2019 IEEE 4th International Conference on Advanced
Robotics and Mechatronics (ICARM)

Unsupervised classification of high-dimension and low-sample data with variational autoencoder based dimensionality reduction

Mohammad Sultan Mahmud and Xianghua Fu

Abstract— **In data mining research and development, one of the defining challenges is to perform classification or clustering tasks for relatively limited-samples with high-dimensions data, also known as high-dimensional limited-sample size (HDLSS) problem. Due to the limited-sample-size, there is a lack of enough training data to train classification models. Also, the 'curse of dimensionality' aspect is often a restriction on the effectiveness of many methods for solving HDLSS problem. Classification model with limited-sample dataset lead to overfitting and cannot achieve a satisfactory result. Thus, the unsupervised method is a better choice to solve such problems. Due to the emergence of deep learning, their plenty of applications and promising outcome, it is required an extensive analysis of the deep learning technique on HDLSS dataset. This paper aims at evaluating the performance of variational autoencoder (VAE) based dimensionality reduction and unsupervised classification on the HDLSS dataset. The performance of VAE is compared with two existing techniques namely PCA and NMF on fourteen datasets in term of three evaluation metrics namely purity, Rand index, and NMI. The experimental result shows the superiority of VAE over the traditional methods on the HDLSS dataset.**

Index Terms— **HDLSS dataset, dimensionality reduction, variational autoencoder, unsupervised classification**

I. INTRODUCTION

In many domains, including bioinformatics, ecology, geology, neuroscience dataset are identified by a large number of features p but a small of samples N (records). These datasets named the high-dimension limited-sample-size (HDLSS) dataset, often addressed $p \gg N$. HDLSS data classification and clustering are both crucial and challenging tasks in data mining and machine learning. High variance and bias are the main concern in dealing with HDLSS dataset. Consequently, simple and highly regularized classification and regression techniques usually adopted [1].

Over the past years, the caution of insufficiently small-sample-size dataset has been flagged [2][3][4]. Many datasets with limited-sample-size are typically too small to support for the split into training and testing or k-fold cross-validation. However, data miners train a classifier model and evaluate classification accuracy. It can be challenging to build a stable and reliable classifier and draw conclusions from such dataset.

*This research is supported by the National Nature Science Foundation of China under grant 61472258.

M. S. Mahmud is with the Big Data Institute, College of Computer and Software Engineering, Shenzhen University, China 518060. (fax: +86-755-26534780; e-mail: sultan@ szu.edu.cn).

X. Fu is with the Big Data Institute, College of Computer and Software Engineering, Shenzhen University, China 518060, also with the Faculty of Arts and Sciences, Shenzhen Technology University, China 518118. (e-mail: fuxh@szu.edu.cn).

The performance of classification and clustering algorithms tends to decay due to a phenomenon called the curse of dimensionality [5]. Hence, dimensionality reduction is an essential desirable requirement for classification and clustering when p is large. Principal components analysis (PCA) and non-negative matrix function (NMF) are the most widely used methods for dimensionality reduction [6]. Efficient dimensionality reduction model using PCA and NMF usually require sufficient data. Unless, traditional dimensionality reduction methods (e.g., PCA, NMF) might be less effective. In the setting of $p > N$, there is considerable employment of PCA. However, in the case of $p \gg N$, PCA's obtained transformed dimension is lower than or equal to sample size ($d \leq N$), there is difficulty in preserving information about the original data in such extremely lower-dimension. Therefore, for HDLSS problems ($p \gg N$), it is clear that the basic formulation of PCA does not achieve.

Recently, deep learning has succeeded in a diversity of fields to extract patterns from high-dimensional space such as image, speech, text, and vision [7][8]. The weakness of deep learning is receiving a large number of training data to guarantee learning precision. Different types of deep architecture proposed to resolve the insufficient samples problem [9][10]. Unsupervised deep learning models such as generative adversarial net (GAN) and variational autoencoder (VAE) have been shown the modeling capability without the demand of data labels. VAE is capable of generating 'blurry' data compared to other generative models, is also stable to train [11]. Moreover, unlike many existing techniques (e.g., PCA), VAE able to reduce the dimension as necessary from the high-dimensional space.

Our recent work [4] presented an empirical investigation that the VAE based dimensionality reduction outperforms other approaches (i.e., PCA, fastICA, FA, NMF, and LDA) in classification and provided some analysis about this observation. It is also observed that all the classifiers and dimensions do not act in the same way for classification accuracy. There are varieties of classification algorithms, but the challenge is an appropriate selection in the application of limited-sample dataset. Since it is challenging to get ideal classifier especially in the case of limited-sample, it is more advantageous to adopt the unsupervised learning framework.

