Paper ID: UGC 48846-928

FEATURE SELECTION ON HIGH DIMENSIONAL DATASETS

Poonam, Prof.Dharmender Kumar, Sunita Beniwal
Guru Jambeshwar University of Science and Technology, Hisar, Haryana

Abstract : High dimensional data contain billions of entries with small numbers of instances and large numbers of attributes. Some problems with high dimensional data are large dimensions, overfitting, class imbalance and outliers which decrease the performance. To handle these problems need a algorithm called feature selection. Feature selection is used to reduce dimensions by removing irrelevant attributes. Mainly three techniques are used for feature selection namely, filter, wrapper and embedded. This paper is about overview of feature selection techniques for high dimensional data. Also describe some existing related work found in this literature.

Keywords: Dimensional Data, Techniques, Embedded

Introduction

Introduction to High Dimensional Data

High dimensional datasetsis a collection ofthousands of entries including different attributes and relational databases. High dimensional datasets have different tuples and relational databases, which need large amount of space and disk storage [1].High dimensional data made up of many different type variables. Application domainof high dimensional databases these days are in medical research, imaging, financial analysis, and many other domains [4].

Examples of high dimensional data in different different fields [5] : Biological Data such as Microarrays, Deep Sequencing - Counts, Micro RNA Expression - Continuous, CGH (Copy Number Variation) - Continuous / Categorical, SNPs (Single Nucleotide Polymorphisms) - Binary / Categorical, Methalaytion - Continuous.Random Fields Data such as Functional MRIs (fMRI),Finance (Time Series Data), Climate Data (Spatial Data, Spatio-temporal Data), Neuroimaging (DTI - Diffusion Tensor Imaging, Calcium-Florescence Imaging, EEG & MEG). Collaborative Filtering Data such as Netflix Movie Rating Data, Amazon, Facebook, Yahoo, Twitter.

Intrinsic characteristics of high dimensional data

Small sample size: The first problem that find when dealing with high dimensional data is related to the small sample size (usually less than 100). A key point related to this is that error estimation is greatly affected by small samples.

Class imbalance: This occurs when a dataset is contain large no.of samples of a class rather than other class. I.e, major difference between numbers of instances of two class. There are some difficulties related to this problem that occur, such as a small sample size.

Data complexity: Data complexity measures are related with represent characteristics of the data. Difficult due to complexity in classification tasks, such as the overlapping among classes, their separability or the linearity of the decision boundaries.

Dataset shift: Another common problem when datasets were originally divided to training and test sets, is the so-called dataset shift. This occurs when the testing (unseen) data experience a phenomenon that leads to a change in the distribution of a single feature, a combination of features, or the class boundaries.

Paper ID: UGC 48846-928

Outliers: An important aspect in the literature is to detect outliers in data samples. In some datasets, there are samples that are incorrectly labeled or identified such as noise.They can exert a negative effect on the selection of relevant attributes [2,3].

Problems while using high dimensional data:

- Large no. of attributes
- Less samples
- Time complex
- Less generality (overfitting)
- Low quality, unreliable
- Redundant and noisy data

Feature Selection

Feature selections, also known as variable selection, attribute selection or variable subset selection. Feature selection can be defined as a process that chooses a minimum subset ofM features from the original set of N features, so that the feature space is optimallyreduced according to a certain evaluation criterion [6]. Feature selection is the process of identifying the mostrelevant attributes and removing the redundant and irrelevantattributes [9].It is the process of selecting a subset of relevant features (variables, predictors, attributes) for use in model construction. Feature selection is an important topic in data mining, especially for high dimensional datasets. Feature selection (also known as subset selection) is a process commonly used in machine learning, wherein subsets of the features available from the data are selected for application of a learning algorithm. The best subset contains minimum number of dimensions that produce maximum accuracy; we discard the remaining, unimportant, irrelevant dimensions or attributes. This is an important stage of preprocessing and used to avoiding the curse of dimensionality [7]. Feature selection has been an important and active field of research area in pattern recognition, machine learning, statistics and data mining communities.. The main goal of feature selection to find a subset that provide higher accuracy in supervised learning [17 synps].

