Journal of Computational Science 26 (2018) 141–146

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Dimensionality scale back in massive datasets using PDLPP

Jasem M. Alostad

College of Basic Education, The Public Authority of Applied Education and Training, Safat 13092, Kuwait

ARTICLE INFO

Article history:
Received 21 December 2017
Received in revised form 11 April 2018
Accepted 19 April 2018
Available online 22 April 2018

Keywords:
Non-orthogonality
Dimensionality reduction
Distance-based locality preserving projections
Normalized mutual information

ABSTRACT

The main objective of this paper is to reduce data dimensionality in high-dimensional feature datasets. It uses an effective distance based Non-integer Matrix Factorization (NMF) method to resolve the problem of data dimensionality. The non-orthogonality arising due to increasing dimensionality is resolved using NMF and an effective distance measurement. This process involves organizing the datasets to form a defined geometric structure since conventional dimensionality reduction principles capture the structured data using a similarity matrix with a distance-based measurement. However, such distance-based measurements cannot fit dynamic data structure to the model and most of the intrinsic structure of the data is ignored. Hence, to avoid this complexity, the proposed method uses Probabilistic Distance Locality Preserving Projections (PDLPP) to structure the dynamic data. The proposed method is evaluated against the conventional methods in terms of its accuracy and normalized mutual information over different test cases. The proposed method increases the performance of learning the patterns in high dimensional data with less computation time. The results demonstrated that the proposed method fits well with static and dynamic data to query the objects in the search space.

1. Introduction

Data reduction reduces the data dimensionality, and retains the data representation. The data reduction is selected to reduce the instances of a given dataset. In spite of many efforts to deal with such instances, data mining algorithms experience severe challenges due to the non-applicability of datasets to large instances. Hence, the computational complexity of the system increases with larger instances and leads to problems in scaling increased storage requirements and clustering accuracy. The other problems associated with larger data instances include: improper association or interaction in the feature space; lack of ability to handle the large datasets with discrete variables; inability to classify the data and poor knowledge generation for a given query; and, finally, poor computation due to missing variables or low dimensional features in high dimensional datasets.

The major aim of this paper is to propose a novel probabilistic distance based LPP method to reduce the dimensionality of the given datasets. This method uses Non-integer Matrix Factorization (NMF) to remove the low dimensional features in the given datasets. This probabilistic distance estimation computes the distances based on the probability of occurrence between the two different data samples placed on a graph node.

The major contribution of our proposed method involves:

1. Estimation of similarity between the data samples using probabilistic distance based representation.
2. We introduce NMI to remove the lower dimensional features so as to reduce the data dimensionality.
3. We have introduced discriminative based LPP to structure the data dimensions using weighted estimation of probabilistic distance estimation.
4. This method has been tested over three different document datasets to validate the accuracy and normalized mutual information (NMF) of DBLPP.

Organization: The paper is further presented as follows: Section 2 examines the related works with high dimensional data and LPP; Section 3 explores Locality Preserving Projection and Section 4 Distance based Similarity Measurement; Section 5 provides modifications to the proposed work using the DLPP method; Section 6 evaluates the proposed work with related datasets; Finally, Section 7 concludes the paper with an indication of future work.

2. Related works

2.1. High dimensional data

This section discusses list of existing methods used to reduce the high dimensional dataset. Data dimensionality is reduced by Simão et al. [1], who proposed principal component analysis (PCA) with bi-cubic interpolation, which allows re-sampling of raw data

E-mail address: jm.alostad@paaet.edu.kw

https://doi.org/10.1016/j.jocs.2018.04.016
1877-7503/© 2018 Elsevier B.V. All rights reserved.

[1]. Tunga [2] used multi-variance product representation and a k-means clustering algorithm to overcome the difficulties of high dimensional data due to high multi-variance [2]. Zhu and Xue [3] proposed an orthogonal distance based method to improve the relationship between the data and to avoid inference of data with high dimensionality [3]. Liu et al. [4] tested the covariance matrix and mean vector in high dimensional datasets using asymptotic null distribution to improve the unbiased asymptotic datasets [4]. Lucas et al. [5] proposed Discriminative Pattern (DP) mining to improve the data characteristics in high dimensional datasets [5]. Zhao et al. [6] proposed improving the accuracy of predicting the high dimensional dataset by using an improved Ando and Li [46] method [6]. Zamora et al. [7] proposed locality-sensitive hashing (LSH) using MinHash with Jaccard similarity approximations, and SimHash with cosine similarity approximation. This method focuses on clustering the high dimensional datasets with incomplete information [7]. Ultsch, and Lötsch [8] proposed emergent self-organizing feature maps to distribute the high dimensional datasets to form a cluster structure [8]. Sang et al. [9] improve the geometric topology of high dimensional data using discretization and a supervised area based chi-square discretization algorithm [9]. Apiletti et al. [10] proposed a Map Reduce-based frequent closed item set mining algorithm (PaMPa-HD) to improve the mining of high dimensional datasets with hidden and non-trivial patterns [10]. Zhou et al. [11] proposed Online Feature Selection based on the Dependency in K nearest neighbors using Rough Set theory to select the high dimensional data features [11]. Lansangan et al. [12] proposed a constrained optimization method with variable selection and dimension reduction in high dimensional data [12]. Jing et al. [13] proposed stratified sampling that samples the features of high dimensional data in a random way [13]. Liu and Li [14] proposed integrated constraint based clustering to address the problems related to the selection, weighting of data dimensions and assignment of data [14]. Cardoso et al. [15] proposed Iterative random projections to increase the data dimensionality and reduce the time complexity over each k-means convergence [15]. Moayedikia et al. [16] proposed a SYMON feature selection method with harmony search and symmetrical uncertainty to avoid misclassification in high dimensional datasets [16]. Pedergnana and García [17] avoid the challenges associated with data intensive algorithms to handle the regression data optimally using a comprehensive systematic approach over high dimensional data [17]. Itoh et al. [18] proposed graph visualization over well-correlated dimensions to construct the user selected subsets [18]. Chang and Yang [47] proposed a novel semi supervised feature selection framework by mining correlations among multiple tasks. This method leverages the knowledge of multiple related tasks and improves the performance of feature selection. Zhihui Li et al. [48] proposed a Linear discriminant analysis (LDA) supervised linear dimensional reduction model that learns low-dimensional representation from high-dimensional feature space using a transformation matrix and preserves the discriminative information through a between-class scatter matrix and reduces the within class scatter matrix. This method reduces the high dimensional datasets more effectively.

