International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-6, Issue-6)

Research Article

June 2017

Optimizing Storage Space for Higher-Dimensional Data Using Feature Subset Selection Approach

Donia Augustine
Dept. of CSE, Thejus Engineering College, Vellarakkad,
Kerala, India

Abstract:

As applications producing data of higher dimensions has increased tremendously, clustering of data under reduced memory became a necessity. Feature selection is a typical approach to cluster higher dimensional data. It involves identifying a subset of most relevant features from the entire set of features. Our approach suggests a method to efficiently cluster higher dimensional data under reduced memory. An N-dimensional feature selection algorithm, NDFS is used for identifying the subset of relevant features. The concept of feature selection helps in removing the irrelevant and redundant features from each cluster. In the initial phase of NDFS algorithm features are divided into clusters using graph-theoretic clustering methods. The final phase of the algorithm generates the subset of relevant features that are closely related to the target class. Features in different clusters are relatively independent. In particular, the minimum spanning tree is constructed to efficiently manipulate the subset of features. Traditionally, feature subset selection research has focused on searching for relevant features. The clustering based strategy of NDFS have a high probability of producing a subset of useful and independent features.

Keywords: CAFS, ANN, FCBF, ACO, MST, NDFS

I. INTRODUCTION

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

The technologies present investigators with the task of extracting meaningful statistical and biological information from high dimensional data. A great deal of data from different domains such as medicine, business, science is high dimensional. Many objects can be represented under high dimensions such as speech signals, images, videos, text documents, hand writing letters and numbers. We often need to analyze large amount of data and efficiently 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 suitably. However due to high dimension of data, direct processing of these data may be very complicated and unstable so that it is infeasible.

Feature selection is one of the most frequent and important techniques in data preprocessing, and has become an indispensable component of the machine learning process. In machine learning studies, feature selection is 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. Irrelevant features are those that provide no useful information, and redundant features provide no more information than the currently selected features. In terms of supervised learning, feature selection gives a set of candidate features with the best commitment among size and evaluation measure. Feature selection algorithms improves learning, either in term of generalization capability, learning speed, or reducing the complexity of the induced model. In the process of feature selection, irrelevant and redundant features or noise in the data may be hinder in many situations, because they are not relevant and important with respect to the class concept. Machine learning methods gets particularly difficult when the number of samples is much less than the features, because the search space will be sparsely populated. Therefore, the model will find it difficult to differentiate between noise and relevant data.

II. RELATED WORK

This section reviews various methods of feature selection based on the machine learning algorithm used. They are categorized as wrapper, filter, and hybrid methods.

In wrapper approach, the searching is an overhead since the searching technique does not have the domain knowledge. In order to overcome the searching time overhead, Inza et al used estimation of Bayesian network algorithm for feature subset selection using naive Bayes and ID3 (Iterative Dichotomiser 3) [21]. Dy Brodley developed a

wrapperbased approach for unsupervised learning using order identification (recognizing the number of clusters in the data) with the expectation maximization (EM) clustering algorithm using maximum likelihood (ML) criterion [22]. IKabir et al developed a wrapper- based constructive approach for feature selection (CAFS) using neural network (NN). In this method, the correlation measure is used to remove the redundancy in the searching strategy for improving the performance of NN [25]. Stein et al proposed an ant colony optimization- based feature selection with wrapper model. In this approach, the ant colony optimization is used as a searching method in order to reduce the searching overhead such as blind search or forward selection or backward elimination searching methods [26]. Further, it is observed that the wrapper-based methods are suffered by the searching overhead, overfitting and have more computational complexity with less generality since they use the supervised learning algorithm for evaluating the generated subsets by the searching method. Therefore, these methods are not suitable choice for the high-dimensional space.

