The Effect of Autoencoders over Reducing the Dimensionality of A Dermatology Data Set

Abdullah Caliskan[1], Hasan Badem[2], Alper Basturk[2], Mehmet Emin Yuksel[1]
[1]Dept. of Biomedical Engineering, Erciyes University, Kayseri, Turkey
{acaliskan, yuksel}@erciyes.edu.tr
[2]Dept. of Computer Engineering, Erciyes University, Kayseri, Turkey
{hbadem, ab}@erciyes.edu.tr

***Abstract*—The effect of using autoencoders for dimensionality reduction of a medical data set is investigated. A stack of two autoencoders has been trained for popular benchmark medical data set for dermatological disease diagnosis. The improvement of the presented approach has been visualized by the Principal Component Analysis method. Results shows that the use of a autoencoders significantly improves the accuracy of dermatological disease diagnosis.**

Keywords*—autoencoder, principal component analysis, dermatology.

I. INTRODUCTION

Dimensionality of data is very important for classification, clustering or visualization problems. Therefore, handling a large number of variables in multivariate data analysis causes a common problem which depends on the dimension of the data. In general, a simpler structure which has the least impact on the result of data analysis is preferred to reduce the data dimensionality. Therefore, it is desired to reduce the dimensionality of a data set while safeguarding the most relevant part of information [1].

Autoencoder (AE) networks have recently been of high research interest regarding data dimensionality reduction. Indeed, autoencoders are capable of reducing the dimensionality of data without needing any prior information about the data [2].

The AE uses an encoder function, which is an adaptive multilayered "encoder" network, to transform the high-dimensional data into a low-dimensional code-like representation of data, and a decoder function which recovers data from the code. At the beginning, the AE is initialized with random weights and trained by decreasing the inconsistency between the original data and its reconstruction. Gradient-based algorithms can be used for optimization of weights by backpropagating error derivatives. Due to the capability to reduce the data dimensionality, the AE is considered as a nonlinear generalization of the Principal Component Analysis (PCA) [3].

In this paper, the effect of reducing the dimension of a data set by using an AE network is investigated. The improvement obtained by using the proposed approach for classification and clustering methods is tested over a popular benchmark entitled *Dermatology Data Set*, which includes 33 attributes and 366 instances [4], and visualized by using the PCA. It is observed that the AE network efficiently reduces the dimension of the data set and effectively improves the classification performance.

The rest of the paper is organized as follows. In section II, the PCA is explained and visualization of data is described. In section III, the autoencoder and its application for reducing the dimension of data are explained. In section IV, experimental results for testing the performance of the AE method are reported. Finally, conclusions are presented in section V.

II. PRINCIPAL COMPONENT ANALYSIS

PCA is a method frequently used for decreasing the size of a data set while retaining its variance. PCA neglects low variance of the data set because it usually means background noise. Therefore, PCA assigns new coordinate axes to the direction where the highest variance is observed. The new axes or variables ranked by the variance are known as principal components. The first principal component (PC 1) shows the direction in which the highest variance of the data set is observed. The second principal component (PC 2) is perpendicular to the first principal component and it represents the second highest variance of the data sets. This pattern can be continued by the other components. Each component carries some amount of information about the data set [5], [6].

Low variance usually represents background noise therefore dimension of the data can be reduced by excluding the components having low variances without losing the relevant information in the data. This process is applied to high dimensional data as a pre-processing step [5], [6].

III. AUTOENCODER

The AE aims at learning the structure of a neural network model which can be considered as implementing the identity function $h_{W,b}(x) \approx x$. The most important purpose of the AE is to acquire the structural attributes of the data through the hidden layer. The AE is an unsupervised learning model trained easily by using the back-propagation algorithm. Therefore, the input data set is also used as the target output data set. The AE comprises two sections entitled as the *encoder* and the *decoder*, respectively. This is illustrated in Figure-1. The input weights of the hidden layer are defined as the *encoder weights* and the output weights are defined as the *decoder weights*. The number of nodes in the hidden layer may be more or fewer than the number of inputs. The outputs of the neurons in the hidden

978-1-5090-2386-8/16/$31.00

layer denote a compressed representation of the features in the input vector when the number of hidden neurons is fewer than that of the inputs. In other words, the AE acts in a similar manner as the principal component analysis [7].

