International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 1, February 2019, pp. 593~601
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i1.pp593-601

Improved probabilistic distance based locality preserving projections method to reduce dimensionality in large datasets

Jasem M. Alostad
College of Basic Education, The Public Authority of Applied Education and Training (PAAET), Kuwait

Article Info

Article history:

Received Jan 21, 2018
Revised Aug 2, 2018
Accepted Aug 18, 2018

Keywords:

Large dimensional datasets
Locality preserving projection
Mutual information
Non-orthogonality

ABSTRACT

In this paper, a dimensionality reduction is achieved in large datasets using the proposed distance based Non-integer Matrix Factorization (NMF) technique, which is intended to solve the data dimensionality problem. Here, NMF and distance measurement aim to resolve the non-orthogonality problem due to increased dataset dimensionality. It initially partitions the datasets, organizes them into a defined geometric structure and it avoids capturing the dataset structure through a distance based similarity measurement. The proposed method is designed to fit the dynamic datasets and it includes the intrinsic structure using data geometry. Therefore, the complexity of data is further avoided using an Improved Distance based Locality Preserving Projection. The proposed method is evaluated against existing methods in terms of accuracy, average accuracy, mutual information and average mutual information.

Corresponding Author:

Jasem M. Alostad,
College of Basic Education,
The Public Authority of Applied Education and Training,
PO.Box 23167, Safat 13092, Kuwait.
Email: jm.alostad@paaet.edu.kw

1. INTRODUCTION

In recent years, large dimensional datasets have been generated in the presence of uncertainty, and they have been increasingly used in several applications like environmental monitoring, sensor networks, data cleaning, moving object management and data integration. The presence of uncertainty in large dimensional datasets is due to imprecise measurement, unreliable data transfer, privacy protection, repeated sampling and so on [1]. These applications create a demand for effective management of large dimensional datasets and their processing, which is the major issue in large database systems [2].

The data reduction [3] in large datasets reduces the data dimensionality and retains the data t. In spite of many ges due to the non-applicability of datasets with 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 [4]. 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 or feature selection [5] in high dimensional datasets [6,7].

There are several dimensionality reduction techniques [8-26, 27] dealing with high-dimensional data [26]. The common strategy in all the literature includes reduction of dimensionality which is based on the variations in their class labels [27]. Hence, to boost the performance of learning in classification systems and to address the above problems, an effective unsupervised model is needed for eliminating the large dimensional datasets. This is usually carried out through the reduction or elimination of unwanted features

from the datasets [28-30]. Feature reduction in clinical data set is discussed in [31] and [32] explains the dimensionality reduction in kernel PCA.

In this paper, we propose an Improved Distance based Locality Preserving Projections (IDLPP) technique for reducing the datasets which possess high dimensionality. The notion of the proposed system was inspired by the idea of LPP. In this paper NMF is used for eliminating the low dimensional features. The distance estimation is computed using a probabilistic distance measurement, which represents the estimation of probability between two different data samples.

The main contributions involve the following:

1. The proposed solution finds the similarity between the data samples using squared distance representation.
2. The low dimensional features are eliminated using NMF.
3. Finally, the proposed IDLPP technique is compared against other LPP methods using accuracy, average accuracy, NMI and average NMI.

The outline of the paper is as follows: Section 2 gives the outline of LPP with the data partitioning technique, NMF metric estimation. Section 3 discusses the similarity measurement based on distance between the nodes. Section 4 evaluates the IDLPP with other LPPN experimentally and the results are discussed. Finally, section 5 concludes the paper.

2. LOCALITY PRESERVING PROJECTIONS

The LPP as an unsupervised method is used as a dimensionality reduction technique in data mining with larger datasets. This method handles the structure of such datasets in a better manner than principle component analysis. Further, the local dataset structure is preserved through the construction of adjacent graphs using the *k* nearest neighbor algorithm.

