Deep Learning for Healthcare Decision Making with EMRs

Zhaohui Liang[†], Gang Zhang[‡], Jimmy Xiangji Huang[†*], Qinming Vivian Hu[♦]
[†] School of Information Technology, York University, Toronto, ON, M3J1P3, Canada
[‡] School of Automation, Guangdong University of Technology, Guangzhou, 510006, China
[♦]Department of Computer Science and Technology, East China Normal University, Shanghai, 200241,China
*Corresponding author: jhuang@yorku.ca

Abstract—**Computer aid technology is widely applied in decision-making and outcome assessment of healthcare delivery, in which modeling knowledge and expert experience is technically important. However, the conventional rule-based models are incapable of capturing the underlying knowledge because they are incapable of simulating the complexity of human brains and highly rely on feature representation of problem domains. Thus we attempt to apply a deep model to overcome this weakness. The deep model can simulate the thinking procedure of human and combine feature representation and learning in a unified model. A modified version of convolutional deep belief networks is used as an effective training method for large-scale data sets. Then it is tested by two instances: a dataset on hypertension retrieved from a HIS system, and a dataset on Chinese medical diagnosis and treatment prescription from a manual converted electronic medical record (EMR) database. The experimental results indicate that the proposed deep model is able to reveal previously unknown concepts and performs much better than the conventional shallow models.**

Keywords-deep learning; unsupervised feature learning; deep belief network; restricted Boltzmann machine; syndrome classification

I. INTRODUCTION

Computer aid technology is playing a significant role in decision-making, cost-effect analysis and efficacy assessment in the healthcare industry. Numerous studies are reported recently either in the theoretical or the empirical aspect, in which knowledge and experience modeling is essential to system performance. However, the complexity of human brains in thinking and cognition is considered as concepts with unknown levels thus it is difficult to be expressed formally [1-3]. Good performance is expected under specific circumstances when well designed and formulated inputs are guaranteed. Good feature representations usually require massive manual efforts and efficient key factors, thus it is difficult to be disseminated to common applications.

The available shallow models are incapable of capturing the underlying concepts and the corresponding relationships. And this weakness causes a performance gap between medical doctors and learning models. Hereby shallow models refer to the models with short paths from input to output, while deep models refer to the models with relatively long processing paths such as support vector machine (SVM) (See Figure 1) and decision tree (DT) [4].

SVM adopts a linear combination of the values in a set of kernel functions with the training data set. It is a shallow model because there are only 3 layers from input to output. DT is also a shallow model with 3 layers with its final output rendering a 1-of-K coding for each path. Each path is considered as a processing node, thus DT turns to be a 3-layer model from the input, the conjunctive normal form induced by a path, and finalizing to a 1-of-K node.

In previous studies, shallow models require well-design feature representations to avoid manual intervention of human experts. So we believe a model with automatic feature learning and classification or regression will be an effective solution. Applying deep architecture to HIS (Hospital Information System) and EMR (Electronic Medical Record) data analysis has two merits. First, deep architecture can express different concept levels that are incapable to be expressed explicitly or formally in a problem domain. Second, deep learning models with multiple-layer networks can simulate the complex procedure of human brains by storing the features as weights of connections between nodes. In our study, we built the decision making system with the multiple-layer neural network (MLNN) deep learning model as illustrated in Figure 1.

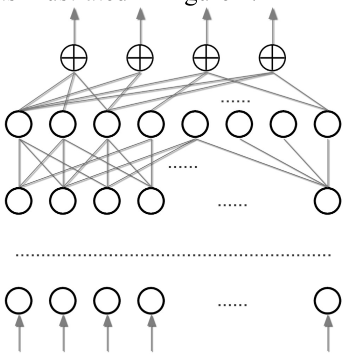

Figure 1. A sample structure of multiple-layer neural network.

