Progressive Redundant and Irrelevant Data Detection and Removal

[1]Prof. Vishal R. Shinde, [2]Abhishek R. Dubey, [3]Megha G. Gadge, [4]Asmita B. Singh

[1]Asst. Professor, [2,3,4]UG Student, [1,2,3,4]Computer Engg. Dept. Shivajirao S.Jondhle College of Engineering & Technology, Asangaon, Maharshatra, India.

[1]*mailme.vishalshinde@gmail.com*, [2]*dubeyabhishek36@gmail.com*, [3]*123megha45@gmail.com*, [4]*asmi0905@gmail.com*

Abstract—Duplicate data detection is the process of recognizing multiple representations of the same data. Today, duplicate detection methods need to process every larger data-sets in short period, along with maintaining the quality of a data-set becomes very difficult. The paper presents two theories, progressive duplicate detection algorithms that significantly increase the efficiency of finding duplicates if the execution time is limited and fast clustering which reduces the irrelevancy of data. The algorithm maximizes the gain of the overall process within the time available by reporting most results much earlier than traditional approaches. Comprehensive tests show that the progressive algorithms along with feature subset selection can double the efficiency over time of traditional duplicate detection and significantly improve upon related work.

Keywords—*Duplicate detection, entity resolution, pay-as-you-go, progressiveness, data cleaning*.

I. INTRODUCTION

Data mining is processing databases to identify patterns and establish relationships. Data mining is the process of analysing large amount of data stored in a data warehouse. Clustering is data mining technique used to place data elements into related groups without advance knowledge of group definition.

Feature selection is achieved automatically in analysis and each algorithm has a set of default techniques for wisely applying feature reduction on database. Feature selection is regularly performed before the model is qualified to automatically prefer the attributes in a dataset that are most likely to be used in that model. In spite of this, you can also manually select parameters to induce feature selection behaviour. In general, feature selection works by measuring a score for each feature and then selecting only that features which have the best score of all. Analysis Services provides multiple techniques for measurement of the scores and the accurate method that is applied in any model depends on the below mentioned factors:

A. The algorithm used in your representation.
B. The data type of the feature.
C. Any factors that you might have set on your representation.

There are three main categories of feature selection algorithms: wrappers, filters and embedded methods. The wrapper technique uses a predictive model to score feature subsets. The error rate of the model gives the score for that subset. Filters are mostly fewer computationally exhaustive than wrapper technique, but they generate a feature set which is not regulated to a specific type of predictive model. Many filters supply a feature ranking relatively than a best feature subset, and the interrupt point in the ranking is chosen through cross-validation. The wrapper methods use the predictive exactness of a determined learning algorithm to determine the integrity of the selected subsets, the accuracy of the machine learning algorithms are generally high. However, the majority of the selected features are partial and the computational complexity is huge.

Clustering of the features by using the graph-theoretic approach to select most representative feature related to target class is done in this paper. For that, the algorithm take Minimum-Spanning-Tree (MST) in Fast clustering-based feature Selection algorithm (FAST). FAST algorithm completes in two steps. First, features are separated into various clusters, next the most effective feature is selected from each cluster.

II. AIM & OBJECTIVE

1. ***AIM***

The paper identifies most duplicate pairs early in the detection process. Instead of reducing the overall time needed to finish the entire process, progressive approaches try to reduce the average time after which a duplicate is found. Due to early termination it yields more complete results on a progressive algorithm than on any traditional approach.

Feature subset selection generally focused on searching relevant features while neglecting the redundant features. The duplicate data detection workflow comprises the three steps such as pair-selection, pair-wise comparison, and clustering.

For a progressive workflow, only the first and last step needs to be modified.

2. ***OBJECTIVES***

A. Try to group a set of points into clusters.
B. Points in the same clusters will be most similar to each other than points in different clusters.
C. To use similarity matrix method.
D. To remove irrelevant and redundant data.
E. To increase predictive accuracy.
F. Improving the efficiency and effectiveness of data retrieval.
G. Using linked datasets as an input dataset.