2.1. Data partitioning

Consider the two sample data x_i and x_j in a large dimensional dataset, which lie at closer proximity. The distance between these two samples is found through the k-nearest neighbor algorithm. This forms an edge between the data samples, and the weights of the two sample data are thus computed as,

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

Assume that the sample set at the parent node is represented as a matrix $X = [x_1, x_2...x_n]^T \in \mathrm{R}^{n\times d}$, where n is the total examples in a sample set. The sample set is divided into two subsets i.e. left and right child node based on a decision, which is represented as: $X_1 \in \mathrm{R}^{n1\times d}$, and $X_2 \in \mathrm{R}^{n2\times d}$. Each instance has its own attributes that is weighted through a combination weighted vector, say, w. This estimates the sample point project point (x) in matrix (X) along the orientation of the weighted vector, say, w: $P(x) = w \cdot x$. After defining the split value p, the matrix (X) is divided into two values X_1 and X_2 based on the projection values $\{P(x), x \in X\}$:

$$\begin{cases} x \in X_1 \ \ if\ P(x) = w \cdot x > p \\ x \in X_2 \ \ if\ P(x) = w \cdot x \le p \end{cases}$$

and p considers the medium (m) of all matrix (X) projections: $p = m = \text{median}\{P(x_i), x_i \in X, i = i_1, i_2, ..., n\}$. The similarity matrix is obtained based on $S = \{S_{ij}\}_{i,j=1}^{N}$ which finds the similarity estimation between the data samples (N).

Assume two different samples x_i and x_j lie in a subspace at closer proximity, then the new data samples y_i and y_j will lie at the new subspace. Therefore, the estimation of projection vector (a) is carried out using the following equation,

$$0.5\sum_{ij}\left(y_i - y_j\right)^2 S_{ij} = 0.5\sum_{ij}\left(x_i a^T - x_j a^T\right)^2 S_{ij} \quad i = 1,2,....,N. \tag{2}$$

where, $y_i = x_i a^T$ has a sample matrix (X).

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

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

where,
D or $D_{ii} = \sum_j S_{ij}$ is the diagonal matrix and
$L = D\text{-}S$is the Laplacian matrix.

The diagonal matrix is further limited to find the objective function of LPP, which is given by the following condition.

$$\begin{aligned} & \arg \min_a Xa^T LX^{\mathrm{T}} a \\ & s.t.\ Xa^T DX^{\mathrm{T}} a = 1 \end{aligned} \tag{4}$$

The optimal projection vector (a) is found by solving the generalized Eigen value problem. The following equation shows the optimal projection vector (a).

$$XLX^{\mathrm{T}} a = \lambda XDX^{\mathrm{T}} a \tag{5}$$

2.2. NMF Metric Estimation

Assume the optimal projection vector (a) is approximated and applied over the features space (G), which represents the features vectors (F) of sample data (X). The feature vector is normalized to $f^T f = 1$ and then gram matrix (FGF) is found for the obtained normalized feature vector using a metric (M).

$$\mathrm{M} = \mathrm{F}^T\mathrm{F},\ s.t.\ u_l^T u_l = 1,\ \forall l = 1,...,q \tag{6}$$

The label information is avoided using a metric (M) that estimates the gram matrix and approximation of the sample data vector over the feature space is used to obtain the metric in feature space i.e. $M = F^T F$.

3. DLPP BASED SIMILARITY MEASUREMENT

The Euclidean distance η between the vectors $X_i = (x_{i1}, x_{i2} \ldots, x_{iD})^T$ and $X_j = (x_{j1}, x_{j2} \ldots, x_{jD})^T$ is given by,

$$\eta = \sqrt{z} = \sqrt{\sum_{d=1}^{D}\left(x_{id} - x_{jd}\right)^2}$$

where, z is the squared distance between the vector X_i and X_j

$$z = \left\| X_i - X_j \right\|_2^2$$
$$z = \sum_{d=1}^{D}\left(x_{id} - x_{jd}\right)^2$$

