Information Gain Measured Feature Selection to Reduce High Dimensional Data

Thee Zin Win, Nang Saing Moon Kham
University of Computer Studies
theezinwin@ucsy.edu.mm, moonkham@ucsy.edu.mm

Abstract

While demand of the massive amount of data to be more effective and efficient mining strategies is increasing significantly, practitioners and researchers are trying to develop scalable machine learning algorithms and strategies in turning mountains of data into nuggets. High dimension of data makes the memory, storage requirements and computational costs increased significantly. Therefore, reducing dimension can mainly improve learning performance. Feature selection, a data preprocessing technique, is effective and efficient to enhance data mining, data analytics and machine learning. Most feature selection algorithms have been trying to eliminate irrelevant features. However, removing only irrelevant features is not enough to get the best insight and patterns. Not only irrelevant features but also redundant features can degrade learning performance. Feature selection methods which can eliminate both irrelevant and redundant features are demanding in high dimensional data analytics. To solve this problem, information gain measured feature selection is presented in this work.

1. Introduction

Nowadays, massive amount of high dimensional data has become abundant in our daily lives. As increasing the massive amount of data demands effective and efficient mining strategies, practitioners and researchers are trying to develop scalable mining algorithms, machine learning algorithms and strategies to be successful data mining in turning mountains of data into nuggets. When applying data mining and machine learning on high dimensional data, there is an important issue which is called curse of dimensionality issue [6]. In addition, learning models tend to over fitting which may lead performance degradation on unseen data because of a large number of features. Moreover, analyzing high dimensional data makes memory storage requirement and computational costs for data analytics heighten significantly. High dimension of feature reduction can be mainly categorized into Feature Extraction and Feature Selection. Feature extraction transforms original high dimensional feature space into a new feature space with low dimensionality. In other words, feature extraction transforms the original features into a reduced set of features which are not interpretable. Feature selection, selecting only the most relevant and least redundant features to target classes, is effective and efficient in data mining, data analytics and machine learning problems. The main purpose of data analytics is to get patterns and models with high accuracy. The accuracy of data analytics mostly depends on the quality of data. To get high quality of data, feature selection is essential because it chooses the most relevant features among features. Data analyzing or data mining the most relevant features can improve mining results with higher accuracy, mining performance within less processing time, less resources such as CPU, Memory. Besides, building simpler models, comprehensible models, improving data mining performance, accuracy and preparing clean and understandable data are some of the objectives of feature selection.

In this section the introduction of high dimensional data analytic and the role of feature selection is presented. Related works of feature selection in high dimensional data are reviewed in section two. In section three, the theoretical background of feature selection and its measures will be discussed. Information Gain to be used in feature selection and some of feature evaluation methods will also be discussed. Then experiments and results will be presented in section four. Finally, conclusion of this work and future work will be discussed.

2. Related Works

In this section, some related works on dimension reduction feature selection of high dimensional data are reviewed. A technical challenge of intrusion detection systems is an example of the curse of high dimensionality. In [14], the authors proposed two feature selection algorithms: modified mutual information-based and linear correlation based. They compared performance of these two algorithms with mutual information-based feature selection method. They used both a linear measure, linear correlation coefficient and a non-linear measure, mutual information. Their intrusion detection system can result with high accuracy especially for remote login and user to remote attacks by experimenting KDDCup99 dataset using Least Squares Support Vector Machine by comparing with their proposed mutual information based feature selection method. In [2], authors investigate an alternative method to feature selection of micro array data, using Markov Blanket Filter to decide on particular feature subsets for each subset cardinality. M.Tan [12] proposed a new adaptive feature scaling scheme which is feature generating paradigm for feature selection of ultrahigh dimensional Big Data. Their proposed method can handle two challenges of feature selection: group-based feature selection with complex structures, and nonlinear feature selection with explicit feature mappings. Hoque introduced greedy feature selection method using mutual information because decreasing in the number of irrelevant features can lead computation time reduction. The mutual information of feature to feature and feature to class is used to determine an optimal set of features [5]. A fast and efficient feature selection method is proposed by Fleuret(2004), which is based on conditional mutual information [3]. His method is compared with C4.5 binary trees and fast correlation based filter feature selection techniques. Their experiments showed that this method with Naïve Bayesian for binary input features is better than support vector machine and boosting. M. Ali [10] analyzed high dimensional gene data, Leukaemia data set by using PCA to reduce the number of attributes. Then Factor Analysis was used to extract the important features. It can extract the important attributes in Leukaemia data. M. Savic and authors [11] proposed Feature Selection Method based on Feature Correlation Networks (FSFCN) with four variants: Fast greedy modularity optimization community detection algorithm, the Louvain algorithm, Walktrap algorithm, Infomap algorithm, feature selection method which is based on complex weighted networks and which can describe the best correlated features in order to reduce high dimension. A dataset including 120 plasma signaling protein features which is related to the diagnosis of Alzheimer disease is used to experiment the method. Classification accuracy is used to compare seven classifiers' performance. L. Liu [9] proposed a fast correlation based feature selection for high dimensional data. The reason why Fast Correlation based Feature Selection (FCBF) is proposed is that feature selection with only removing irrelevant feature is not enough for high dimensional data. The redundant features in the remaining features are still degrading the learning performance. Authors use correlation measure to find out redundant features. However, this FCBF can evaluate feature correlation with only WrapperSubset Evaluator. In [1], authors proposed framework that used filter and wrapper based technique to support prediction process in future study. In this work, FCBF searching is used because it was capable of evaluation the worth of feature subsets by weighing the individual predictive ability through their redundancy degree between features. FCBF is a kind of filter approach. Unfortunately, the size of selected features after applying FCBF is still large. Therefore, they choose to use wrapper approach to reduce redundant features again.

