International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 6 No. 1 Aug. 2014, pp. 75-81

http://www.ijisr.issr-journals.org/

EFFICIENT AND EFFECTIVE SUBSET SELECTION PROCESS BASED ON CLUSTERING ALGORITHM

K. Revathi and T. Kalai Selvi

Computer Science and Engineering, Erode Sengunthar Engineering College,
Anna University Chennai, Tamilnadu, India

ABSTRACT: Process with high dimensional data is enormous issue in data mining and machine learning applications. Feature selection is the mode of recognize the good number of features that produce well-suited outcome as the unique entire set of features. Feature selection process constructs a pathway to reduce the dimensionality and time complexity and also improve the accuracy level of classifier. In this paper, we use an alternative approach, called affinity propagation algorithm for effective and efficient feature selection and clustering process. The endeavor is to improve the performance in terms accuracy and time complexity.

KEYWORDS: Classification, Data mining, Feature selection, Feature clustering.

1 INTRODUCTION

Data mining is the route of ascertaining the interesting knowledge from hefty amounts of information repository. Noisy, incomplete, inconsistent records are humdrum properties of huge real world databases and data warehouses. To handle this type of errors, data preprocessing techniques are extremely essential for producing good quality result. Feature selection, also known as attribute subset selection is similar to preprocessing technique, used for dimensionality reduction; improve the classifier accuracy, removing irrelevant and redundant data. Feature selection techniques are categorized into four types: the Filter, wrapper, Embedded, and hybrid methods [1]. Filter method [11], [12] is momentous selection when we use large number of features. Filter the features using ranking based approach. Wrapper method [2], [14] is used to estimate the integrity of the selected subset features by using predictive accuracy of machine learning algorithm [1], [14] which provides greatest accuracy. Embedded method [14] has grand efficiency than remaining methods, it works with training process. Hybrid method is the mixture of filter and wrapper method. Thus, we will focus on the wrapper method in this paper.

Several conventional feature selection algorithms are available but we will focus on application of cluster analysis for more effective feature selection process. Cluster analysis is the progression of grouping similar objects into one class. For produce an optimal result, affinity propagation algorithm have been studied and used in this paper.

In general affinity propagation is most flexible and simple clustering algorithm. It works through the concept of "information passing" between data objects. Key benefit of this algorithm is low error; maintain high speed and prominently no need to compute the number of clusters before executing the algorithm.

The proposed feature selection process based on affinity propagation algorithm produce optimal subset of features with high accuracy and minimum time requirement.

The rest of the paper is organized as follows: in section II, we demonstrate the related work. In section III, we analysis about the process of existing work. In section IV, we summarize the proposed work with comparison analysis. In section V represent the conclusion about this paper.

Corresponding Author: K. Revathi

2 RELATED WORK

Dimensionality reduction, identifying and removing irrelevant and redundant features are done with the process of feature selection. Feature selection is same as the data preprocessing technique for producing best possible subset of features.

There are several algorithms and schemes for feature selection process are obtainable, Relief is well known and good feature estimator. Using relief algorithm [1], [4], [5] estimate the quality of the feature subset but it successfully remove irrelevant features only, does not consider about the redundant features.

The Mutual Information [1], [6] is another method for determine the dependence of pair of features and feature with target class. In [3], M. Dash and H. Liu et al, focuses on inconsistency measure for feature selection process with various search strategies, each strategies correspond to different algorithms, such as

- Focus: Exhaustive search
- ABB: Complete search
- SetCover: Heuristic search
- LVF: probability search
- QBB: Hybrid search.

Figure 1 shows the flow diagram of feature selection process.

Steps for feature selection process:

Invention: process with original dataset, produce a candidate subset.

Estimation: evaluation is done using candidate subset.

