Comparison of Dimension Reduction Techniques on High Dimensional Datasets

Kazim Yildiz[1], Yilmaz Camurcu[2], and Buket Dogan[1]
[1]Deparment of Computer Engineering, Marmara Unıversity, Turkey
[2]Department of Computer Engineering, Fatih Sultan Mehmet Waqf University, Turkey

Abstract: *High dimensional data becomes very common with the rapid growth of data that has been stored in databases or other information areas. Thus clustering process became an urgent problem. The well-known clustering algorithms are not adequate for the high dimensional space because of the problem that is called curse of dimensionality. So dimensionality reduction techniques have been used for accurate clustering results and improve the clustering time in high dimensional space. In this work different dimensionality reduction techniques were combined with Fuzzy C-Means clustering algorithm. It is aimed to reduce the complexity of high dimensional datasets and to generate more accurate clustering results. The results were compared in terms of cluster purity, cluster entropy and mutual info. Dimension reduction techniques are compared with current Central Processing Unit (CPU), current memory and elapsed CPU time. The experiments showed that the proposed work produces promising results on high dimensional space.*

Keywords: *High dimensional data, clustering, dimensionality reduction, data mining.*

Received October 23, 2014; accepted December 21, 2015

1. Introduction

High dimensional datasets are frequently used in many applications, such as image processing, in biology for computational and global climate research. Classical clustering algorithms become inefficient when they apply to higher dimensional data, because of the growing amount of data. It is a really difficult problem to progress effective clustering methods for high dimensional datasets. The result of clustering process is not qualitative and also clustering operation takes a long time, especially in the data that have a high number of attributes. The dimension reduction methods have been applied in order to get more realistic and faster results. Dimensionality reduction techniques have been used to get better the computation time and give correct results [47].

This study aimed to compare Principal Component Analysis (PCA), Laplacian Eigenmaps, Fast Maximum Variance Unfolding (FastMVU), Isometric Mapping (Isomap), Landmark Isomap (L-Isomap), t distributed Stochastic Neighbor Embedding (t-SNE) and Stochastic Neighbor Embedding (SNE) methods in the dimension reduction of Abalone, Milliyet and British Broadcasting Corporation (BBC) datasets before the clustering process. Cluster analysis is one of the difficulties of the high dimensional clustering process. Clustering [29, 32] means to divide data into significant or available groups. It has been widely used in different applications [23, 29]. In many application areas dimension reduction is performed as a preprocessing step. It decreases undesired features of high dimensional space and holds the most significant characteristic of a dataset and omitted the outlier points [18]. PCA [16, 30] is used for selection of appropriate dimensions with Singular Value Decomposition (SVD) [20] is a popular approach for numerical attributes. SVD has been used by Latent Semantic Indexing (LSI) to project textual documents [12] in knowledge retrieval. For a probabilistic model SVD is shown the optimal solution for word occurrence [13]. Random projections have also been used for selecting appropriate subspaces [10, 14].

In a previous study which was done by Bilgin and Camurcu [6], the performances of Opossum, Graclus, PSpace+Graclus algorithms were compared on Milliyet and BBC data. Lee, Abbott and Araman found that the clustering process was obtained with this study was more certain and more decisive compared to one in Euclidean space [34]. Bilgin and Camurcu [5] made filtering outliers, density based clustering and visualizing on high dimensional space. Fern and Bradley [17] used PCA and Random Projection for building cluster assembly on high dimensional dataset. The cluster assembly that was built by Random Projection was more satisfied than by PCA. Teng *et al.* [42] compared Local Tangent Space Analysis and PCA algorithms on datasets in the way of visualization that nonlinear dimension reduction techniques have an influence on microarray datasets. Yang and Pedersen [46] compared four well-known dimension reduction techniques, Document Frequency, Random Projection [3], Latent Semantic Indexing [7], Independent Component Analysis [27] for the document clustering task using five benchmark datasets [40]. Davidson proposes the graph-driven constrained dimension reduction by a linear projection approach that gave a weighted graph attempts to find a series of dimensions that are linear combinations of the old dimensions [11]. Ture *et al*. [43] show NN which have higher percentages of explaining variances than classical methods could be used for dimension reduction. Shi and Luo [38] compared the performance of PCA and Isomap algorithms. They have used these algorithms for visualization and clustering of cancer tissue samples. Somwang and Lilakiatsakun used [39] PCA and fuzzy adaptive resonance theory map for feature selection in anomaly traffic detection. Zhou *et al*. [48] has prepared the Manifold Elastic Net (MEN) which is a unified framework for sparse dimension reductio. Izakian and Abraham [28] has suggested a hybrid fuzzy clustering method based on Fuzzy C-Means (FCM) and Fuzzy Particle Swarm Optimization (FPSO). They have tried out the algorithm on real-world data sets which cover low, medium and high dimensions. Özsen and Ceylan [37] is used an artificial immune system for the data reduction process. They have compared the performance with the FCM algorithm on breast cancer and diabetes datasets. Jun *et al*. [31] have proposed a method that using dimension reduction and K-Means clustering. It uses support vector clustering and silhouette measure. Krawczak and Szkatuła [33] have proposed a new method on multidimensional datasets for dimension reduction. The original datasets have been formed with nominal representation. Their representation retained information for classification and clustering process.

