International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 03, Issue 09, September 2014 ISSN 2319 - 4847

Survey on Feature Selection in High-dimensional data via Constraint, Relevance and Redundancy

N.Anitha[1] and S.Deepa[2]

[1]PG Scholar, Department of CSE, KalaignarKarunanidhi Institute of Technology, Coimbatore, Tamil Nadu – 641062.

[2] Assistant professor, Department of CSE, KalaignarKarunanidhi Institute of Technology, Coimbatore, Tamil Nadu – 641062.

ABSTRACT

Machine learning provides tools by which large qualities of data can be automatically analyzed. Fundamental to machine learning is feature selection. Feature selection by identifying the most salient features of learning, focuses a learning algorithm on those aspects of the data most useful for analysis for feature prediction. Feature selection as a preprocessing step to machine learning, has been effectively in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving comprehensibility. Most feature selection methods focus on finding relevant features for optimizing high-dimensional data. Dimensionality reduction is a significant task dealing with high dimensional data. Semi supervised clustering aims to improve the clustering performance by considering the pair-wise constraints. Semi supervised feature selection method is most efficient for finding the relevant features, eliminating the redundant features.

Keywords:- Feature selection, Dimensionality reduction, Semi supervised, Constraint, Relevance, and Redundancy.

1. INTRODUCTION

Feature selection has been an active research area in pattern reorganization, statistics and data mining communication. Nowadays a rapid growth of high dimensional data such as digital images, gene expression microarrays, dimensionality reduction has been a fundamental tool for many data mining tasks. According to weather supervised information is available or not, existing dimensionality reduction methods can be roughly categorized into supervised ones and unsupervised ones. Fisher Linear Discriminant (FLD) [10] is an example of supervised dimensionality reduction methods; while Principle Component Analysis (PCA) [10] is an example of unsupervised dimensionality reduction methods. In general domain knowledge can be expressed in diverse forms, such as class labels, pair-wise constraints or other prior information. We focus on domain knowledge in the form of pair-wise constraints, i.e. pairs of instances known as belonging to the same class (must link constraints) or different classes (cannot-link constraints). Pair-wise constraints arise naturally in many tasks such as image retrieval. In those applications considering the pair-wise constraints in more practical than trying to obtain class labels, because the true labels may not be known a priori, while it could be automatically easier for a user to specify whether some pairs of instances belonging to same class or not. Moreover, the pair-wise constraints can derive from labeled data but not vice versa. Furthermore, unlike class labels, the pair-wise constraints can sometimes be automatically obtained without human intervention [11].

1.1 Supervised Feature selection

In machine learning, the classification task described above in commonly referred to as supervised learning. In supervised learning there is specific set of classes and objects are labeled with the appropriate class. The learning is to generalize from the training objects that will enable the novel objects to be identified as belonging to one of the classes. Figure 1 represents the class labels mentioned in the dataset.

Instance #	Features				Class
	Outlook	Temperature	Humidity	Wind	
1	sunny	hot	high	false	Don't play
2	sunny	hot	high	true	Don't Play
3	overcast	hot	high	false	Play
4	rain	mild	high	false	Play
5	rain	cool	normal	false	Play
6	rain	cool	normal	true	Don't Play
7	overcast	cool	normal	true	Play
8	sunny	mild	high	false	Don't Play
9	sunny	cool	normal	false	Play
10	rain	mild	normal	false	Play
11	sunny	mild	normal	true	Play
12	overcast	mild	high	true	Play
13	overcast	hot	normal	false	Play
14	rain	mild	high	true	Don't Play

Figure 1 The golf dataset

1.2 Unsupervised Feature selection

Unsupervised learning knows how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. It can be used to cluster the input data in classes on the

basis of their statistical properties. The labeling can be carried out even if the labels are not available for small number of objects representative of the desired classes. Two classes of methods have been suggested for unsupervised learning. A density estimation technique explicitly builds statistical models of how underlying causes could create the input. Feature extraction techniques try to extract statistical regularities directly from the inputs.

1.3 Semi Supervised Feature selection

Traditional classifiers use only labeled data to train. Labeled instances however often difficult, expensive or time consuming to obtain, as they require the efforts of experienced human annotators. Meanwhile unlabeled data may be relatively easy to collect, but there has been few ways to use them. Semi supervised learning addresses this problem by using large amount of unlabeled data, together with the labeled data to build better classifiers. Because semi supervised learning requires less human effort and give higher accuracy, it is of greater interest both theory and practice. Semi supervised learning is a process of finding a better classifier from both labeled and unlabeled data. The methodology can be used to adapt to a variety of situations by identifying as opposed to specify the relationship between the labeled data and unlabeled data.

1.4 Feature selection in statistical and pattern recognition

In pattern recognition, feature selection can have an impact on the economics of data acquisition and on the accuracy and complexity of the classifier. Feature selection has been shown to improve the comprehensibility of extracted knowledge. Many statistical methods for evaluating the worth of feature subsets based on the characteristics of the training data are only applicable to numeric features.