Generally, features are characterized [7] as:

1.Relevant: These are features which have an influence on the output and their role can notbe assumed by the rest. Output is directly affected by relevant attributes.

Irrelevant: Irrelevant features are defined as those features not having any influence on the output, and whose values are generated at random for each example.

Redundant: A redundancy exists whenever a feature can take the role of another. Two attributes have same influence on output are redundant.

Feature selection techniques are used for three reasons:

- Simplification of models to make them easier to interpret,
- Shorter training times,
- Enhanced generalization by reducing overfitting.

Advantage of Feature Selection

- Dimension reduction.
- Reduction of computational cost.
- Remove redundant, irrelevant or noisy data.
- Improve classification accuracy.

Feature selection methods

Main principles of feature selection methods: filter method, wrapper method, embedded method. In addition, feature selectionmethods may also be divided into univariate and multivariate types. Univariate methods evaluate each feature independentlyof other features andprovide rank to individual feature (a single attribute), a drawback of Univariate methods is that can not remove redundancy.This can be overcome by multivariate techniques that combine featuredependencies to some degree, demanding more computational resources and cost [3,11]. Also multivariate filters evaluate a subset of features. Search strategy required for Feature subset generation for multivariatefilters. Various search strategies used for feature subset generation: forward selection,backward elimination, bidirectional selection, and heuristic feature subset selection [11].

Filter method

In the filter approach, attributes or features are selected according to the intrinsic characteristics of attributes. It works as a preprocessing step.The filter model works on general characteristics of the training data to select some relevant features without involving any learning algorithm [9].Filter methods can rank individual features or evaluate entire feature subsets [8].Filter methods observe the goodness of single attribute or subsets by evaluating only the statisticalmeasures of data. Different evaluation criteria for filter method namely, distance, information, dependency and consistency [3]. Various filter methods used for high dimensional data are given below:

Information gain (IG): It is more general technique based on entropy. Entropy-baseddiscretization method is generally used for geneexpression data [3]. Information gain is also called as the asymmetric dependency coefficient (ADC). The goal of these methods is to find a good approximation of the conditional distribution, $P(Cj/\mathrm{F})$, whereF is the overall feature vector and C is the class label. [6,9]. **T-Test:** It compares theactual difference between two means in relation to thevariation in the data. The t-test iswidely used to measure the relevance of a attribute [10]. **Correlation Feature Selection (CFS):** CFS is a simple multivariate filter algorithm. The purpose of CFS is to find subsets of relevant featuresthat are highly correlated with the class and uncorrelated witheach other. The rest of irrelevant features should be ignoredbecause they will have low correlation with the class [3,9].**Fast Correlation-Based Filter (FCBF):** FCBF evaluate attributes based on symmetricaluncertainty (SU) . Symmetricaluncertainty is defined as the ratio between the information gain and the entropy of two attributes. FCBF was designed for high-dimensionality data and effective in removing both irrelevant and redundant features [3]. **Chi-Square Test:** Chi-square is based on statistical test commonly used to compare observed data with expect data according to a specific hypothesis. **ReliefF:** Also known asthe consistency-based filter method. It is an extension of the original Relief algorithm. ReliefF adds the ability to dealing with multiclass problems [3]. Relief-F is an instance-based feature selection methodwhich evaluates a feature by how well its value distinguishessamples that are from different groups but are similar to eachother [9].For each feature *X*, Relief-F selects a random sampleand k of its nearest neighbors from the same class and each ofdifferent classes [3,9].ReliefF is more robust and capable of dealing with incomplete and noisy data [3]. **mRMR (minimum Redundancy Maximum Relevance):** It works based on two criteria relevance and redundancy. Features are selected that have the highest relevance with the target class and are also lowest redundancy. In other words it selects features that are maximum dissimilar to each other. Both optimization criteria (maximum-relevance and minimum-redundancy) are based on mutual information [3]. **One-R:** One-R is a simple algorithmrules based on one feature only. Itbuilds one rule for each attribute in the training data and thenselects the rule with the smallest error. It treats allnumerically valued features as continuous and uses a straightforward method to dividethe range of values into several disjoint intervals. It handles missing values by treating"missing" as a legitimate value[6,9].

