

Individual Evaluation-Based Attribute Diminution Algorithm For High Dimensional Data

Author: G. Abinaya¹, P.Sangeetha²

Affiliation: ¹Department of CSE, P.A College of Engineering and Technology, Coimbatore, Tamilnadu

² Department of CSE, P.A College of Engineering and Technology, Coimbatore, Tamilnadu

ABSTRACT

Individual Evaluation-Based Attribute Diminution algorithm involves finding a reduced set with the essential attribute which produces as the original set of attribute. The attribute diminution is performed by removing the irrelevant and redundant attributes. The Individual Evaluation-Based Attribute Diminution algorithm (IEAD) removes the irrelevancy by identifying relevant attribute to the target, and eliminating the rest and finally formed as weight graph. The graph is converted to MST using Boruvka's algorithm. From that resultant set the representative attributes are selected by reducing MST into forest using Symmetric Uncertainty measures. The image, microarray and text dataset are taken as input to the IEAD algorithm and reduced set given as the input to classifiers called the Naïve Bayes and the decision tree and compared with original set of attributes. The IEAD produces smaller subsets of attributes yet increases the performances.

Keywords

Attribute diminution, Individual Evaluation, unsupervised method.

1.INTRODUCTION

Attribute selection is the important as size of dataset grows with attributes. major approaches involved in attribute selection for clustering. The first is a filter approach which assesses and selects attributes without any explicit using any clustering algorithm. The second is a wrapper approach, uses a clustering algorithm as an evaluator for finding attribute subsets and may use different search strategies for choosing the subsets. The final approach tries to combine clustering and attribute selection by adding the attribute selection strategy into the clustering algorithm. Any attribute weighting scheme can be used into an attribute selection method by simply applying a weight threshold to the computed attribute weights.

Irrelevant attributes, along with redundant attributes, severely affects the performance. so, attribute subset diminution method identify and eliminate irrelevancy and redundancy information.

By having that problem, we given a balanced algorithm which can handle both irrelevant and redundant attributes, and obtain a better attribute subset.

Relevant attributes are who's which have correlation with target concept necessary in resultant reduced set. Thus, notions of attribute redundancy and attribute relevance are normally in terms of attribute correlation and attribute-target concept correlation. The Symmetric uncertainty (SU) is calculated using the normalized mutual information to the entropies of attribute values.

We can have attribute diminution framework of the two Combined Sub-components of irrelevant attribute and surplus attribute elimination. The first component obtains attributes relevant to the target concept and removes irrelevant attributes, and the next removes surplus attributes from relevant ones via choosing representatives attribute from different clusters. Individual Evaluation –Based Attribute Diminution Algorithm (IEAD). IEAD eliminates both irrelevant and redundant attributes. The IEAD algorithm works as follows. In the first step, irrelevant attributes are eliminated. As a next step, proceeding with resulting graph is difficult to process so convert into Tree using Boruvka's Algorithm. And the last step is to eliminate the redundant attribute by finding forest of tree attributes necessary to target classes is selected. Then attributes in different clusters are independent and same clusters are dependent.

The paper is organized as mentioned below, Section I defines need of Attribute selection and introduces their usage, Section II presents existing work, Section III presents proposed work. Section IV presents results with different data and V at last, states the possible follow-ups to this work draws the conclusions.

2. RELATED WORK

Numerous Irrelevant and Redundant elimination schemes have been proposed previously. Those schemes either eliminate irrelevant or redundant data. Our IEAD eliminates both the redundant and irrelevant data with better efficiency. FCBF [2], CMIM [3], and CFS [4] are algorithms which removes redundant attributes alone. In attribute selection Relief algorithm will provides the ranking to each attribute to

differentiae attribute related to target. It will remove only repetitive attributes alone. [6]. Relief-F [7],[8] extends Relief, scheme removes irrelevant data alone. But it cannot rectify redundant attributes.