The filter-based approaches are independent of the supervised learning algorithm therefore offer more generality and they are computationally cheaper than the wrapper and embedded approaches. For processing the high-dimensional data, the filter methods are suitable rather than the wrapper and embedded methods. Generally, the process of feature selection aimed at choosing the relevant features. The best example is Relief [31] that was developed with the distance-based metric function that weights each feature based on their relevancy (correlation) with the target-class. However, Relief is in effective as it can handle only the two-class problems and also does not deal with redundant features. The modified version of the Relief known as ReliefF [32] can handle the multi-class problems and deal with incomplete and noisy datasets too. However, it fails to remove the redundant features. Holte developed a rule based attribute selection known as OneR which forms one rule for each feature and selects the rule with the smallest error [33]. Lei Yu in their work proposed a feature selection algorithm which is specially used for high dimensional data which is called as fast correlation base filter. This algorithm is for re- moving irrelevant and redundant data. They applied FCBF, ReliefF, CorrF, and ConSF on four datasets and recorded the running time and number of features selected. Then they applied C4.5 and NBC classification on the data. Peng et al proposed a mutual information-based maxrelevancy min-redundancy (MRMR) feature selection. To identify the feature relevancy, the mutual information is computed between the individual feature and target-class, and to identify the redundant feature, the mutually exclusive condition is applied [35]. Lin Tang introduced an information theory-based conditional infomax feature extraction (CIFE) algorithm to measure the class-relevancy and redundancy for feature selection [39]. Further, it is observed that the filter-based methods are computationally better than the wrapper and embedded methods. Therefore, the filter-based methods can be a suitable choice for high-dimensional space. The filter-based methods achieve high generality since they do not use the supervised learning algorithm.

The hybrid methods are the combination of filter and wrapper-based approaches. Bermejo et al developed a hybrid feature selection method known as filter-wrapper approach. In this approach, they used a statistical measure to rank the features based on their relevancy then the higher ranked features are given to the wrapper method so that the number of evaluations required for the wrapper method is linear. Thus, the computational complexity is reduced using hybrid method for medical data classification [45]. Xie et al developed a hybrid approach for diagnosing the erythemato-squamous diseases. In this approach, F-score measure is used to rank the features to identify the relevant features (filter approach). The significant features are selected from the ranked features with the sequential forward floating search (SFFS) and SVM (wrapper method) [47]. Kannan Faez presented a hybrid feature selection framework. In this approach, ant colony optimization (ACO)-based local search (LS) is used with the symmetric uncertainty measure to rank the features [48]. Naseriparsa et al proposed a hybrid method using information gain and genetic algorithm-based searching method combined with a supervised learning algorithm [50]. Huda et al developed a hybrid feature selection method by combining the mutual information (MI) and artificial neural network (ANN) [51]. Gunal presented a hybrid feature selection method by combining filter and wrapper method for text classification. In this method, information gain measure is used for ranking the significant features and the genetic algorithm is used as the searching strategy with support vector machine [52]. Further, it is observed that the hybrid methods are computationally intensive than the filter methods since they combine the wrapper and filter methods and have less generality compared to the filter methods since they use the supervised learning algorithm in feature selection process. These hybrid methods take more computational time than the filter-based methods.

III. PROPOSED SYSTEM

With the aim of choosing a subset of good features with respect to the target concepts, it was found that feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Feature selection is also useful as part of the data analysis process, as it shows which features are important for prediction, and how these features are related. Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other."

The system architecture shown in figure 4.1 is composed of two connected components, irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset. The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit sophisticated.

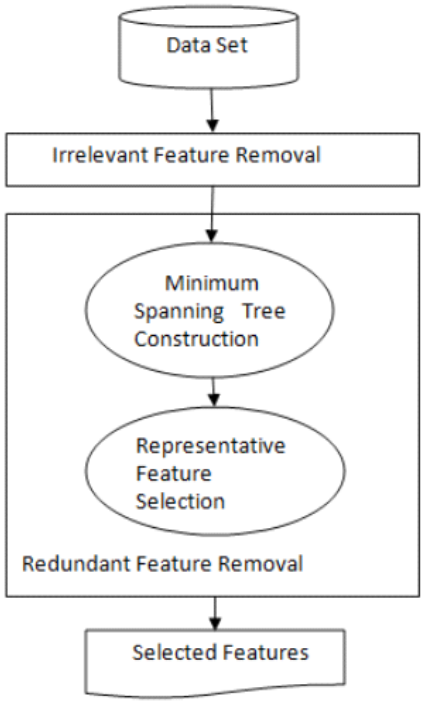

Figure 1: System Architecture

In particular, the system adopt the minimum spanning tree (MST) based clustering algorithms, because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice. Based on the MST method, we propose a N-Dimensional Feature Selection (NDFS) algorithm to effectively reduce the problem with higher dimensions.