As mentioned before this study focused on reliable HDLSS data classification. Based on the above analysis, to avoid these phenomena, this study presents an unsupervised classification technique on the HDLSS dataset. In particular, we employed VAE for dimensionality reduction and K-means clustering applied to the latent space (low-dimension) of VAE. Then, the clustering result matches with the original class levels. We demonstrate the performance of the

978-1-7281-0064-7/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: UNIVERSITY OF MISSISSIPPI. Downloaded on August 28,2020 at 20:51:31 UTC from IEEE Xplore. Restrictions apply.

presented approach on different types of HDLSS dataset such as biological, image, mass spectrometry, etc.

II. Literature Review

The analysis of HDLSS data is also vital for scientific discoveries in many areas. Over the past decades in literature, there are many methods proposed for HDLSS data analysis. When dealing with HDLSS data, the overfitting and high-variance gradients are the main challenges for the majority of the model. In the past, a significant study has been performed on HDLSS asymptotic theory, where the sample size N is fixed or $N/d \to 0$ as the data dimension $d \to \infty$ [12][13][5]. Also, investigated different types of geometric representations of HDLSS data, and pointed inconsistency characteristics of the sample eigenvalues and eigenvectors in the HDLSS setting in [14].

Supervised or classification methods are frequently used for HDLSS data analysis. Most of the achievement in classification show that larger samples and low-dimension can enhance the performance of classifiers. However, adequate large samples are required to build a classifier model with excellent generalization ability, expected that work equally fit for the train and independent test dataset. Hence, the classification technique does not suit with the small-sample dataset. Clustering (unsupervised) techniques are employed in HDLSS analysis to avoid classification overfitting (training and validation data). Clustering task is an exploratory data analysis. Accordingly, the explorer might have little or no prior information about the data domain and parameters for successful clustering. Moreover, dimensionality reduction has been extensively considered as a powerful tool to analyze HDLSS data. Most widely applied dimensionality reduction method is principal component analysis (PCA). PCA has been used in the classification and clustering of correlated microarray genes expression and RNA-sequence data [15][16][6].

In the literature, most papers studied PCA as a dimensionality reduction method, though it is more valuable for data visualization in high-dimensional contexts [17][18]. Moreover, PCA conquers the dimensionality of the data linearly that not able to extract nonlinear relationships of data [19]. NMF is another widely used tools for high-dimensional data analysis. NMF has also been implemented for genes clustering, microarray data, and protein sequence recognition [20][21]. PCA is deterministic whereas NMF is stochastic, so NMF seems to be more fit for HDLSS data analysis than PCA.

Over the year, deep leaning (DL) techniques have succeeded in state-of-the-art performance in applications with large-sample-size. Nevertheless, latterly, few attempts have been devoted to applying DL to the HDLSS problem [7][8][22]. DL also suffer overfitting on HDLSS problems. 'Dropout' method proposed to prevent overfitting by efficiently decrease the parameters of the full-connection layer, and resolve the difficulty of limited-samples [22][9]. Moreover, transfer learning based deep convolutional neural network (CNN) developed to resolve the issue of the limited-sample dataset [10].

III. Method

This research focuses on two features. Firstly, exclusively applying a deep generative model, variational autoencoder (VAE) to investigate the dimensionality reduction ability. Secondly, applying unsupervised classification, clustering technique on HDLSS dataset instead of classification (supervised model) to avoid overfitting. However, conventional classification techniques cannot cope with HDLSS dataset because of their limited-samples to build and test a classifier. The work process is illustrated in Fig. 1.

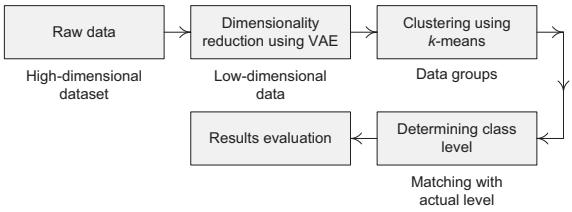

Figure 1. Detail workflow.

Step 1. Determine the required feature space (latent space) for dimensionality reduction and employ the VAE model on the original features (to all the features).

Step 2. Remove the levels and apply 'sensible' clustering on the VAE's obtained lower-dimensional generative space (latent space) by considering the number of classes as the number of clusters.

Step 3. Determining the class level from the clustering result and compare with ground-truth.

Clustering is unsupervised learning. The basic task of exploratory data analysis is that groups similar data in a cluster and dissimilar data in separate clusters. K-means is broadly used and one of the topmost data mining techniques [23]. In this study, simple K-means clustering applied to avoid overfitting phenomena and to investigate dimensionality reduction ability (quality) of VAE on the HDLSS dataset.