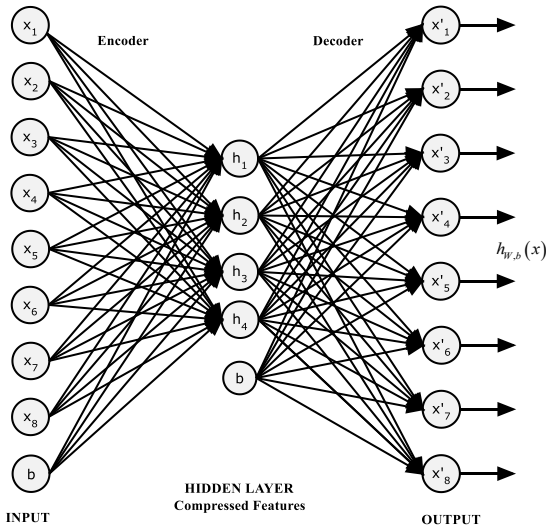

Figure 1: A model of an autoencoder network

Let $\{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$ represent the input vectors applied to the AE, where $x^{(i)}$ is the i^{th} attribute.

The activation value of the i^{th} neuron is represented as $a_i^{(2)}$ and given by

$$a_i^{(2)} = f\left(W_{ij}x_j^{(1)} + b_i^{(2)}\right) \tag{1}$$

where W_{ij} is weights between input layer and hidden layer.

The mean activation value of j^{th} neuron for all data set is symbolized with $\hat{p}_j$ and given by (2).

$$\hat{p}_j = \frac{1}{m}\sum_{i=1}^{m}\left[a_j^{(2)}\left(x^{(i)}\right)\right] \tag{2}$$

The mean activation values of each neuron in the hidden layer is desired to be equal to p which is termed as the *sparsity parameter*. The sparsity parameter is set to nearly $p = 0.05$. The training is achieved by minimizing the objective function of the backpropagation algorithm, which corresponds to updating equation (3) [8].

$$J_{sparse}(W,b) = J(W,b) + \beta\sum_{j=1}^{s_2} KL(p \parallel \hat{p}_j) \tag{3}$$

where s_2 is the number of neurons in the hidden layer. Also $J(W,b)$ is defined as follows:

$$J(W,b) = \frac{1}{n}\sum_{k=1}^{n} e_k^2 + \frac{\lambda}{2}\|W\|; \tag{4}$$

where e_k is the error between the model and the data set, and λ is the regularization term (also known as the *weight decay term*). This term is used to prevent overfitting.

$\sum_{j=1}^{s_2} KL(p \parallel \hat{p}_j)$ is a penalty term added to the objective function and β is described as the *sparsity penalty term*. Here $KL()$ is the Kullback-Leibler (KL) divergence and is calculated as fallows [8]:

$$KL(p \parallel \hat{p}_j) = plog\frac{p}{\hat{p}_j} + (1-p)log\frac{1-p}{1-\hat{p}_j} \tag{5}$$

IV. REDUCING THE DIMENSIONALITY OF THE DERMATOLOGY DATA SET

In this section, the AE method is applied for reducing the dimensionality of the dermatology data set. First, the dermatology data set used in the experiments is described in detail. Secondly implementation of the AE and its application to data are explained. Finally, performance improvement obtained by using the AE method is visualized by utilizing the PCA method.

A. Dermatology Data Set

It is known that acanthosis and parakeratosis can be observed in different degrees in almost all dermatology diseases. Fibrosis of the papillary dermis can be seen for chronic dermatitis. For lichen planus, melanin incontinence, disappearance of the granular layer, vacuolization and damage of the basal layer, saw-tooth appearance of retes and a band-like infiltrate are diagnostic features. Exocytosis may be seen in lichen planus, pityriasis rosea and seboreic dermatitis diseases. Psoriasis is diagnosed over thinning of the suprapapillary epidermis and clubbing of the rete ridges. Follicular horn plug and perifollicular parakeratosis are usually indicators of the pityriasis rubra pilaris disease.