3. Feature Selection

Feature selection is a process of removing the irrelevant and redundant features from a dataset in order to improve learning performance in terms of accuracy and time to build the model. Feature selection can be categorized into three parts: Supervised, Semi-supervised and Unsupervised. Supervised feature selection for classification or regression problems and unsupervised feature selection for clustering problems are designed respectively. Because of the lack of label information to evaluate feature importance in clustering, data similarity and local discriminative information are used to define feature relevance. With respect to selection methods, feature selection methods can be classified into Wrapper, Filter and Embedded methods. Wrapper methods depend on the predefined learning algorithms' performance to evaluate the quality of selected features. The quality of selected features depends on learning algorithms used in selecting features. Filter methods do not depend on learning algorithms performance. Certain characteristics of data are their resources to assess the importance of features. There are two steps in individual feature evaluation Filter method. In the first step, feature importance is ranked by their feature relevance which is evaluated according to their evaluation criteria. In the second step, high ranked features are selected by filtering low ranked features by comparing with threshold.

Typically, a feature selection process consists of four basic steps, namely, subset generation, subset evaluation, stopping criterion, and result validation [13]. In the first step, a candidate feature subset will be chosen based on a given search strategy. In the second step, subset will be evaluated according to certain evaluation criterion. In the third step, step one and two will stop when it meets the stopping criterion. In the final step, the chosen subset will be validated by using some validation methods such as domain knowledge or validation set. There are different search strategies to generate candidate feature subsets. Some of them are complete search, heuristic search and random search. In individual evaluation, features are ranked according to their importance by using some measures such as distance, information, dependency and consistency. In subset evaluation, it finds for a minimum subset of features that satisfies some goodness measure.

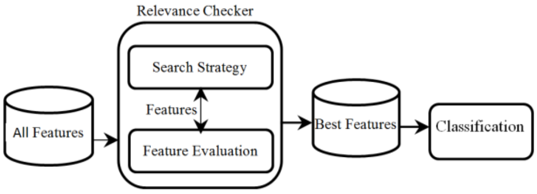

Figure 1. General Framework of Relevant Feature Selection

3.1 Information Gain

Information Gain (IG) is an entropy-based feature evaluation method, and it is also widely used in the field of machine learning. IG measures how much information a feature gives facts about the target class. Based on target class, IG is able to detect the features with the most information. Features with high IG are strongly relevant to target class and they are usually selected to get the best classification results. However, IG is not able to eliminate redundant features. Therefore, we still need to filter out redundant features. IG is derived from Entropy, as described in equations. Entropy is used to measure uncertainty of a class by using the probability of a certain event or attribute. IG is inversely proportional to Entropy. The information gain usually depends on two facts: how much information was available before knowing the attribute value and after knowing the attribute value. The maximum value of IG for multi classes is 1.