Some existing works using the filter approach for high dimensional data can be found in the earliest years. M. Yasodha and P. Ponmuthuramalingam [11], proposed gene ranking technique which use is T-Score for Lymphoma and Leukemia dataset. Accuracy and execution time consider as evolution metrics with LDA classifier. Observed by their study that the planned technique provides improved results by their correctness in percentage and outcome in the reduced time with LDA classification.Pinar Yildirim [9], presented a research work on comparison between filter based feature selection methods for Hepatitis dataset.Selected features are fed into four classifiers naive bayes , J48 IBK , Decision Table. Overall result evaluated by the authors Computed Naïve Bayes and Decision Table classifiers have higher accuracy rates on the hepatitis dataset than the others after the application of feature selection methods. Also Consistency Subset, Info Gain Attribute Eval, One-R Attribute Eval and Relief Attribute Evalmethods performed better results than the others.C.Lavanya, et al. [10], used T-Test, Chi-Square Test and Information gain feature selection methods for cancer gene to achieve improved classification performance.This paper presents the Naive Bayes algorithm for the classification task. Learned from this study is that there is no single best feature selection technique across different datasets and improvement in generalization performance.

Wrapper method

In the wrapper approach, to score or rank the feature subset a learning algorithm is used based on the some resultant predictive power, and an optimal feature subset is searched for a specific classifier. The wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected [9]. As for each new subset of features,the wrapper model needs to learn a hypothesis.It tends to find features better suited to the predetermined learning algorithm resulting in superior learning performance. Wrappers are more computationally expensive and complex than the filter model.Wrappers depend on the resource demands of the modelling Algorithm due to this they are much slower than filters in finding good subsets[8,9]. Working of typical wrapper method a search process is applied on the whole data set. This classifier is trained only with the attributes which found in search. After that the goodness of each gene subset is evaluatedby the accuracy achieved by the specific classifier [3]. Example of typical wrapper method - evaluated classical wrapper search algorithms and BIRS. Sequential selection algorithms and Evolutionary search algorithms are two main types of Wrapper methods [25]. Some wrapper feature selection methodsare :**SFS**- Sequential forward selection, ASFS is thewell-known simplest, greedy search algorithm. SFS is a recuressive process, starts with anempty selection of attributes and, in each round, it adds each unused attribute of the given example set.The performance is estimated using the cross validation for each added attribute. Attributes added to the selection which givingthe highest performance [15].**Genetic algorithms (GA)** is used to find theoptimum in to solve problems like arrangement and assignment. Genetic Algorithms (GA) is an optimization technique based on Heuristic approach. GA is an Evolutionary, optimization and a population-based technique which finds the optimal solution in the process of natural selection and crossover. GA uses a iterative process to produce a new population by manipulating one population using functions crossover and mutation [17,18]. **Sequential backward selection (SBS)** works in the opposite direction to SFS. It starts with thefull set of attributes and, in each round, it removes each remaining attribute of the given example set which giving least decreasing performance. Two advantage of this method first, it can discard several features and second, it allows for backtracking [15]. **Simulated annealing (SA)** The algorithm is run separately for each class resulting in the feature subset for that class [20].Optimize selection (evolutionary) is a method which selects the most relevant attributes of the givenexample set

[15].Other methods are plus L minus R, Beam search, Randomized hill climbing, Estimation of distribution algorithms, [12].