A. Variational autoencoder (VAE) model

VAE [24][25] is an unsupervised deep learning model and uses a latent representation z of given data x with stochastic variables, replaces traditional autoencoder (see Fig. 2). Autoencoder is deterministic, is trained by decreasing reconstruction error. Instead, VAE is stochastic that learn the distribution of features over samples (used variational inference). The main idea of variational inference is to pose the inference by the approach, it as an optimization problem. VAE performs these modeling by learning two distinct parameters, mean and standard deviation. The model utilizes a Kullback-Leibler (KL) divergence, to minimize the difference between the approximate and true latent posteriors. In this work, we aim to build a VAE that compresses high-dimensional features and reveals a relevant latent space.

A variational autoencoder consists of an encoder, a decoder, and a loss function.

978-1-7281-0064-7/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: UNIVERSITY OF MISSISSIPPI. Downloaded on August 28,2020 at 20:51:31 UTC from IEEE Xplore. Restrictions apply.

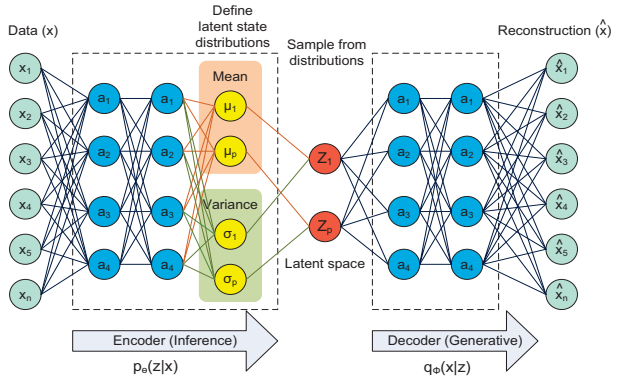

Figure 2. VAE framework [24][25].

Encoder. The encoder is a neural network, compresses data x into a latent representation z (lower-dimensional space). It holds weight and bias parameter θ. The encoder 'encodes' the data which is N-dimensional into a latent (hidden) representation z which is much less than N dimension. The latent representation is stochastic; encoded output parameter is a Gaussian probability density $p_\theta(z|x)$.

Decoder. The decoder is another neural network, receives the latent variable z as input and reconstruct output $\hat{x}$ from the parameters of a probability distribution of the data, and has weight and bias parameter ϕ. The decoder reproduces the data is denoted by $q_\phi(z|x)$. It performs from a lower to a higher dimension. Information loss calculated using the regeneration log-likelihood, $\log q_\phi(z|x)$. This measure affirms how efficiently the decoder 'decodes' the real-valued numbers in z into N real-valued numbers.

Loss function. The loss function is a negative log-likelihood with a regularizer because of no global representation that is participated by all data points. This loss term can decompose into only phases that depend on a particular data point l_i. The total loss is $\sum_{i=1}^{n} l_i$ for nl data points. The loss function l_i data point x_i is:

$$l_i(\theta, \phi) = \underbrace{-E_{(z\sim p_\theta(z|x_i)}[\log q_\theta(x_i|z)]}_{\text{reconstruction loss}} + \underbrace{KL(p_\theta(z|x_i)||p(z)}_{\text{regularizer}}$$

In the above equation, the first term is expected reconstruction error or negative log-likelihood of the ith data point. The second term is a regularizer, in the objective is the KL divergence of the encoder's distribution $p_\theta(z|x)$ and $p(z)$. This divergence defines how close probability density functions q and p. KL loss penalizes the model if the encoder outputs are different than a standard normal distribution.

IV. Experiments and Discussions

A. Datasets

In the experiment used fourteen (14) high-dimensional low-sample ($p \gg N$) datasets that obtained from the Arizona State University repository[1]. The characteristics of the datasets are presented in Table I.

TABLE I. Datasets at a glance.

SI	Dataset	Sample (N)	Feature (p)	Class (K)	Keywords
	Biological				
1	ALLAML	72	7129	2	binary, continuous
2	CARCINOM	174	9182	11	multi-class, continuous
3	CLL_SUB_111	111	11340	3	multi-class, continuous
4	GLI_85	85	22283	2	binary, continuous
5	GLIOMA	50	4434	4	multi-class, continuous
6	NCI9	60	9712	9	multi-class, discrete
7	PROSTATE_GE	102	5966	2	binary, continuous
8	SMK_CAN_187	187	19993	2	binary, continuous
9	TOX_171	171	5748	4	multi-class, continuous
	Face image				
10	ORLRAWS10P	100	10304	10	multi-class, continuous
11	PIXRAW10P	100	10000	10	multi-class, continuous
12	WARPAR10P	130	2400	10	multi-class, continuous
13	WARPPIE10P	210	2420	10	multi-class, continuous
	Other				
14	ARCENE	200	10000	2	binary, continuous

B. Experiment design

The experiment focused on two distinct types of model. Firstly, M_1: without dimensionality reduction, ensures that all the features are used for clustering. Secondly, M_2: with dimensionality reduction, selected different choice of dimension (i.e., $50,100, \ldots, 500$). We implement our proposed technique VAE and two existing dimensionality reduction techniques PCA and NMF are compared. Experimental environment is: Intel(R) Core i3-23S0M, CPU speed 2.30 GHz, RAM 4.0 GB, an x64-based processor; the code implementation is based on Keras and Tensorflow framework.