The formula of Entropy to analyze more than two classes is shown below.

$$H(X) = -\sum_{i=1}^{K} P(x_i)\log_2 P(x_i)$$

K is the number of classes. And IG of a feature X and the class labels Y is calculated as follow.

$$\text{IG(X, Y)} = H\ (\text{X}) - H\ (\text{X|Y})$$

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j)\ \log_2(P(x_i|y_j))$$

H(X) is Entropy of X and H (X|Y) is Entropy of X after observing Y. Since IG is a filter technique, it can scale well with the high dimensionality data. It is also applicable on several classifiers due to being classifier independent. In this work, after analyzing the improvement in the classification efficiency, the effect of information gain feature selection on overall classifier performances will be compared with other feature selection methods.

3.2. Removing Irrelevant and Redundant features

The filter-based approaches are independent of the supervised learning algorithm. Therefore, they 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. Relief [7] that was developed with the distance-based metric function weights each feature based on their relevancy (correlation) with the target-class. However, Relief is ineffective as it can handle only the two-class problems and also cannot solve features redundancy because both the features which are highly relevant to the target class and the features which are highly relevant to other features will be selected by Relief. The modified version of the Relief known as ReliefF [8] can handle the multi-class problems and deal with incomplete and noisy datasets. However, removing the redundant features is still essential. As the irrelevant features should be excluded, the redundant features should also be omitted because they can degrade the learning performance, accuracy and speed. We propose the framework to select the most relevant and less redundant features in high-dimensional data. In this work, there is something to consider why choosing relevant features come first. Why can finding redundant features not come first? The answer is it cannot come first because relevant features can be missed if redundant features are initially selected. In classification, relevancy of features to target class impacts the accuracy of classification. Therefore, we conduct finding relevant features primarily and removing redundant features from these relevant features after primary task of selecting relevant features. The proposed feature selection approach consists of two mainly parts. The first one is finding relevancy of features with target class. The second one includes calculating and removing the redundant features by comparing IG values of selected relevant features. In evaluating the relevance of features on target class is used individual evaluation because it is more suitable than subset evaluation in feature selection of high dimensional big data. The proposed algorithm can be described by the following procedure:

1) Initialization: set
F ←Initial set of all features
S ←Relevant empty set
C ← Target class
R ← Redundant empty set
2) Computation of the Information Gain of the features with the target class
Compute IG(f_i;C) for each feature (f_i∈F) and rank them in descending order.
3) Selection of the relevant features
Find the features fi that maximizes IG of features to target class, by selecting higher ranked features by comparing with threshold.
Set F ←F-f_i , S ←f_i
4) Selection of redundant features
Comparing the IG of features of relevant feature set (f_i ∈S)
S ← S- f_i , removing redundant features from relevant features.
Set R ←S
5) *Output the set containing the selected most relevant and less redundant features*
Output ←R

3.3 Ranking features

Weka Ranker ranks attributes through their individual evaluations. It is specially used in conjunction with attribute evaluators such as ReliefF, Gain Ratio, Entropy etc. There are many options to rank attributes in Ranker. The generateRanking is a constant option because Ranker is only for generating attribute rankings. The numToSelect is to specify the number of attributes to preserve. The default value (-1) indicates that all attributes are to be retained. We can use either this option or a threshold to reduce the attribute set by setting the number of attributes to be selected. The threshold is set by which attributes can be discarded by comparing the attributes rank values. Default value results in no attributes being discarded. We can use either this option or numToSelect limiting the number of features to be selected in order to reduce the attribute set. We can also specify a set of attributes to ignore by the startSet.

4. Experiment and Results

Experiment is conducted on WEKA 3.8, Windows 7 32 Ultimate with 4 GB memory and Core i7 CPU. After experimenting, it is observed that the feature ranking-based methods using the statistical measures or information measures to weight the individual feature only by observing the relevancy between the individual feature and the target-class as in table (4) and (5). Hence, these methods take less runtime as shown in table (2) and (3) but fail to remove the redundant features. The feature ranking-based methods follow a filter-based approach since these methods do not involve any supervised learning algorithm to evaluate the significance of the features. Consequently, these methods are independent of the supervised learning algorithm hence they achieve more generality and less computational complexity. Thus, the feature ranking-based methods can be a good choice for selecting the significant features from the high-dimensional space. Datasets used in experiment are shown in table (1).