The squared distance η is estimated between any two vectors X_i, $X_j \in X$ *with* one or more missing datasets. Hence, we assume that vector X_i and vector X_j are independent. Since the squared distance η is a transform of vector X_i and vector X_j, the squared distance is regarded as a random variable. This takes into account the missing datasets, which are modeled below. Consider the squared distance η as a non-negative function, where the expected distance is given in terms of a Probability Density Function $p(\eta)$,

$$E[\eta] = \int_0^{\infty} P(n)\eta\, d\eta$$

The statistical model is used to resolve the above squared integral function, which is given as,

$$z = \sum_{d=1}^{D} \phi_d^2$$

Assume a component, say x_{id} or x_{jd}, is missing in the given data space, then the value of z is considered as the summation of squared random variables (φ^2). Depending on [18], the distribution of summed φ^2 is assumed to be Gamma function iff PDF of the random variables φ is given by,

$$p(\phi) = h(\phi)|\phi|^{2\alpha-1} \exp\{-\beta\phi^2\}$$

where α and β are the distribution parameters and the value of a random variable is assigned to a constant (ζ), which is given by

$\forall\varphi : h(\varphi) + h(-\varphi) = \zeta$.

Assume $z$is a Gamma distribution that reasonably chooses a Nakagami [12] distribution for the expected value η. The random variable is considered as a Nakagami function i.e. φ~Nakagami(m, Ω), which is obtained by using $\sqrt{\phi} \sim Gamma(\alpha, \beta)$.

The Nakagami distribution is a function of two parameters (shape and spread) that models the scattered datasets and reaches the receiver through multiple paths. Based on the assumption η~Nakagami(m, Ω), the expected value of the squared distance i.e. E(η) is given as:

$$E[\eta] = \frac{\Gamma(0.5 + m)}{\Gamma(m)} \sqrt{\frac{\Omega}{m}}$$

where m is the shape function of the Nakagami distribution and Ω is the spread function of the Nakagami distribution, which is a Gamma function.

4. EXPERIMENTAL EVALUATION

The proposed IDLPP method is tested against three datasets, namely 20 Newsgroups data shown in Figure 1. Reuters 21578 data shown in Figure 2 and R52 data shown in Figure 3. Initially, the data is preprocessed using the trunc5 stemmer technique and POS Tagger technique. Then the stop word removal technique is used to remove the stop words and remaining words are accepted based on mutual information. The dataset sample selection is given in Table 1.