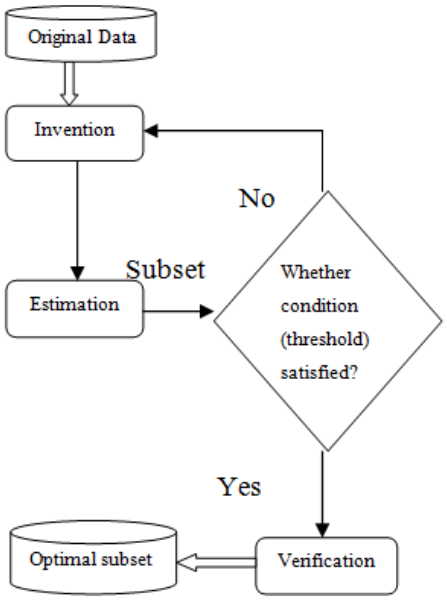


Fig. 1.Feature selection process

Verdict process: Compare the selected subset with checking criteria (threshold value)

Substantiation: cross validation is performed for select optimal subset of features.

In [8], Z. Zhao and H.Liu et al, Proposed a method to handle the feature order problem with help of INTERACT algorithm. Based on the interaction (correlation) between features, achieve the feature selection process. For increase se the efficiency of the performance of feature selection, cluster analysis very important concept. Grouping the data objects into clusters as like tree structure, also known as dentrogram is the idea of hierarchical clustering, [1], [9] it is classified into two types: Agglomerative (Bottom up approach) and divisive (Top down approach) algorithm. R. Butterworth

et al, use the AGENS for making the dentrogram of votes dataset, this structure helps to understand the features and corresponding relative importance.

In [1], Qinbao Song et al, proposed a new FAST algorithm that gain more accuracy and reduce time complexity than traditional feature selection algorithm like, FCBF, Relief, CFS, FOCUS-SF, Consist and also compare the classification accuracy with prominent classifiers. Graph-theoretic clustering and MST based approach is used for ensure the efficiency of feature selection process.

Classifiers plays vital roles in feature selection operation since accuracy of selected features are measured using the progression of classifiers. The following classifiers are utilized to classify the data sets [1], [8], Naïve Bayes: it works under Bayes theory and is based on probabilistic approach and yet then offers first-rate classification output. C4.5 is the successor of ID3 [1] support of decision tree induction method. Gain ratio, gini index information gain are the measures used for the process of attribute selection. Simplest algorithm is IB1 (instance based) [1]. Based on the distance vectors, it performs the classification process. RIPPER [1] is the rule based technique, it make a set of rules for the purpose of classify the data sets. Classifier is one of the evaluation parameter for measuring the accuracy of the process.

3 EXISTING WORK

Out of many feature selection algorithms, FAST algorithm is one of the most effective feature selection algorithms. It works based on the following terms [1]:

- T- Relevance
- F-Correlation
- R-Feature
- F-Redundancy

For irrelevant preprocess calculate the symmetric uncertainty between feature and the target concept, based on that threshold value removing irrelevant features. Then apply the graph theoretic clustering [1] method for grouping the relevant features. For best result use the minimum spanning tree concept, it follows two steps:

1. Construction of tree with minimum weight.
2. Partitioning the tree, each tree represents the cluster of features.

FAST algorithm gains the [1] performance in terms of (i) accuracy, (ii) runtime, (iii) Proportion of selected features than traditional algorithms like ReliefF, CFS, FOCUS-SF and Consist with respect to different classifiers.

4 PROPOSED WORK

In this section, we present the analysis of proposed algorithm and process. Irrelevant and redundant data highly affect the performance of the mining process. Feature selection is the process of recognize and eliminate the unrelated and redundant features for improving the classification accuracy, reduce the dimensionality etc…

The main issue in the FAST algorithm is construction of minimum spanning tree because it takes more time to building and partitioning the tree for selecting most optimal features. To overcome this type of issue we choose alternative algorithm for reducing running time and also improving the accuracy level of features.

In our proposed implements, Semi supervised learning technique has detains the features. Semi supervised learning is a machine learning pattern in which the model is constructed using both labeled and unlabeled data for training typically a small amount of labeled data and a large amount of unlabeled data. Affinity propagation algorithm is one of the most important clustering algorithms with high speed and low error. This algorithm is really suitable for selecting most appropriate features. For improving the accuracy level we utilize the some similarity measures such as, Jaccard similarity measure and cosine similarity measure.