III. LITERATURE SURVEY

1] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast clustering based feature subset selection algorithm for high dimensional data", In proceedings of the IEEE Transactions n Knowledge and data engineering, 2013. : The paper expressed the FAST grouping based calculation is compelling and productive. Proficiency alludes to time required to seek a specific list of capabilities through vast information and powerful worries with quality or precision of the chose highlight set. Quick uses MST for highlight choice. It works in two unique steps including dividing information into bunches and after that finding fitting list of capabilities. The paper likewise clarified wrapper, channel, half and half systems and its impediments. The paper has tested FCBF, Relief F, CFS, Consist, and FOCUS-SF strategies on 35 distinctive datasets and infer that FAST calculation is more successful than all others. The careful working of FAST calculation is clarified in this paper. As FAST likewise deals with evacuating immaterial information, it is more helpful in finding precise results.
Merits: Improve the performance of classifiers
Demerits: Required more time

2] L. Yu and H. Liu, "Feature Selection for High Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.: Information mining of high dimensional information is enormous issue. Highlight choice from high dimensional information is fundamentally centered around uprooting superfluous components i.e. not identified with the required hunt. Be that as it may, it is additionally hard to uproot immaterial information. The paper gives an investigation of highlight repetition in high-dimensional information and proposes a novel correlation based way to deal with highlight determination inside of the channel model. Established direct connection uproots highlights with almost zero straight relationship to the class and decrease repetition among chose highlights. It utilizes FCBF calculation.
Merits: Fast, Remove noisy and irrelevant data
Demerits: Unable to handle large volumes of data

3] M. Dash, H. Liu, and H. Motoda, "Consistency Based Feature Selection," Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109, 2000.: The paper expressed that the component determination calculations utilizes different measures to decide the helpfulness and viability of item. The paper principally focused on consistency measure for highlight choice. The paper has clarified the consistency, its properties and correlation with other present measures for highlight choice. Likewise, the paper concentrated on how to compute irregularity measure, separation measure, consistency measure, data measure for better hunt.
Merits: Accuracy is high
Demerits: Computational complexity is large

4] Yijun Sun, SinisaTodorovic "Local Learning Based Feature Selection for High Dimensional Data Analysis" IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 32, No. 9, Sept.2010: The paper expresses the apportioning of complex non-straight issues into basic nearby direct sets with the assistance of neighborhood learning. At that point through numerical investigation and machine learning, highlight importance is gotten all around. The calculation is utilized as a part of high dimensional information. In this calculation, initial step is to figure the edge by separation capacity i.e. to begin with discover two neighbors of each example, one from closest hit class and other from closest miss class. Through this edge, the commotion or unimportance of elements can be gotten. The calculation needs to process edge and remove work locally inevitably. at that point it finds shrouded variables. It additionally states real issue with RELIEF calculation that the closest neighbors of a given example are predefined in the first component space, which ordinarily yields mistaken closest hits and misses in the vicinity of plentiful unimportant elements.

5] A New Clustering Based Algorithm for Feature Subset Selection (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5272-5275: The paper contains the primary thought of the FAST grouping based component determination calculation and its progression for working of calculation. The paper has additionally clarified the principle challenge that if more than one component is joint and it suit the objective element so it can be dealt with as pertinent. Highlight interface is the new test for distinguishing the pertinent component. Quick calculation removes just focused on elements out of numerous elements. These elements don't quantify the insignificant and excess information in light of the fact that immaterial and repetitive information influences the fitness and adequacy of the calculation. It additionally clarifies appropriated grouping and time unpredictability of demur's calculation.

6] Ding cheng Feng, Feng Chen, and Wenli Xu "Efficient Leave-One-Out Strategy for Supervised Feature Selection" TSINGHUA SCIENCE AND TECHNOLOGY

ISSNll1007- 0214ll09/10llpp629 635 Volume 18, Number 6, December 2013: The paper demonstrates to choose upgraded highlight sets on the premise of grouping and in addition characterization. The paper has clarified eager in reverse disposal calculation for better advancement. It expresses that finding related element is more troublesome than to discover unimportant information sets or non-related component. Henceforth forget one procedure might be more useful and savvy. The constraint of eager in reverse end calculation has overcome through this leave-one out procedure.

