ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 12, December 2013

Survey: Effective Feature Subset Selection Methods and Algorithms for High Dimensional Data

K.Revathi, T.Kalai Selvi

Abstract— **Feature selection is the mode of recognize the good number of features that fabricate well-suited outcome as the unique entire set of features. Feature Extraction is the special form of dimensionality reduction where feature selection is the subfield of feature extraction. Feature selection algorithms essentially have two basic criteria named as, time requirement and quality. The core idea of feature selection process is improve accuracy level of classifier, reduce dimensionality; speedup the clustering task etc., this paper mainly focuses on Comparison of various techniques and algorithms for feature selection process.**

Index Terms— **Classifiers, Feature Selection, Feature Clustering, Mutual information**.

I. INTRODUCTION

A. Data mining

Data mining is the route of ascertaining the interesting knowledge from hefty amounts of information repositories or data warehouses. Data mining tasks are specified by its functionalities that tasks are classified into two forms: 1. Descriptive mining tasks: Portray the general properties of the data. 2. Predictive mining tasks: Perform the implication on the current data order to craft prediction.

Data mining Functionalities are:

- Characterization and Discrimination
- Mining Frequent Patterns
- Association and Correlations
- Classification and Prediction
- Cluster Analysis
- Outlier Analysis
- Evolution Analysis

Out of these functionalities our paper focuses on the cluster Analysis.

B. Cluster Analysis

Clustering is the progression of grouping similar objects into one class. A cluster is an assembly of data objects that are similar to one another within the identical cluster and are dissimilar to the objects in other clusters. Document clustering (Text clustering) is closely related to the concept of data clustering. Document clustering is a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering.

C. Data Preprocessing

Data preprocessing is used to improve the efficiency and ease of the mining process. Whenever we want to extract some data from the data warehouse that data may be incomplete, inconsistent or contain noisy because data warehouse collect and store the data from various external resources.

Data preprocessing Techniques are:

- *Data cleaning*: Attempt to fill in missing values, smooth out noise, correct inconsistencies in the data. Cleaning techniques are binning, regression etc.,
- *Data Integration and Transformation*: Merging of data from multiple data sources, these sources may include multiple database, data cubes or flat files. Transformation is the process of consolidate the data into another form, it includes aggregation, generalization, normalization and attribute construction.
- *Data Reduction*: Techniques can be applied to obtain a reduced version of the

data set that is much smaller in quantity but maintains the integrity of original data, which contains following strategies: data cube aggregation, attribute subset selection dimensionality reduction etc.,

- *Concept hierarchy Generation*: Concept hierarchies can be used to condense the data by collection and replacing low level concept with high level concept.

The rest of the paper is organized as follows: in Section II, we present about the feature selection. In Section III, we describe about the classification. In Section IV, we briefly review the related work. In Section V, will introduce comparative study on various algorithms and its graph representation. Finally in section VI,we summarize our conclusion.

II. FEATURE SELECTION

Feature selection is similar to data preprocessing technique. It is an approach of identifying subset of features that are mostly related to target model. The main aim is to remove irrelevant and redundant features, it is also known as attribute subset selection. The purpose of feature selection is to increase the level of accuracy, condense dimensionality; shorter training time and enhances generalization by reducing over fitting. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points).

Steps in a Feature Selection Method:

- *Invention Procedure:* Produce candidate subset from original feature set.
- *Estimation Function:* Estimate the candidate subset.
- *Evaluation:* Compare with user defined threshold value.
- *Verification Method:* Test out whether the subset is valid.

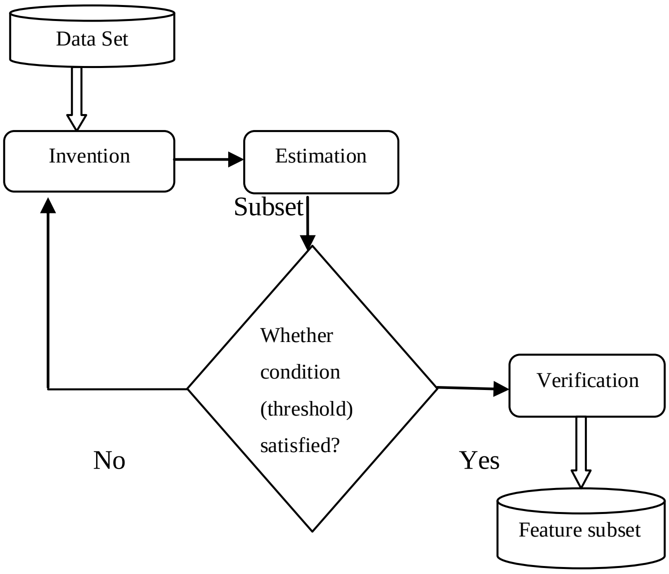


Fig : Steps involved in feature selection

III. CLASSIFICAION

Classification is the process of predicts categorical labels. It is used to classify the data based on the training set and the values in a classifying attribute. The basic classification techniques are Decision tree induction; Bayesian classification, and Rule-based classification etc., Classification task plays an important role in clustering process. Classification is performed via following two step processes: **1.** Model Construction**:** Describing a set of predetermined classes. Model is represented as classification rules, decision trees and mathematical formulae. **2.** Model Usage**:** It is used to estimate the accuracy of model. Accuracy rate is the percentage of test set samples that are correctly classified by the model.

