ISSN:2348-2079

Volume-6 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Dimensionality Reduction and Data Partitioning with Feature Hybridization Scheme

[1]**Mrs. M.S. Vinu,** Assistant Professor,
[1]Department of Computer Science and Engineering,**VSB College of Engineering and Technical Campus, Coimbatore.**
Mail ID: vinuja@gmail.com

Abstract

Data mining and machine learning methods are applied to extract knowledge from large databases. Dimensionality is the key issue in the data mining and machine learning applications. The high dimensional data analysis requires huge computational resources and processing time. The performance and accuracy are reduced with reference to the irrelevant, noisy and redundant features. Dimensionality methods are applied for better visualization, data compression, noise removal, understandability and generalization factors. Text mining, web mining, image processing and bioinformatics applications are build with dimensionality reduction methods.

Dimensionality reduction is carried out with two models Feature Selection (FS) and Feature Extraction (FE). Feature selection discovers the suitable features from the original set of features. The feature extraction method transforms the original set of features into required form. The compound feature generation (CFG) model integrates the feature selection and extraction methods to fetch the original and transformed features. The Minimum Projection error Minimum Redundancy (MPeMR) framework is build with Unified iterative algorithm to fetch features in supervised and unsupervised cases.

The Compound Feature generation (CFG) method is build with pairs of features in minimum projection error and redundancy estimation process. The feature hybridization scheme is build to combine the original and transformed features with generalized matching criteria. The feature integration operation is improved with diverse feature count based models. The data partitioning process is carried out on the dimensionality reduced data with K-Means clustering algorithm.

Index Terms: High Dimensional Data, Dimensionality Reduction, Feature Selection, Feature Extraction and Data Clustering Methods.

1. Introduction

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

The high dimensionality of data poses challenges to learning tasks such as the curse of dimensionality. In the presence of many irrelevant features, learning models tend to over fitting and become less comprehensible. Feature selection is one effective means to identify relevant features for dimension reduction. Various studies features can be removed without performance deterioration. The training data can be either labeled or unlabeled, leading to the development of supervised and unsupervised feature selection algorithms.

To date, researchers have studied the two types of feature selection algorithms largely separately. Supervised feature selection determines feature relevance by evaluating feature's correlation with the class, and without labels, unsupervised feature selection exploits data variance and separability to evaluate feature relevance. In this paper, we endeavor to investigate some intrinsic properties of supervised and unsupervised feature selection algorithms, explore their possible connections, and develop a unified framework that will enable us to (1) jointly study supervised and unsupervised feature selection algorithms, (2) gain a deeper understanding of some existing successful algorithms, and (3) derive novel algorithms with better performance. To the best of our knowledge, this work presents the first attempt to unify supervised and unsupervised feature selection by developing a general framework. The chasm between supervised and unsupervised feature

selection seems difficult to close as one works with class labels and the other does not. If we change the perspective and put less focus on class information, both supervised and unsupervised feature selection can be viewed as an effort to select features that are consistent with the target concept. In supervised learning the target concept is related to class affiliation, while in unsupervised learning the target concept is usually related to the innate structures of the data. Essentially, in both cases, the target concept is related to dividing instances into well separable subsets according to different definitions of the separability. The challenge now is how to develop a unified representation based on which different types of separability can be measured [1]. Pairwise instance similarity is widely used in both supervised and unsupervised learning to describe the relationships among in- stances. Given a set of pairwise instance similarities S, the separability of the instances can be studied by analyzing the spectrum of the graph induced from S.

For feature selection, therefore, if we can develop the capability of determining feature relevance using S, we will be able to build a framework that unifies both supervised and unsupervised feature selection. Based on spectral graph theory, we present a unified framework for feature selection using the spectrum of the graph induced from S. By designing different S's, the unified framework can produce families of algorithms for both supervised and unsupervised feature selection.

2. Related Work

This section introduces the related work in the areas of 1) temporal mining of social media; 2) event detection and forecasting; 3) supervised and unsupervised learning; and 4) multitask learning.

