Unsupervised Feature Learning Using Domain Knowledge Based Autoencoder

Hyunjae Kim[1], and Byeng D. Youn[1]

[1]*Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul, 08826, Republic of Korea*
secutus07@snu.ac.kr, bdyoun@snu.ac.kr

ABSTRACT

Autoencoder is an unsupervised feature engineering technique, which is an emerging technique in the prognostics and health management (PHM) domain. However, since neural network based techniques such as auto-encoder have many hyper-parameters belonging to the number of hidden units, the number of layers and activation functions, substantial efforts are required in a heuristic way to make the autoencoder learn proper features. In this paper, we propose a novel method to regularize an auto-encoder by exploiting domain knowledge, such as mechanical engineering expertise. The proposed autoencoder learns robust features in a fast and efficient manner, therefore resulting in minimal consideration for the hyper-parameters. In the proposed method, some of the hidden units of the autoencoder are forcibly pre-trained by back-bone signals, such as 1X sinusoidal wave of vibration, and the remaining hidden units efficiently learn the features of fault signals by minimizing the redundancy of learned features. The domain knowledge based regularization reduces the degree of freedom (DOF) of the autoencoder model as well as guide the model to learn the more physically reasonable features. Various fault data measured from a journal bearing rotor testbed are used for demonstration of the proposed method.

1. INTRODUCTION

As part of efforts to find fault-related features, feature engineering based on unsupervised learning has been actively studied in the PHM domain [1]. Unsupervised learning has its main purpose in determining latent variables of high dimensional data and determining latent variables having a strong correlation with target task, such as fault diagnosis. Unsupervised learning treats two stages of feature extraction and selection, which have been used in the conventional PHM domain, as an automated sense without knowledge of the relevant domain. This is an inexpensive and convenient method compared to existing methods, but there is also a blind spot. If the intention of the user is not sufficiently reflected in the learning method, the possibility of learning incorrect latent variables increases. This is especially true in neural network based learning methods such as autoencoder because the high complexity and nonlinearity of the neural network can result in learning results falling into the local minima or overfitting the data.

Various studies have been conducted to overcome the limitations of unsupervised learning. Overfitting for a learning model means that the model only works well for training data and not for new inputs, and the methods for resolving it are called regularization. Methods for regularization of neural networks have been studied, including a norm penalty to limit network capacity, drop-out to exclude some units from training randomly, and early stopping to stop training on appropriate epochs before overfitting. Especially, there are special regularization methods only work for autoencoder, such as, sparse autoencoder, contractive autoencoder, and denoising autoencoder.

The regularization of the autoencoder shown in the above paragraph is implemented so that the autoencoder can obtain useful expressions (latent variables) for data obtained from various domains in general viewpoint. However, from the perspective of regularization imposing prior knowledge in order to obtain "useful expressions" from data [6], more knowledge of target data and tasks can be used for better feature learning. For example, some latent variables are imposed on a neural network in advance for the training. When a dataset has a set of optimal latent variables, the best feature learning method results we can expect is to identify all the latent variables. However, the latent variables obtained by the autoencoder vary depending on the network architecture, learning late, and class balance in dataset. However, if the user is able to use a useful expression, i.e. knowing part of the latent variable, and using it to force the autoencoder to search for the remaining latent variables, the learned latent variables will be obtained more optimally.

One concrete example of the idea proposed above is that, in many engineering systems, fault signals have somewhat similar characteristics to normal signals. For example, if a stationary vibration signal generated by a specific system is measured close to a sine wave when the system is normal, the sine wave is distorted or a ripple is added when the system is faulty. The point is that it is important to teach the autoencoder some explicit latent variables (in this case, normal - sinusoidal signal) beforehand to learn all the latent variables of the rest of the fault signals. This is like the problem of finding axes of the feature space where latent variables exist with single axis is already known (assuming this axis is close to truth). In this case, we will define latent variables more accurately than finding all the axes from the beginning.

In order to implement the proposed concept to autoencoder, in this paper, we pre-train autoencoder using only normal data with some hidden units are deactivated thereafter, train all classes of data including faulty data with all hidden units are activated. This method captures the latent variables existing in the normal data prior to the main training and then induces additional hidden units which newly activated in the main training to learn the latent variables existing in the faulty data. This will be explained in more detail in Section 3.

The remaining of this paper are organized as follows. Section 2 reviews the basic principles and behavior of autoencoder and a notion of manifold learning. Section 3 describes the proposed learning method of the autoencoder and provides a simple example and a condition for this method to work well. Section 4 shows and analyzes the performance of the proposed method for normal and fault data in real engineering systems. Finally, we conclude with a discussion in Section 5.

REFERENCES

Alain, G., & Bengio, Y. (2014). What regularized auto-encoders learn from the data-generating distribution. *Journal of Machine Learning Research, 15*(1), 3563-3593.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2007). Analysis of representations for domain adaptation. *Advances in neural information processing systems, 19*, 137.

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. *ICML Unsupervised and Transfer Learning, 27*, 17-36.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on pattern analysis and machine intelligence, 35*(8), 1798-1828.

Bengio, Y., Paiement, J.-F., Vincent, P., Delalleau, O., Le Roux, N., & Ouimet, M. (2003). Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Mij, 1*, 2.

Bengio, Y., Yao, L., Alain, G., & Vincent, P. (2013). *Generalized denoising auto-encoders as generative models.* Paper presented at the Advances in neural information processing systems.