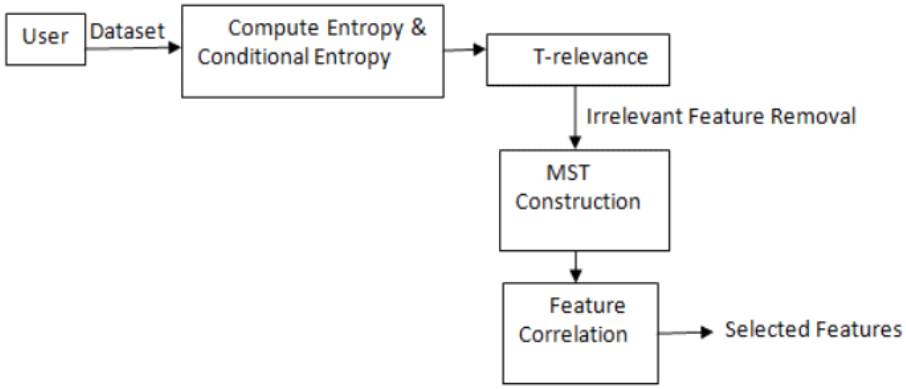

Figure 2: Sytem Flow Diagram

Figure 4.2 shows the structural flow of the proposed NDFS system. The system initially divides the features into clusters by using graph-theoretic clustering methods. Further the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features. Features in different clusters are relatively independent, the clustering-based strategy of NDFS has a high probability of producing a subset of useful and independent features. The algorithm efficiently utilizes the memory required to store the higher dimensional data using the concept of minimum spanning trees.

ALGORITHM

Input : D(F1,F2,....Fm,C)- the given dataset.
Output : S- Selected feature subset
θ - the T-relevance threshold

• Part1 : Irrelevant Feature Removal
1. for i=1 to m do
2. T-Relevance=SU(F_i ,C)
3. if T-Relevance > θ then
4. S= S∪F_i;
• Part2 : Minimum spanning tree construction
5. G= NULL; // G is a complete graph

6. for each pair of features $F'_i, F'_j \subset S$ do
7. F-Correlation = $SU(F'_i , F'_j)$
8. Add F'_i or F'_j to G with F-Correlation as the weight of the corresponding edge;
9. minSpanTree = Kruskal(G); //Using Kruskal Algorithm to generate the MST

• Part3 : Tree Partition and Representative Feature Selection
10. Forest= minSpanTree
11. for each edge $E_{ij} \in$ Forest do
12. if $SU(F'_i, F'_j) < SU(F'_i,C) SU(F'_i, F'_j) < SU(F'_j,C)$ then
13. Forest= Forest- E_{ij}
14. S= Φ
15. for each tree $T_i \in$ Forest do
16. F^j_r = argmax $F'_k \in T_i$ $SU(F'_k,C)$
17. S= S∪(F^j_r);
18. return S

The algorithm can be expected to be divided into 3 major parts:
1. The first part is concerned with removal of irrelevant features;
2. The second part is used for removing the redundant features and
3. The final part of the algorithm is concerned with representative feature selection.

The data set 'D' with 'm' features F= (F1,F2,...Fm) and class 'C', compute the T- Relevance as symmetric uncertainty, $SU(F_i ,C)$ value for every feature ($1 \leq i \leq m$). The symmetric uncertainty is defined as follows:

$$SU(F_i ,C) = [2 * Gain(F_i |C)] / [H(F_i) + H(C)] \text{ where}$$

$$Gain(F_i|C) = H(F_i) - H(F_i /C) \text{ or } H(C) - H(F_i |C)$$

H(F i) andH(F i |C) denotes the feature entropies and conditional entropies respectively. In general they are given by:

$$H(X) = - \sum_x P(x)\log_2 P(x)$$

$$H(X|Y) = - \sum_y P(y) \sum_x P(x|y)\log_2 P(x|y)$$

where, p(x) is the probability density function and p (x|y) is the conditional probability density function.

TIME COMPLEXITY

The major amount of work for NDFS algorithm involves the computation of SU values for T- Relevance and F Correlation, which has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity O(m) in terms of the number of features m. Assuming k($1 \leq k \leq m$) features are selected as relevant ones in the first part, when k = 1, only one feature is selected. Thus, there is no need to continue the rest parts of the algorithm, and the complexity is O(m). When $1 < k \leq m$, the second part of the algorithm firstly constructs a complete graph from relevant features and the complexity is O(elogk), and then generates a MST from the graph using Kruskals algorithm whose time complexity is O(elogk). The third part partitions the MST and chooses the representative features with the complexity of O(k). Thus when $1 < k \leq m$, the complexity of the algorithm is O(m+elogk).Thus, NDFS has a better runtime performance with high dimensional data.