Temporal mining of social media: In recent years, much attention has been paid to this area, which focuses on modeling the temporal pattern such as evolutional publish sentiment, dynamic topic, online collaborative environments and information diffusion. Tan et al. proposed two topic models that leverage lexicon based knowledge to characterize the variations of the public sentiment. Zhao et al. developed a framework that can track themes of targeted domain dynamically utilizing the heterogeneous links such as co-occurrence, friendship, authorship, and replying. Guan et al. proposed a method for locating appropriate expert on relevant knowledge by modeling and identifying people's knowledge based on their web activities [3]. Zhang et al. leverage triadic structures to investigate the formation of other neighboring links triggered by "following" links.

Event detection: A large body of work focuses on the identification of ongoing events, including earthquakes, disease outbreaks and other types of events [4]. In general, these researchers use either classification or clustering to extract tweets of interest and then examine the spatial, temporal, or spatiotemporal burstiness the extracted tweets. Instead of forecasting events in the future, these approaches typically uncover them only after their occurrence.

Event forecasting: Most research in this area focuses on temporal events and ignores the underlying geographical information. This approach is generally used for events such as the forecasting of elections, stock market movements, disease outbreaks and crimes [5]. These studies can be grouped into three categorizes: 1) Linear regression model, where simple features, such as tweet volumes, are utilized to predict the occurrence time of future events; 2) Nonlinear models, where more sophisticated features such as topic-related keywords are used as the input to build forecasting models using existing methods such as support vector machines or LASSO and 3) Time series-based methods, where methods such as autoregressive models are used to model the temporal evolution of event-related indicators. Few existing approaches can provide true spatiotemporal resolution for predicted events.Wang et al. developed a spatiotemporal generalized additive model to characterize and predict spatio-temporal criminal incidents, but their model requires demographic data. Ramakrishnan et al. built separate LASSO models for different locations to predict the occurrence of civil unrest events. Zhao et al. also designed a new predictive model based on topic models that jointly characterize the temporal evolution for both the semantics and geographical burstiness of social media content.

Supervised approaches: They involve considering a set of stationary terms whose distribution can be learned from historical data [6]. For example, LASSO regression methods estimate a sparse predictive model based on a predefined set of keyword terms for each location that predicts the probability of an ongoing event in this location in each predefined time interval. Similarly, burst detection methods search for geographic regions the aggregated counts of certain predefined terms are abnormally high compared with the counts for the same terms outside those cities. For example, Sakaki et al. utilize spatiotemporal Kalman filtering, which is similar to space-time burst detection, to track the geographical trajectory of hot spots of tweets related to earthquakes.

Unsupervised approaches: They utilize a set of dynamic terms that could be different in different time intervals, and apply unsupervised learning techniques for event detection [9]. The dynamic query expansion method (DQE) iteratively expands a predefined set of seed terms using current tweets to identify and rank new terms that are relevant to ongoing events, then retain the top terms and tweets containing these terms for further modeling. Clustering-based methods search

for novel spatial clusters of documents or terms using predefined similarity metrics, such as cosine similarity and social similarity for documents, or auto-correlations and co occurrences for terms.

Multi-task learning: Multi-task learning (MTL) models multiple related tasks simultaneously to improve generalization performance. Many MTL approaches have been proposed in the past. Evgeniou et al. proposed a regularized MTL that constrained the models of all tasks to be close to each other. The task relatedness can also be modeled by constraining multiple tasks to share a common underlying structure, e.g., a common set of features, or a common subspace [7]. MTL approaches have been applied in many domains, including computer vision and biomedical informatics. To the best of our knowledge, ours is the first work that applies MTL for civil unrest forecasting.

3. Compound Feature Generation Scheme

In recent years, high dimensional data sets have become very common in machine learning and data mining applications. Processing of such data sets requires huge computational time and resources. Moreover, with the presence of irrelevant, redundant and noisy features, the performance of the learning algorithm degrades. It is crucial to reduce the dimensionality of the data to improve both the efficiency and effectiveness of most of the data mining algorithms [2] . Also it is important for better visualization, data compression, noise removal, improved understanding ability and generalization of the learning algorithms. Traditional and state-of-the-art dimensionality reduction methods fall into two categories: feature selection and feature extraction. These approaches have been successfully applied in many real applications, such as Image processing, text categorization, bioinformatics, etc. Feature selection aims at finding a subset of most useful features from the original set of features, whereas feature extraction methods provide combinations original features.