C. VAE design

VAE designed with the following structure: input encoded to d features and reconstructed back to the original dimensions ($d = 50, 100, \ldots, 500$). The VAE network trained with an 'adam' optimizer included 'rectified linear units' and 'sigmoid' activation in the encoding and decoding stage, respectively. We conducted a parameter sweep over batch size $50, 100, 120,$ and 200; epochs $100, 150, 200,$ and 300; learning rates $0.005,\ 0.001,\ 0.0015,$ and 0.0025; warmups (k) 0.01,$0\ .05, 0.001,$ and 0.0005, respectively. k controls the KL divergence loss offers to learn. In general, training was moderately stable for parameter combinations, however, was inferior for larger batches, especially with low learning rates. The best parameter succession based on validation loss obtained batch size 100, learning rate 0.0005, and 150 epochs. Training stabilized nearly 120 epochs.

D. Evaluation criteria

We used three well-known external cluster evaluation measures namely purity, rand index (RI), and normalized mutual information (NMI).

E. Analysis and discussions

Table II presents the summarized results for each dataset (without dimensionality reduction and an average of 10 reduced dimensions). From Table II, it is observed that apart from two datasets (i.e., GLI_85 and SMK_CAN_187), for all

[1] http://featureselection.asu.edu/

Authorized licensed use limited to: UNIVERSITY OF MISSISSIPPI. Downloaded on August 28,2020 at 20:51:31 UTC from IEEE Xplore. Restrictions apply.

other datasets VAE performs better than other techniques in terms of purity, RI, and NMI.

Note that, for HDLSS data PCA's transformed dimension cannot be greater than sample-size. Therefore, obtain transformed dimension by PCA is equal or less than samples ($d =< N$). In the clustering, we use eleven dimensions (d) values: $50, 100, ..., 500$, and all original features. An empirical analysis of d values is provided in Fig. 3, 4, and 5.

Fig. 3 depicts the purity of different techniques of different dimensions on the experimental datasets. The purity of VAE with d-values are approximately similar in most cases. From this fact, we can confirm that there is sufficient condition of the low-dimensional transformed consistency for VAE. We also present the rand score (RI) and NMI on each dataset in Fig. 4 and 5, respectively. From Fig. 4 and 5, it is clear that VAE performs better than the PCA and NMF in the term of rand index, and NMI on the datasets used in this study.

According to the experimental results shown in the Fig. 3, 4, and 5, it is proven that dimensionality reduction provides higher performance than the use of all features, and VAE performs significantly better than PCA and NMF.

TABLE II. AVERAGE PERFORMANCE OF DIFFERENT TECHNIQUES ON THE FOURTEEN DATASETS (HIGHER VALUE IS BETTER).

		Purity				**Rand index (RI)**				**NMI**			
		M_1	M_2			M_1	M_2			M_1	M_2		
SI	**Datasets**		PCA	NMF	VAE		PCA	NMF	VAE		PCA	NMF	VAE
1	ALLAML	0.71	0.65	0.65	**0.88**	0.58	0.54	0.54	**0.80**	0.09	0.08	0.04	**0.49**
2	CARCINOM	0.67	0.64	0.27	**0.75**	0.91	0.91	0.68	**0.94**	0.32	0.65	0.20	**0.76**
3	CLL_SUB_111	0.53	0.54	0.48	**0.61**	0.55	0.58	0.43	**0.59**	0.19	0.32	0.06	**0.24**
4	GLI_85	0.65	**0.71**	0.69	**0.71**	0.54	0.58	0.57	**0.59**	**0.20**	0.07	0.03	0.13
5	GLIOMA	0.60	0.58	0.52	**0.67**	0.73	**0.76**	0.67	**0.76**	0.49	**0.51**	0.41	0.46
6	NCI9	0.43	0.35	0.26	**0.53**	0.81	0.78	0.42	**0.85**	0.44	0.37	0.27	**0.52**
7	PROSTATE_GE	0.58	0.58	0.57	**0.61**	0.51	0.51	0.51	**0.52**	0.02	0.02	0.05	**0.06**
8	SMK_CAN_187	0.52	**0.56**	0.52	0.54	**0.50**	**0.50**	**0.50**	**0.50**	0.00	0.01	**0.02**	0.01
9	TOX_171	0.44	0.45	0.31	**0.56**	0.68	0.65	0.34	**0.71**	0.16	0.28	0.08	**0.33**
10	ORLRAWS10P	0.76	0.73	0.36	**0.82**	0.94	0.93	0.75	**0.95**	**0.85**	0.79	0.43	**0.85**
11	PIXRAW10P	0.81	0.77	0.44	**0.89**	0.95	0.94	0.80	**0.97**	0.90	0.84	0.55	**0.91**
12	WARPAR10P	0.32	0.29	0.27	**0.41**	0.84	0.82	0.69	**0.85**	0.29	0.27	0.24	**0.44**
13	WARPPIE10P	0.31	0.32	0.32	**0.64**	0.82	0.82	0.67	**0.90**	0.33	0.33	0.30	**0.67**
14	ARCENE	0.34	0.65	0.57	**0.66**	**0.55**	0.54	0.51	**0.55**	**0.09**	0.08	0.03	**0.09**
	Score (Out of 14)	0	2	0	**13**	2	2	1	**14**	3	1	1	**12**