IV. CONCLUSION

NDFS proposes a feature selection algorithm for high dimensional data. The algorithm includes (i) irrelevant features removal (ii) construction of a minimum spanning tree (MST) from, and (iii) selecting the representative features. Feature subset selection should be able to recognize and remove as much of the unrelated and redundant information. In the proposed algorithm, a cluster will be used to develop a MST for faster searching of relevant data from high dimensional data. Removal of irrelevant features reduces the volume of data to be processed thereby efficiently handling data of higher dimensions. NDFS algorithm will obtain the best proportion of selected features, the best runtime, and the best classification accuracy. Overall the system will be effective in generating more relevant and accurate features which can provide faster results.

In future more challenging domains with more features and a higher proportion of irrelevant ones will require more sophisticated methods for feature selection. For the future work, exploring different types of correlation measures for better halting criteria and inventing more intelligent techniques for selecting an initial set of features from which to start the search will improve efficiency without sacrificing useful feature sets.

REFERENCES

[1] Inza, I, Larranaga, P, Etxeberria, R Sierra, B(2000), Feature subset selection by Bayesian network-based optimization, Artificial intelligence, vol. 123, no. 1, pp.157-184.
[2] Dy, JG Brodley, CE(2000), Feature subset selection and order identification for unsupervised learning, proceedings In Proceedings of the Seventeenth International Conference on Machine Learning, pp. 247-254.
[3] Kabir, MM, Islam, MM Murase, K (2010), A new wrapper feature selection approach using neural network, Neurocomputing, vol. 73, no. 16, pp.3273- 3283.

[4] Stein, G, Chen, B, Wu, AS Hua, KA (2005), Decision tree classifier for network intrusion detection with GA-based feature selection, Proceedings of the forty-third ACM Annual Southeast regional conference, Kennesaw, GA, USA, vol. 2, pp. 136-141.
[5] Kira, K Rendell, LA(1992), A practical approach to feature selection,Proceedings of the ninth international workshop on Machine learning, Aberdeen, Scotland, UK (pp.249-256).
[6] Kononenko, I(1994), Estimating attributes: analysis and extensions of RELIEF, Proceeding of European Conference on Machine Learning, Catania, Italy, pp. 171-182.
[7] Holte, RC(1993), Very simple classification rules perform well on most commonly used datasets, Machine learning, vol.11, no.1, pp.63-90.
[8] Peng, H, Long, F Ding C(2005), Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no.8, pp.1226-1238.
[9] Lin, D Tang, X(2006), Conditional infomax learning: an integrated framework for feature extraction and fusion, Proceeding of ninth European Conference on Computer Vision, Graz, pp. 68-82.
[10] Bermejo, P, Gamez, J Puerta, J(2008), On incremental wrapper-based attribute selection: experimental analysis of the relevance criteria, Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, France, pp.638-645.
[11] Xie, J, Xie, W, Wang, C Gao, X(2010), A Novel Hybrid Feature Selection Method Based on IFSFFS and SVM for the Diagnosis of Erythemato-Squamous Diseases, Proceedings of Workshop on Applications of Pattern Analysis, Cumberland Lodge, Windsor, UK, pp. 142-151.
[12] Kannan, SS Ramaraj, N (2010), A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm, Knowledge-Based Systems, vol. 23, no. 6, pp.580-585.
[13] Naseriparsa, M, Bidgoli, AM Varaee, T(2013), A Hybrid Feature Selection method to improve performance of a group of classification algorithms, International Journal of Computer Applications, vol. 69, no. 17, pp. 0975-8887.
[14] Huda, S, Yearwood, J Stranieri, A(2011), Hybrid wrapper-filter approaches for input feature selection using maximum relevance-minimum redundancy and artificial neural network input gain measurement approximation (ANNIGMA), Proceedings of the Thirty-Fourth Australasian Computer Science Conference, Australia, vol. 113, pp. 43-52.
[15] Gunal, S(2012), Hybrid feature selection for text classification, Turkish Journal of Electrical Engineering and Computer Sciences, vol. 20, no.2, pp.1296-1311.