While, both of incomplete data such as irrelevant attributes, redundant attributes affect the mining of related answers to the queries. So it should be eradicated [16], [5], [15],[22],[10].FAST [1], does not eliminate irrelevant and redundant data effectively. Then, Fast Correlated Based Filter ([2], [9]) uses the filter method for eliminating repetitive and not relevant attributes without considering correlation factor of each attribute. Correlation Feature Selection [4] measures the relevancy by correlation factors with respect to target though it is not effective in removing unrelated attributes. Conditional Mutual Information Maximization [3] criterion that does select a feature which adds more information to MI.

The methods exist for clustering such as partitioning method, partitions the grammatical words into different clusters. The user should provide the k-value in advance [10] [11]. Hierarchical methods BIRCH starts partitioning attributes hierarchically by forming tree structures, suitable only for spherical shapes Butterworth et al. [13]. It is deployed for word selection and text classification ([10], [11], [12], [16], [17],[18]and [19]). Agglomerative is high computational cost word clusters. Hulle [14] proposed a hybrid feature subset selection algorithm combines both filter and wrapper methods.

Density-based clustering has three representative methods, namely, DBSCAN [23] randomly choose the unvisited node and based on no. of objects together finds the core objects recursively those parameter settings are difficult, OPTICS does not produce data clustering directly, and DENCLUE includes noisy and incomplete data. Grid based clustering embed the space needed to partition into the cells. STING [20] explores statistical information stored in the grid cells, produces lower quality of clusters. CLIQUE [21] combines both grid- and density-based approach for clustering data space with high-dimension. The traditional clustering methods are based on the distance method while, not suitable for clustering data described with more than many attributes called high dimension data. Bi-Clustering is used for clustering gene expression type data struggles a lot for processing more data [24].

3. INDIVIDUAL EVALUATION-BASED ATTRIBUTE DIMINUTION ALGORITHM

Attribute selection is important processing in current environment as the dimensions grow. The attribute diminution is important in attribute selection. The attribute diminution can be done by removing both irrelevant and redundant attribute. There are many algorithms which eliminate either irrelevant or redundant alone. We proposed

an IEAD algorithm which eliminates simultaneously. There exist three sub works to carry out the task.

1. Irrelevant data is eliminated from original data.

2. weighted graph is converted to MST to reduce the processing of heavily dense data using Boruvka's Algorithm.

3. elimination of redundant data.

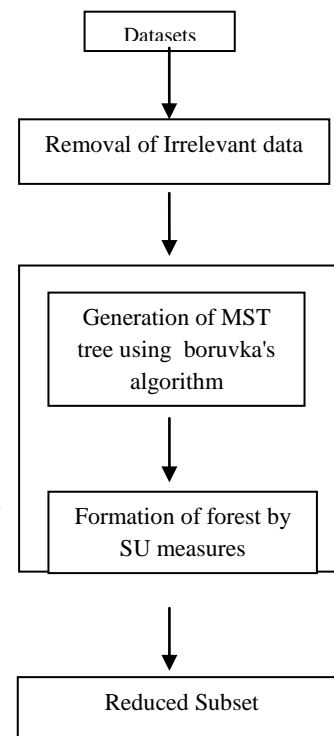


Fig 1: IEAD Framework

The eliminating the irrelevant attribute and the relevant attributes (see Figure 1). And then converting the graph into MST using Boruvka's algorithm. Since processing is difficult. Then from the MST the redundant attributes are eliminated.

Algorithm 1

Input: IEAD (*dataset, threshold*)

Part1: Irrelevant data removal

1. for(*i=0; i<n, 1++*)
2. if($P(F_i) < 0$) then
3. remove F_i .
4. end if
5. end for

Part2: Redundant data elimination

Function 1:

//generate Minimum Spanning Tree

Input: A connected weighted graph G

Output: T is MST of G.