1.5 SSDR

Here we formulate semi-supervised dimensionality reduction as follows. Given a set of data samples X = [x_1 , x_2, ...,x_n] together with some pair-wise must-link constraints (M) and cannot-link constraints (C), find a set of projective vectors **W** = [w_1 , w_2 ,....., w_d], such that the transformed low dimensional representations $x_i = W^T x_i$ can preserve the structure of the original data set as well as pair-wise constraints M and C, i.e. instances involved by C should be far in the low dimensional space.

Define the objective function as maximizing J (w)

$$J(w) = \frac{1}{2n_c} \sum_{(x_i,x_j)\in C} (x_i - x_j)^2 \qquad (1)$$

Where xi and xj are data samples

1.6 Feature Selection Algorithms

It is simply analyzable or has properties that create approximate optimization easier, for instance, Sub modularity (Nemhauser et al., 1978; Guestrin et al., 2005). Several algorithms rework (1) into a continuous drawback by introducing weights on the scale (Weston et al., 2000; Bradley and Mangasarian, 1998; photographer et al., 2003; Neal, 1998). These ways perform well for linearly divisible problems. For nonlinear issues, however, the improvement sometimes becomes non-convex and a neighborhood optimum doesn't essentially offer smart options. Greedy approaches, like forward choice and backward elimination, are typically won't to tackle it directly. Forward choice tries to extend alphabetic character (T) the maximum amount as doable for every inclusion of options, and backward elimination tries to realize this for every deletion of options (Guyon et al., 2002). Though forward choice is computationally a lot of economical, backward elimination provides higher options normally since the options are assessed inside the context of all others gift.In principle, the Hilbert-Schmidt independence criterion can be employed for feature selection using a weighting scheme, forward selection or backward selection, or even a mix of several strategies. As we shall see, several specific choices of kernel function will lead to well known feature selection and feature rating methods. Note that backward elimination using HSIC (BAHSIC) is a filter method for feature selection. It selects features independent of a particular classifier. Such decoupling not only facilitates subsequent feature interpretation but also speeds up the computation over wrapper and embedded methods.

2. CLUSTERING

Cluster may be a range of comparable objects classified along. It can also be outlined because the organization of dataset into homogeneous and/or well separated teams with relevance distance or equivalently similarity live. Cluster is associate aggregation of points in check area such the space between any 2 points in cluster is a smaller amount than the space between any 2 purposes within the cluster and any point not in it. There square measure 2 kinds of attributes related to bunch, numerical and categorical attributes. Numerical attributes square measure associated with ordered values like height of an individual and speed of a train. Categorical attributes square measure those with unordered values like reasonably a drink and whole of automobile.

Clustering is available in flavors of

- Hierarchical
- Partition (non Hierarchical)

In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a chain of partitions takes place, which may run from a particular cluster containing every objects to n clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which continue by sequence of fusions of the n objects into groups, and divisive methods, which divide in objects sequentially into finer groupings.

2.1 K-Means Clustering

Unsupervised K-means learning algorithms that solve the well known clustering problem. The procedure follows to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a running way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.

3. RELATED WORKS

3.1 Constraint selection for feature selection

The aforesaid scores recorded necessary ends up in sure application scenarios; they need some limitations:

3.1.1 Laplacian score: This score investigates the variance of the information additionally to the section protective ability of the options. Hence, a "good" feature for this score is that the one at that 2 neighboring examples record shut values. However, this score doesn't cash in on the background info (the CL constraints in particular), that area unit provided to guide the educational method. Additionally, the neighborhood selection isn't clearly outlined, that is, the range of (k) selections has important effects on results.

3.1.2 Constraint score: Utilizing few labels of knowledge, this score recorded higher results than Fisher score that employs the entire labeled in feature choice method. Even so, this score has many drawbacks:

1. It just exercises the labeled data in the feature selection; such vital restriction may mislead the learning process, especially in a semi-supervised context, where the unlabeled party is normally larger than the labeled one.
2. As this score depends merely on the chosen constraint subset. The choice of constraints is still a problematic issue, which could derogate the performance of the feature selection process.

In order to overcome the listed restrictions, we propose an approach that will:

- Deploy a constraint selection in order to select the coherent subset of pair-wise constraints extracted from the labeled data.
- Utilize the data structure in the definition of the neighborhood between examples.

3.1.3 Constraint Selection

While it had been expected that totally different constraints sets would contribute a lot of or less in up bunch accuracy, it had been found that some constraints sets really decrease bunch performance. it had been discovered that constraints will have unwell effects even once they square measure generated from the information labels that square measure accustomed assess accuracy, thus this behavior isn't caused by noise or errors within the constraints. Instead, it's a result of the interaction between a given set of constraints and also the formula being employed. Thus it's a lot of vital to understand why do some constraint sets increase bunch accuracy whereas others don't have any impact or perhaps decrease accuracy. For that, the authors in [7] have outlined 2 vital measures, informativeness and coherence, that capture relevant properties of constraint sets. These measures offer insight into the impact a given constraint set has for a selected unnatural bunch formula. During this paper, we tend to solely use the coherence live that is freelance of any learning. Nowadays feature selection research has focused on searching for relevant features. Although some recent work has pointed out the existence and effect of feature redundancy (koller and sahami, 1996; kohavi and john, 1997; Hall, 2000), there is little work on explicit treatment of feature redundancy.