The data set contains 366 vectors in total. Each vector in the data set includes 34 features obtained from a patient. Each feature has either nominal (discrete) or linear (continuous) value having different weights showing the relevance to the diagnosis. If the family history of a patient is available, its value is set to 1, otherwise set to 0. The age of a patient is evaluated in terms of years. Clinical and histopathological features are categorized relatively in the range of $0-3$ [9], [10]. Detailed description of the data set is given in Table I [4], [10].

Table I: Detailed distribution of Dermatology Data Set

Value of Class	Labels of Class	Number of Instances
1	Psoriasis	112
2	Seboreic dermatitis	61
3	Lichen planus	72
4	Pityriasis rosea	49
5	Cronic dermatitis	52
6	Pityriasis rubra pilaris	20

B. The Application of the Autoencoder

The application of the AE to dermatology data set is performed with the following steps:

First, features in the data set under investigation is statistically normalized so as to make its mean 0 and variance 1. In order to obtain a regularisation before starting the analysis of the data set, the tuning parameters of the two AEs used in the classification experiments are chosen as in Table II. Then, statistically normalized features are applied to the input of the first AE. Following this, the code output of first AE is applied to the input of the second AE and the code output of the second AE is obtained. Finally, the PCA method is utilized to visualize the code outputs of the two AEs to observe the improvement obtained by using the two AEs. The obtained plots are shown in Figure 2.

Table II: Setting parameters

Autoencoder 1		Autoencoder 2	
Number of Neuron	400	Number of Neuron	6
Sparsity (p)	0.2	Sparsity (p)	0.4
Lambda (λ)	0.003	Lambda (λ)	0.003
Beta (β)	1	Beta (β)	4
Training Algorithm	LFBGS	Training Algorithm	LFBGS
Iteration	400	Iteration	400
Input Normalized?	Yes	Input Normalized?	Yes

Figure 2a shows the first two principal components of the data set. It is clearly seen from this figure that only two out of six classes are distinguishable from each other. This is an undesirable situation regarding classification. The classification performance in Figure 2b is significantly improved by using the first AE, which shows the first two principal components of the codes obtained at the output of first AE. The classification performance is further improved in Figure 2c, which shows the first two principal components of the codes obtained at the output of second AE. Hence, the three plots in Figure 2 reveal that the use of the two AEs for preprocessing the data before classification significantly improves the classification performance.

V. CONCLUSION

The use of AEs to preprocess a data set for reducing its dimensionality and improving classification performance is investigated. The performance of the proposed approach is tested on a popular medical data set entitled as The Dermatology Data Set available in the literature. It is seen that the AE networks efficiently reduce the dimension of the data set and effectively improve the classification performance.

REFERENCES

[1] F. Saadaoui, P. R. Bertrand, G. Boudet, K. Rouffiac, F. Dutheil, A. Chamoux, "A Dimensionally Reduced Clustering Methodology for Heterogeneous Occupational Medicine Data Mining,"*in IEEE Transactions on NanoBioscience,* vol.14, no.7, pp.707-715, Oct. 2015

[2] G. E. Hinton, R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks,"*Science,* vol. 313, pp. 504-507, July 2006.

[3] S. Huang, D. Yang, G. Yongxin, X. Zhang, "Combined supervised information with PCA via discriminative component selection,"*SInformation Processing Letters*, Vol. 115, no. 11, pp. 812-816, November 2015.

[4] M. Lichman. (2013). UCI Machine Learning Repository. [Online] *University of California, Irvine, School of Information and Computer Sciences*. Available: http://archive.ics.uci.edu/ml

[5] J. E. Jackson, *A User's Guide to Principal Components*. John Wiley and Sons, 1991.

[6] I. T.Jolliffe, *Principal Component Analysis, 2nd edition*, Springer, 2002.