M_1: Without dimensionality reduction M_2: With dimensionality reduction

Bold values are best comparative to other methods

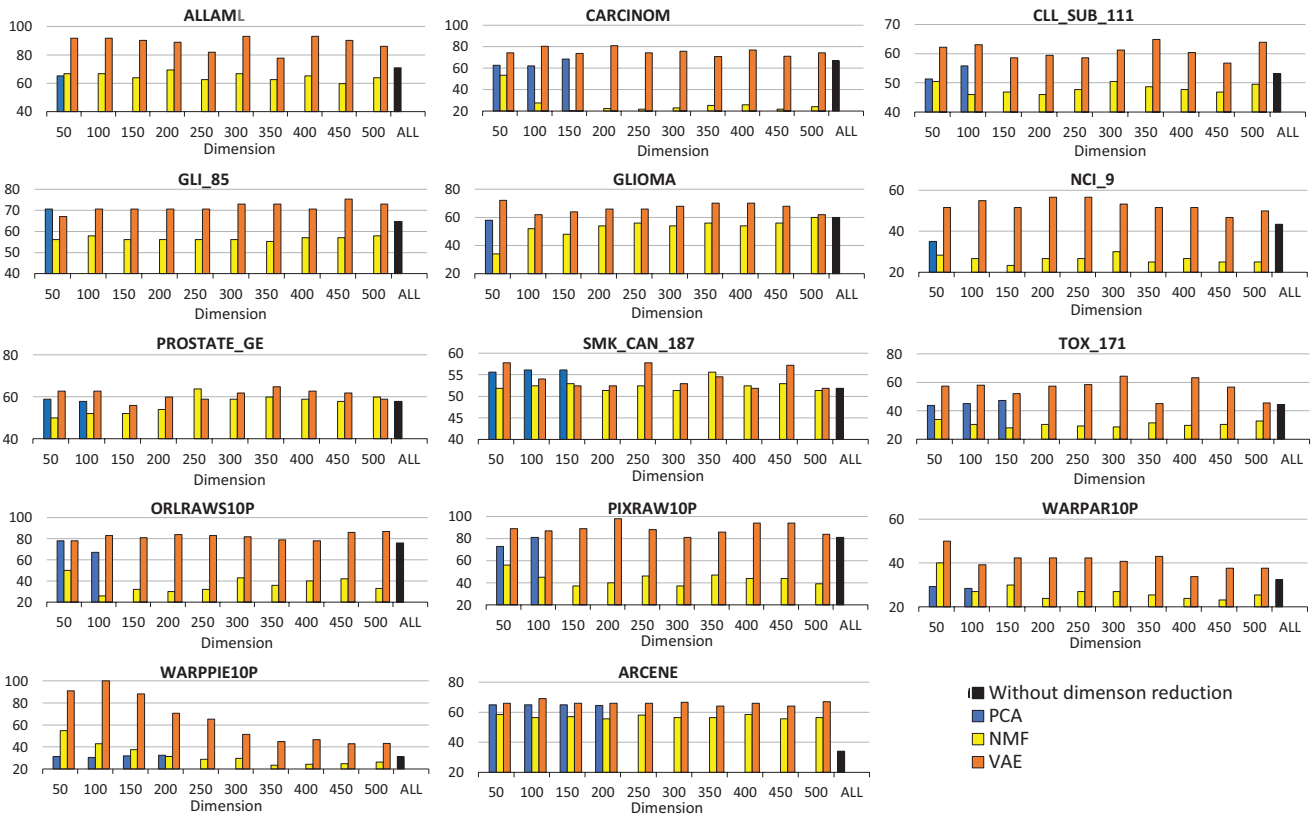

Figure 3. Detail purity of different techniques on the fourteen datasets (higher value is better).