Some existing works using the wrapper approach for high dimensional data can be found in the earliest years. RattanawadeePanthonga et al. [15]Theyworked on methods sequential forward selection (SFS), sequential backward selection (SBS) and optimize selection (evolutionary) basedon ensemble algorithms namely Bagging and AdaBoost.Selected attributes are performed using twoclassifiers; Decision Tree and Naïve Bayes using Thirteen datasets containing different numbers of attributes anddimensions. Their study shows that the search techniqueusing SFS based on the bagging algorithm using Decision Tree obtained better results in accuracythan other methods.M. Wanderley et al. [16] presentedevolutionary wrapper method (GA-KDE-Bayes). It uses a non-parametric density estimationmethod and a Bayesian classifier. The authors state that non-parametric methods are a good alternative for scarce and sparsedata, such as the bioinformatics problem. Sung-Sam Hong et al. [17] worked on feature selection using genetic algorithm. This study analyzed about feature selection method to increase performance with lower computational complexity. They designed a new geneticalgorithm to extract features in text mining based on TF-IDF which is used to reflect document-term relationships in feature extraction. The authors obtained improved AverageAccuracy and F1-measure of clustering using FSGA.Sharma et al. [19] proposed an algorithm called successive feature selection (SFS). It have the drawback that aweakly ranked gene that could perform well in terms of classification with an appropriate subset of genes will be left outof the selection. Trying to overcome this shortcoming, the proposed SFS consists of first partitioning the features into smallerblocks. Once the top features from each of the blocks are obtained according to their classificationperformance, they arecompared in order to obtain the best feature subset.

Embedded method

When a feature selection is embedded into a learning algorithm and optimized for it, it is called an embedded method [13]. The main disadvantage of the filter method is it does not interact with the classifier and the wrapper model comes with an expensive computational cost. Embedded methods an intermediate solution, to rank features which use the core of the classifier to establish a criteria [3]. The embedded method incorporate feature selections a part of training process and are usually specific to given learning algorithm. It is efficient than other methods [24]. These methods allowsinteractions with the learning algorithm for feature selection andalso the computational time is smallerthan wrapper methods [25]. This embedded method roughly categorized into three types, namely pruning method, built-in mechanism and regularization models. **Pruning based method**, in this initially all the features are taken into the training process for building the model and the features which have less correlation coefficient value are removed recursively using the support vector machine (SVM). In the **Built-in mechanism-based** feature selection method, a part of the training phase of the C4.5 and ID3 supervised learning algorithms are used to select the features. In the **Regularization method**, fitting errors are minimized using the objective functions and the features with near zero regression coefficients are eliminated [25]. The following are some other embedded feature selection algorithms. Decision trees, Weighted naïve bayes, Feature selection using the weighted vector of SVM [12].

Liang et al. [26] integrated multiple data sources and described the Multi-Source k-Nearest Neighbor (MS-k NN) algorithm for function prediction, which finds k-nearest neighbors of a query protein based on different types of similarity measures and predicts its function by weighted averaging of its neighbors' functions. Cao et al. [27] proposed a novel fast feature selection method based on multiple Support Vector Data Description (SVDD) and applies it to multi-class microarray data.Wang et al. [28]proposed a First Order Inductive Learner (FOIL) rule based feature subset selection algorithm, calledFRFS.

Hybrid

To combine the advantages of both methods filter and wrapper, hybrid model have been proposed to deal with high dimensional data. These algorithms mainly focus on combining filter and wrapper algorithms to achieve best possible performance with a particular learning algorithm.Hybrid method comes with similar time complexity of filter algorithms [9,24].Hybrid methods usually combine two or more feature selection algorithms in a sequentialmanner [3].