In the past few decades, these two approaches have been studied extensively. All the studies have been done separately or independently. Although the ultimate aim of both the approaches is to improve the efficiency of a learning algorithm, none of the feature selection methods provide even a single combination of features may be more informative than the original features. A feature extraction approach provides transformed features, where each transformed feature is a combination of all original features, and no original feature appears among the transformed features. If these two approaches can be integrated in a systematic way, to provide reduced set with both types of features, they could complement each other. Two synthetic data sets in 3D are provided to show the effectiveness of having both types of features. For these data, when the dimension is reduced from three to two, one original and one combination of features (here the combination does not involve all of the original features) produce better representation, than having either two original features or two transformed features. So, there must exist methods where the final result will be a few original features and a few linear combinations.

Many algorithms for feature selection/extraction have been suggested in the literature. The main idea of feature selection is to choose a subset of original features by eliminating features with little or no predictive information. With respect to whether the label information is available, different methods for feature selection can be divided into supervised, unsupervised, or semi supervised algorithms. In supervised feature selection algorithms, important features are determined by estimating their correlation with the class labels or their performance in prediction. Unsupervised feature selection algorithms select features by exploiting data variance or distribution. In a semi-supervised feature selection algorithm, small amount of labeled data is used as additional information to improve the performance of the unsupervised feature selection algorithm. Based on different selection strategies used, methods for feature selection can be categorized into three groups, filter, wrapper and embedded methods. Filter algorithms evaluate features using certain statistical criteria and independent of any classifier. On the contrary, wrapper methods select a set of features based on a selection criteria with respect to a given classifier, such as: Bayes, Knn, SVM. Wrapper methods in general are more computationally expensive and hence, for real-life applications with large data sets, the filter model is more popular. The wrapper model has been empirically proven to be superior, in terms of classification accuracy, to a filter model. Finally, the embedded method achieves model fitting and feature selection simultaneously. In addition, feature selection algorithms can also be categorized as subset selection algorithms returns a subset of selected features or feature weighting algorithms returns weight corresponding to each feature.

Feature extraction, linearly or non-linearly, transforms the original high dimensional data to a low dimensional data. The objective of feature extraction is to find an appropriate transformation that maps the original D-dimensional space to a new d-dimensional feature space, where $d << D$. According to the availability of the class label information, feature extraction methods are categorized into supervised or unsupervised methods. They are also broadly divided into linear and non-linear methods. Linear feature extraction seeks a meaningful low dimensional subspace in a high dimensional input space by linear transformation. Among all the linear feature extraction methods, the most well

known are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). PCA seeks a transformation to produce uncorrelated and orthogonal principal components and LDA produces a transform while preserving as much class discriminatory information as possible. Other unsupervised feature extraction methods are:

Factor Analysis (FA), projection pursuit, Independent Component Analysis (ICA), etc. Some of the well known supervised feature extraction methods are: Maximum Margin Criterion (MMC), Angular Liner Discriminant Embedding (ALDE), etc. Transformed features of these methods usually contain all the original variables in their linear combinations which may be difficult to interpret.

To overcome this drawback, sparse principal component analysis (SPCA), sparse linear discriminant analysis (SLDA) is introduced to produce modified principal components which just contain a few original variables. However, unlike PCA, sparse PCA cannot guarantee that different principal components are uncorrelated. Some methods which study feature selection and extraction together exists in the literature. A general transformation-based dimensionality reduction algorithm has been converted in to a feature selection formulation. A joint framework to do feature selection and subspace learning simultaneously based on using L2;1- norm on the projection matrix, which leads to selecting relevant features and learning transformation simultaneously.

All the methods listed above provide reduced set either with original or transformed features, but not both of them. We propose to bridge the gap between feature selection and extraction approaches, which exists as one provides original and the other provides transformed features. We study these two methods together with the aim of obtaining a reduced feature set to contain both kinds of features. An approach for dimensionality reduction where linear combinations of features are considered, and orthogonality is maintained on selected linear combinations of features and original features is suggested. We also present an approximation algorithm under this framework.

4. K-Means Clustering Algorithm

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets, so that the data in each subset share some common trait - often proximity according to some defined distance measure [8]. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. It is possible to guarantee that homogeneous clusters are created by breaking apart any cluster that is unhomogeneous into smaller clusters that are homogeneous.

The K Means clustering algorithm is applied for the postulated in the nineteen sixties. For a *m* attribute problem, each instance maps into a *m* dimensional space. The cluster centroid describes the cluster and is a point in *m* dimensional space around which instances belonging to the cluster occur. The distance from an instance to a cluster center is typically the Euclidean distance though variations such as the Manhattan distance are common. As most implementations of K-Means clustering use Euclidean distance.

