others can efficiently erase unnecessary features while ignoring duplicates. Our suggested Clustering Based Feature Selection method is in the second set. Finding pertinent traits has historically been the aim of feature subset selection research.

Using a distance-based criterion function, a property like relief is weighted based on how well it can identify instances under various targets. Relief fails to eliminate redundant qualities because it is possible that two predictive but closely related traits will be equally weighted. Relief-F enhances Relief by giving it the capacity to handle noisy and imperfect data sets, as well as multiclass issues, but it is unable to identify duplicate features. The architecture of the proposed work is shown in Figure 1.

ISSN: 0975-3583,0976-2833 VOL08, ISSUE 03, 2017

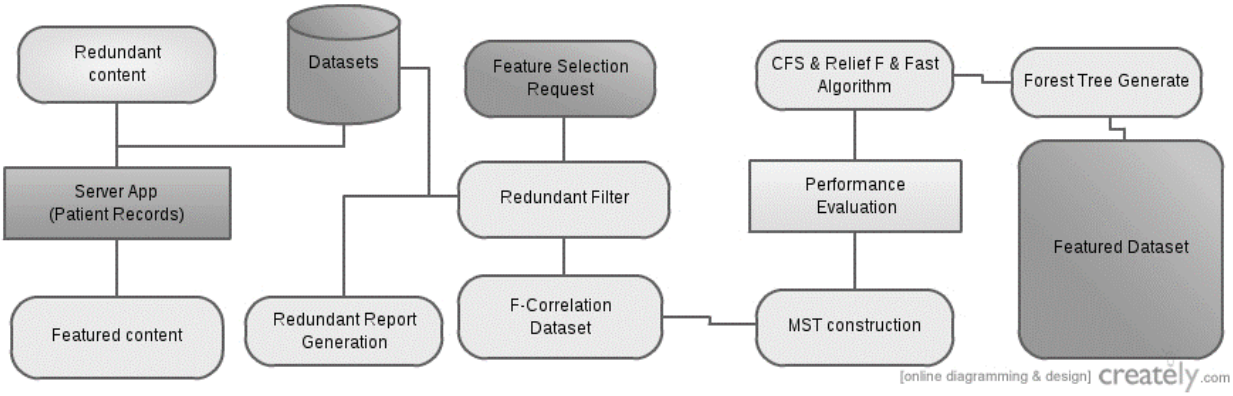


**Figure 1: Proposed system**

The proposed system's benefits include:

Highly linked (predictive of the class) yet uncorrelated (not predictive of one another) qualities in good feature subsets.

• By skillfully handling both irrelevant and redundant features, you may obtain a respectable feature subset.

• Each of the six strategies significantly reduces the number of original qualities chosen by choosing only a tiny portion of them.

• The null hypothesis of the Friedman test is that the runtime of all feature selection methods is the same.

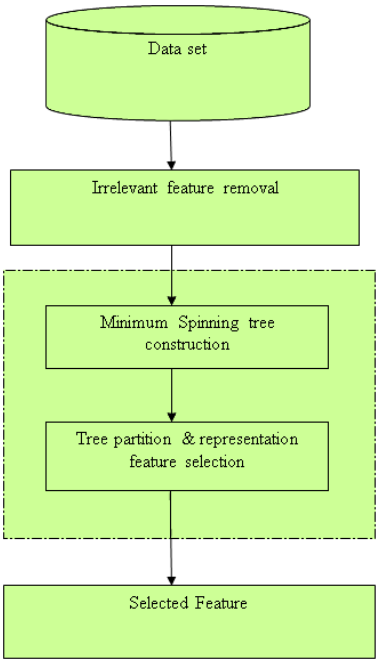


**Figure 2: Work flow**

ISSN: 0975-3583,0976-2833 VOL08, ISSUE 03, 2017

## 4. RESULT & DISCUSSION

An empirical inquiry is used to assess the effectiveness and efficiency of the Clustering Based Feature Selection technique. Four popular classifier types—the probability-based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER—are compared to FSUC and a number of illustrative feature selection algorithms, including FCBF, ReliefF, CFS, Consist, and FOCUS-SF—both before and after feature selection.