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). *Algorithms for hyper-parameter optimization.* Paper presented at the Advances in neural information processing systems.

Chen, M., Weinberger, K. Q., & Blitzer, J. (2011). *Co-training for domain adaptation.* Paper presented at the Advances in neural information processing systems.

Chen, M., Xu, Z., Weinberger, K., & Sha, F. (2012). Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683*.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization.* Paper presented at the Advances in neural information processing systems.

Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., . . . Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research, 17*(59), 1-35.

Glorot, X., Bordes, A., & Bengio, Y. (2011a). *Deep Sparse Rectifier Neural Networks.* Paper presented at the Aistats.

Glorot, X., Bordes, A., & Bengio, Y. (2011b). *Domain adaptation for large-scale sentiment classification: A deep learning approach.* Paper presented at the Proceedings of the 28th international conference on machine learning (ICML-11).

Goodfellow, I., Lee, H., Le, Q. V., Saxe, A., & Ng, A. Y. (2009). *Measuring invariances in deep networks.* Paper presented at the Advances in neural information processing systems.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). *Generative adversarial nets.* Paper presented at the Advances in neural information processing systems.

Gopalan, R., Li, R., & Chellappa, R. (2011). *Domain adaptation for object recognition: An unsupervised approach.* Paper presented at the Computer Vision (ICCV), 2011 IEEE International Conference on.

Goroshin, R., & LeCun, Y. (2013). Saturating auto-encoders. *arXiv preprint arXiv:1301.3577*.

Im, D. J., Kim, C. D., Jiang, H., & Memisevic, R. (2016). Generating images with recurrent adversarial networks. *arXiv preprint arXiv:1602.05110*.

Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M. (2014). *Semi-supervised learning with deep generative models.* Paper presented at the Advances in neural information processing systems.

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient backprop *Neural networks: Tricks of the trade* (pp. 9-48): Springer.

Makhzani, A., & Frey, B. (2013). K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.

Makhzani, A., & Frey, B. J. (2015). *Winner-take-all autoencoders.* Paper presented at the Advances in neural information processing systems.

Ming Harry Hsu, T., Yu Chen, W., Hou, C.-A., Hubert Tsai, Y.-H., Yeh, Y.-R., & Frank Wang, Y.-C. (2015). *Unsupervised Domain Adaptation with Imbalanced Cross-Domain Data.* Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering, 22*(10), 1345-1359.

Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A. Y. (2007). *Self-taught learning: transfer learning from unlabeled data.* Paper presented at the Proceedings of the 24th international conference on Machine learning.

Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., & Glorot, X. (2011). Higher order contractive auto-encoder. *Machine Learning and Knowledge Discovery in Databases*, 645-660.

Rifai, S., Vincent, P., Muller, X., Glorot, X., & Bengio, Y. (2011). *Contractive auto-encoders: Explicit invariance during feature extraction.* Paper presented at the Proceedings of the 28th international conference on machine learning (ICML-11).

Sanguinetti, G. (2008). Dimensionality reduction of clustered data sets. *IEEE Transactions on pattern analysis and machine intelligence, 30*(3), 535-540.

Skolidis, G. (2012). Transfer learning with Gaussian processes.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929-1958.

Srivastava, N., Mansimov, E., & Salakhutdinov, R. (2015). *Unsupervised Learning of Video Representations using LSTMs.* Paper presented at the ICML.

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res, 10*, 66-71.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). *Extracting and composing robust features with denoising autoencoders.* Paper presented at the Proceedings of the 25th international conference on Machine learning.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research, 11*(Dec), 3371-3408.

Yu, D., & Deng, L. (2011). Deep Learning and Its Applications to Signal and Information Processing [Exploratory DSP. *IEEE Signal Processing Magazine, 28*(1), 145-154. doi:10.1109/msp.2010.939038

Zhang, C., Zhang, L., & Ye, J. (2012). *Generalization bounds for domain adaptation.* Paper presented at the Advances in neural information processing systems.

BIOGRAPHIES

Hyunjae Kim received his B.S. degree from Seoul National University, Seoul, Republic of Korea, in 2012. He is a Ph.D. student in Seoul National University. His research topic is battery thermal and power management. He received three awards including the IEEE PHM Data Challenge Competition Winner (2014) and the PHM Society Data Challenge Competition Winner (2014, 2015).

Byeng D. Youn received the B.S. degree from Inha University, Incheon, South Korea, in 1996, the M.S. degree from KAIST, Daejeon, Republic of Korea, in 1998, and the Ph.D. degree from the University of Iowa, Iowa City, IA, USA, in 2001. He is an Associate Professor of mechanical and aerospace engineering at Seoul National University (SNU), Seoul, Republic of Korea. Before joining SNU, he was an Assistant Professor in the Department of Mechanical Engineering, University of Maryland, College Park. His research goal is to develop rational reliability and design methods based on mathematics, physics, and statistics for use in complex engineered systems, mainly focused on energy systems. His current research includes reliability-based design, prognostics and health management (PHM), energy harvester design, and virtual product testing. Dr. Youn's dedication and efforts in research have garnered substantive peer recognition resulting in four notable awards including the ASME IDETC Best Paper Awards (2001 and 2008), the ISSMO/Springer Prize for a Young Scientist (2005), the IEEE PHM Competition Winner (2014), the PHM society Data Challenge Competition Winner (2014, 2015), etc