Authorized licensed use limited to: UNIVERSITY OF MISSISSIPPI. Downloaded on August 28,2020 at 20:51:31 UTC from IEEE Xplore. Restrictions apply.

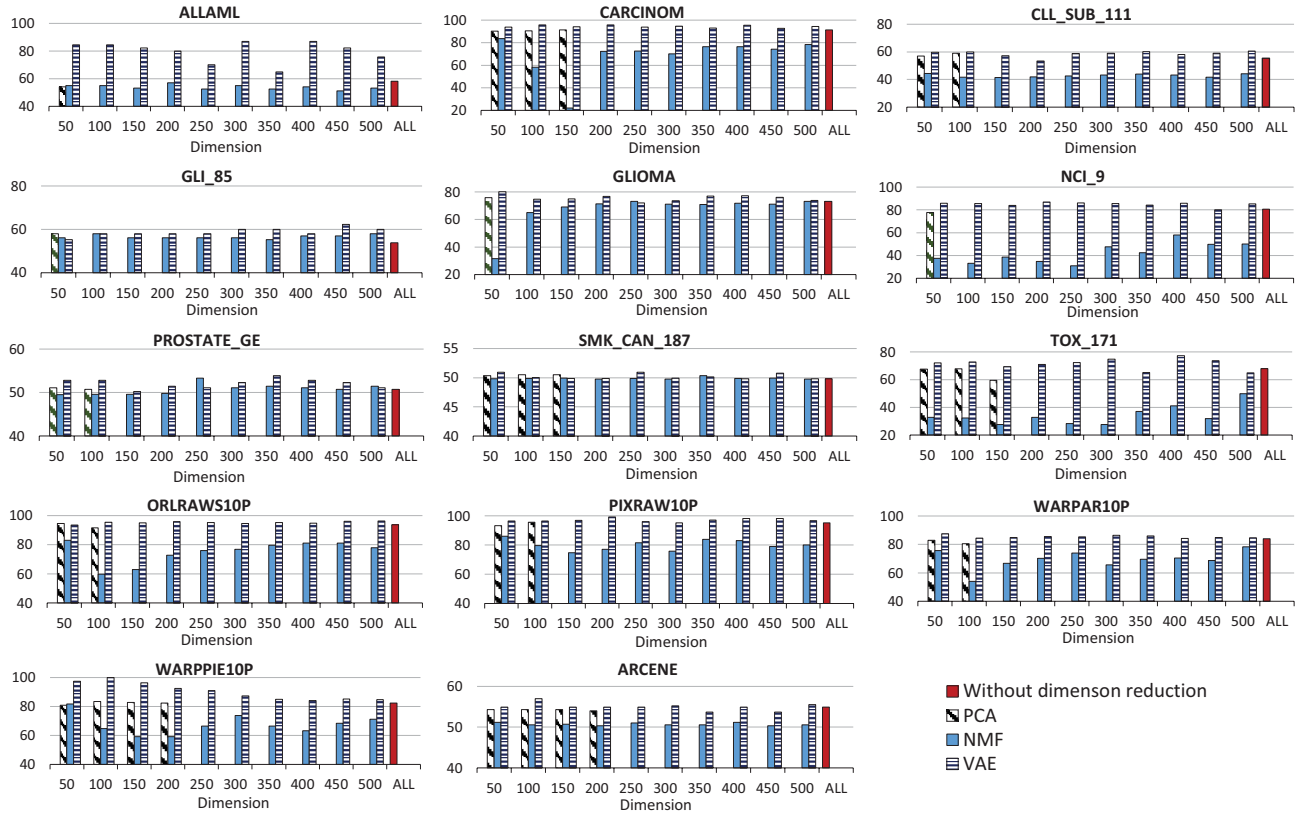

Figure 4. Detail RI of different techniques on the fourteen datasets (higher value is better).