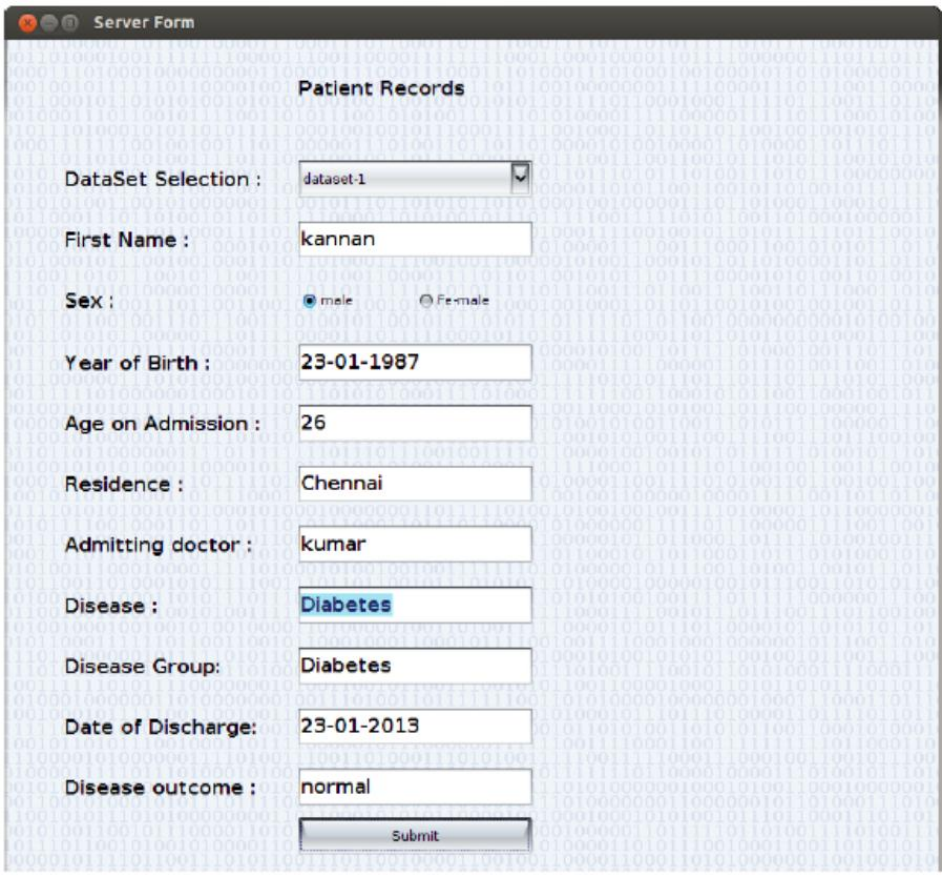


**FIGURE 3: PATIENT RECORDS**

ISSN: 0975-3583,0976-2833 VOL08, ISSUE 03, 2017

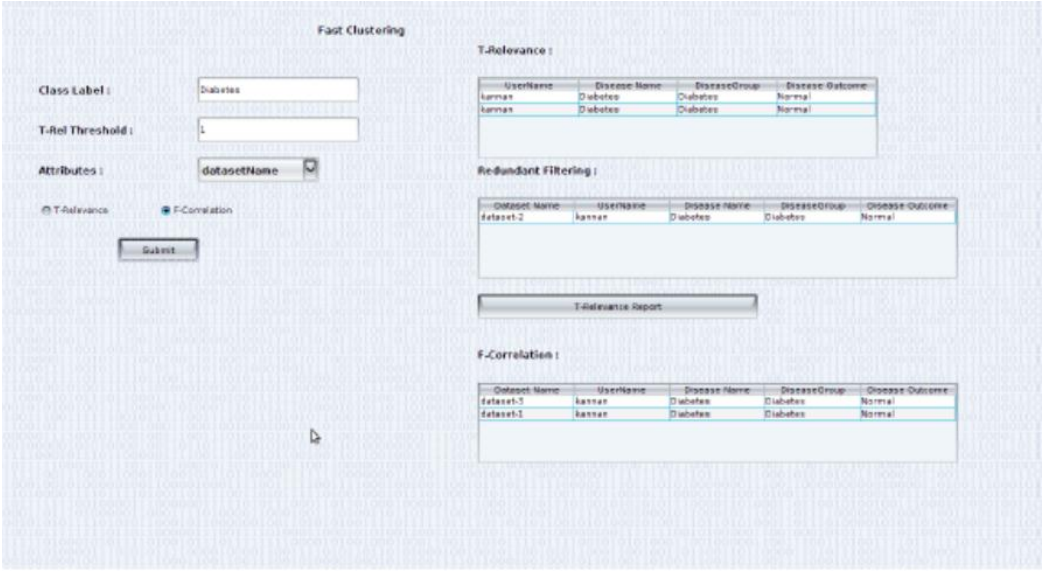


**FIGURE 4: Clustering based feature selection list**

The results show that the Clustering Based Feature Selection not only delivers fewer subsets of features but also enhances the performance of the four kinds of classifiable data on 35 publically available real-world high-dimensional image, microarray, and text data.

## CONCLUSION

In this study, we provide a novel clustering-based feature subset selection approach for high-dimensional data. The MST is divided, irrelevant features are used to build a minimum spanning tree, and representative features are selected. A cluster is produced by the suggested algorithm's characteristics. Because each cluster is treated as a single feature, the dimensionality is significantly reduced. A cluster is produced by the suggested algorithm's characteristics. The handling of each cluster as a single feature considerably reduces the dimensionality.

## REFERENCES

1. F e rri, F., P. P u d i l. Comparative Study of Techniques for Large-Scale Feature Selection. – Pattern Recognit. Pract. IV, Vol. 1994, 1994, pp. 403-413.
2. P u di l, P., J. N ovo v i č o v á, J. K i t t le r. Floating Search Methods in Feature Selection. – Pattern Recognit. Lett., Vol. 15, November 1994, No 11, pp. 1119-1125.
3. D oak, J. An Evaluation of Feature Selection Methodsand Their Application to Computer Security. CSE-92-18, 1992. 82 p.
4. Y u, L., H. Liu. Efficient Feature Selection via Analysis of Relevance and Redundancy. – J. Mach. Learn. Res., Vol. 5, 2004, No Oct, pp. 1205-1224.