The paper is organized as follows. In section 2 general system of dimension reduction process, clustering algorithms and dimension reduction methods which are used in this study introduced. Section 3 describes the experimental results on high dimensional dataset. Finally conclusions are reported.

2. Materials and Methods

2.1. General System of Dimension Reduction Process

The block schema of a dimension reduction process that is used in this study can be seen in Figure 1. Initially for loading dataset one of the test data set is chosen. After that, there are two ways of clustering process. The first way is clustering data without dimension reduction process. Another way is, applying one of the dimension reduction techniques to the data with required parameters, and then the reduced data is clustered by FCM algorithm. In this paper, Fuzzy C-means algorithm was applied to high dimensional datasets for clustering. Firstly, dimensions of the data were reduced by the dimensional reduction techniques which has described below. The obtained new low dimensional data has clustered with FCM algorithm. The clustering result was compared by using the cluster purity, cluster entropy and mutual information values respectively. Additionally, for the computer resources usage measurement, the quantity of the memory and CPU that is used during the dimensional reduction process, and the time that passed on the dimension reduction process were compared.

2.2. Clustering

Cluster analysis explains relationship between objects based on the knowledge in the data. Its aim that the data points in a class should be related to one and unrelated to the data points in other classes. Clustering is a illustrative assignment that search for intimate stacks of data points with values of their sizes [29, 32]. In statistics [1], pattern recognition [15, 19] and machine learning [8, 35] and many other applications clustering techniques have been analyzed frequently. In this study, it was stated about traditional clustering algorithm that is Fuzzy C-Means.

2.2.1. Fuzzy C-Means

It is the most known and commonly used form of fuzzy partitioning clustering techniques. FCM algorithm was proposed by Dunn in 1973 and was improved by Bezdek in 1981 [26]. FCM method allows objects that could belong to two or more clusters. Due to fuzzy logic principle, each data is assigned to each cluster with a membership value that is between 0 and 1. The total affiliation values of a data, which belongs to the all classes, should be “1”. The probability of belonging to a cluster is related to the distance between the object and cluster. So the probability of an object belonging to a closer cluster is higher than the distant one.

2.3. Dimension Reduction

Dimension reduction is an algorithmic technique for reducing the dimensionality of data. From a programmers point, a d-dimensional array of real numbers, after applying this technique, is represented by a much smaller array. With the increase of data size, generally only a few numbers of dimensions are directly related to the clusters. But the irrelevant dimension of data can cause to very much noise and this also causes to conceal the data that will be discovered [22]. Hence so far the most important one is data to become sparse when the dimension increases. A dataset which has a certain number of data points, becomes sparse in an exponential way as the dimension number increases [4]. For dimensionality reduction process numerous methods have been proposed. In [18] dimension reduction techniques have been discussed.

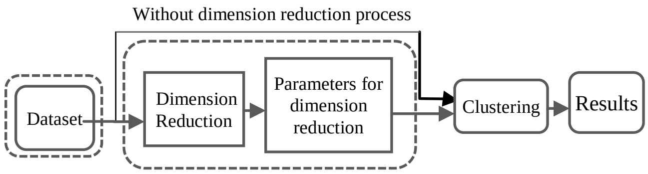

Figure 1. General diagram of dimension reduction process.

2.4. Overview of Dimension Reduction Techniques

Details of seven different dimension reduction techniques, Principal Component Analysis, Laplacian, Fast Maximum Variance Unfolding (MVU), Isometric Mapping, Landmark Isometric Mapping, Stochastic Neighbor Embedding, t-distributed Stochastic Embedding are introduced.