Table 1. Datasets used in experiments

Dataset	Source	Attribute type	Attributes	Instances	Classes
Arrhythmia	UCI	Categorical, Integer, Real	279	452	13
mfeat	UCI	Integer, Real	217	2000	10

We use Info Gain attribute evaluation method with Ranker search method in Weka. Info gain evaluates the worth of an attribute by measuring the information gain of this attribute with respect to the class. Ranker search methods rank attributes by their individual assessments and they are used by combining with attribute evaluators such as ReliefF, GainRatio and Entropy. Either the number of attributes to retain can be specified to reduce the attribute set. Threshold value can also set to reduce or discard the attribute set. Correlation based attribute evaluation calculates the Pearson's correlation measures of features and class [4]. Principal Components Analysis executes analysis and transformation of data in conjunction with Ranker search. Its choosing enough eigenvectors to account percentage of variance in the original data accomplishes dimensionality reduction. Therefore, we selected PCA in our experiment to compare with information based feature selection methods. ReliefF evaluates the value of an attribute with repeated sampling an instance and it can operate on both discrete and continuous data. We use subset evaluator, Classifier which estimates attribute subsets on training data in order to compare individual attribute evaluators.

5. Conclusion and Discussion

We present theoretical analysis of Information measures. Moreover, we propose framework for relevance and redundancy of feature reduction and conduct some experiment to compare information based measures such as correlation, entropy or uncertainty. We observe that information based measures can evaluate the most relevant features. We also witnessed that the feature ranking-based methods are better than subset-based methods in terms of computational complexity, time and accuracy of classification and that feature relevance evaluation using Information Gain measure can enhance the quality of selected features because features selected using mutual information, information gain, gain ratio and entropy can improve the classification accuracy. Therefore, we chose to use individual evaluation based on information measure instead of subset evaluation to reduce high dimension of big data. Table (2) and (3) also shows that individual evaluation will take less than subset evaluation because of reducing iterations in feature selection because we selected features by removing low rank features and using user defined threshold. We will continue implementing our proposed work on big data classification, with hundreds and thousands of features in future.

Table 2. Time taken by Arrhythmia dataset

FS Algos	Search Methods	Time(Seconds) to build model			
		NB	IBK	J48	Random Forest
Cfs	Genetic	0.1	0.01	0.77	2.51
Cons	Genetic	0.02	0.01	0.18	2.06
Cfs	PSO	0.18	0	0.88	2.45
Cons	PSO	0.02	0	0.2	1.85
IG	Ranker	0.15	0	1.57	3.15
SU	Ranker	0.26	0	1.43	3.04
SUSs	FCBF	0.06	0	0.43	2.24
GR	Ranker	0.15	0	1.57	3.15
CorA	Ranker	0.15	0	1.57	3.15
RelF	Ranker	0.12	0	0.95	2.42
PCA	Ranker	0.05	0	0.31	1.79
All	All	0.15	0	1.57	3.15

Table 3. Time taken by mFeat dataset

FS Algos	Search Method	Time (Seconds) to build model			
		NB	IBK	J48	Random Forest
Cfs	Genetic	0.03	0	0.19	0.69
Cons	Genetic	0.02	0	0.33	0.8
Cfs	PSO	0.03	0	0.16	0.63
Cons	PSO	0.03	0.02	0.19	0.67
IG	Ranker	0	0	0.2	0.75
SU	Ranker	0.05	0	0.55	0.97
SUSs	FCBF	0	0	0.1	**0.57**
ChiS	Ranker	0.03	0.01	0.23	0.68
GR	Ranker	0.02	0	0.13	0.62
SVM	Ranker	0.06	0	0.5	0.94
CorA	Ranker	0.06	0	0.52	0.95
CV	Ranker	0.05	0.02	0.42	0.87
RelF	Ranker	0.03	0	0.53	0.91
PCA	Ranker	0.03	0	0.2	0.62
All	All	0.05	0	0.55	0.97

Table 4. Accuracy by mFeat dataset

FS Algos	Search Method	No of Selected features	Naïve Bayes	Ibk(KNN)	J48	Random Forest
Cfs	Genetic	113	93.55%	96.25%	88.30%	96.80%
Cons	Genetic	20	85.30%	92.95%	83.30%	92.65%
Cfs	PSO	109	93.70%	95.90%	89.30%	96.55%
Cons	PSO	18	84.65%	92.50%	82.80%	91.90%
IG	Ranker	217	92.65%	95.85%	88.95%	96.55%
SU	Ranker	211	92.65%	95.95%	88.80%	96.70%
SUSs	FCBF	39	93.50%	96.15%	88.45%	97.00%
GR	Ranker	217	92.65%	95.85%	88.95%	96.55%
CorA	Ranker	217	92.65%	95.85%	88.95%	96.55%
RelF	Ranker	149	92.00%	96.00%	88.85%	96.85%
PCA	Ranker	31	90.35%	94.95%	85.65%	95.25%
All	All	217	92.65%	95.85%	88.95%	96.55%