5. G h ey a s, I. A., L. S. S m i t h. Feature Subset Selection in Large Dimensionality Domains. – Pattern Recognit., Vol. 43, January 2010, No 1, pp. 5-13.

ISSN: 0975-3583,0976-2833 VOL08, ISSUE 03, 2017

6. Y a ng, Y., J. O. P e d e r se n. A Comparative Study on Feature Selection in Text Categorization. – In: Proc. of 14th International Conference on Machine Learning, ICML'97, 1997, pp. 412-420.
7. J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2012.
8. D. A. P. Kalpit and ocusweG. Soni, "Comparative analysis of k-Means and k-Medoids algorithm on Iris dataa", Int. J. Comput. Intell. Res, vol. 13, no. 5, pp. 899-906, 2017.
9. M. Wei, T. W. S. Chow and R. H. M. Chan, "Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation", Entropy, vol. 17, no. 3, pp. 1535-1548, 2015.
10. M. N. Nisha, S. Mohanavalli and R. Swathika, "Improving the quality of clustering using cluster ensembles", 2013 IEEE Conf. Inf. Commun. Technol. ICT 2013, pp. 88-92, 2013.
11. C. C. Aggarwal, Data Mining: The Textbook, Switzerland:Springer International Publishing, 2015.
12. I. Quinzán, J. M. Sotoca and F. Pla, "Clustering-based feature selection in semi-supervised problems", ISDA 2009 - 9th Int. Conf. Intell. Syst. Des. Appl, no. January, pp. 535-540, 2009.
13. N. Arbin, N. S. Suhaimi, N. Z. Mokhtar and Z. Othman, "Comparative analysis between k-means and k-medoids for statistical clustering", Proc. AIMS 2015 3rd Int. Conf. Artif. Intell. Model. Simul, pp. 117-121, 2016.
14. J. Mao, Y. Hu, D. Jiang, T. Wei and F. Shen, "CBFS: A clustering-based feature selection mechanism for network anomaly detection", IEEE Access, vol. 8, pp. 116216-116225, 2020.