2.2. Reviews related to locality preserving projections (LPP)

This section discusses the conventional LPP methods used to improve the feature extraction and dimensionality reduction. Lu et al. [19] and Huang and Zhuang [22] proposed a feature extraction using matrix exponential discriminant LPP. This method avoids a small sample size, which is lower than sample dimensions using symmetric matrix exponentials. Wang et al. ([37] proposed exponential LPP, Shikkenawis and Mitra [20] proposed Extended LPP with a supervised variant to discover the low dimensional manifold to reduce data dimensionality [20]. Xu et al. [21] proposed coupled LPP to attain higher classification rates to preserve the low dimensional manifold [21]. Jiang et al. [23] proposed anchor graph based LPP, Zhang et al. [24] proposed sparse discriminative representation based LPP, Zheng et al. [25] regression with sparse penalty LPP with compressive sensing theory, Chen et al. [26] and Xu et al. [39] proposed 2D discriminant LPP using L_1 matrix norm maximization, Dornaika and Assoum [27] parameter less LPP with affinity matrix, supervised LPP, Orthogonal LPP [27], Zhang et al. [28] proposed sparsity LPP and α-regularization sparsity LPP, Wang et al.[29] proposed novel level set with shape priors using LPP, Guo et al. [30] proposed a least square support vector machine with kernel LPP, Chen et al. [31] proposed optimal LPP using singular Eigen computation, Lu et al. [32] proposed discriminant LPP using a maximum margin criterion, He et al. [33] proposed statistics LPP with statistics pattern analysis using non-Gaussian properties, Wong et al. [34] proposed supervised optimal LPP (SOLPP) and normalized Laplacian SOLPP, Zhang, et al. [35] proposed graph optimized LPP using a novel DR algorithm, Chao et al. [36] LPP using local binary and class regularization patterns, Lu and Tan [38] proposed improved discriminant LPP, Rangarajan [42] proposed diagonal and secondary diagonal LPP, Weng and Shen [40] proposed multivariate time series LPP and, finally, Yu [41] proposed degradation assessment LPP for improving the dimensionality features in a given dataset, which can either be textual or image-based.

These are the conventional works intended to reduce the data dimensionality in large databases. It is seen that these works effectively reduces the data dimensionality by eliminating the unwanted instances from large database. Hence, the proposed method involves a design of new variant in LPP, which removal the unwanted instances from large databases.

3. The locality preserving projection

The LPP is an unsupervised technique to reduce the dimensionality of the data in large mining datasets. The manifold structures of large dimensional data sets are handled better than in principle component analysis. The adjacent graph is constructed using a k nearest neighbor algorithm to preserve the local structure of the datasets.

Consider a sample data x_i and x_j are located at the nearest distance using k-nearest neighbor, then an edge is added between the sample data and hence the weights between x_i and x_j are computed using,

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}} \tag{1}$$

The similarity matrix is obtained based on $S = \left\{S_{ij}\right\}_{i,j=1}^{N}$, which is used to relate the similarity between the samples (N). When the two samples, namely, x_i and $x_{j,}$ lie as neighbors in the original subspace, then the new samples y_i and y_j lie close in the new subspace. Hence, the projection vector (a) can be calculated as,

$$0.5\sum_{ij}\left(y_i - y_j\right)^2 S_{ij} = 0.5\sum_{ij}\left(x_i a^T - x_j a^T\right)^2 S_{ij} \tag{2}$$

where, $y_i = x_i\, a^T$ with a sample matrix (X), where $i = 1,2,\ldots.., N$.

Further, the diagonal matrix (D) is multiplied by Eq. (2) to obtain the following relation using a Laplacian matrix, which is given by,

$$\begin{aligned} &0.5\sum_{ij}\left(y_i - y_j\right)^2 S_{ij} \\ &= \sum_{i}\left(x_i a^T D_{ij} a x_i^T\right) - \sum_{ij}\left(x_i a^T S_{ij} a x_j^T\right) \\ &= X\left(D - S\right) a^T X^T a \\ &= X a^T X L a \end{aligned} \tag{3}$$