Pca, [25] constructs a low-dimensional representation of the multidimensional space. It is possible to describe as much of the variance in the data with the use of Pca. Principal components that were obtained from high dimension are independent. Thus the structure of dependence between variables is removed. Laplacian, preserve regional features of the manifold for finding a low-dimensional space [2]. The pairwise distances between near neighbors are used for the local properties. It computes a low dimensional space of the data with the nearest neighbors is minimized. Fast MVU defines a neighborhood graph of the data which is used to retain pairwise distances. Fast MVU is separated from Isomap which explicitly enterprises to unfold data manifolds [44]. Isomap is a graphic based dimension reduction technique. In this technique the points that are close to each other, remain close again after the dimension reduction process. And the points with far distances protect their distances again [41]. L-Isomap uses a subset that has been selected randomly of points from the high dimension to build the low dimensional space. Samples of the subset are not selected as signs which are located on the map by operating the derived embedding vectors [9]. SNE, is an iterative technique that enterprises to keep the pairwise intervals between the data points in the low-dimensional space [24]. t-SNE, is effective in getting the local structure of high dimensional data and also preparing then demonstrating the clusters that have different scales on global structures [45].

3. Results

This study was presented with Intel(R) Core(TM) 2 Duo CPU P8700 3.0 GHZ 64 bit operating system and 4 GB RAM on computer. For the software environment, MATLAB software was used. In the application, Abalone, Milliyet and BBC data sets were used. The dataset features were given in Table 1. Abalone dataset has features about oyster which are formed under three classes; female, male and asexual shell [36] that consists of 4177 instances and 8 attributes.

Milliyet dataset has 3 classes; economy, politics and sports that are gathered in Milliyet newspaper internet archive [6] that consists of 1695 terms in Turkish from 1455 news articles.

BBC dataset is gathered from the BBC internet archive and class names are business, entertainment, politics, sports and technology [21]. This dataset contains 9635 terms from BBC news website which has 2225 news articles.

Table 1. Feature of data sets.

Dataset Name	Instances	Attributes	Class Nunber
Abalone	4177	7	3
Milliyet	1455	1695	3
BBC	2225	9635	5

The results obtained without dimension reduction process on high dimensional datasets are shown in the Table 2. It can be seen that Fuzzy C-Means algorithm is not so much efficient in high dimensional space.

Table 2. Clustering results without dimension reduction.

	Data Sets								
	Abalone			Milliyet			BBC		
	Purity	Entropy	Mutual Info	Purity	Entropy	Mutual Info	Purity	Entropy	Mutual Info
FCM							1.0000	0.0000	
	0.6954	0.2298		1.0000	0.0000		0.5203	0.2111	
	0.9940	0.0054	0.7966	0.0113	0.0462	0.4366	0.0000	1.0000	0.5481
	1.0000	0.0000		1.0000	0.0000		0.2954	0.2238	
							0.8079	0.1070	

The results obtained by using clustering algorithms after dimension reduction process on datasets, are shown in the Table 3. According to the results, it can be concluded that the clustering results are more successful when the clustering process is made on datasets, obtained by using dimension reduction methods. And also it can be seen that all clusters on datasets are predicted almost in an accurate way by using Fuzzy C-Means algorithm.

Table 3. Clustering result after dimension reduction.