Jaccard Similarity Measure: To find the similarity between sample sets, Jaccard similarity measure most appropriate measure for improving the classification accuracy. It is defined as the amount of intersection is divided by amount of union of the sample sets.

$$J(R, S) = \frac{| R \cap S |}{| R \cup S |}$$

Where, R and S are the two sample sets.

| R **|, |** S **|** are the cardinality of R S, it represents the elements in the sets R, S.

| R ∩ S | is the value of the intersection between two sets R and S**.** | R U S | is the value of the union between two sets R and S. J (R, S) is the jaccard index value,used to calculate the similarity between two sample sets. This type of measure mainly used to improve the performance of the classification process.

For example consider two sets with numbers, R = {0, 2, 5, 7} and S = {1, 7, 8, 9, 0}. Calculate how similar are R and S?

$$J(R, S) = \frac{| R \cap S |}{| R \cup S |}$$

$$= \frac{| \{0, 7\} |}{| \{0, 1, 2, 5, 7, 8, 9\} |}$$

$$= 2 / 7 = 0.2857$$

The above method is the process of determine the similarity range between two sample sets. Apply this concept to documents, it contains bag of words.

Consider the following example for establishing jaccard coefficient value for two documents.

R= {beautiful flower}

S= {lotus national flower}

$$J(R, S) = \frac{| R \cap S |}{| R \cup S |}$$

$$= \frac{| \{flower\} |}{| \{beautiful, lotus, national, flowers\} |}$$

$$= 1/ 4 = 0.25$$

For efficient estimation of accuracy min hash method has been used, it is represented as,

$D_J = 1-J(R, S)$

$= 1- 0.25 = 0.75$

Where, D_J denotes the accurate distance between two sets (i.e. Training set and original file).

COMPARISION ANALYSIS

The following graphs represent the results of different algorithms in the following terms:

- Classification Accuracy
- Runtime

Classification Accuracy: It represents the quality of selected subset of features. Also known as effectiveness.

Runtime: It shows the execution time of the entire process. It also denoted as efficiency.