Some existing works using the hybrid approach for high dimensional data can be found in the earliest years. Shutao Li, et al.[3] proposed a hybrid gene extraction method by using two standard feature extraction methods, namely the T-test method and kernel partial least squares (KPLS). First, T-test method is used to filter irrelevant and noisy genes. Then KPLS is used to extract features with high information content. The features extracted in the first step are further filtered by using KPLS. Finally, the extracted features are fed into a classifier. K-nearest neighbor classifier (k-NN), Feedforward neural network (NN), Support vector machine (SVM) used for this. Cross validation is used to test accuracy. The proposed method, attains the best classification accuracy. The proposed gene extraction method proved to be a reliable gene extraction method.Eric P. Xing ,et al.[22] proposed a hybrid approach of filter and wrapper approaches to feature selection. They make use of a sequence of simple filters, culminating in Markov Blanket filter. They compare between the result using cross validation. This paper also define regularization methods as an alternative to feature selection. They used a Gaussian classifier, a logistic regression classifier and a nearest neighbor classifier in their study and observed that feature selection curves are generally associated with smaller error.Wei Luoetal. [23]designeda hybrid two-step feature selection method by combining modified t-test and PCA . Used three cancer data sets, namely the lymphoma data set, the SRBCT data set, and the ovarian cancer data set.SVM classifier used by them to solve the problem of cancer classification. They first used a modified t-test method to remove the genes irrelevant to the classification, and after that, extract principal components from the top-ranked genes based on the first step (select 100 genes). The state that results in all the three data sets show their two-step methods is able to achieve 100% accuracy with much fewer genes than other published results.

Conclusion

High dimensional data are very complex in use because of its small sample size and large attribute size. Many problems while using High dimensional data are overfitting, outliers, low accuracy, class imbalance. To remove these problems and to make data more reliable feature selection are used. Feature selection is the process of find a subset of relevant attributes by removing irrelevant and redundant attributes. Optimal features subset selected using feature selection which increase the overall performance. The main goal of feature selection is to improve clarification performance and reduce dimensionality. Feature selection method are categorized into three, filter, wrapper and embedded. This paper represents techniques used for high dimensional data with some related work.

Paper ID: UGC 48846-928

References

[1] Vinod S. Bawane , Shireesh P. Bhoyar , Manish P. Tembhurkar . "A Review on High Dimensional Data Visualization" *International Journal of Emerging Trends in Engineering and Development*, Vol.3, Issue 4, pp. 878-884, ISSN 2249-6149, May, 2014.

[2] VerónicaBolón-Canedo, Noelia Sánchez-Maroño, Amparo Alonso-Betanzos. "Introduction to High-Dimensionality" in *Introduction to High-Dimensionality,* Springer International Publishing pp. 1-12, 2015

[3] V. Bolon-Canedo , N. Sanchez-Marono , A. Alonso-Betanzos , J.M. Benitez , F. Herrera . " A review of microarray datasets and applied feature selection methods" . *Information Sciences* vol. 282, pp. 111–135, 2014.

[4] Michel Verleysen. "Learning high-dimensional data", *Limitations and Future Trends in Neural Computation*, IOS Press, pp. 141-162, 2003.

[5] Genevera I. Allen "Examples of High-Dimensional Data" , Statistical Learning: High-Dimensional Data , January 2011.

[6] JasminaNovakovic, PericaStrbac, DusanBulatovic . "Toward optimal feature selection using ranking methods and classification algorithms" *Yugoslav Journal of Operations Research* , Number 1, pp. 119-135, 2011.

[7] L.Ladha ,T.Deepa. "Feature selection methods an algorithms" *International Journal on Computer Science and Engineering (IJCSE)* , Vol. 3 ,No. 5, May 2011.

[8] A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications" ,*IEEE* ,pp. 1200-1205, 2015.

[9] Pinar Yildirim . "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease" *International Journal of Machine Learning and Computing,* Vol. 5, No. 4, August 2015

[10] C.Lavanya, M.Nandihini, R.Niranjana, C.Gunavathi . "Classification of Microarray Data Based On Feature Selection Method" , *International Journal of Innovative Research in Science, Engineering and Technology* Volume 3, Special Issue 1, February 2014.