1. Initialize at first forest T to be a set of one-vertex trees, each vertex of the graph.
2. While $T \leq 1$ component:
3. for each component C of T:

4. Begin with an empty set of edges S
5. for each vertex v in C:
6. Find the cheapest edge from v to a vertex beyond of C, and add it to S
7. Add the cheapest edge in S to T

Function 2:

//Finding Forest

1. Forest=T(minimum spanning tree)
2. for each edge of T
3. $SU(F_i, F_j) < SU(F_i, C)$ or $SU(F_j, C)$
4. Remove the edge $SU(F_i, F_j)$
5. Form new forest
6. End for

Output: Reduced subset S

The Reduced subset obtained as output of algorithm 1. It is stored in repository. Afterwards when queries are made, the proposed system uses answers the query quickly.

4. RESULTS

The environment in which the prototype application developed is JSE (Java Standard Edition) 6.0, Eclipse IDE that run in Windows 7 OS. PC with 2.9x GHz processor and 2 GB RAM is used. When doing an evaluation of a proposed framework, the performances are analyzed based on a metrics such as Data size and processing speed of algorithm. The numbers of attributes used in IEAD has attributes vary from 5,000 to 10,000 sectors.

TABLE I. Dataset for processing of IEAD

Sl.no	Original data	Relevant data	Non-redundant data
1	100	80	50
2	150	120	80
3	130	100	60
4	160	130	100

The data sets tested from the domains are as follows: Image, and Microarray data of UCI repository. The size of data taken as input is reduced by removing the redundant and irrelevant data as shown in above TABLE I. This TABLE I shows the proof of reduced by data counting the no of remaining data before and after processing. Thus the performance of the reduced resultant data is increased when compared to performance of original data. The Accuracy is evaluated by

$$Accuracy = \frac{\text{Data for elimination of Irrelevant \& redundant}}{\text{Total no of data}} \tag{1}$$

The Table I shows the no of original data taken as input for first step. Second column gives the result after the elimination of Irrelevant data. The third column shows the elimination after redundant data. The accuracy of Dataset1, Dataset2, Dataset3, and

Dataset4 are 0.5, 0.53, 0.46 and 1.6 respectively. Output is shown below in graph Fig.2.

TABLE II. Datasets of IEAD

Dataset	Name of the dataset
Dataset1	Iris Flower dataset
Dataset2	DNA dataset
Dataset3	Leukemia dataset
Dataset4	Brain Tumor

The Table II shows the types of data taken for Processing of IEAD algorithm.

The dataset1, dataset2, dataset3, and dataset4 are Iris(image), DNA(microarray), Leukemia(microarray), Brain Tumor(microarray) respectively.

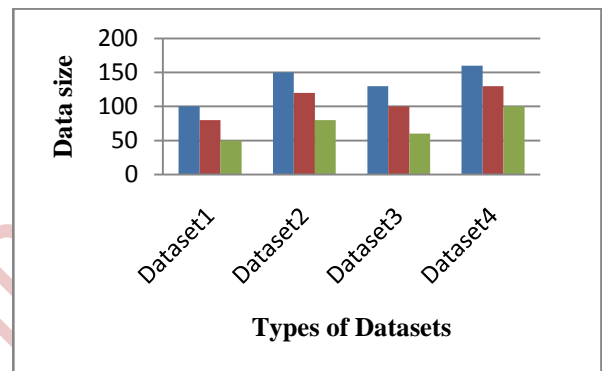


Fig 2: Out of Individual Evaluation-Based Attribute Diminution algorithm.

5. CONCLUSION

Considering the constraint brought by eliminating irrelevant and redundant data from huge dataset from UCI repository, it efficiently process the data. The proposed system shows that the decreases the size and run-Time comparatively with the existing dataset and algorithm respectively. Algorithm can be extended to text domain data set and then efficiency of the algorithm is calculated by using different classifiers in future.

6. REFERENCES

[1] A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data (Periodical style), *IEEE Trans. Knowledge and Data Engineering Q. Song, J. Ni, and G. Wang* vol. 25, no. 1, pp. 1-14, 2013.