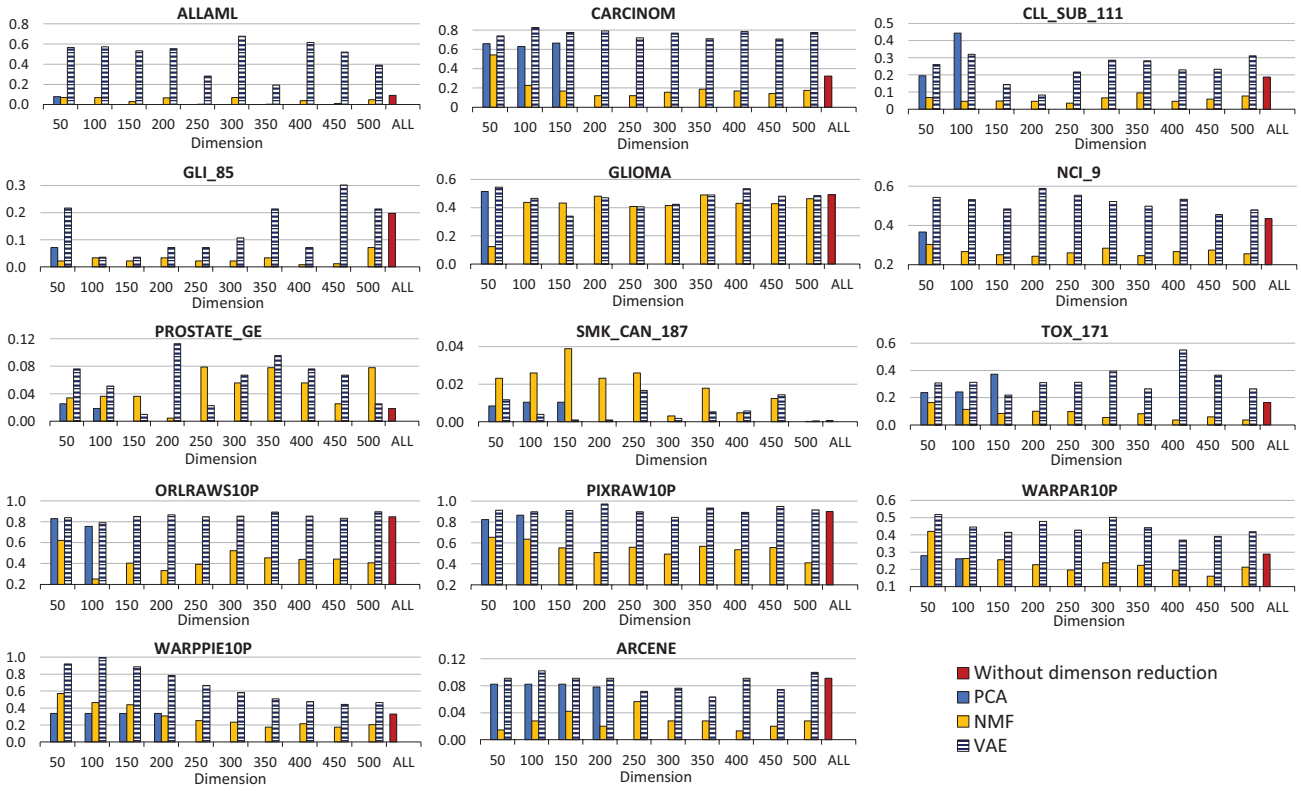

Figure 5. Detail NMI of different techniques on the fourteen datasets (higher value is better).

Authorized licensed use limited to: UNIVERSITY OF MISSISSIPPI. Downloaded on August 28,2020 at 20:51:31 UTC from IEEE Xplore. Restrictions apply.

F. Statistical analysis

A nonparametric test, sign test is performed to compare the statistical significance of VAE over the PCA, NMF, and without dimensionality reduction method.

Sign test: The right-tailed (one-tailed) sign test is conducted in the 95% significance level ($\text{i.e.}, \alpha = 0.05$). Fig. 6 depicts the sign test result on the fourteen HDLSS datasets in term of used evaluation measures purity, RI, and NMI. In the figure, the first three bars display the z-value (test statistics value) for VAE versus other techniques whereas the fourth bar shows the **z-ref** value. If calculated z-value is higher than the **z-ref** value, then it means that the performances of VAE versus the other techniques are statistically significant. From Fig. 6, it is clear that the result obtained by VAE is significantly better than without dimensionality reduction, PCA, and NMF overall in datasets.

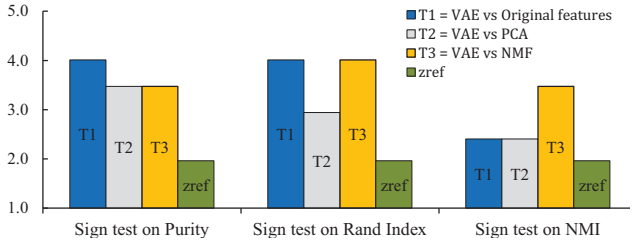

Figure 6. Sign test of VAE on the 14 datasets.

V. Conclusion

This study suggests the effectiveness of variational autoencoder based dimensionality reduction and unsupervised classification to deal with HDLSS data. We have investigated the dimensionality reduction ability of VAE on HDLSS data. An empirical analysis is shown on the fourteen (14) HDLSS datasets and compared to the two (2) other traditional techniques namely PCA and NMF. The experimental results indicate that VAE performs better than traditional techniques in term of all three evaluation metrics on the used datasets in this study. A statistical test, nonparametric sign test also conducted on the results of techniques that also demonstrate the significance of VAE over the existing techniques (e.g., PCA, NMF) on the HDLSS dataset.