DATA FLOW DIAGRAM

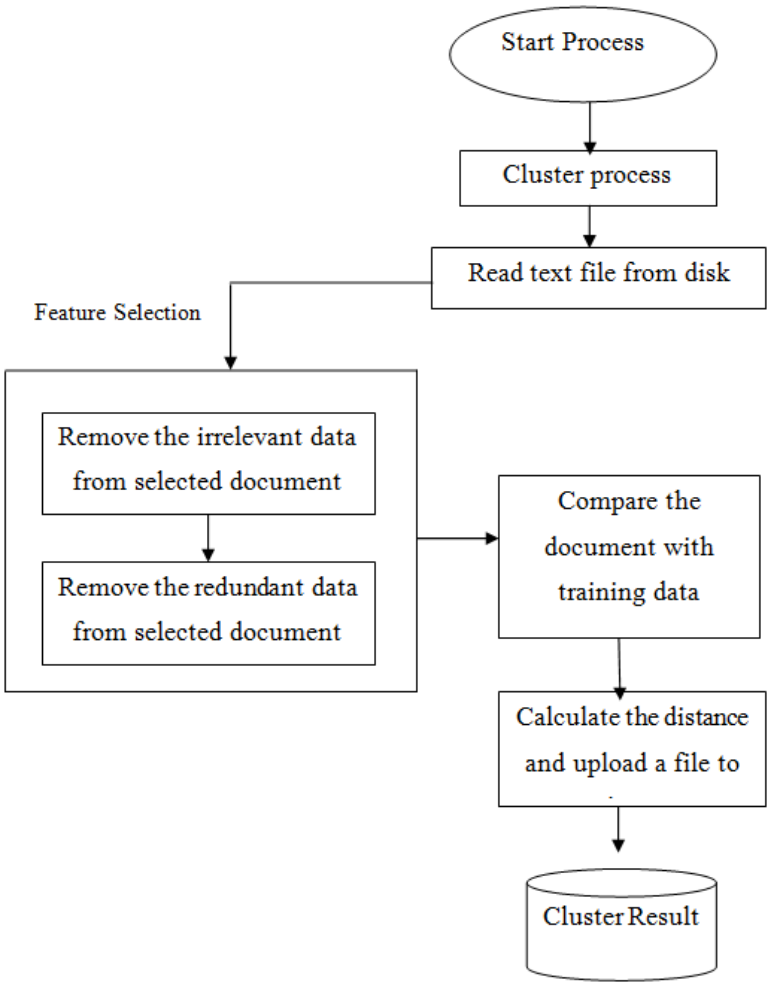
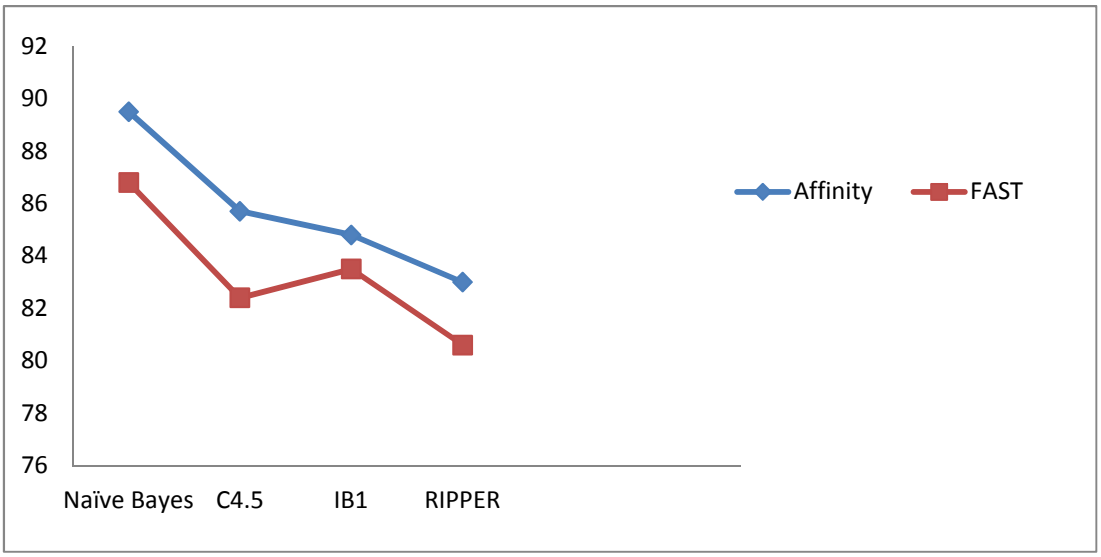
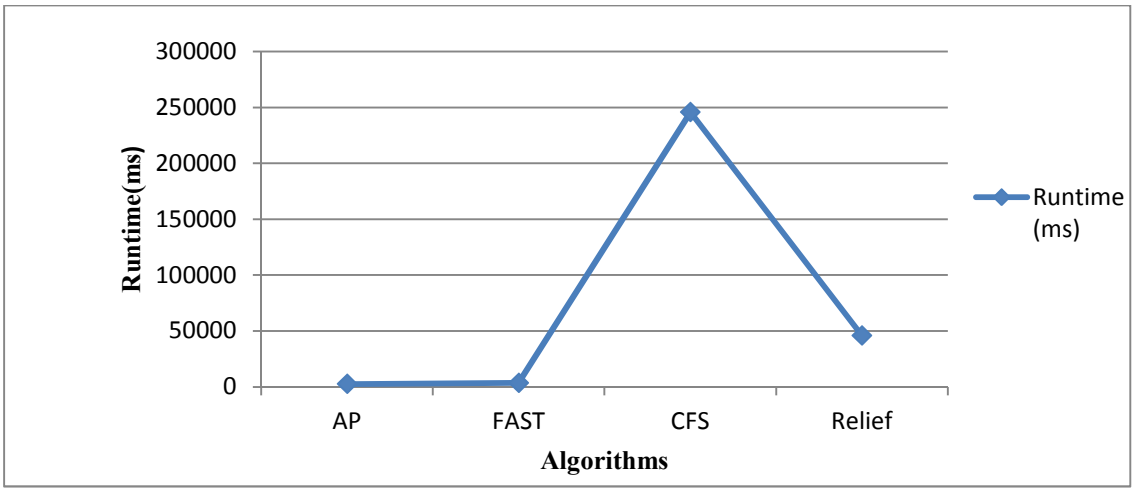