3.2 Feature Relevance

Based on a review of previous definitions of feature relevance, john, kohavi and pfleger classified features into three disjoint categories, namely, strongly relevant, weakly relevant and irrelevant features (John et al., 1994). Let F be a full set of features, F_i a feature, and $S_i = F- \{F_i\}$. These categories of relevance can be formalized as follows.

Definition 3.2.1 (Strong relevance) A feature Fi is strongly relevant if

$$P(C \mid F_i, S_i) \neq P(C \mid S_i). \quad (2)$$

Definition 3.2. 2 (Weak relevance) A feature Fi is weakly relevant if

$$P(C \mid F_i , S_i) = P(C \mid S_i) \quad (3)$$

Corollary 3.2. 1 (Irrelevance) A feature Fi is irrelevant if

$$\forall S_i' \subseteq S_i, P(C \mid F_i, S_i') = P(C \mid S_i') \quad (4)$$

3.3 Existing approaches of Feature Relevance

As mentioned earlier, there exist two major approaches in feature selection: individual evaluation and subset evaluation. Individual evaluation, also known as feature weighting/ranking (Blum aand Langley, 1997; Guyon and Elisseeff, 2003), assesses individual features and assigns them weights according to their degree of relevance. A subset of features is often selected from the top of a ranking list, which approximates the set of relevant features.

3.4 Feature Redundancy

Feature Redundancy naturally correlated to feature correlation measures. Correlation is widely used in machine learning and statistics for relevance and redundancy analysis. It is implementing the largest dependency condition; we have a tendency to 1st derive constant kind, known as minimal-redundancy-maximal-relevance criterion (mRMR), for first-order progressive feature choice. Then, we have a tendency to gift a two-stage feature choice algorithmic rule by combining mRMR and different a lot of refined feature selectors (e.g., wrappers). This enables U.S. to pick a compact set of superior options at terribly low price. We have a tendency to perform in depth experimental comparison of our algorithmic rule and different strategies exploitation 3 completely different classifiers (naive Thomas Bayes, support vector machine, and linear discriminate analysis) and 4 completely different knowledge sets (handwritten digits,

arrhythmia, NCI neoplastic cell lines, and cancer tissues). The results ensure that mRMR ends up in promising improvement on feature choice and classification accuracy.

3.5 Correlation Measures

There exist broadly two types of measures for the correlation between two random variables: linear and non – linear. Of the linear correlation, the most well known measure is linear correlation coefficient.

For a pair of variables (X,Y), the linear correlation coefficient ρ is given by

$$\rho = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2}\sqrt{\sum_i (y_i - \overline{y_i})^2}} \quad (5)$$

Where (x_i)¯ is the mean of X, and (y_i)¯ is the mean of Y. The value of ρ lies between -1 and 1, inclusive.

4. CONCLUSION

In this paper it describes about the Feature selection based on the relevance and redundancy analysis. A Feature selection method for high dimensional data can be extracted by partitioning the data, which overcome the irrelevant and non redundant features. SSDR algorithm is implemented for dimensionality reduction. To identify the relevant features constrained Laplacian algorithm is more efficient via spectral graph analysis and eliminating redundant features by correlation measures via maximum spanning tree method for optimized and accurate feature selection.

REFERENCES

[1] Z. Zhao and H. Liu, Spectral Feature Selection for Data Mining (Data Mining and Knowledge Discovery Series).BocaRaton, FL, USA: Chapman and Hall-CRC, 2012.
[2] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by local linear embedding,” Science, vol. 290, no. PP.5500,2323–2326, Dec. 2000.
[3] B. Scholkopf, A. Smola, and K. R. Muller, “Nonlinear component analysis as a Kernel Eigenvalue problem,” NeuralComput., vol. 10, no. 5, pp. 1299–1319, 1998.
[4] I. Guyon and A. Elisseeff, “An introduction to variable and fea-ture selection,” J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Mar. 2003.
[5] K. Benabdeslem and M. Hindawi, “Constrained Laplacian score for semi-supervised feature selection,” in Proc, ECML-PKDD, Athens, Greece, 2011, pp. 204–218.
[6] K. Allab and K. Benabdeslem, “Constraint selection for semi-supervised topological clustering,” in Proc. ECML-PKDD, Athens, Greece, 2011, pp. 28–43.
[7] I. Davidson, K. Wagstaff, and S. Basu, “Measuring constraint-set utility for partitional clustering algorithms,” in Proc. ECML/PKDD, 2006.
[8] M. Hindawi, K. Allab, and K. Benabdeslem, “Constraint selection based semi-supervised feature selection,” in IEEE ICDM, Vancouver, BC, Canada, 2011, pp. 1080–1085.