- ✓ **Strength of the K-Means:**
  - Relatively efficient: O (t kn), where n is of objects, k is of clusters and t is of iterations. Normally, k, t << n
  - Often terminates at a local optimum
- ✓ **Weakness of the K-Means:**
  - Applicable only when mean is defined; what about categorical data?
  - Need to specify k, the number of clusters, in advance
  - Unable to handle noisy data and outliers
- ✓ **Variations of K-Means usually differ in:**
  - Selection of the initial k means
  - Dissimilarity calculations
  - Strategies to calculate cluster means
- ✓ **Partitioning Methods**
  - Reallocation method - start with an initial assignment of items to clusters and then move items from cluster to cluster to obtain an improved partitioning
  - It involves movement or "reallocation" of records from one cluster to other to create best clusters. It uses multiple passes through the database fastly.
  - Single Pass method - simple and efficient, but produces large clusters and depends on order in which items are processed
  - The database must be passed through only once in order to create clusters

General algorithm for a Single-Pass technique:

Step 1: Read in a record from the database and determine the cluster that is best fits into.

Step 2: If the nearest cluster is still pretty for away to create a new cluster with this new record in it.

Step 3: Read in the next record.

Reallocation Method

Algorithm 1:

Step 1: Select K data points as the initial representatives.

Step 2: for i = 1 to N, assign item xi to the most similar centroid

Step 3: for j = 1 to K, recalculate the cluster centroid Cj

Step 4: Repeat steps 2 and 3 until there is (little or) no change in clusters

Algorithm 2:

Step 1: Pre-select the number of clusters desired

Step 2: Randomly pick a record to become the center or "seed" for each of these clusters

Step 3: Go through the database and assign each record to the nearest cluster.
Step 4: Recalculate the centers of the clusters.
Step 5: Repeat steps 3 and 4 until there is a minimum or no change in clusters.

5. Conclusion

The high dimensional data values are processed with dimensionality reduction schemes for mining operations. A new strategy for Dimensionality Reduction with the aim of providing reduced set with both original and combinations of features is studied. For this purpose, a framework MPeMR is adapted to generate orthogonal compound features by minimizing both projection error and redundancy [10]. An iterative approximation method under the proposed framework for compound feature generation, without losing orthogonality property, is also utilized. The clustering process is carried out using the K-Means clustering algorithm with feature selection and extraction process. The clustering process is improved with accuracy levels in feature based data partitioning process.

References

1. Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu and Naren Ramakrishnan,"Feature Constrained Multi-Task Learning Models for Spatiotemporal Event Forecasting", IEEE Transactions on Knowledge and Data EngineeringMay 2017
2. Sreevani and C. A. Murthy "Bridging Feature Selection and Extraction: Compound Feature Generation", IEEE Transactions on Knowledge and Data Engineering, April 2017.
3. Z. Zhao, L. Wang, H. Liu and J. Ye, "On similarity preserving feature selection," Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 3, pp. 619–632, 2013.
4. D. Wang, F. Nie and H. Huang, "Global redundancy minimization for feature ranking," Knowledge and Data Engineering, IEEE Transaction on, 2015.
5. D. Cai, C. Zhang and X. He, "Unsupervised feature selection for multi-cluster data," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010, pp. 333–342.
6. M. Qian and C. Zhai, "Robust unsupervised feature selection," in Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press, 2013, pp. 1621–1627.
7. Z. Xu, I. King, M. R.-T. Lyu and R. Jin, "Discriminative semisupervised feature selection via manifold regularization," Neural Networks, IEEE Transactions on, vol. 21, no. 7, pp. 1033–1047, 2010.
8. K. Benabdeslem and M. Hindawi, "Efficient semi-supervised feature selection: Constraint, relevance, and redundancy," Knowledge and Data Engineering, IEEE Transactions on, vol. 26, no. 5, pp. 1131–1143, 2014.
9. Q. Song, J. Ni and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 1, pp. 1–14, 2013
10. S. Liu, L. Feng, and H. Qiao, "Scatter balance: an angle-based supervised dimensionality reduction," Neural Networks and Learning Systems, IEEE Transactions on, vol. 26, no. 2, pp. 277–289, 2015