Similar applications of deep learning model to medical data analysis are reported in recent years. One typical instance is a histopathological image analysis system for histopathological diagnosis on the images. It applied a function mapping set of pixels to a set of labels to indicate the degree of existence patterns of the observed diseases. There are several steps and potential concepts to when examine a histopathological image designed by experts . The model performance is highly relied on what concepts the original image are expressed. [5]. The second example is a study on knowledge patterns of Chinese medicine (CM) diagnosis. It analyzed the data retrieved from ERM on outpatient visit and treatment records by CM doctors in plain text. The Restricted Boltzmann Machine (RBM) is selected

978-1-4799-5669-2/14/$31.00 ©2014 IEEE

as the deep learning model in the study. The result showed that RBM well obeys the principles of deep learning when the standard terminology (ICD-10) is introduced as a potential concept in the model. Thus it implies that models with deep architectures are more robust than conventional shallow ones. [6]. Another study focused on CM diagnosis patterns classification. The LEVIS Hypertension TCM dataset is anayzed by a model with pre-processing and feature-level information fusion. The result indicated that feature representation can enhance the target model performance [7].

The above examples reflect that a good feature representation induced from basic features is critical for target model building. The deep models are valuable to observe complex potential concepts in medicine by working closely to human brain, and they are capable of obtain ability to discover and express concepts in the problem domain. However, the application of deep models is blocked by some challenges. For example, the back propagating (BP) algorithm was proved unreliable to converge with random network parameter initialization. And over-fitting, a common problem in machine learning reduces the model performance of unseen samples. Thus in this work, we need to apply the appropriate strategies to a deep learning model to overcome the above restrions.

II. Related Work

Deep learning has been a noticeable topic in machine learning since 2006 when Bengio reviewed the motivation of deep learning and summarized the famous algorithms for deep learning [8]. Its application covers image retrieval [9], music informatics [10], web search [11], and natural language processing [12]. Regarding the algorithms, Hinton reported that deep supervised nets can be trained by adding an unsupervised pre-training by RBM for parameter initialization [13]. Lee proposed to scale unsupervised learning of hierarchical generative models for large data sets with high dimensions [14]. The above works reflect that kernel methods can be applied to deep learning model as a regularizer at the output layer or a part of the model.

In the applications of health studies, Zhang proposed a decision tree with kernel mapping to model the efficacy assessment on acupuncture [15], and he applied a RBM model to the analysis of EMR of Chinese medicine [6]. The above work indicated that deep models perform well in EMR and HIS data analysis with abstract and complex experience.

III. Deep architecture for medical data analysis

A. *Unsupervised feature learning*

We propose a modified version of deep belief network (DBN) as the main deep model for the study. The deep model acts as an unsupervised concept extractor for the original data samples. A DBN is treated as a neural network with many hidden layers (See Figure 1). In order to overcome the difficulty to train multiple layers networks, we use Restricted Boltzmann Machine (RBM) to perform layer-wise training, which is proved effective [16]. RBM is based on the assumption of Boltzmann distribution between observed and hidden variables generalized to Gaussian distributions [17], which set up the foundation of our model.

For pair-wise training between nodes of input layers and hidden layers, a conditional Gaussian distribution is used. Eq. (1) and (2) show the conditional probabilities.

$$p(x_i|h) = N\left(b_i + \sigma_i \textstyle\sum_j h_i\omega_{ij}, \sigma_i\right) \qquad (1)$$

$$p\left(h_j = 1\middle|X\right) = g\left(b_i + \textstyle\sum_i \omega_{ij}x_i/\sigma_i\right) \qquad (2)$$

where N stands for Gaussian distribution and g is the logistic function: $g(a) = \frac{1}{1+\exp(-x)}$

x is containing all input variables, in which the ith element is denoted as x_i. w_{ij} is the weight of edge connecting an input node i and an hidden node j. b_i is a bias for the distribution of x_i. The marginal distribution of the observed feature vector x is expressed as Eq. (3):

$$p(x) = \textstyle\sum_h\left(\frac{\exp(-E(x,h))}{\int_u \sum_g \exp(-E(u,g))du}\right. \qquad (3)$$