Table 5. Accuracy by Arrhythmia dataset

FS Algos	Search Method	Selected Features	Naïve Bayes	Ibk(KNN)	J48	Random Forest
Cfs	Genetic	80	77.92%	66.88%	56.49%	73.38%
Cons	Genetic	139	67.53%	54.55%	70.78%	70.78%
Cons	PSO	279	66.23%	50.00%	69.48%	70.78%
Cfs	PSO	24	73.38%	52.60%	70.13%	74.68%
SU	FCBF	12	70.80%	99.78%	85.40%	100%
IG	Ranker	279	74.03%	61.04%	70.78%	77.27%
SU	Ranker	279	74.03%	61.04%	70.78%	77.27%
CorA	Ranker	42	68.83%	54.55%	71.43%	70.13%
RelF	Ranker	154	67.53%	55.19%	68.18%	69.48%
CV	Ranker	149	69.48%	51.95%	69.48%	70.13%
ChiS	Ranker	280	68.83%	59.09%	66.23%	75.32%
GR	Ranker	139	72.08%	53.90%	64.94%	72.73%
PCA	Ranker	19	62.34%	50.65%	62.99%	66.23%
All Features		280	68.83%	59.09%	66.23%	75.32%

References

[1] A. Shahiri, W. Husainmm, 2017. A proposed framework on hybrid feature selection techniques for handling high dimensional

educational data, AIP Conference Proceedings, October.

[2] P. Xing, Eric & Jordan, Michael & Karp, Richard. (2001). Feature Selection for High-Dimensional Genomic Microarray Data. 18th International Conference on Machine Learning.

[3] Fleuret F. 2004, Fast binary feature selection with conditional mutual information, Journal of Machine Learning Research (JMLR) Volume 5, 12/1/2004, Pages 1531-1555.

[4] Hall, MA 1999, Correlation-based feature selection for machine learning, Ph.D. thesis, The University of Waikato, New Zealand.

[5] Hoque, Nazrul & Bhattacharyya, Dhruba K & Kalita, Jugal. (2014). MIFS-ND: A mutual information-based feature selection method. Expert Systems with Applications. 41. 6371–6385. 10.1016/j.eswa. 2014.04.019.

[6] J. Li and H.Liu, 2016, Challenges of Feature Selection for Big Data Analytics, Article in Intelligent Systems, IEEE 32(2), November.

[7] Kira, K & Rendell, LA 1992, A practical approach to feature selection, in Proceedings of the ninth international workshop on Machine learning, Aberdeen, Scotland, UK (pp. 249-256).

[8] Kononenko, I 1994, Estimating attributes: analysis and extensions of RELIEF in the Proceeding of European Conference on Machine Learning, Catania, Italy, pp. 171-182.

[9] Lei Yu leiyu, Huan Liu, 2003. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution(FCBF), Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC.

[10] M. Ali, S. Ahmed, J. Ferzund, A. Mehmood, A. Rehman, 2017. Using PCA and Factor Analysis for Dimensionality Reduction of Bio-informatics Data, in the International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 8, No. 5.

[11] M. Savic, V. Kurbalija, I. Ivanovic, Z.Bosnic, 2017. A Feature Selection Method based on Feature Correlation Networks(FSFCN).

[12] M.Tan, I.Tsang, L.Wang, 2014. Towards Ultrahigh Dimensional Feature Selection for Big Data, Journal of Machine Learning Research, 15(Apr):1371−1429.

[13] S. Wang, J. Tang, H.Liu, 2016. Feature Selection, in Book of Encyclopedia of Machine Learning and Data Mining, Springer Science+ Business Media, New York.

[14] Amiri, Fatemeh, R.Yousefi, et. l, (2011). Mutual information-based feature selection for intrusion detection systems. J. Network and Computer Applications. 34. 1184-1199. 10.1016/j.jnca.2011. 01.002.