7] Houtao Deng, George Runger "Feature Selection via Regularized Trees" The 2012 International Joint Conference on Neural Networks (IJCNN), IEEE, 2012: The creator proposed tree regularization system for highlight subset choice. This model or structure gives compelling results as tree models can manage variables of numerical and clear cut, distinctive scales between variables, missing information, collaborations and nonlinearities and so forth. The calculation incorporates the
Part of tree from first hub up to last. At that point the woodland trees are produced which are alluded as regularized irregular backwoods (RRF) and regularized helped arbitrary (R Boost). This system can manage either solid or feeble classifiers.

8] SriparnaSaha "Feature selection and semi-supervised clustering using multiobjective optimization" SpringerPlus 2014, 10.1186/2193-1801-3-465.: The paper expressed the multi-target enhancement to conquer the issue of programmed highlight choice and semi-directed grouping. group focuses and components are encoded as a string. The expressed system documented multi-objective mimicked tempering (AMOSA) is used to identify the proper subset of elements, fitting number of bunches and in addition the suitable apportioning from any given information set. The recreated strengthening (SA) system has impediments that it gives single answer for highlight choice after single run. Consequently, for multi-dimensional information, it is not valuable to discover list of capabilities.

9] R.Munieswari,"A Survey on Feature Selection Using FAST Approach to Reduce High Dimensional Data", IJETT, Volume 8 Number 5- Feb 2014 : The paper expressed the overview on various element determination strategies or calculations including wrapper, channel, quick grouping calculation, half breed approach and help calculation. At that point reorder the group as indicated by max weight. At the point when the bunch weight crossed the limit esteem, the specific group is taken as list of capabilities. Every single other strategy is concentrated on in past and different papers said above. The paper have likewise expressed the FAST calculation with utilization of MST and gives point of preference of FAST over every other system.

10] JesnaJose,"Fast for Feature Subset Selection Over Dataset" International Journal of Science and Research (IJSR), Volume 3 Issue 6, June 2014: The wrapper, filter importance examination models are studied in this paper. Examination of importance and excess arranged the elements on pertinence premise as whether is of class solid significance, frail pertinent or insignificant. Solid significance of a component shows that the element is constantly vital for an ideal subset; it can't be uprooted without influencing the first contingent class appropriation. Frail significance proposes that the element is not generally essential but rather might get to be important for an ideal subset at specific conditions. Superfluity demonstrates that the component is a bit much by any means.

IV. EXISTING SYSTEM

Feature subset selection generally focused on searching relevant features while neglecting the redundant features. A good example of such feature selection is Relief, which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function [9]. But, Relief is ineffective in removing redundant features as the two predictive but highly correlated features are likely to be highly weighted. Relief-F [6] is an extension of the traditional Relief. This method supports working with noisy and incomplete data sets and to deal with multi-class problems, but is still it is ineffective in identifying redundant features. However, along with irrelevant features, redundant features also do affect the speed and accuracy of all the probable learning algorithms, and thus should are also important to be eliminated.

This is because:
A. Irrelevant features do not contribute to the predictive accuracy.
B. Redundant features do not rebound to getting a better predictor.

The wrapper methods make use of predictive accuracy of a predetermined learning algorithm to determine the effectiveness of the selected subsets[7]. The accuracy of the learning algorithms is mostly very high. The however the generality of the selected features is limited and the computational complexity is very large. Thus, the wrapper methods are computationally expensive and tend to over fit on small feature training sets. Wrapper uses a search algorithm for searching through the space of possible features and evaluates individual subset by running a model on the subset. The filter methods [3] are independent of the learning algorithms, and have good generality. Computational complexity is low, but the accuracy of such learning algorithms is not guaranteed.