Classification methods:

Bayesian Classification: Bayesian classifiers are statistical classifiers used to predict class membership probabilities. It is also known as naïve Bayesian classifier based on Bayes theorem. Compare to other classifiers it have the minimum error rate.

Decision tree Induction: Decision trees are constructed in a top-down recursive divide-and-conquer method. It consists of three algorithms such as ID3 (Iterative Dichotomiser), C4.5 (successor of ID3), CART (Classification and Regression Trees). The procedure employs an attribute selection measure such as gini index, information gain and gain ratio. Attribute selection measure [1] is used to separates the original data set (D) into individual classes.

Rule Based Classification: A rule-based classifiers uses a set of rules for classification task. This method effectively produces the subset of features using different heuristic techniques.

IV. RELATED WORK

A. Minimum-Spanning Tree

A minimum spanning tree (MST) [1] is an undirected, connected, weighted graph is a spanning tree of minimum weight. A tree is an acyclic graph. The idea is to start with an empty graph and try to add edges one at a time, the resulting graph is a subset of some minimum Spanning tree. Each graph has several spanning trees. This method is mainly used to make the appropriate feature subset clustering but it take time to construct the cluster.

Applications:

- Design of computer networks and Telecommunications networks
- Transportation networks, water supply networks, and electrical grids.
- Cluster analysis
- Constructing trees for broadcasting in computer networks.
- Image registration and segmentation
- Handwriting recognition of mathematical expression

B. Graph Clustering

Graph-theoretic clustering methods have been used in many applications. The general graph-theoretic [1] clustering is simple to compute a neighborhood graph of instances, and then delete any edge in the graph that is much longer/shorter (it's based on some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. The complete graph G reflects the correlations among all the target-relevant features.

C. Consistency Measure

Consistency measure focuses to locate the optimal subset of related feature for improve the overall accuracy of classification task and deduce the size of the dataset. This method [2], [3] based on inconsistency rate over the dataset for a given feature set. Apply the consistency measure to feature selection task, first they calculate the inconsistency rate IR(S). Inconsistency rate is less than user threshold value then the subset (S) is known as consistent. Consistency Measure use different Search Strategies such as Exhaustive, Complete, Heuristic, Probabilistic, and Hybrid. This method is monotonic, fast and Suitable for remove irrelevant and redundant features.

Exhaustive Search: Focus: Focus is one of the earliest algorithms within machine learning. It starts with an empty set and carries out breath-first search until it finds a minimal subset that predicts pure classes

Complete Search: ABB: ABB is Automatic Branch and Bound, extensions of Branch & Bound method. ABB algorithm having its bound set to the inconsistency rate of the original feature set.

Heuristic Search: Set Cover: Set Cover exploits the observation that the problem of finding a smallest set of consistent features is equivalent to 'covering' each pair of examples that have different class labels.

Probabilistic Search: LVF: Las Vegas Filter algorithm adopts the inconsistency rate as the evaluation measure. It generates feature subsets randomly with equal probability, and once a consistent feature subset is obtained that satisfies the threshold inconsistency rate.LVF is fast in reducing the number of features in the early stages and can produce optimal solutions.

Hybrid Search: QBB: QBB is hybrid of LVF and ABB. QBB (Quick Branch and Bound) is a two phase algorithm that runs LVF in the first phase and ABB in the second phase

D. Relief Algorithm

Relief is well known and good feature set estimator. Feature set estimators evaluate features individually. The fundamental idea of Relief algorithm [4], [5] is estimate the quality of subset of features by comparing the nearest features with the selected features. With nearest hit (H) from the same class and nearest miss (M) from the different class perform the evaluation function to estimate the quality of features. This method used to guiding the search process as well as selecting the most appropriate feature set. Relief estimates are better than usual statistical attribute estimates, like correlation or covariance because it consider attribute interrelationships. It is better to use a reduced set of features.

E.Mutual Information

Consecutive features are grouped into clusters, and replaces into single feature. The clustering process based on the nature of data. This paper shows the information [7] about feature grouping, feature clustering and functional modeling. Select the features by relevance estimation which is calculated using Mutual Information of two variables A and B, it is defined as the uncertainty reduction of B when A is known. In this method, first collect the possible candidate subset from the original data set, and then applies the forward search is used to choose the most relevant features. The process is then iterated until the discover the optimal feature subset.