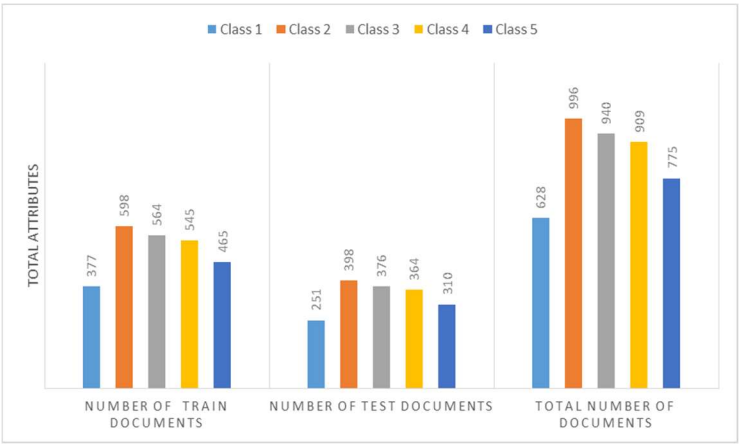

Figure 1. Attributes of 20 newsgroups dataset

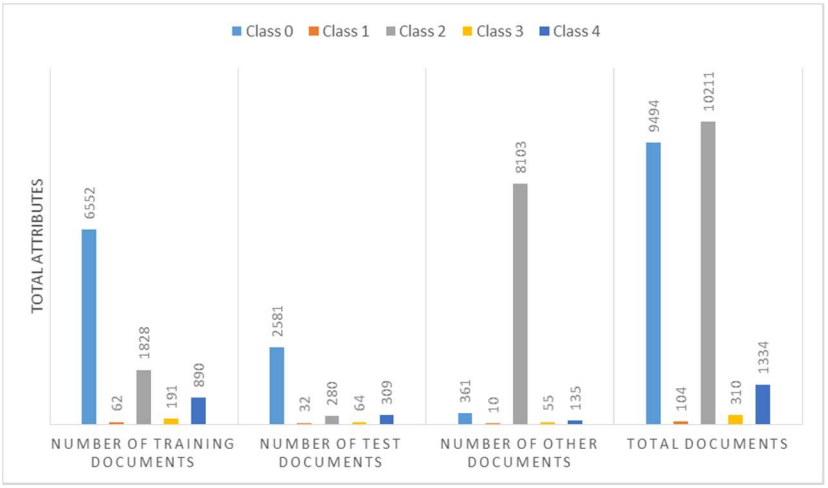

Figure 2. Attributes of reuters 21578 dataset

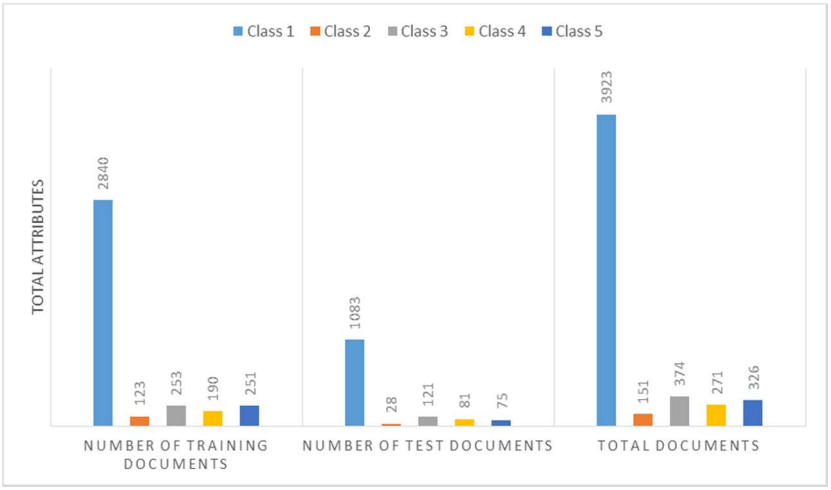

Figure 3. Attributes of R52 dataset

Table 1. Dataset Sample Selection

Samples	R52 dataset	20 News Group dataset	Reuters 21578 dataset
Sample 1	7	7	6
Sample 2	7	6	7
Sample 3	6	7	7
Sample 4	5	8	7
Sample 5	7	8	5
Sample 6	8	7	5
Sample 7	5	7	8
Sample 8	7	5	8
Sample 9	10	5	5
Sample 10	5	5	10
Sample 11	5	10	5
Sample 12	10	10	0
Sample 13	10	0	10
Sample 14	0	10	10
Sample 15	5	15	0
Sample 16	5	0	15
Sample 17	0	15	5
Sample 18	20	0	0
Sample 19	0	0	20
Sample 20	0	20	0

4.1. Result discussion

Figure 4 shows the results of accuracy between IDLPP and existing LPP methods in relation to 20 samples. Figure 5 shows the results of average accuracy between IDLPP and existing LPP methods in relation to three datasets i.e. 20 news groups, Reuters 21578 and R52 datasets. The result shows that the proposed method obtains a higher accuracy rate than other methods. The discarding of irrelevant feature vectors from the dataset using the proposed method is efficient and more robust than other existing LPP methods, which is evident from the results. Figure 6 shows the results of NMI between IDLPP and existing LPP methods in relation to 20 samples. Figure 7 shows the results of average NMI between IDLPP and existing LPP methods in relation to three datasets i.e. 20 news group, Reuters 21578 and R52 datasets. The proposed method obtains higher NMI than other methods, which is due to the effective reduction of redundant data samples from the larger datasets. The use of NMF helps to reduce the feature vector and the use of distance based measurement reduces the distance between the dataset samples.