V. PROBLEM STATEMENT

Certain problems occur in the detection process and have several use cases such as a user may have limited or unknown time for data cleansing, also user may have little knowledge

Sr. No	Paper Name	Author Name	Technology/ Algorithm	Advantages	Disadvantages
01	A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data	Qinbao Song, Jingjie Ni and Guangtao Wang	FAST	1.Improve the performance of the classifiers. 2.The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.	--
02	Fast Correlation Based Filter (FCBF) with a Different Search Strategy	Baris Senliol1 , Gokhan Gulgezen1 , Lei Yu2 and Zehra Cataltepe1	FCBF	1. FCBF compares only individual features with each other. 2. Select fewer features with higher accuracy. 3. Feature selection is a pre-processing step to machine learning which is effective in reducing dimensionality, removing irrelevant data.	1.Cannot detect some features. 2. Four features generated by four Gaussian functions and adding 4 additional redundant features, FCBF selected only 3 features.
03	An efficient k-means clustering algorithm: analysis and implementation	T. Kanungo	K-MEAN	1. If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. 2. K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular. 3. Fast, robust and easier to understand.	1. Fixed number of clusters can make it difficult to predict what K should be. 2. Does not work well with non-globular clusters. 3. Different initial partitions can result in different final clusters. 4. It does not work well with clusters (in the original data) of Different size and Different density.
04	A Simulated Annealing-Based Multi objective Optimization Algorithm: AMOSA	Sanghamitra Bandyopadhyay	Simulating Annealing	1.Accuracy, Useful for small datasets.	1.Single feature for single turn.
05	Clustering Approach to Collaborative Filtering Using Social Networks	Emir Cogoand Dzenana Donko	Filter Approach	1. Suitable for very large features.	1.Accuracy is not Guaranteed.

about the given data. It is not possible to eliminate several factors of duplicate detection such as effectiveness and scalability due to database size. The problem here is no one cares about the database maintenance with ease manner. The systems like Distortion and Blocking algorithm [8], which creates an individual area for each and every word from the already selected transactional database, those are collectively called as dataset, which will be suitable for a set of particular words, but it will be problematic for the set of records, once the user get confused the data can't be recovered back.

Irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that it provides mostly information which is already present in another feature.

Progressive Sorted Neighborhood Method (PSNM) is based on the traditional sorted neighborhood method that increase the efficiency of duplicate data detection with limited execution time. This methods detect only duplicate records serially and not removes that records.

The wrapper methods use the predictive accuracy technique of an already determined learning algorithm to calculate the usefulness of the selected subsets, which makes the work faster. The FAST algorithm [1] research has focused on searching for relevant features. The most 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.

VI. COMPARATIVE ANALYSIS

Table:6.1*- Comparative Analysis*

VII. PROPOSED SYSTEM

The paper proposes a clustering algorithm, FAST for high dimensional data. The algorithm includes

A. Irrelevant features removal

B. Construction of a minimum spanning tree (MST).

C.Partitioning the MST and selecting the representative features.

Feature subset selection algorithm should be able to recognize and remove as much of the unrelated and redundant information. In this proposed algorithm, a cluster will be used to develop a MST for faster searching of relevant data from high dimensional data. Each cluster will be treated as a single feature and thus volume of data to be processed is drastically reduced to small size. FAST clustering will obtain the best proportion of selected features, the best runtime along with the best classification accuracy. Overall the system will be effective in generating more relevant and accurate features which can provide faster results.

VIII. MATHEMETICAL MODEL

$H(X)$ is the entropy of a discrete random variable X. Let (x) be the prior probabilities for all values of X, then (X) is defined by

$$H(X) = -\sum_{x \in X} p(x) \log 2 p(x)$$

Gain (X | Y) determines the amount by which the entropy of Y decreases. It is given by,

$$\text{Gain } (X|Y) = H(X) - H(X|Y)$$

$$= H(Y) - H(Y|X)$$

Where H(X | Y) is the conditional entropy and is calculated as,

$$H\left(\frac{X}{Y}\right) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x) \log_2 p(x)$$

Where, X is a Feature and Y is a Class.