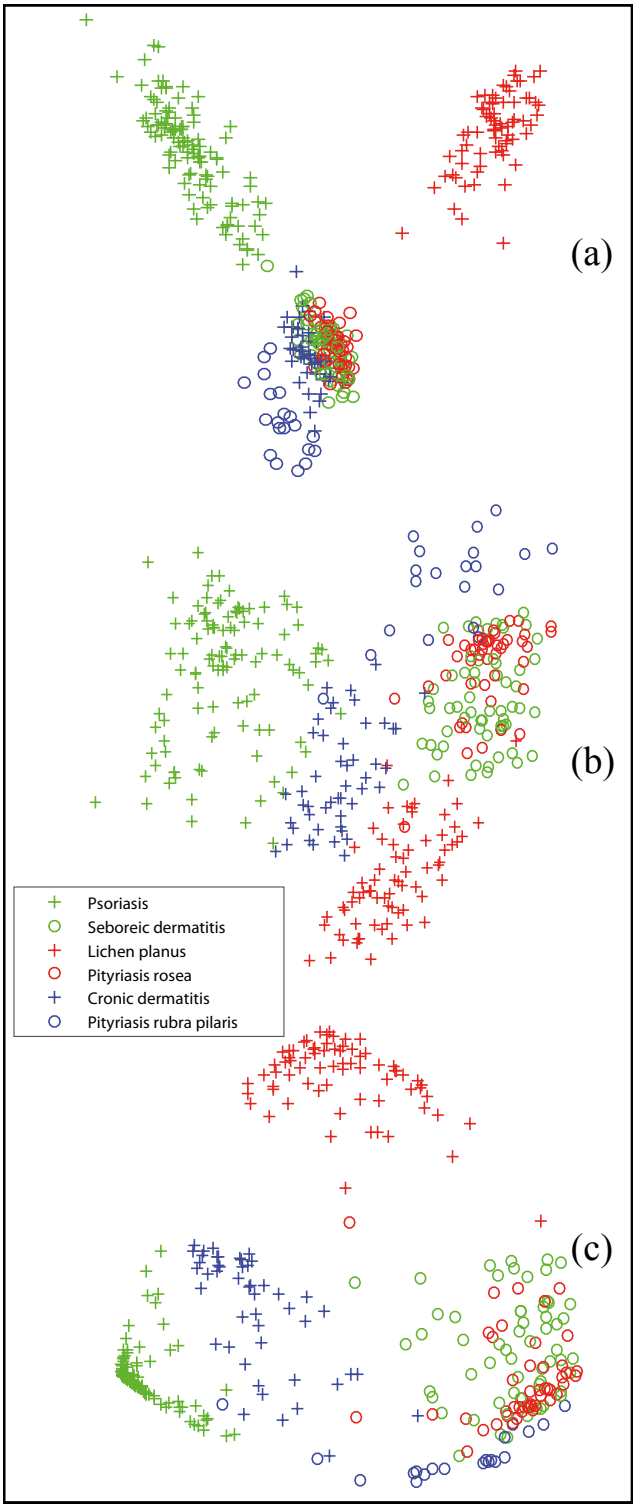

Figure 2: Plot of a) visualized only normalized Data set, b) visualized output of first Autoencoder c) visualized output of second Autoencoder

[7] P. Baldi, I. Guyon, G. Dror, V. Lemaire, G. Taylor, D. Silver. "Autoencoders, unsupervised learning and deep architecture,". *JMLR:work shop on unsupervised and transfer learning*, 2012, pp.37-50

[8] Sparse Autoenkoder. (2013), [Online] *Unsupervised Feature Learning and Deep Learning Tutorial*. Available: http://deeplearning.stanford.edu/wiki/index.php/Autoencoders_and_Sparsity.

[9] K. Polat, S. Güneş, "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems,"*Expert Systems with Applications*, Vol. 36, No. 2, Part 1, pp. 1587-1592, March 2009,

[10] H. A. Güvenir,,G. Demiröz,, N. Ilter, "Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals,". *Artificial intelligence in medicine*, Vol. 13 No.3, pp. 147-165, 1998