International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 03, Issue 09, September 2014 ISSN 2319 - 4847

Literature Review on Feature Subset Selection Techniques

Mr. Swapnil R Kumbhar[1], Mr.Suhel S Mulla[2]

[1]Persuing M.E.in Computer Science & Engg. at Annasaheb Dange College of Engineering & Technology,Ashta,Sangli.

[2] Persuing M.Tech.in Electorincs (Digital System).at College Of Engineering, Pune

ABSTRACT

Feature selection process involves identifying a subset of features that produces similar results as the original entire set of features. Feature selection, also known as attribute subset selection is similar used for dimensionality reduction; improve the classifier accuracy; removing irrelevant and redundant features. It is a research area of great practical significance and has been developed and evolved to answer the challenges due to data of increasingly high dimensionality. The main task in feature subset selection is removal of irrelevant features and eliminating redundant feature in accordance to getting quality feature subset. The feature selection algorithm is constructed with the consideration of efficiency and effectiveness point of view. The efficiency related to time required to find the subset of features. The effectiveness concerns with quality of subset produced. The objective of feature selections are improving the prediction performance of the predictors, providing accurate and immediate result. This paper review summarizes the various techniques of feature subset selection

Keywords: Feature Selection, Redundancy, Relevance, Clustering, FCBF, ReliefF.

1. INTRODUCTION

Data Mining

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful
and ultimately understandable patterns in data.

Feature Selection

In machine learning and statistics, feature selection also known as attribute selection or variable subset selection. It is the process of selecting a subset of relevant features for model construction. The logic behind using a feature subset selection technique is that the data contains many redundant or irrelevant features, Redundant features are those which provide duplicate information, and irrelevant features provide no useful information in any context. 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. The use of feature selection in analyzing DNA microarrays, where there are many thousands of features, and a few tens to hundreds of samples. Feature selection techniques provide three main benefits when constructing predictive models: improved model interpretability, training time is shorter and enhanced by reducing over fitting.

Categories of Feature Subset Selection Algorithm

The feature subset selection algorithms are generally categories into main four categories as Filter, Wrapper, Embedded and Hybrid Method.

Wrapper Method

Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model.

Filter Method

Filter methods use a proxy measure instead of the error rate to score a feature subset. This measure is chosen to be fast to compute, whilst still capturing the usefulness of the feature set. Common measures include the Mutual Information, Pearson product-moment correlation coefficient, and inter/intra class distance. Filters are usually less computationally intensive than wrappers, but they produce feature set which is not tuned to a specific type of predictive model.

Embedded Method

Embedded method is a catch-all group of techniques which perform feature selection as part of the model construction process. One other popular approach is the Recursive Feature Elimination algorithm, commonly used with Support

Vector Machines to repeatedly construct a model and remove features with low weights. These approaches tend to be between filters and wrappers in terms of computational complexity. The embedded method incorporate feature selections a part of training process and are usually specific to given learning algorithm hence it is efficient than other methods.

Hybrid Method

Hybrid method is combination of filter and wrapper methods. It uses a filter method to reduce the search space that will be considered by subsequent wrapper. They mainly focus on combination of filter and wrapper methods in accordance to
achieve best performance with particular learning algorithm with similar time complexity of the filter methods.

2. BACKGROUND

This paper mainly focuses on feature selection technique which is an important research area in data mining. Feature subset selection can tremendously affect on performance of predictive model, the feature selection algorithm is constructed with the consideration of efficiency and effectiveness point of view. The efficiency related to time required to find the subset of features. The effectiveness concerns with quality of subset produced. A database consists of several dimensions or attributes. Finding the feature subset that produces similar result as that of original entire set of features is key and challenging task. Several researchers present different techniques for features subset selection.

3. RELATED WORK

1. Efficient Feature Selection via Analysis of Relevance and Redundancy

This paper[1] propose a new framework of feature selection which avoids implicitly handling feature redundancy and turns to efficient elimination of redundant features via explicitly handling feature redundancy. Relevance definitions divide features into strongly relevant features, weakly relevant features and irrelevant features; redundancy definition divides weakly relevant features into redundant and non redundant ones. Thus produces the final subset. Its advantage is decoupling relevance and redundancy analysis and allows a both efficient and effective way in finding a subset that approximates an efficient subset. It uses C-correlation for relevance analysis and both C & F correlations for redundancy analysis.

2. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution

This correlation based filter approach is making use of symmetric uncertainty method. Symmetric uncertainty measure how much a feature is related to another feature. This involves two steps: first how to decide whether a feature is relevant to the class or not; and second how to decide whether such a relevant feature is redundant or not when considering it with other relevant features. The solution to the first question can be using a user- defined threshold SU value, as the method used by many other feature weighting algorithms (e.g., Relief). The answer to the second question is more complicated because it may involve analysis of pair wise correlations between all features (named F-correlation), which results in a time complexity of O(N2) associated with the number of features N for most existing algorithms. To solve this problem, FCBF algorithm is proposed. FCBF means Fast Correlation-Based Filter Solution [2]. This algorithm involves two steps. First step is select relevant features and arrange them in descending order according to the correlation value. Second step is remove redundant features and only keeps predominant ones. The linear correlation & information Theory is used. It propose new algorithm for good feature selection with less time complexity.

3. Feature Selection through Clustering

The author [3] introduces an algorithm for feature selection that clusters attributes using a special metric of Barthelemy-Montjardet distance and then uses a hierarchical clustering for feature selection. Hierarchical algorithm generates clusters that are placed in cluster tree which is commonly known as dendrogram. The dendrogram of resulting cluster hierarchy to choose the most representative attributes. Clustering's are obtained by extracting those clusters that are situated at given height in this tree.

4. Features Election for High-Dimensional data a Pearson Redundancy Based Filter

The author [4] introduces an algorithm for filtering information based on the Pearson χ^2 test approach has been implemented and tested on feature selection. This is useful for high dimensional data where no sample set is large. This test is frequently used in biomedical data analysis and used only for nominal (discretized) features. This algorithm has only one parameter, statistical confidence level that two distributions are identical. Empirical comparisons with four other features selection algorithms (FCBF, CorrSF, ReliefF and ConnSF) are done to find quality of feature selected. This algorithm work fine with the linear SVM classifier. PRBF (Pearson's Redundancy Based Filter) algorithm is important similar to other correlation-based filters and much lower than ReliefF

5. Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection

In this paper [5], author describes the advantages and disadvantages of filter and wrapper methods for feature selection and proposes a new hybrid algorithm that uses boosting and incorporates some of the features of wrapper methods into a fast filter method for feature selection. Empirical results are reported on six real-world datasets from the UCI

repository, showing that hybrid algorithm is competitive with wrapper methods while being much faster, and scales well to datasets with thousands of features.

6. A Feature Set Measure Based on Relief

The author [6] presents Relief is a well known and good feature set estimator. Feature selection methods try to find a subset of the available features to improve the application of a learning algorithm. Many methods are based on searching a feature set that optimizes some evaluation function. Feature set estimators evaluate features individually. On artificial datasets, the proposed feature set measure based on relief can be better than the wrapper approach to guide a common feature selection search process. This method is compared with a consistency measure, and the highly reputed wrapper approach in this paper. The main disadvantage of this system is, it measure low accuracy of the search process.

4. CONCLUSION

This paper provides a literature review on different types of existing Feature Selection Techniques. Feature selection is a term commonly used in data mining to describe the tools available for reducing inputs data to a manageable size for processing and analysis. Feature subset selection implies not only dimensionality reduction but it gives way towards improving classifier accuracy which is helpful in many applications as well. Feature selection techniques have wide variety of applications in data mining, digital image processing, and apparent need in many bioinformatics applications. The study of different techniques leads to conclusion that there is need of effective and efficient method for handling redundancy in high dimensional datasets.

REFERENCES

[1] L. Yu and H. Liu, “Efficient Feature Selection via Analysis of Relevance and Redundancy” J.Machine Learning Research, vol. 10,no. 5, pp. 1205-1224, 2004

[2] L. Yu and H. Liu, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution”Proc.20th Intl Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.

[3] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, “On Feature Selection through Clustering” Proc. IEEE Fifth Intl Conf. Data Mining, pp. 581-584, 2005.

[4] J. Biesiada and W. Duch, “Features Election for High-Dimensional data a Pearson Redundancy Based Filter” Advances in Soft Computing, vol. 45, pp. 242-249, 2008.

[5] S. Das, “Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection” Proc. 18thIntl Conf. Machine Learning, pp. 74-81, 2001

[6] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, “A Feature Set Measure Based on Relief” Proc. Fifth Intl Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

AUTHOR

Mr.Swapnil R Kumbhar, Persuing M.E.in CSE at Annasaheb Dange College Of Engineering & Technology,Ashta,Sangli. My Research interests are Data Mining, Information Communication Technology and Information Security.

Mr.Suhel S Mulla, Persuing M.Tech.in Electorincs (Digital System).at College Of Engineering, Pune. His research interests are Embedded System, Fuzzy Logic,Digital Design.