F.Hierarchical clustering

Hierarchical clustering is a procedure of grouping data objects into a tree of clusters. It has two types: 1) Agglomerative approach is a bottom up approach; the clustering processes starts with each object forming a separate group and then merge these atomic group into larger clusters or group, until all the objects are in a single cluster. 2) Divisive approach is reverse process of agglomerative; it is top down approach, starts with all of objects in the same cluster. In each iteration, a clusters split up into smaller clusters, until a termination condition holds. Tree formations (dendrogram) symbolize the process of hierarchical clustering. AGENS [8], [9] is stands for AGlomerative NESting application of agglomerative hierarchical clustering method. It is single linkage approach initially places each object into a cluster of its own and the similarity between two clusters is determined by the resemblance of the closest pair of data points belonging to different clusters.

G. Feature selection methods

Evaluation functions are used to measure the goodness of the subset. Feature subset selection method is categorized into four types: Embedded, Filter, Wrapper, and Hybrid. Wrapper method is used to calculate the integrity [1] of the selected subset features by using predictive accuracy of machine learning algorithm which provides greatest accuracy of learning algorithms but it has more expensive. Filter is significant choice when the selected feature is very large [1], [11]. It is the independent of learning algorithm and has the low computational complexity. Hybrid method is the integration of filter and wrapper method is worn better [1] performance of learning algorithms.

V. COMPARISION OF VARIOUS ALGORITHMS

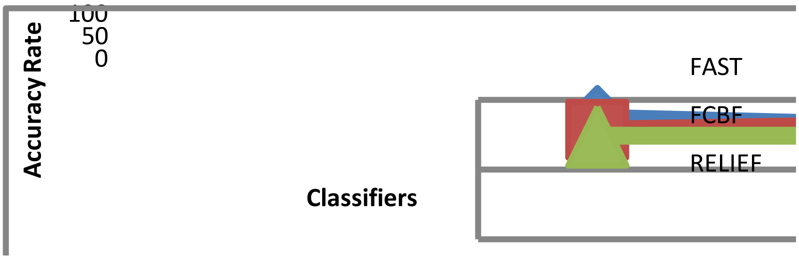
In this section, we present the comparison results in terms of

- Classification accuracy (Accuracy of the selected features)
- Runtime (time to obtain the feature subset),
- Proportion of selected features (ratio of the number of features selected by a feature selection algorithm).

.

Classification Accuracy:

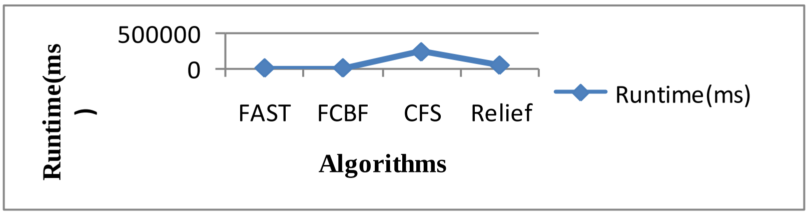
FAST Algorithm has better classification accuracy according to the four classifiers, such as Naive Bayes, C4.5, IB1, and RIPPER.



Graph: Classification Accuracy

Runtime:

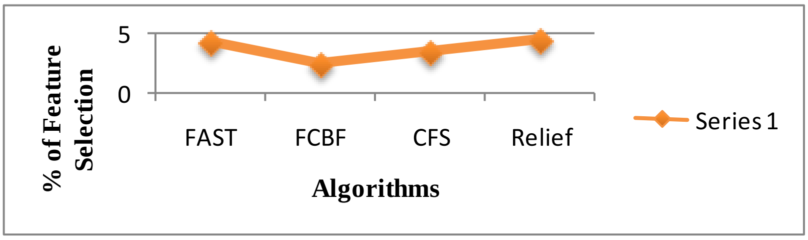
Runtime of FAST algorithm is much faster than other subset evaluation based algorithms, named as FAST, FCBF, CFS and Relief.



Graph: Runtimes

Proportion of selected Features:

FAST algorithm obtains best proportion of selected features compare to other algorithms, named as FAST, FCBF, CFS and Relief



Graph: Proportion of Feature Selection

Comparison of various algorithms and techniques are discussed as follows:

S.NO	**Techniques (or)Algorithms**	**Advantages**	**Disadvantages**
1.	FAST Algorithm	Improve the performance of classifiers	Required more time
2.	Consistency Measure	Fast, Remove noisy and irrelevant data	Unable to handle large volumes of data
3.	Wrapper Approach	Accuracy is high	Computational complexity is large
4.	Filter Approach	Suitable for very large features	Accuracy is not guaranteed
5.	Agglomerative linkage algorithm	Reduce Complexity	Decrease the Quality when dimensionality become high
6.	INTERACT Algorithm	Improve Accuracy	Only deal with irrelevant data
7.	Distributional clustering	Higher classification accuracy	Difficult to evaluation
8.	Relief Algorithm	Improve efficiency and Reduce Cost	Powerless to detect Redundant features

Tab: Comparison of various techniques and algorithms

VI. CONCLUSION

Feature selection method is an efficient way to improve the accuracy of classifiers, dimensionality reduction, removing both irrelevant and redundant data. In this paper, we have made a comparative study of various feature selection methods and algorithms. Discussions of those techniques are reviewed and the benefits and drawbacks of feature selection methods and algorithms are summarized.

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.

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 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 4 papers in reputed journals 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.