ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 9, September 2015

A Detailed Analysis of Feature Subset Selection Techniques for Classification Problems

L.Jegatha Deborah, P.Vijayakumar

[1,2] Department of Computer Science and Engineering, University College of Engineering Tindivanam, Melpakkam, India- 604 001

E-mail: vijibond2000@gmail.com, blessedjeny@gmail.com

Abstract

Data preprocessing is the most preliminary step to be done for high dimensional data. Enormous challenging issues are predominant in the preprocessing stages especially in the data dimensions. The high dimensional dataset in real time applications poses to be a very severe problem for data mining applications. This paper analyzes in detail the several factors of high dimension data problems. Moreover, the detailed analysis on feature subset selection techniques is the core interest of this paper. A detailed survey had been conducted in the feature subset selection techniques for classification algorithms. The final outcome of this survey is that several research gaps had been identified in order to solve the issues occurring in the classification algorithms sue to high dimension dataset.

Keywords : data preprocessing, high dimension data, classification algorithm, feature subset selection, data mining.

Introduction:

Data preprocessing is an important and the most preliminary task to be accomplished in data mining.[18] Data mining task such as classification and clustering poses several challenges while handling high-dimensional data [3,4]. So the dimensionality of the data can be reduced by dimensionality reduction technique. Dimensionality reduction can be performed using two approaches namely feature extraction and feature selection [6,8-9]. Feature extraction selects new features from the combination of new features. Feature extraction algorithms include Principle Component Analysis(PCA), Linear Discriminant Analysis(LDA) and Canonical Correlation Analysis(CCA) [19]. Feature selection chooses a subset of valuable features from a larger dataset. Feature selection aims to remove the irrelevant and redundant features from the dataset. Feature selection process reduces the number of features which makes

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

the data analysis easier and minimizes storage requirements. The four major steps of feature selection are subset generation, subset evaluation, stopping criteria and result validation.[18]

The feature selection algorithms can be categorized as supervised, semi-supervised and unsupervised feature selection algorithms based on labelling information. In supervised feature selection , label information is used to select distinguished features. But labelling data is a time-consuming and costly process [12-14]. Semi-supervised feature selection have a mixture of smaller labelled data and larger unlabelled data. This is called "small-labelled sample problem". Unsupervised feature selection does not have the label information and hence it is harder to find the discriminative features. Supervised feature selection algorithms are further classified as filter, wrapper, embedded and hybrid methods.[17] The filter method selects the subset of features without the help of learning algorithm. This method works well for large number of features and their computational complexity is low. The wrapper method utilizes the predictive accuracy of predetermined learning algorithm to determine the value of the feature subsets selected. In this method the accuracy of learning algorithms is high when compared to filter method but the computational complexity is also high. The hybrid method combines both the filter and wrapper methods. First the filter method is used to select significant feature subsets and then the wrapper method selects the most promising features among the significant feature subsets. In the embedded method the feature selection process is included in the classifier itself. The various fields of research under feature selection includes statistical pattern recognition, machine learning, data mining and statistics. [16].

The major objective of this paper is to discuss the different feature subset algorithms with respect to efficiency, effectiveness, subset size of the selected features and the classification accuracy. This paper is organized as, section 2 provides a detailed analysis on the different feature subset selection algorithms. The tabular format of the survey is also depicted in the same section. Section 3 gives the concluding remarks of this analysis work.

2. Study and Analysis of Feature Subset Selection Algorithms

2.1 A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data.[1]

Qinbao Song et.al[2013] proposed an algorithm called fast clustering based feature selection algorithm (FAST). This algorithm works in two phases, where the first phase involves the removal of irrelevant features and in the second phase it removes the redundant features by selecting the most promising features from different clusters. The different clusters are formed by constructing minimum

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 9, September 2015

spanning tree and then partitioning the tree. This algorithm is very efficient and produces smaller subset of features when compared to other feature selection algorithms.