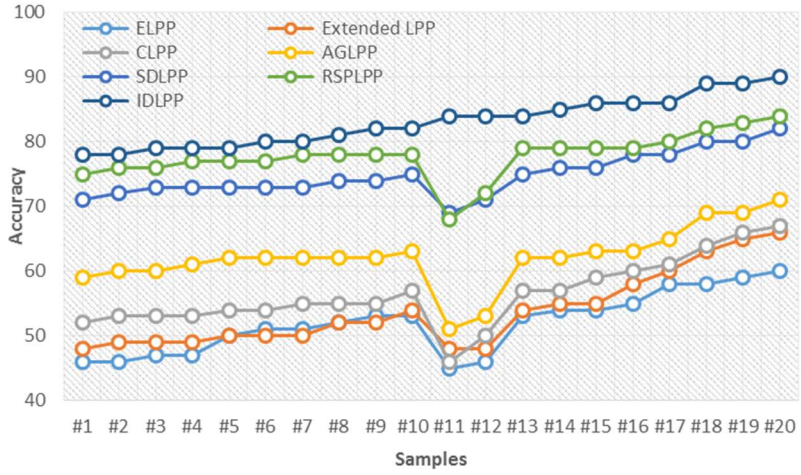

Figure 4. Results of accuracy using IDLPP and other LPP methods

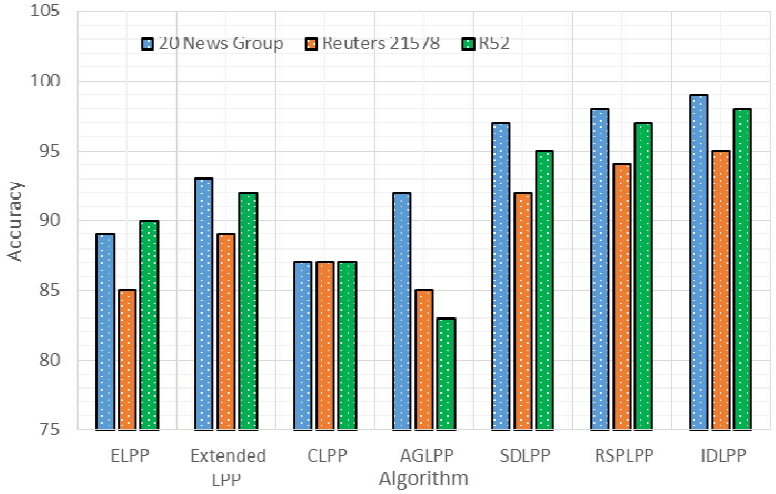

Figure 5. Results of average accuracy using IDLPP and other LPP methods

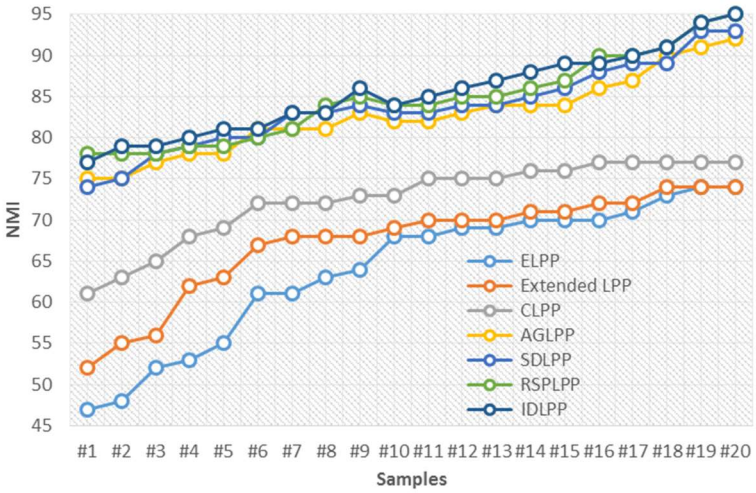

Figure 6. Results of NMI using IDLPP and other LPP methods

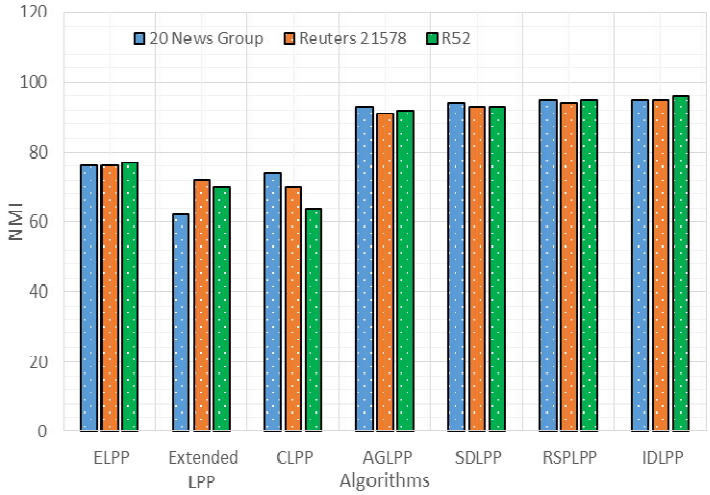

Figure 7. Results of average NMI using IDLPP

Further, the proposed method and other existing methods have been tested over UCI datasets. Figure 8 shows the classification accuracy of UCI datasets. The total number of instances, classes and dimensions are listed in Table 2 for evaluation. The estimation of classification accuracy between the proposed and existing methods has been tested and the result shows that the proposed method obtains higher classification accuracy than the other methods. This demonstrated the efficacy of the proposed method.

Table 2. UCI Dataset Sample

Dataset	No. of Instances	No. of Classes	No. of Dimensions