The symmetric uncertainty (SU) is defined as follows,

$$SU(X,Y) = \frac{2 \times \text{Gain}\left(\frac{X}{Y}\right)}{H(X) + H(Y)}$$

Given that (X, Y)be the symmetric uncertainty of variables X and Y, the relevance T-Relevance between a feature and the target concept [8] C, the correlation F- Correlation between a pair of features, the feature redundancy F-Redundancy and the representative feature R- feature of a feature cluster can be defined as follows. T-Relevance - The relevance between the feature Fi ∈ F and the target concept is referred to as the T-Relevance of Fi and C, and denoted by SU (Fi,C). If SU (Fi,C) is greater than a predetermined threshold θ,Symmetric Uncertainty of each Feature is greater than the T-Relevance threshold is checked.

Algorithm : FAST

inputs: D($F1, F2, ..., Fm, C$) - the given data set
θ- the T-Relevance threshold.
output: S - selected feature subset [1].
//==== Part 1 : Irrelevant Feature Removal
1 for i = 1 to m do
2 T-Relevance = SU (Fi, C)
3 if T-Relevance >θthen
4 S = S ∪ {Fi};
//==== Part 2 : Minimum Spanning Tree Construction
5 G = NULL;
6 for each pair of features {F'_i, F'_j}⊂ S do
7 F-Correlation = SU ($F'_i,'_j$)
8 $Add\ F'i\ and/or\ F'j\ to$ with F-Correlation as the weight of the corresponding edge;
9 minST = Prim (G); // Prim Algorithm to generate the minimum spanning tree
//==== Part 3 : Tree Partition and Representative Feature Selection
10 Forest = minST
11 for each edge $E_{i,j}$∈Forest do
12 if SU ($F'_i,'_j$) < SU(F'_i, C) ∧SU(F'_i, F'_j) <SU(F'_j, C) then
13 Forest = Forest − E_{ij}
14 S = ϕ
15 for each tree T_i∈Forest do
16 F^j_R=argmaxF_k∈TSU(F'_k, C)
17 S = S ∪ {F^j_R};
18 return S

IX. SYSTEM ARCHITECTURE

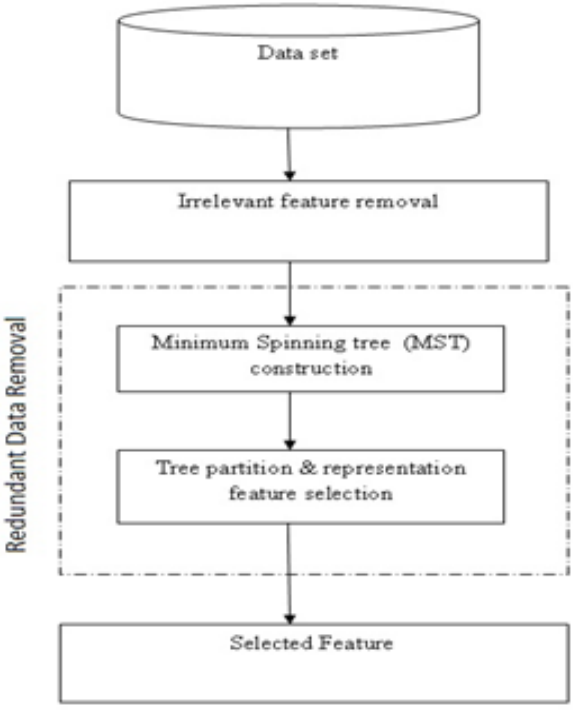

Fig:1- *System Architecture*

A. **LOAD THE DATASET**

According to the system architecture, first load the dataset into the process. The dataset has to be pre-processed for eliminating absent values, noise and outliers.

B. **GENERATE SUBSET PARTITION**

Generating the partitions is the step to divide the whole dataset into various partitions, which will be able to classify and identify for irrelevant & redundant features. A strategy for reviewing the quality of model simplification is to divide the data source.