2.2 Feature selection based on class-dependent densities for high-dimensional binary data. [2]

Kashif Javed et.al[2012] proposed a feature ranking algorithm called class-dependent density-based feature elimination (CPDE) for binary data sets. This algorithm uses a measure called diff-criterion to calculate the relevance among the features. This measure is used to assign weight to the features for feature ranking and the algorithm uses a classifier to get the final subset. The framework uses two-stage feature selection algorithms by combining feature ranking with feature subset selection algorithms. The first stage gives the reduced initial feature subset with good classification accuracy and second stage selects the final subset with the help of two feature subset selection algorithms. However the framework does not address the effectiveness and efficiency of their proposed algorithm.

2.3 Feature Subset Selection and Ranking for Data dimensionality Reduction.[5]

Hua-Liang Wei et.al[2005] proposed Forward Orthogonal Search by Maximizing the Overall Dependency algorithm(FOS-MOD) for feature selection and ranking. In their algorithm, the features are selected one by one with the criteria that the selected features must represent the characteristics of the overall features. To find the dependency between the features squarred correlation function is used. The selected features are ranked based on the sum of error reduction ratio. The squarred correlation function addresses only the linear dependency between the features. The algorithm produces efficient and effective feature subsets but failed to address the subset size of the selected features.

2.4 Unsupervised feature selection using feature similarity.[7]

Pabitra Mitra et.al[2002] proposed an unsupervised feature selection algorithm to measure the similarity between two features for removing the redundant features. They developed a similarity measure called maximum information compression index to measure the similarity between two features in order to select feature relevance for feature selection. This measure is a linear dependency measure which is used to reduce the redundant features. The authors addressed the efficiency of the algorithm but the effectiveness and the subset size of the selected features are not addressed.

2.5 Feature Selection Via Discretization.[10]

Feature selection is a dimensionality reduction technique to remove the irrelevant and redundant attributes from the data set. The authors developed the feature selection with a help of discretization technique. The algorithm used for discretization is chi square which

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 9, September 2015

converts the numerical attributes to discrete and also removes the irrelevant and redundant features .The authors have done the analysis on three real world data sets such as Iris data, Breast cancer data, and Heart disease data. However they failed to address the effectiveness and subset size of selected features in their work.

2.6 Text clustering with feature selection using statistical data.[11]

Yanjun Li et.al[2008] contributed a supervised feature selection algorithm called CHIR for text clustering. This algorithm uses a statistical measure called chi2 and also it measures the dependency between the term and the class category to be positive or negative. CHIR algorithm chooses the features which have very effective positive dependency to the corresponding category. In addition to that the author have found a new text clustering algorithm called Text Clustering with Feature Selection(TCFS). TCFS uses CHIR iteratively for selection of relevant terms and then it performs the clustering process based on the selected terms. The work focused on improving the clustering accuracy but failed to address the subset size of the terms selected.

2.7 Efficient semi-supervised feature selection: constraint, relevance, and redundancy.[15]

Khalid Benabdeslem et.al[2014] proposed an architecture which encompasses constraint selection, feature relevance and redundancy analysis for semi-supervised feature selection. Instead of class labels which give the prior knowledge about the data, we go for pair wise constraints such as must-link and cannot-link constraints to divide the features into different subsets. The constraint selection is done by measuring the coherence between the two constraints. Constraint laplacian score is used to select the relevant features Redundancy analysis is performed to eliminate the redundant features from the features selected based on maximum spanning tree method. The algorithm proposed here focuses the efficiency and not the effectiveness and subset size of the feature selection.