In Eq. (3), the term E (x, h) is an energy function indicating the energy transmission lost between input layer and hidden layer. And in accordance with Hinton's method [25], we have:

$$\textstyle\sum(x,h) = \sum_i \frac{(x_i-b_i)^2}{2\sigma_i^2} - \sum_{ij} h_i\omega_{ij}x_i/\sigma_i \qquad (4)$$

Then we choose the parameters containing σ_i, b_i that can maximize the gradient according to ω, as shown in Eq. (5):

$$\Delta w_{ij} = r \times \frac{\partial \log p(x)}{\partial \omega_{ij}} \qquad (5)$$

$$= r \times \left(E_{data}\left[x_ih_j/\sigma_i\right] - E_{model}\left[x_ih_j/\sigma_i\right]\right)$$

Then we can perform layer-wise training between the input layer and the nearest hidden layer, and other adjacent hidden layers. Figure 2 illustrates the idea mentioned above.

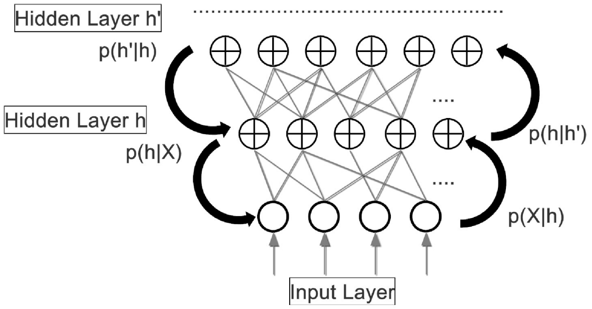

Figure 2. General framework of layer-wise pre-training.

We hereby merely perform unsupervised learning. A series of different feature representations of the original features can be obtained by increasing the number of hidden layers with Eq. (1) to (5). For each pair of adjacent layers, we maximize p $(x|h)$ and p $(h|x)$, where x stands for the input features and h stands for the corresponding hidden layer. Note that x can be original features or the output of a lower hidden layer. We can get different concept representations by encoding the input. Since the conditional probabilities are maximized, the information of the original feature representation is kept as much as possible.

B. *Supervised fine-tune for deep learning model*

After the network is trained to obtain different feature representations, the network parameters (i.e. weights) can be tuned to a supervised manner. Augmented by the previous unsupervised feature learning, the network training will converge fast even if the number of layers is large. In the output layer, it requires a classification or regression model to accept the features encoding by hidden layers and generate final decisions in either real outputs or 1-of-K coding. For a multiple classification problem, a softmax activation function is often used. We hereby use a model such as SVM to replace the softmax function because SVM achieves maximal margin classification in training data set. Suppose the output vector of the last hidden layer is v, then the softmax activation function can be expressed as:

$$p_i = \frac{\exp\left(\sum_k h_k W_{ki}\right)}{\sum_j \exp\left(\sum_k h_k W_{kj}\right)} \qquad (6)$$

When a standard SVM is replaced, the BP training procedure can still work. The objective function of SVM is:

$$\min_w (1/2\omega^T w) + C \sum_{d=1}^{|D|} \max(1 - \omega^T x_n t_n, 0) \qquad (7)$$

In a BP update step, a partial derivative with respect to ω is needed. From Eq. (7) we can write down the derivative as:

$$\frac{\partial G(\omega)}{\partial h_r} = -C t_r \omega(I(1 > \omega^T h_r t_r)) \qquad (8)$$

Eq. (7) can be directly plugged in the standard BP training procedure. Being different from a softmax function, SVM is a more powerful model when the kernel function is adjusted to support nonlinear mapping.

IV. Model Evaluation

A. *Data sets for testing*

The proposed algorithm is tested by two data sets. The first data set ($D1$) is a clinical set retrieved from EMR. It records the personal and disease history, symptoms, and treatment strategies by senior doctors. The data were manually input and indexed with standard diagnosis of ICD-10 by certificated medical doctors. The original contents are in plain text composing of natural personal language sentences and some standard TCM terms.