	Data Sets								
	Abalone			**Milliyet**			**BBC**		
	Purity		**Mutual Info**	**Purity**	**Entropy**	**Mutual Info**	**Purity**	**Entropy**	**Mutual Info**
Pca+FCM							1.0000	0.0000	
	1.0000	0.0000		1.0000	0.0000		0.8561	0.0826	
	0.6691	0.2446		0.1841	0.2836		0.0000	1.0000	
	1.0000	0.0000	0.7852	1.0000	0.0000	0.4699	0.4716	0.2202	0.6720
							1.0000	0.0000	
Isomap+ FCM							1.0000	0.0000	
	1.0000	0.0000		1.0000	0.0000		0.6243	0.1827	
	0.6713	0.2434		0.3626	0.3348		0.6882	0.1597	0.6375
	1.0000	0.0000	0.7289	0.7065	0.2234	0.5059	1.0000	0.0000	
							0.2967	0.2239	
L-Isomap+ FCM							0.9215	0.0467	
	0.6908	0.2325		1.0000	0.0000		0.1683	0.1863	
	0.9940	0.0054		0.1473	0.2568		1.0000	0.0000	0.7069
	1.0000	0.0000	0.7289	1.0000	0.0000	0.4577	0.4344	0.2250	
							1.0000	0.0000	
Laplacian+ FCM							1.0000	0.0000	
	0.9892	0.0096		0.9762	0.0213		0.1269	0.1627	
	1.0000	0.0000	0.8517	0.8640	0.1149	0.8154	0.4988	0.2155	0.7611
	0.8671	0.1125		1.0000	0.0000		1.0000	0.0000	
							0.9825	0.0107	
FastMVU+ FCM							0.9078	0.0545	
	1.0000	0.0000		0.5594	0.2958		0.9870	0.0079	
	0.9947	0.0047	0.9415	0.9433	0.0500	0.4446	1.0000	0.0000	0.8826
	0.9659	0.0304		1.0000	0.0000		0.9178	0.0489	
							0.9376	0.0375	
SNE+FCM							0.0941	0.1381	
	1.0000	0.0000		0.0000	1.0000		0.0000	1.0000	
	0.8248	0.1445		0.4831	0.3199		1.0000	0.0000	
	1.0000	0.0000	0.8266	1.0000	0.0000	0.4837	0.0000	1.0000	0.1522
							0.1147	0.1543	
t-SNE+ FCM				0.0000	1.0000		0.8725	0.0739	
	0.9846	0.0138		0.6376	0.2641		0.9792	0.0127	0.8665
	1.0000	0.0000	0.9001	0.9546	0.0403	0.5315	1.0000	0.0000	
	0.9378	0.0547		1.0000	0.0000		0.8610	0.0800	

The highest mutual information values were obtained by applying the combinations that are shown in Table 4.

Table 4. The highest mutual information values according to the dimension reduction and clustering.

Data Set	Dimension Reduction	Clustering Algorithm
Abalone	FastMVU	FCM
Milliyet	Laplacian	FCM
BBC	FastMVU	FCM

As can be seen in Table 4, the best result has been taken with FastMVU and FCM technique in the application of Abalone data set. On Milliyet data set, by using Laplacian with FCM technique has become the best value for mutual info. The best results were obtained by using the FastMVU with the FCM technique on BBC data set. On Abalone, the best results were obtained by using the FastMVU and FCM techniques together. The best mutual info value was obtained with the combination of Laplacian and FCM techniques on Milliyet data set. The optimum result of the BBC dataset was performed by using FCM and FastMVU techniques. Figure 2 shows the CPU values (%) that are used in the dimension reduction process on three data sets. On Abalone data set, the dimension reduction technique that uses the highest CPU value is t-SNE and the lowest one is PCA. The highest CPU value is used by t-SNE and the lowest one is used by PCA on Milliyet. On BBC data set, the dimension reduction technique that uses the highest CPU value is t-SNE and the lowest one is PCA. All of the data set, the dimension reduction technique that uses the highest CPU value is t-SNE and the least one is PCA.

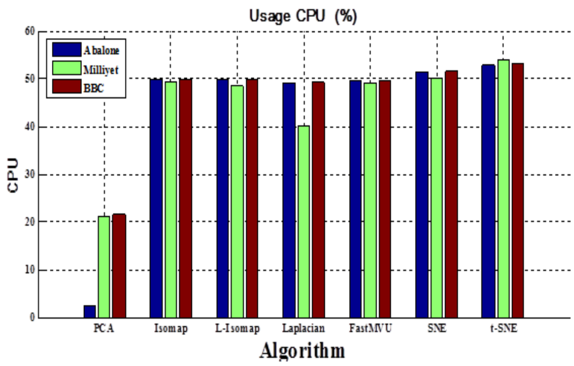

Figure 2. CPU Values (%) that are used in dimension reduction process.

Figure 3 shows the amount of used memory in the dimension reduction process on three datasets. On Abalone data set, the dimension reduction technique that uses the highest memory is Isomap and the lowest one is PCA. On Milliyet dataset, the dimension reduction technique that uses the highest memory is Isomap and the lowest one is PCA. On BBC dataset, the dimension reduction technique that uses the highest memory is Laplacian and the lowest one is SNE.

