A Survey on Feature Selection Methods for High Dimensional Data

Swati V. Jadhav
M.E. Pursuing
Dept. of Computer Engineering
Shah & Anchor Kutchhi Engineering College, Mumbai
jadhav.swati30@gmai.com

Vishwakarma Pinki
Assistant Professor
Dept. of Computer Engineering
Shah & Anchor Kutchhi Engineering College, Mumbai
vishwakarmapp@yahoo.com

Abstract:-Feature selection is one of the most and important technique in preprocessing of high dimensional datasets. In order to remove irrelevant features and get relevant feature subset to achieve objectives of classification and clustering. This paper introduces overview of high dimensional data, their issues, concept of feature selection, process, various feature selection methods and their comparison
.

Keywords:- Feature selection is one of the most and important technique in preprocessing of high dimensional datasets. In order to remove irrelevant features and get relevant feature subset to achieve objectives of classification and clustering. This paper introduces overview of high dimensional data, their issues, concept of feature selection, process, various feature selection methods and their comparison.

*****

1. INTRODUCTION

With the rapid growth of computational biology and e-commerce applications, high dimensional data becomes very common. The mining of high dimensional data is an urgent problem in day today life. Data mining is the extraction of hidden predictive information from large database, is a powerful new technology to help companies focus on the most important information in their data warehouses. Data mining incorporated many techniques such as machine learning, pattern recognition, database and data warehouse systems, visualization, algorithms, high performance computing, and many application domains.

Another name for data mining is the knowledge discover process, it typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation and knowledge representation.

We present a multidimensional view of data mining. The major dimensions are data, knowledge, technologies, and applications.

Data mining functionalities are:

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

1.1 High Dimensional Data

The technologies present investigators with the task of extracting meaningful stastical and biological information from high dimensional data. A great deal of data from different domains such as medicine, finance, science is high dimensional.

Many objects can be represented with high dimensional such as speech signals, images, videos, text documents, hand writing letters and numbers. We often need to analyze large amount of data and process them. For e.g. need to identify person fingerprints, certain hidden patterns and images, to trace objects from videos. To complete these tasks, we develop the systems to process data. However due to high dimension of data, the system directly processing them may be very complicated and unstable so that it is infeasible.

1.1.1 Challenges in High Dimensional

Curse of Dimensionality: It is phenomena that arise when analyzing and organizing data in high dimensional spaces that do not occur in low dimensional such as three dimensional space in every day. Therefore, in order to process high dimensional data in the system dimensionality reduction becomes necessary [2]. Effect of High dimensionality on distance measures in Euclidian spaces: For any point in high dimensional space the expected gap between Euclidian distance to the closest neighbor and that to farthest point shrinks as the dimensionality grows.

Visualization: It is difficult to visualize and understand as it is high dimensional data.

1.2 Feature Selection

In machine learning and statistics feature selection also known as variable selection, attribute selection or variable subset selection. It is the process of detecting relevant features and removing irrelevant, redundant or noisy data [1].

1.2.1 Two Approaches for Feature Selection

- Individual Evaluation: The weight of an individual feature is assigned according to its degree of relevance.
- Subset Evaluation: candid feature subsets are constructed using search strategy.
-

1.2.2 Feature Selection process

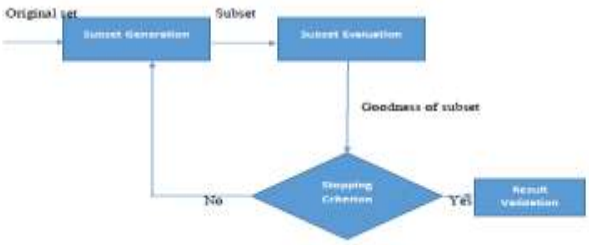

Figure 1. Four key steps for the feature selection process [9]

Subset generation is a heuristic search in which each state specifies a candid subset for evaluation in search space. In this process successor generation decides the search starting point, which influences the search direction and search organization is responsible for feature selection process with specific strategy such as sequential search, random search.

Evaluation criteria determine the goodness of the candid subset of features. The criteria can be of two types:

- Independent Criteria: without involving any mining algorithm it evaluates the goodness of features. The criteria namely distance measures, probability of error measures, consistency measures.