HDLSS data classification severe overfitting and high-variance gradients, and clustering have shown to be a valid selection when facing these problems. Besides, an efficient dimensionality reduction method is essential for HDLSS analysis. Instead of PCA while applying VAE can decrease the dimensions as to fit from the high-dimensional dataset that improves the performance. This study combines the advantages of both unsupervised dimensionality reduction and unsupervised classification.

Further investigation is required to generalize the finding. It sounds interesting to make the approach applicable to semi-supervised clustering. In future work, we intend to find an efficient dimension selection method (use of appropriate d from p). Moreover, the reliability of the HDLSS data classification can be improved in meta or ensemble model.

References

[1] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, 2nd ed., Springer Series in Statistics. New York Inc., 2008.
[2] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 94:1-94:45, 2017.
[3] L. I. Kuncheva and J. J. Rodrguez, "On feature selection protocols for very low-sample-size data," *Pattern Recog.*, vol. 81, pp. 660-673, 2018.
[4] M. S. Mahmud, X. Fu, J. Z. Huang, and M. A. Masud, "High dimensional limited-sample biomedical data classification using variational autoencoder," in *Australasian Conference on Data Mining (AusDM 2018)*, pp. 30-42.
[5] J. Lv, "Impacts of high dimensionality in finite samples," *The Annals of Statistics*, vol. 41, no. 4, pp. 2236-2262, 2013.
[6] K. Yata and M. Aoshima, "Principal component analysis based clustering for high-dimension, low-sample-size data," *arXiv*, 2015.
[7] B. Liu, Y. Wei, Y. Zhang, and Q. Yang, "Deep neural networks for high dimension, low sample size data," in *Proc. of the 26th Int. Joint Conf. on Art. Intellig. (IJCAI'17)*, 2017, pp. 2287-2293.
[8] T. Yue and H. Wang, "Deep learning for genomics: A concise overview," *arXiv*, 2018.
[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929-1958, 2014.
[10] W. Zhao, "Research on the deep learning of the small sample data based on transfer learning," *Bioinformatics*, vol. 1864, no. 1, 2017.
[11] A. Lamb, V. Dumoulin, and A. Courville, "Discriminative Regularizationfor Generative Models," *arXiv*, 2016.
[12] P. Hall, J. S. Marron, and A. Neeman, "Geometric representation of high dimension, low sample size data," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 3, pp. 427-444, 2005.
[13] K. Yata and M. Aoshima, "Effective pca for high-dimension, low sample-size data with noise reduction via geometric representations," *J. Multivar. Anal.*, vol. 105, no. 1, pp. 193-215, 2012.
[14] S. Jung and J. S. Marron, "Pca consistency in high dimension, low sample size context," *The Annals of Statistics*, vol. 37, no. 6B, pp. 4104-4130, 2009.
[15] D. Mishra, R. Dash, A. K. Rath, and M. Acharya, *Feature Selection in Gene Expression Data Using Principal Component Analysis and Rough Set Theory*. New York, NY: Springer New York, 2011.
[16] M. Sato-Ilic, "Structural classification based correlation and its application to principal component analysis for high-dimension low-sample size data," in *IEEE Int. Conf. on Fuzzy Syst.*, 2012, pp. 1-8.
[17] J. O. Ramsay and B. W. Silverman, Applied functional data analysis: methods and case studies. New York, USA: Springer, 2002, vol. 77.
[18] D. Shen, H. Shen, H. Zhu, and J. S. Marron, "The statistics and mathematics of high dimension low sample size asymptotics," *Stat Sin.*, vol. 26, no. 4, pp. 1747-1770, 2016.
[19] A. Gupta, H. Wang, and M. Ganapathiraju, "Learning structure in gene expression data using deep architectures, with an application to gene clustering," in *IEEE Int. Conf. Bio. & Biomed.*, 2015, pp. 1328-1335.
[20] A. D. Pascual-Montano, P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo, and R. D. Pascual-Marqui, "bionmf: a versatile tool for non-negative matrix factorization in biology," *BMC Bioinformatics*, vol. 7, pp. 366-366, 2006.
[21] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics and Data Analysis*, vol. 52, no. 1, pp. 155-173, 2007.
[22] P. Danaee, R. Ghaeini, and D. Hendrix, "A deep learning approach for cancer detection and relevant gene identification," *Pacific Symposium on Biocomputing*, vol. 22, pp. 219-229, 2017.
[23] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1-37, 2007.
[24] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd Int. Conf. on Learning Representations (ICLR2014)*, 2014.
[25] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. of the 31st Int. Conf. on Mach. Learn. (ICML'14)*, Vol. 32, 2014, pp. 1278-1286.

Authorized licensed use limited to: UNIVERSITY OF MISSISSIPPI. Downloaded on August 28,2020 at 20:51:31 UTC from IEEE Xplore. Restrictions apply.