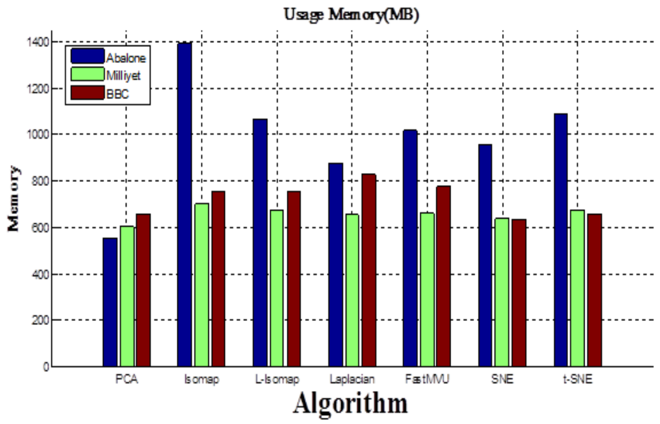

Figure 3. Used memory in dimension reduction process.

Elapsed CPU time as a minute, that is required during the dimension reduction process on datasets is shown in Figure 4. The longest elapsed CPU time is Isomap and the shortest elapsed CPU time is PCA algorithm on Abalone dataset. The longest elapsed CPU time is SNE and the shortest elapsed CPU time is a Laplacian algorithm on Milliyet data set. The longest elapsed CPU time is t-SNE and the shortest elapsed CPU time is PCA algorithm on BBC dataset.

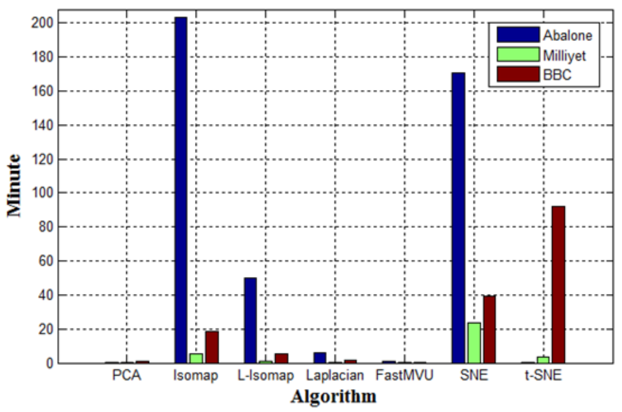

Figure 4. Elapsed CPU time.

4. Conclusions

We have presented experimental results on PCA, Isomap, L-Isomap, Laplacian, FastMVU, SNE and t-SNE algorithms for dimensionality reduction of real-world datasets. The criteria are the cluster purity, cluster entropy and mutual info for comparing different methods this process. Also, these techniques were compared on behalf of CPU usage, memory usage and the dimension reduction process period.

From the experimental results, it was observed that, the most efficient algorithms among the dimensional reduction algorithms are Laplacian, FastMVU and t-SNE algorithms. In addition, when the dataset dimension reduced its purity and mutual information increases. The t-SNE algorithm, on Abalone dataset has the most CPU usage rate and the PCA has the lowest CPU usage rate. During the dimension reduction process on Abalone dataset on our Matlab application, the Isomap algorithm has the most memory usage rate and the least one is PCA algorithm and the PCA algorithm has the least elapsed time, the algorithm with the longest elapsed CPU time is Isomap. On Milliyet dataset, the t-SNE algorithm has the highest CPU usage rate and the lowest CPU usage rate is in the PCA algorithm. During the dimension reduction process on Milliyet dataset on Matlab, the Isomap algorithm has the highest memory usage rate and the lowest one is PCA algorithm. The Laplacian algorithm has the least period, the most one is SNE algorithm.

The t-SNE algorithm has the most CPU usage rate and the lowest CPU usage rate is in the PCA algorithm for BBC data set. During the dimension reduction process of the BBC data set of MATLAB, the highest memory usage rate is on the Laplacian and the lowest memory usage rate is on SNE algorithm and the Fast MVU algorithm has the least period, the most one is t-SNE algorithm. The proposed work produces promising results for clustering on high dimensional space. In subsequent studies the method that will be developed for reducing the required time during dimension reduction and amount of memory. Genetic algorithm will be used for choice of the best subspace that represents the data obtained from high dimensional data sets.

References

[1] Arabie P. and Hubert L., *An Overview of Combinational Data Analysis*, Clustering and Classification, 1996.

[2] Belkin M. and Niyogi P., "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," *in Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, British Columbia, pp. 585-591, 2001.