- Dependent Criteria: It involve predetermine mining algorithm for feature selection to select features based on the performance of mining algorithm applied to the selected subset of features.

Stopping Criterion is used to stop the selection process. There are some general stopping criteria:

- When the search completes

- Deletion or addition of features to the subset.

Result validation: Feature selection method must be validating by carrying out different tests and comparison with previously established results or comparisons with the result of competing methods using artificial datasets, real world datasets or both.

2. RELATED WORK

With respect to the filter feature selections, the application of cluster analysis has been demonstrated in this paper. They stated the FAST clustering based algorithm is effective and efficient. The algorithm works in two steps in the first step features are divided into clusters by using graph theoretic clustering methods. In the second step the most representative feature is strongly related to target classes is selected from each cluster to form a subset of features. The framework of proposed feature composed of the two connected components of irrelevant feature removal and redundant feature elimination. FAST algorithm uses minimum spanning tree based method to cluster features. They have experimented FCBF, Relief F, CFS, Consist, Focus-SF techniques on 35 different datasets and conclude that the FAST algorithm is effective than all others [1]. A new FR algorithm termed as class dependent density based feature elimination (CDFE) for high dimensional binary data sets. CDFE uses filtrapper approach to select a final subset. For data set having hundreds of thousands of features. Feature selection with FR algorithm is simple and computationally efficient but redundant information may not be removed. FSS algorithm analyses the data for redundancies but may become computationally impractical on high dimensional datasets. They address these problems by combining FR and FSS methods in the form of two stage feature selection algorithm. CDFE not only presents them with feature subset good in terms of classification but also relieves them from heavy computation. Two FSS algorithms are employed in second stage to test the two stage feature selection idea. Instead of using threshold value CDFE determines the final subset with the help of classifier [2]. The framework developed to perform feature selection for graph embedding in which a category of graph embedding method is cast as least squares regression problem. In contrast to filter methods, wrapper methods are application dependent. The embedded method encapsulates the feature selection into sparse regression method termed as LASSO. In this framework a binary feature selector is introduced to naturally handle the feature cardinality in the least squares formulation. The resultant integral programming problem is then relaxed into a convex quadratic ally constraint quadratic program (QCQP) learning problem which can be efficiently solved via a sequence accelerated proximal gradient (AGP) methods. The proposed framework is applied to several is embedding learning problems including supervised, unsupervised and semi supervised graph embedding. The graph embedding suffers from two weakness that is it is hard to interpret the resultant features when using all dimensions for embedding and the original data inevitably contains noisy feature could make graph embedding unreliable and noisy [3]. To find nearest neighbor matching, the two algorithms are most efficient the randomized k-d forest and a new algorithm the priority search k-means tree. Also proposed new algorithm for matching binary features by searching multiple hierarchical

clustering trees. They show that the optimal nearest neighbor algorithm and its parameter depend on the data set characteristics and describe an automated configuration procedure for finding the best algorithm to search a particular data set. They have been released as an open source library called fast library for approximate nearest neighbors (FLANN) into openCV and is now one of the most popular libraries for nearest neighbor matching [4]. They presented novel concept predominant correlation and propose a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis [5]. They presented filter-wrapper hybrid method (FWHM) to optimize the efficiency of feature selection. FWHM is divided into two phase, which orders these features according to reasonable criterion at first, then selected best features based on final criterion. These experiments on benchmark model and engineering model prove that FWHM has better performance both in accuracy and efficiency more than conventional methods [6]. A new hybrid algorithm that uses boosting and incorporates some of the features of wrapper methods into a fast filter method. For feature selection results are reported on six world datasets and hybrid method is much faster and scales well to datasets with thousands of features [7]. The definitions for irrelevance and for two degrees of relevance incorporated in this paper. The features selected should depend not only on the features and the target concept but also on the induction algorithm. A method is described for feature subset selection using cross validation that is applicable to any induction algorithm and experiments conducted with ID3 and C4.5 on artificial and real datasets [8].

3. FEATURE SELECTION METHODS

Many feature selection methods have been proposed in Figure 2.