Anneal	898	5	90
Breast Tissue	106	6	9
Colic	368	2	60
Hepatitis	155	2	19
House	232	2	16
Hypothyroid	368	2	60
Promoter	106	2	57
Sonar	208	2	60
Wdbc	569	2	30
Wine	178	3	13

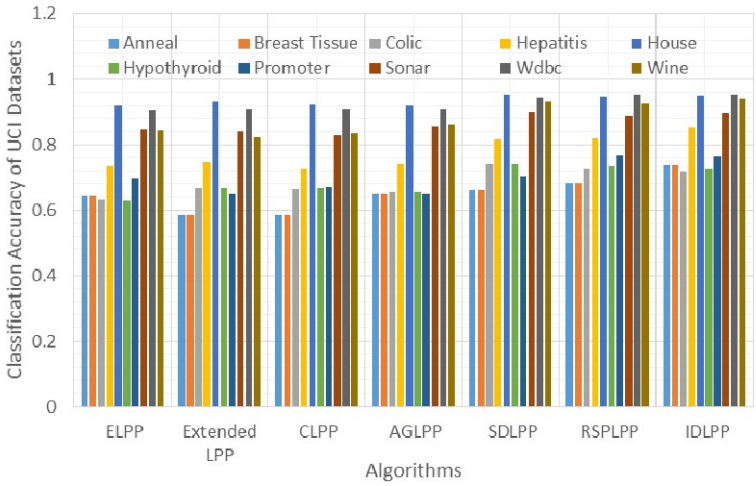

Figure 8. Classification Accuracy of UCI Datasets

5. CONCLUSION

In this paper, an IDLPP method is presented that increases the rate of accuracy and Mutual Information over large dimensional text datasets to retrieve the results effectively for the given queries. The distance measurement has been carried out in a probabilistic way in IDLPP between the sample data vector and this reveals that there is a hidden geometric pattern. It also reduces high dimensional irrelevant samples in large datasets and the geometric information of the datasets is preserved and this has increased the robustness. The results show that that the IDLPP method yields an improved rate of accuracy and an improved rate of NMI over other LPP methods and it is an improved method to preserve the locality projections.