[3] Berry M., Dumais S., and O'Brien G., "Using Linear Algebra for Intelligent Information Retrieval," *Society for Industrial and Applied Mathematics*, vol. 37, no. 4, pp. 573-595, 1995.

[4] Bilgin T., Three new Frameworks for the Design and Application of Visual Data Mining in High Dimensional Space, PhD thesis, Marmara University, 2007.

[5] Bilgin T. and Camurcu Y., "A Modified Relationship based Clustering Framework for Density based Clustering and Outlier Filtering on High Dimensional Datasets," *in Proccedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Nanjing, pp. 409-416, 2007.

[6] Bilgin T. and Camurcu Y., "A Clustering Framework for Unbalanced Partitioning and

Outlier Filtering on High Dimensional Datasets," *in Proccedings of East European Conference on Advances in Databases and Information Systems*, Varna, pp. 205-216, 2007.

[7] Bingham E. And Mannila H., "Random Projection in Dimensionality Reduction: Applications to Image and Text Data," *in Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data Mining*, California, pp. 245-250, 2001.

[8] Cheeseman P. and Stutz J., "Bayesian Classification (AutoClass): Theory and Results," *in proccedings of Advances in Knowledge Discovery and Data Mining*, Menlo Park, pp. 153-180, 1996.

[9] Chen Y., Crawford M., and Ghosh, J., "Improved Nonlinear Manifold Learning for Land Cover Classification Via Intelligent Landmark Selection," *in Proceedings of IEEE International Conference on Geoscience and Remote Sensing Symposium*, Denver, pp. 545-548, 2006.

[10] Dasgupta S., "Experiments with Random Projection," *in Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, San Francisco, pp. 143-151, 2000.

[11] Davidson I., "Knowledge Driven Dimension Reduction for Clustering," *in Proceedings of International Joint Conference on Artificial Intelligence*, California, pp. 1034-1039, 2009.

[12] Deerwester S., Dumais S., Furnas G., Landauer T., and Harshman R., "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990.

[13] Ding C., "A Similarity-Based Probability Model for Latent Semantic Indexing," *in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, California, pp. 58-65, 1999.

[14] Drineas P., Frieze A., Kannan R., Vempala S., and Vinay V., "Clustering in Large Graphs and Matrices," *in Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*, Maryland, pp. 291-299, 1999.

[15] Duda R. and Hart P., *Pattern Classification and Scene Analysis*, Wiley, 1973.

[16] Duda R., Hart P., and Stork D., *Pattern Classification*, Wiley, 2000.

[17] Fern X. and Brodley C., *Cluster Ensembles for High Dimensional Clustering: An Empirical Study*, Journal Machine Learning Research. 2004.

[18] Fodor I., *A Survey of Dimension Reduction Techniques*, Lawrence Livermore National Laboratory, 2002.

[19] Fukunaga K., *Introduction to Statistical Pattern Recognition 2ed*, Academic Press,1990.

[20] Golub G. and Van Loan C., *Matrix computations*, Johns Hopkins University Press, 1996.

[21] Greene D. and Cunningham P., "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," *in Proceedings of the 23rd International Conference on Machine Learning*, Pennsylvania, pp. 377-384, 2006.

[22] Han J. and Kamber M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[23] Hartigan J., *Clustering Algorithms*, John Wiley and Sons, 1975.

[24] Hinton G. and Roweis S., *Stochastic Neighbor Embedding*, Advances in Neural Information Processing Systems, 2003.

[25] Hotelling H., "Analysis of a Complex of Statistical Variables in to Principal Components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417-441, 1933.

[26] Höppner F., Klawonn F., Kruse R., and Runkler T., *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*, John Wiley, 2000.

[27] Hyvärinen A. and Oja E., "Independent Component Analysis: Algorithms and Applications," *Neural networks*, vol. 13, no. 4-5, pp. 411-430, 2000.

[28] Izakian H. and Abraham A., "Fuzzy C-means and Fuzzy Swarm for Fuzzy Clustering Problem," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1835-1838, 2011.

[29] Jain A. and Dubes R., *Algorithms for Clustering Data*, Prentice-Hall, 1988.

[30] Jolliffe I., *Principal Component Analysis*, John Wiley and Sons, 2005.

[31] Jun S., Park S., and Jang D., "Document Clustering Method Using Dimension Reduction and Support Vector Clustering to Overcome Sparseness," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3204-3212, 2014.

[32] Kaufman L. and Rousseeuw P., *Finding Groups in Data: an Introduction to Cluster Analysis*, Wiley Online Library,1990.