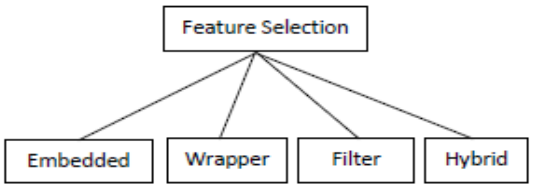

Figure 2. Categories of Feature Selection

The above figure shows the categorization of selection methods studied in machine learning applications the Embedded, Wrapper, Filter and Hybrid.

3.1 Embedded Method

The embedded method incorporates feature selection process as a part of training process. They are specific to learning algorithms. It also captures feature dependencies. It uses independent criteria to decide optimal subsets for cardinality. Examples are classification trees, random forests and methods based on regularization technique [1], [9].

Advantage:

- Computationally less than wrapper method.

Disadvantage:

- It is Specific to learning machine.

3.2 Wrapper Method

This method uses the predictive accuracy of a predetermined algorithm to determine the goodness of the selected subsets. Evaluation uses the criteria related to classification algorithm [1], [9].

Advantage:

- The accuracy of learning algorithm is high.

Disadvantages:

- This method is computationally expensive.
- High risk of overfitting on small training sets.

3.3 Filter Method

This method is independent of learning algorithm. The filter method works well when the number of features are large [1], [9].

Advantages:

- Easily scale to very high dimensional datasets.
- Computationally simple and fast.

Disadvantages:

- They ignore the interaction with classifier.
- They are often univariate or low-variate.

3.4 Hybrid Method

The hybrid method is combination of filter and wrapper methods. It mainly focuses on combining filter and wrapper methods to achieve best suitable performance with a particular learning algorithm with similar time complexity of filter methods. The wrapper methods are computationally expensive and tend to fit on small training set [1].

4. CONCLUSION

Thus, we conclude that the different techniques used for feature selection are found. The high dimensional data, feature selection process is studied and comparison of various feature selection methods are shown by following table.

Table 1: Comparison of different feature selection methods

Algorithm	Advantage	Disadvantage
Embedded Approach	Less Computation	Specific to learning machine
Wrapper Approach	High Accuracy	Computationally

		expensive
Filter Approach	Suitable for very large features	Accuracy is not guaranteed
Relief Algorithm	Improve efficiency, reduces cost	Powerless to detect redundant features
FAST Algorithm	Efficient, effective	Takes more time

5. REFERENCES

[1] Qinbao Song, Jingjie Ni, Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge & Data Engineering, vol.25, no. 1, pp. 1-14, Jan. 2013

[2] Kashif Javed, Haroon A. Babri, Maureen Saeed, "Feature Selection Based on Class-Dependent Densities for High-Dimensional Binary Data", IEEE Transactions on Knowledge & Data Engineering, vol.24, no. 3, pp. 465-477, March 2012

[3] Marcus Chen, Ivor W. Tsang, Mingkui Tan, Tat Jen Cham, "A Unified Feature Selection Framework for Graph Embedding on High Dimensional Data", IEEE Transactions on Knowledge & Data Engineering, vol.27, no. 6, pp. 1465-1477, June 2015

[4] Marius Muja and David G. Lowe: *"Scalable Nearest Neighbor Algorithms for High Dimensional Data"***.** Pattern Analysis and Machine Intelligence (PAMI), Vol. 36, 2014

[5] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Learning, 2003.

[6] Hu Min, Wu Fangfang, "Filter-Wrapper Hybrid Method on Feature Selection", GCIS, 2010, 2010 Second WRI Global Congress on Intelligent Systems, 2010 Second WRI Global Congress on Intelligent Systems 2010, pp. 98-101, doi:10.1109/GCIS.2010.235

[7] Das S, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection," Proc. 18th Int'l Conf. Machine Learning, pp. 74-81, 2001.

[8] G.H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," Proc. 11th Int'l Conf. Machine Learning, pp. 121-129, 1994.

[9] Vipin Kumar, Sonajharia Minz: Feature Selection: A literature Review**.** Smart CR 4(3): 211-229 (2014)

[10] Molina L.C., Belanche L. and Nebot A., Feature selection algorithms: A survey and experimental evaluation, in Proc. IEEE Int. Conf. Data Mining, pp 306-313, 2002