The second data set (D2) is retrieved from HIS with 908 records about hypertension patients. There are 167 features in each record, including 129 symptoms from inspection, auscultation and olfaction, inquiry and palpation, 22 laboratory indicators, and 16 common indexes. Note that hypertension is categorized into primary hypertension and secondary hypertension. In D2 the objects are all primary hypertension. The original data can be downloaded in [18].

B. *Results of evaluation*

Two shallow models are applied for comparison. The first one is a standard SVM without deep architecture-based feature learning (M2). The second one is a decision tree model (M3). Both of them are widely used shallow models. In SVM, we used the default parameter setting and RBF kernel with Euclidean distance function. Every data sample is labeled in both datasets. In order to test the three models effectively, the whole data set is divided into the training and the test set with a ratio of 3 to 7. The zero-one loss is used to evaluate the model performance. The whole set is randomly divided for ten times and their means and variances are recorded.

TABLE I. COMPARISON OF DIFFERENT MODELS FOR D1 (%)

	M1	M2	M3
Arthralgia Syndrome	**84.17**	74.98	63.04
Acne	**86.02**	76.68	69.80
Epilepsy	69.67	**71.57**	68.91
Tinnitus & deafness	**86.18**	73.15	72.93
Abdominal pain	**80.28**	72.86	74.19
Allergic rhinitis	69.05	65.84	**75.09**
Neck & shoulder pain	**72.85**	71.11	65.52
Cervical spondylosis	**78.48**	61.42	73.59
Cough	**87.11**	72.12	73.10
Facial paralysis	**87.26**	58.64	63.25
Traumatic brain injury	**70.31**	63.54	62.38
Migraine	**87.38**	58.92	69.97
Ankylosing Spondylitis	**87.10**	59.94	79.19
Insomnia	**77.19**	74.47	66.81
Headache	**83.81**	71.90	71.71
Flaccidity Syndrome	**69.98**	64.34	64.48
Stomachache	75.86	**77.00**	75.03
Asthma	**86.23**	58.69	65.10
Palpitation	**83.64**	66.77	70.12
Lumbocrural pain	**87.15**	65.63	73.98
Urticaria & Rubella	**80.77**	73.31	77.82

M1, M2 and M3 respectively stand for the proposed model (DBN + SVM), standard SVM and decision tree. The best result in each row is highlighted. It is obvious that M1 has the best results out of the three models. One explanation for the result is that we barely used plain text instead of some high-level concepts as model input in the evaluation. Shallow model cannot capture key features of the problem domain. The successful cases relied on some professional feature extraction operation before training. Therefore, deep models are less expensive and more powerful.

In D2, as all cases are EMR records of primary hypertension, the method by Li [7] is followed to classify the CM syndromes of primary hypertensions. By applying the same configuration, the classification performance of six data sets generated from $D2$ is evaluated. Two criteria (i.e. average precision and coverage) are used for evaluation and the result indicates the proposed model has the best performance. (See TABLE Ⅱ& TABLE Ⅲ)

TABLE II. COMPARISON OF AVERAGE PRECISION FOR D2 (%)

	M1	M2	M3
Inspection	**84.32**	81.11	79.40
Tongue	**83.29**	78.94	78.55
Inquiry	79.57	**80.90**	78.48
Palpation	**79.94**	76.34	75.67
Others	**80.45**	79.02	78.93
Fusional	**85.23**	82.31	81.08

TABLE III. COMPARISON OF COVERAGE FOR D2 (%)

	M1	M2	M3
Inspection	**41.32**	39.11	39.15
Tongue	**42.38**	40.94	38.01
Inquiry	39.31	**40.12**	38.99
Palpation	**41.94**	36.34	38.50
Others	**42.34**	39.13	39.42
Fusional	**43.16**	40.61	41.23

Finally, the sparsity of the model is assessed. As there are numerous nodes in the network, it needs to identify whether only a small number of nodes are activated by a special kind of signal. To illustrate the outcome, we plot the number of non-zero weights of each hidden layer of dataset D1 at different iteration times. (See Figure 3)