[33] Krawczak M. and Szkatuła G., "An Approach to Dimensionality Reduction in Time Series," *Information Sciences*, vol. 260, pp. 15-36, 2014.

[34] Lee S., Abbott A., and Araman P., "Dimensionality Reduction and Clustering on Statistical Manifolds," *in Proceedings of Conference on Computer Vision and Pattern Recognition*, Minneapolis, pp. 1-7, 2007.

[35] Michalski R. and Stepp R., "Learning from Observation: Conceptual Clustering," *Machine Learning*, Berlin, pp. 331-363, 1983.

[36] Nash W., Sellers T., Talbot S., Cawthorn A., and Ford W., Available from: http://archive.ics.uci.edu/ml/datasets/Abalone, Last Visited, 1994.

[37] Özşen S. and Ceylan R., "Comparison of AIS and Fuzzy C-means Clustering Methods on the Classification of Breast Cancer and Diabetes Datasets," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 22, pp. 1241-1254, 2014.

[38] Shi J. and Luo Z., "Nonlinear Dimensionality Reduction of Gene Expression Data for Visualization and Clustering Analysis of Cancertissue Samples," *Computers in Biology and Medicine*, vol. 40, no. 8, pp. 723-732, 2010.

[39] Somwang P. and Lilakiatsakun W., "Anomaly Traffic Detection Based on PCA and SFAM," *The International Arab Journal of Information Technology*, vol. 12, no. 3, pp. 253-260, 2015.

[40] Tang B., Shepherd M., Heywood M., Luo X., Kegl B., and Lapalme G., "Comparing Dimension Reduction Techniques for Document Clustering ," *in Proceedings of Conference of the Canadian Society for Computational Studies of Intelligence*, Victoria, pp. 292-296, 2005.

[41] Tenenbaum J., Silva V., and Langford J., "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, 2000.

[42] Teng L., Li H., Fu X., Chen W., and Shen I., "Dimension Reduction of Microarray Data based on Local Tangent Space Alignment," *in Proceedings of Fourth IEEE Conference on Cognitive Informatics*, Irvine, pp. 154-159, 2005.

[43] Ture M., Kurt I., and Akturk Z., "Comparison of Dimension Reduction Methods using Patient Satisfaction Data," *Expert Systems with Applications*, vol. 32, no. 2, pp. 422-426, 2007.

[44] Van der Maaten L., and Hinton G., "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.

[45] Van der Maaten L., *An Introduction to Dimensionality Reduction using Matlab,* Faculty of Humanities and Sciences, 2007.

[46] Yang Y. and Pedersen J., "A Comparative Study on Feature Selection in Text Categorization," *in Proceedings of the 14th International Conference on Machine Learning*, San Francisco, pp. 412-420, 1997.

[47] Zhang X., Wang J., Fan Z., and Li B., "Spatial Clustering with Obstacles Constraints using Ant Colony and Particle Swarm Optimization," *in Proceedings of Emerging Technologies in Knowledge Discovery and Data Mining*, Nanjing, pp. 344-356, 2007.

[48] Zhou T., Tao D., and Wu X., "Manifold Elasticnet: a Unified Framework for Sparse Dimension Reduction," *Data Mining and Knowledge Discovery*, vol. 22, no.3, pp. 340-371, 2011.

Kazim Yildiz received a Ph.D and Msc in Electronic and Computer Education (2014) and (2010) respectively in Marmara University. He received a B.Sc. degree in Computer and Control Education from Marmara University. From August 2009 to August 2015, Kazim Yildiz worked as a research assistant. From August 2015 he has been working as an Assistant Professor in Computer Engineering department of Technology Faculty. His current research areas are digital image processing, high dimensional data mining and thermal imaging.

Buket Dogan received the MS and PhD degrees in Computer-Control Educationfrom Marmara University in 2001 and 2006, respectively.From 1999 to 2007 she worked as a research assistant. She has been working as an Assistant Professor in Computer Engineering department of Technology Faculty. Her research interests include data mining, image processing and adaptive web based educational systems.

Yilmaz Camurcu received the PhD degree in computer education fromMarmara University, Istanbul in 1996. He is a professor of Computer Engineering in the Faculty of Engineeringand Architecture at Fatih Sultan Mehmet Waqf University. He is a member of ACM. His current research interests are data mining, intelligent tutoring systems, and medical image processing.