[11] A. Jovic,K. Brkicand N. Bogunovic. "A review of feature selection methods withapplications" *IEEE*, may 2015.

[12] S.VanajaK.Rameshkumar. "Analysis of Feature Selection Algorithms on Classification: A Survey ", *International Journal of Computer Applications* ,Vol. 96, pp. 28-35, No.17, June 2014.

[13] KokaneVina ,Lomte Archana . "Feature Selection for High Dimensional and Imbalanced Data- A Comparative Study", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* ,ISSN: 2278-1323, pp. 3800-3804, Vol. 3, Issue 11, November 2014.

[14] M. Yasodha and P. Ponmuthuramalingam . "A fast and efficient feature selection algorithm for microarray gene expression and classification" , *ARPN Journal of Engineering and Applied Sciences* ,Vol. 10, No. 4, March 2015.

[15] RattanawadeePanthonga,AnongnartSrivihok. "Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm",*Information Systems International Conference (ISIC)*,Published by Elsevier B.V.,pp. 162 – 169, 2015.

[16] M. Wanderley, V. Gardeux, R. Natowicz, A. Braga."Ga-kde-bayes: an evolutionary wrapper method based on non-parametric density estimationapplied to bioinformatics problems", *21st European Symposium on Artificial Neural Networks-ESANN*, pp. 155–160, 2013.

[17] Sung-Sam Hong, Wanhee Lee, and Myung-Mook Han. "The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification", *Int. J. Advance Soft Compu. Appl,* Vol. 7, No. 1, ISSN 2074-8523, March 2015.

[18] BabatundeOluleye, Armstrong Leisa, JinsongLeng, Diepeveen Dean. "A Genetic Algorithm-Based Feature Selection", *International Journal of Electronics Communication and ComputerEngineering,*Vol. 5, Issue 4, ISSN : 2278–4209, July 2014.

[19] A. Sharma, S. Imoto, S.Miyano."A top-r feature selection algorithm for microarray gene expression data", *IEEE/ACM Trans. Comput.Biol. Bioinformatics(TCBB)* 9 (3) pp. 754–764, 2012.

[20] V. Susheela Devi "Class Specific Feature SelectionUsing Simulated Annealing", *Springer International Publishing Switzerland*, pp. 12–21, 2015.

[21] Shutao Li, Chen Liao, and James T. Kwok."Gene Feature Extraction Using T-Test Statistics and Kernel Partial Least Squares" pp. 11-20 , 2006.

[22] Eric P. Xing ,Michael I. Jordan , Richard M. Karp **.** " Feature Selection for High-Dimensional Genomic Microarray Data" pp. 601-608, 2001.

[23] Wei Luo, Lipo Wang, Jingjing Sun. "Feature Selection for Cancer Classification Based on Support Vector Machine", Volume: 4 Pages: 422 - 426, 2009.

[24] Mr. Swapnil R Kumbhar, Mr.Suhel S Mulla."Literature Review on Feature SubsetSelectionTechniques", *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, ISSN 2319 - 4847, Vol. 03, pp. 231-233, Issue 09, September 2014.

[25] Rabia Aziz, C.K. Verma, and Namita Srivastava. "Dimension reduction methods for microarray data: a review",*AIMS Bioengineering*, Vol. 4, Issue 2,pp. 179-197, March 2017.

[26] Cao J, Zhang L, Wang B,"A fast gene selection method for multi-cancer classification using multiple support vector data description." *J Biomed Inform,* pp. 381–389, 2015.

[27] Lan L, Djuric N, Guo Y. "MS-kNN: protein function prediction by integrating multiple data sources", *Bioinformatics,* 14: S8, 2013.

[28] G. Wang, Q. Song, B. Xu, Y. Zhou. "Selecting feature subset for high dimensional data via the propositional foil rules" *Pattern Recognition,* vol. 46, Issue 1, 199–214, 2013.