REFERENCES

[1] Li, J., and Deshpande, A, "Ranking Continuous Probabilistic Datasets," *Proceedings of the VLDB Endowment*, 3(1-2), 638-649, 2010.
[2] Cormode, G., and Garofalakis, M, "Histograms and Wavelets on Probabilistic Data," *IEEE Transactions on Knowledge and Data Engineering*, 22(8), 1142-1157, 2010.
[3] Wiharto, W., Kusnanto, H., and Herianto, H, "System Diagnosis of Coronary Heart Disease Using a Combination of Dimensional Reduction and Data Mining Techniques: A Review," *Indonesian Journal of Electrical Engineering and Computer Scienc*e, 7(2), 514-523, 2017.
[4] Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., and Alzheimer's Disease Neuroimaging Initiative, "Multimodal Classification Of Alzheimer's Disease and Mild Cognitive Impairment," *Neuroimage*, 55(3), 856-867, 2011.
[5] Mashhour, E. M., El Houby, E. M., Wassif, K. T., and Salah, A. I, "Feature Selection Approach based on Firefly Algorithm and Chi-square," *International Journal of Electrical and Computer Engineering (IJECE)*, 8(4), 2018.
[6] Chen, L. F., Liao, H. Y. M., Ko, M. T., Lin, J. C., and Yu, G. J, "A New LDA-Based Face Recognition System Which Can Solve The Small Sample Size Problem," *Pattern recognition*, 33(10), 1713-1726, 2000.
[7] Liu, M., Miao, L., and Zhang, D, "Two-Stage Cost-Sensitive Learning for Software Defect Prediction," *IEEE Transactions on Reliability*, 63(2), 676-686, 2014.
[8] Simão, M., Neto, P., and Gibaru, O, "Using Data Dimensionality Reduction for Recognition of Incomplete Dynamic Gestures," *Pattern Recognition Letters*, DOI: 10.1016/j.patrec.2017.01.003.
[9] Tunga, B, "A Hybrid Algorithm With Cluster Analysis In Modelling High Dimensional Data," Discrete Applied Mathematics, 2017.
[10] Zhu, R., and Xue, J. H, "On the Orthogonal Distance to Class Subspaces for High-Dimensional Data Classification," *Information Sciences*, 417, 262-273, 2017.
[11] Liu, Z., Liu, B., Zheng, S., and Shi, N. Z, "Simultaneous Testing of Mean Vector and Covariance Matrix for High-Dimensional Data," *Journal of Statistical Planning and Inference*, 188, 82-93, 2017.
[12] Lucas, T., Silva, T. C., Vimieiro, R., and Ludermir, T. B, "A New Evolutionary Algorithm For Mining Top-K Discriminative Patterns In High Dimensional Data," *Applied Soft Computing*, DOI: 10.1016/j.asoc.2017.05.048.
[13] Zhao, S., Zhou, J., and Li, H, "Model Averaging with High-Dimensional Dependent Data," *Economics Letters*, 148, 68-71, 2016.
[14] Zamora, J., Mendoza, M., and Allende, H, "Hashing-Based Clustering in High Dimensional Data," *Expert Systems with Applications*, 62, 202-211, 2016.