C. **IRRELEVANT FEATURE REMOVAL**

Useful way for sinking dimensionality, eliminating inappropriate data, rising learning accuracy, and civilizing result comprehensibility. The inappropriate feature removal is directly once the right relevance quantify is defined or chosen, while the unnecessary feature elimination is a bit of complicated.

D. **MST CONSTRUCTION**

This can be shown by an example, suppose the Minimum Spanning Tree shown in Fig.2 is produced from a complete graph G. In organize to cluster the features, the algorithm first go across all the six edges, and then decided to remove the edge ($F0$, $F4$) because its weight ($F0$,4)=0.3 is lesser than both $SU(F0,C)$=0.5and$SU(F4,C)$=0.7. This constructs the MST is grouped into two clusters denoted as ($T1$) and ($T2$).To generate MST Prim's Algorithm has been used in this paper.

E. **SELECTED FEATURES LIST & CENTROID CLUSTERING**

Ultimately it includes for final feature subset. Then calculate the accurate/relevant feature. These features are relevant and most useful from the entire set of dataset provided for feature subset selection. In centroid-based clustering method, clusters are denoted by a central vector, which might not essentially be a member of the data set.

X. EXPECTED OUTPUT

Fig:2- Result After Clustering

Above figure (fig:10.1) shows the output of cluster formation by the feature subset selection algorithm. These clusters will be provided as an input for creating MST by Prims and Kruskal algorithm.

XI. ADVANTAGES

A. Less training set along with less memory will occupy semi supervised process.
B. Alike data would not be missed in cluster data by using pair wise constrain technique.
C. Overlapping avoid by using maximum margin cluster process
D. Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.
E. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset which is the biggest advantage.
F. Mostly, all the six algorithms achieve significant reduction of dimensionality by selecting only a small portion of the original features.
G. The null hypothesis of the Friedman test proves that all the feature selection algorithms are equivalent in terms of runtime.

XII. CONCLUSION

We have tried to implement "Thorsten Papenbrock, Arvid Heise, and Felix Naumann, " PROGRESSIVE DUPLICATE DETECTION ", May 2015". According to implementation the algorithm includes eliminating irrelevant features, producing a minimum spanning tree from relative ones, and partitioning the MST and selecting most useful features formed as new clusters from the selected most useful feature list. In the proposed algorithm, a cluster consists of features and every cluster is treated as a distinct feature and thus dimensionality is radically reduced.

REFRENCES

[1] Thorsten Papenbrock, Arvid Heise, and Felix Naumann, " Progressive Duplicate Detection ", in proceedings of the IEEE Transactions Knowledge and data engineering, May 2015.

[2] Karthikeyan.P, High Dimensional Data Clustering Using Fast Cluster Based Feature Selection, Int. Journal of Engineering Research and Applications, March 2014, pp.65-71.

[3] B.Swarna Kumari, M.Doorvasulu Naidu, Feature Subset Selection Algorithm for Elevated Dimensional Data By using Fast Cluster, In International Journal Of Engineering And Computer Science Volume 3 Issue Page No. 7102-7105, 7 July, 2014.

[4] Sumeet Pate, E.V. Ramana, A Search Engine Based On Fast Clustering Algorithm for High Dimensional Data, International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 3, Issue 10, October 2014.

[5] Comparative study of various clustering techniques with FAST, International Journal of Computer Science and Mobile Computing, Volume 3, Issue 10, October 2014.

[6] Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T., Numerical recipes in C. Cambridge University Press, Cambridge, 1988.

[7] 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.

[8]A fast clustering-based feature subset selection algorithmMr. Akshay S. Agrawal1, Prof. Sachin Bojewar21 P.G. Scholar, Department of Computer Engg., ARMIET, Sapgaon, (India)2Associate Professor, VIT, Wadala, (India)

[9] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast clustering based feature subset selection algorithm for high dimensional data", in proceedings of the IEEE Transactions Knowledge and data engineering, 2013.