[2] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, (Presented Conference Paper style)," presented at the 20th Int'l Conf. Machine Learning, vol. 20, no. 2, pp. 856-863, 2003.

- [3] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information, (Periodical style)," *IEEE Trans. Neural Networks J. Machine Learning Research*, vol. 5, pp. 1531-1555, 1994.
- [4] M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning, (Dissertation style)", Univ. of Waikato, 1999.
- [5] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," in *Artificial Intelligence, (Book style)*, vol. 97, nos. 1/2, pp. 273-324, 1997.
- [6] D. Koller and M. Sahami, "Toward Optimal Feature Selection, (Presented Conference Paper style)", presented at the Proc. Int'l Conf. Machine Learning, pp. 284-292, 1996.
- [7] I. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF, (Presented Conference Paper style)", presented at the Proc. European Conf. Machine Learning, pp. 171-182, 1994.
- [8] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," in *Machine Learning Research (Book style)*, vol. 10, no. 5, pp. 1205-1224, 2004.
- [9] F. Pereira, N. Tishby, and L. Lee, "Distributional Clustering of English Words, (Presented Conference Paper style)", presented at the Proc. 31st Ann. Meeting on Assoc. for Computational Linguistics, pp. 183-190, 1993.
- [10] L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification, (Presented Conference Paper style)", presented at the Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.
- [11] I.S. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification," in *J. Machine Learning Research, (Book style)*, vol. 3, pp. 1265-1287, 2003.
- [12] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, "On Feature Selection through Clustering (Presented Conference Paper style)", presented at the Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 581-584, 2005.
- [13] G. Van Dijck and M.M. Van Hulle, "Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis, (Presented Conference Paper style)", presented at the Proc. Int'l Conf. Artificial Neural Networks, 2006.
- [14] M.A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, (Presented Conference Paper style)", presented at the Proc. 17th Int'l Conf. Machine Learning, pp. 359-366, 2000.
- [15] J. Sheinvald, B. Dom, and W. Niblack, "A Modelling Approach to Feature Selection, (Presented Conference Paper style)", presented at the Proc. 10th Int'l Conf. Pattern Recognition, vol. 1, pp. 535-539, 1990.
- [16] J. Souza, "Feature Selection with a General Hybrid Algorithm," (Dissertation style), Univ. of Ottawa, 2004.
- [17] G. Van Dijck and M.M. Van Hulle, "Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis, (Presented Conference Paper style)", presented at the Proc. Int'l Conf. Artificial Neural Networks, 2006.
- [18] G.I. Webb, "Multiboosting: A Technique for Combining Boosting and Wagging," in *Machine Learning (Book style)*, vol. 40, no. 2, pp. 159-196, 2000.
- [19] E. Xing, M. Jordan, and R. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data, (Presented Conference Paper style)", presented at the Proc. 18th Int'l Conf. Machine Learning, pp. 601-608, 2001.
- [20] J. Yu, S.S.R. Abidi, and P.H. Artes, "A Hybrid Feature Selection Strategy for Image Defining Features: Towards Interpretation of Optic Nerve Images, (Presented Conference Paper style)", presented at the Proc. Int'l Conf. Machine Learning and Cybernetics, vol. 8, pp. 5127-5132, 2005.
- [21] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, (Presented Conference Paper style)", presented at the Proc. 20th Int'l Conf. Machine Learning, vol. 20, no. 2, pp. 856-863, 2003.
- [22] L. Yu and H. Liu, "Efficiently Handling Feature Redundancy in High-Dimensional Data, (Presented Conference Paper style)", presented at the Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '03), pp. 685-690, 2003.
- [23] L. Yu and H. Liu, "Redundancy Based Feature Selection for Microarray Data, (Presented Conference Paper style)", presented at the Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 737-742, 2004.
- [24] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy, in *J. Machine Learning Research (Book style)*, vol. 10, no. 5, pp. 1205-1224, 2004.