Fig. 2.Data Flow Diagram



Graph 1: Classification accuracy



Graph 2: Runtime

5 CONCLUSION

In this paper, we have discussed an alternative clustering based feature selection algorithm. This algorithm effectively removes the irrelevant and redundant features for dimensionality reduction. Grouping the features for selecting optimal set of features based on the semi supervised learning method. We have compared the performance of the proposed algorithm with existing FAST algorithm in terms of accuracy and runtime. Main benefit of proposed algorithm is high speed and low error. The motive of this process is increasing the accuracy level of classifiers and reducing the runtime of the algorithm.

REFERENCES

[1] Qinbao Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," IEEE Transaction on Knowledge and Data, Engineering, Vol. 25, No. 1, January 2013.

[2] 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.

[3] M. Dash and H. Liu, "Consistency- Based Search in Feature Selection," Artificial Intelligence, vol. 151, nos. 1/2, pp. 155-176, 2003.

[4] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[5] H. Liu, H. Motoda, and L. Yu, "Selective Sampling Approach to Active Feature Selection," Artificial Intelligence, vol. 159, nos. 1/2, pp. 49-74, 2004.

[6] Battiti,"Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.

[7] C.Krier, D.Francois, F. Rossi, and M. Verleysen, "Feature Clustering and Mutual Information for the Selection of Variables in Spectral Data," Proc. European Symp. Artificial Neural Networks Advances in Computational Intelligence and Learning, pp. 157-162, 2007.

[8] Z. Zhao and H. Liu, "Searching for Interacting Features," Proc. 20th Int'l Joint Conf. Artificial Intelligence, 2007

[9] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 581-584, 2005.

[10] L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103,1998.

[11] 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.

[12] B. Raman and T.R. Ioerger, "Instance-Based Filter for Feature Selection," J. Machine Learning Research, vol. 1, pp. 1-23, 2002.

[13] J. Biesiada and W.Duch, " Feature Election for High-Dimensional data a Pearson Redundancy Based Filter," Advances in soft Computing, vol.45,pp.581-584,2005.

[14] Sivia Cateni, Valentina Colla and Marco Vannucci, "A Genetic Algorithm based Approach for Selecting Input variables and Setting Relevant Network Parameters of a SOM- based classifier", International Journal of Simulation Systems, Science and Technology, vol.12.

[15] Asha Gowda Karegowda, A.S. Manjunath and M.A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection", International Journal of Information Technology and Knowledge Management, vol.2, No2, pp.271-277

Revathi.K received B.E degree in Computer science and Engineering from Vivekananda Institute of Engineering and Technology for Women, Namakkal. She is currently pursuing Master degree in Department of Computer science and Engineering at Erode Sengunthar Engineering College Erode. She has published 2 papers in reputed journal and 3 papers in various national conferences. Her Research interest includes Data Warehousing and Data Mining

.

Kalai Selvi.T received M.E degree in Computer Science and Engineering from Mahendra Engineering College , Namakkal and presently she is working as Assistant Professor (Selection Grade-I) in Department of Computer science and Engineering at Erode Sengunthar Engineering College, Erode. She has published 5 papers in reputed journal and more than 10 papers in various National and International conferences. Her Research interests include Cloud Computing. She is a Life member of ISTE and CSI.