[15] Ultsch, A., & Lötsch, J, "Machine-Learned Cluster Identification in High-Dimensional Data," *Journal of biomedical informatics*, 66, 95-104, 2017.
[16] Sang, Y., Qi, H., Li, K., Jin, Y., Yan, D., and Gao, S, "An Effective Discretization Method For Disposing High-Dimensional Data," *Information Sciences*, 270, 73-91, 2014.
[17] Apiletti, D., Baralis, E., Cerquitelli, T., Garza, P., Pulvirenti, F., and Michiardi, P, "A Parallel MapReduce Algorithm to Efficiently Support Itemset Mining on High Dimensional Data," *Big Data Research*, 10, 53-69, 2017.
[18] Zhou, P., Hu, X., Li, P., and Wu, X, "Online Feature Selection for High-Dimensional Class-Imbalanced Data," *Knowledge-Based Systems*, 136, 187-199, 2017.
[19] Lansangan, J. R. G., and Barrios, E. B, "Simultaneous Dimension Reduction and Variable Selection in Modeling High Dimensional Data," *Computational Statistics & Data Analysi*s, 112, 242-256.
[20] Jing, L., Tian, K., & Huang, J. Z, "Stratified Feature Sampling Method For Ensemble Clustering of High Dimensional Data," *Pattern Recognition*, 48(11), 3688-3702, 2015.
[21] Liu, X., and Li, M, "Integrated Constraint Based Clustering Algorithm for High Dimensional Data," *Neurocomputing*, 142, 478-485, 2014.
[22] Cardoso, Â., and Wichert, A, "Iterative Random Projections for High-Dimensional Data Clustering," *Pattern Recognition Letters*, 33(13), 1749-1755, 2012.
[23] Moayedikia, A., Ong, K. L., Boo, Y. L., Yeoh, W. G., and Jensen, R, "Feature Selection for High Dimensional Imbalanced Class Data Using Harmony Search," *Engineering Applications of Artificial Intelligence*, 57, 38-49, 2017.
[24] Pedergnana, M., and García, S. G, "Smart Sampling And Incremental Function Learning for Very Large High Dimensional Data," *Neural Networks*, 78, 75-87, 2016.
[25] Itoh, T., Kumar, A., Klein, K., and Kim, J, "High-Dimensional Data Visualization By Interactive Construction of Low-Dimensional Parallel Coordinate Plots," *Journal of Visual Languages & Computing*, 2017.
[26] Ando, T., and Li, K. C, "A Model-Averaging Approach For High-Dimensional Regression," *Journal of the American Statistical Association*, 109(505), 254-265, 2014.
[27] Qiao, L., Chen, S., and Tan, X, "Sparsity Preserving Projections With Applications To Face Recognition," Pattern Recognition, 43(1), 331-341, 2010.
[28] Liu, M., Zhang, D., and Chen, S, "Attribute Relation Learning for Zero-Shot Classification," *Neurocomputing*, 139, 34-46, 2014.
[29] Jiang, W., and Chung, F. L, "A Trace Ratio Maximization Approach to Multiple Kernel-Based Dimensionality Reduction," Neural Networks, 49, 96-106, 2014.
[30] Liu, M., and Zhang, D, "Sparsity Score: A Novel Graph-Preserving Feature Selection Method," *International Journal of Pattern Recognition and Artificial Intelligence*, 28(04), 1450009, 2014.
[31] Srividya Sivasankar, Sruthi Nair, M.V. Judy, "Feature Reduction in Clinical Data Classification using Augmented Genetic Algorithm", *International Journal of Electrical and Computer Engineering (IJECE)* Vol. 5, No. 6, December 2015, pp. 1516-1524.
[32] Muhammad Kusban, Adhi Susanto, and Oyas Wahyunggoro, "Combination a Skeleton Filter and Reduction Dimension of Kernel PCA-Based on Palmprint Recognition" *International Journal of Electrical and Computer Engineering (IJECE)* Vol. 6, No. 6, December 2016, pp. 3255-3261.

BIOGRAPHIES OF AUTHORS

Dr. Jasem M.Alostad received his Ph.D from The University of York, United Kingdom in 2006, MS from the Monmouth University, New Jersey, USA in 1996 and B.S (Computer Science) from Western Kentucky University, KY, USA in 1990. He is currently the Director of Computer and Information Centre in The Public Authority of Applied Education and Training (PAAET), Kuwait. He has more than 10 years of experience in both academics and management. He has authored more than 25 technical research papers published in leading journals and conferences from the ACM, Elsevier, Springer etc. His current research includes Software Engineering, HCI, Data Mining, Data Analysis, Big Data Cloud Computing and Internet of things (IoT).