S. No	Technique /Algorithm	Supervised/ Unsupervised/ Semi-supervised	Merits	Demerits
1	Fast clustering based feature selection algorithm(FAST)	Supervised feature selection	Efficient and produces smaller subset of features	Subset size
2	Class Dependent Density based Feature	Supervised feature selection	Reduction in subset size and	Efficiency and effectiveness

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 9, September 2015

	Elimination(CDFE)		Classification accuracy	of algorithm
3	Forward Orthogonal Search by Maximizing the Overall Dependency algorithm(FOS-MOD)	Unsupervised feature selection	Efficient and ranks the feature based on sum of error reduction ratio	Does not address Subset size of features
4	Maximal Information Compression index measure	Unsupervised feature selection	Reduces computational time and effective	Does not address Subset size of features
5	Chi2 for discretization	Supervised feature selection	Efficient	Does not address the effectiveness and subset size of selected features.
6	CHIR and Text Clustering with feature selection(TCFS)	Supervised feature selection	Improves the clustering accuracy	Does not address Subset size of terms selected
7	Constrained Semi-Supervised Feature Selection with Redundancy Elimination(CSFSR)	Semi-supervised feature selection	Effective and produces smaller subset of features	Efficiency

Conclusion

This paper analyzes in detail the various feature selection techniques used for classification algorithms. The major objective of this technique is to reduce the number of dataset available for data processing. The various research papers are identified for their merits and demerits. The future work of this paper is to develop a new feature subset selection technique for classification algorithm especially for dynamic and real time dataset.

References

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

2. Kashif Javed,Haroon..A.babri,mehreen saeed, "Feature selection based on class-

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 9, September 2015

dependent densities for high-dimensional binary data",IEEE Transactions on Knowledge & Data Engineering 2012 vol.24 Issue No.03 - March 2012.

3. Pradipta Maji, "A Rough Hypercuboid Approach for Feature Selection in Approximation Spaces", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.

4. Neil Mac Parthala´ in, Qiang Shen, and Richard Jensen, "A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 3, March 2010.

5. Hua-Liang wei and Stephen A.Billings, "Feature Subset Selection and Ranking for Data dimensionality Reduction",IEEE Transactions on Pattern Analysis and Machine intelligence,VOL.29,No.1,January 2007.

6. Zechao Li, Jing Liu, Yi Yang, Xiaofang Zhou, and Hanqing Lu," Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 9, September 2014.

7. Pabitra Mitra, C.A Murthy and Sankar K.Pal,"Ünsupervised Feature Selection using feature Similarity",IEEE Transactions on Pattern Analysis and Machine Intelligence,VOL.24,No.3,March 2002.

8. Mark A. Hall and Geoffrey Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", IEEE transactions on knowledge and data engineering, vol.15, no. 6, november/december 2003.

9. Patrenahalli M. Narendra And Keinosuke Fukunaga, "A Branch and Bound Algorithm for Feature Subset Selection", IEEE Transactions On Computers, Vol. C-26, No. 9, September 1977.

10. Huan Liu, Member, IEEE, and Rudy Setiono, "Feature Selection via Discretization", IEEE Transactions On Knowledge And Data Engineering, Vol. 9, No. 4, July/August 1997.

11. Yanjun Li, Congnan Luo, and Soon M. Chung, "Text Clustering with Feature Selection by Using Statistical Data", IEEE Transactions On Knowledge And Data Engineering, Vol. Xx, No. Yy, 2008.

12. Hyunjin Yoon, Kiyoung Yang, and Cyrus Shahabi," Feature Subset Selection and Feature Ranking for Multivariate Time Series" , The IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 9, September 2005.

13. Jialei Wang, Peilin Zhao, Steven C.H. Hoi, and Rong Jin, "Online Feature Selection and Its Applications", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 3, March 2014.

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 9, September 2015

14. Jiliang Tang and Huan Liu, "An Unsupervised Feature Selection Framework for Social Media Data" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 12,December 2014.

15. Khalid Benabdeslem and Mohammed Hindawi, "Ëfficient Semi-supervised Feature Selection: Constraint, Relevance and Redundancy" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 5, May 2014.

16. 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.

17. Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97, 273–324.