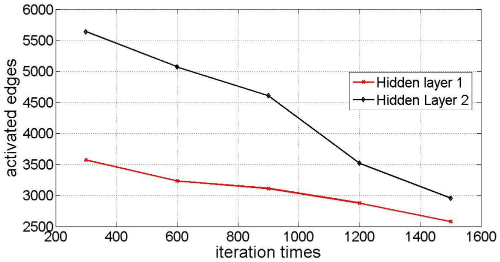

Figure 3. Number of activated edges of the network, 10^3

In Figure 3, the numbers of edges with non-zero weights of the highest two hidden layers are plotted. It shows that during the training procedure, the number of non-zero weights in the network reduces dramatically. An edge with zero weight means there is no connection between the corresponding nodes. Since a network encodes some concepts of a problem domain, the sparsity is preferred for good performance and explanation.

V. Conclusion and Discussions

A deep learning algorithm is proposed for medical knowledge modeling and assistance of decision making. Deep models are able to automatically disclose abstract feature representation from low level inputs, which makes machine learning models capable of learning abstract concepts. We proposed to apply deep belief network for unsupervised feature extraction, and then perform supervised learning through a standard SVM. The results confirm that our proposed deep model is promising in knowledge modeling for data from medical information systems such as EMR and HIS.

Regarding its application in the healthcare industry, deep learning is valuable for knowledge and experience modeling. The healthcare decision-making procedure is originated from numerous features from observation, inspection or previous cases. The concepts or decision rules cannot be directly and formally expressed. The cost for building a knowledge base or expert system is unaffordable, as it requires massive efforts of human experts. Deep learning model can do a large body of work for us to discover concept representation. Different deep learning models can be designed and implemented for different medical data analysis tasks. And the disclosed features can reveal some unknown or unexpressed knowledge during diagnosis procedure, which may improve the understanding of medical experience.

Acknowledgment

This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] J. van Dijk and J. Frens, "Being there, doing it: The challenge of embodied cognition for design," in *Proceedings of the 8th ACM Conference on Creativity and Cognition*, ser. C&C '11. New York, NY, USA: ACM, 2011, pp. 443–444.

[2] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. New York, NY, USA: Cambridge University Press, 2014.

[3] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 19, pp. 4–19, Sep. 2014.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[5] G. Zhang, J. Yin, Z. Li, X. Su, G. Li, H. Zhang, "Automated skin biopsy histopathological image annotation using multi-instance representation and learning," *BMC Med Genomics*, vol. 6, Suppl 3, S10, Nov. 2013.

[6] G. Zhang, J. Yin, Z. Li, Z. Liang, and W. Fu, "Deep learning for acupuncture point selection patterns based on veteran doctor experience of chinese medicine," in Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), ser. BIBMW '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 396–401.

[7] G. Li, S. Yan, M. You, S. Sun, and A. Ou, "Intelligent zheng classification of hypertension depending on ml-knn and information fusion," *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 837245, 5 pages, 2012.

[8] Y. Bengio, "Learning deep architectures for ai," *Found. Trends Mach.Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009.

[9] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 153–162.

[10] E. J. Humphrey, J. P. Bello, and Y. Lecun, "Feature learning and deep architectures: New directions for music informatics," *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 461–481, Dec. 2013.

[11] P.S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, ser. CIKM' 13. New York, NY, USA: ACM, 2013, pp. 2333–2338.

[12] R. Socher, Y. Bengio, and C. D. Manning, "Deep learning for nlp (without magic)," in *Tutorial Abstracts of ACL 2012*, ser. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 5–5.

[13] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[14] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 609–616.

[15] G. Zhang, Z. Liang, J. Yin, W. Fu, and G.-Z. Li, "A similarity based learning framework for interim analysis of outcome prediction of acupuncture for neck pain," *Int. J. Data Min. Bioinformatics*, vol. 8, no. 4, pp. 381–395, Sep. 2013.

[16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[17] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics." in *ICML*, L. Getoor and T. Scheffer, Eds. Omnipress, 2011, pp. 681–688.

[18] G. Li, "Levis hypertension tcm database," http://levis.tongji.edu.cn