Asian Journal of Information Technology 15 (15): 2632-2637, 2016
ISSN: 1682-3915

Ensemble Deep Learning for Multi Label Classification in the Design of Clinical Decision Support System

[1]D. Senthilkumar and [2]S. Paulraj
[1]Department of Computer Science and Engineering, University College of Engineering, Anna University, Tiruchirappalli, Tamil Nadu, India
[2]Department of Mathematics, College of Engineering, Anna University, Tamil Nadu, India

Abstract: Design of Clinical Decision Support System (ICDSS) is a challenging task in which a set of symptoms is possible to have distinct diseases and each disease has the possibility to have different categories or labels, therefore the usage of Multi-Label Classification (MLC) is required in the DD. MLC refers to the problem where each instance is associated with more than one class labels. Multi-Label Data (MLD) are high dimensional and deteriorates the performance of the classifier in terms of diagnostic accuracy. Classification of MLD is a very challenging task by existing methods and need a systematic approach. In this study, an efficient feature selection with ensemble Deep Learning (DL) algorithm for handling MLC problems is proposed. The effectiveness of the proposed Multi-Label Ensemble Deep Learning (MED) algorithm is investigated with two publicly available ML medical data using various evaluation measures. The MED results significant improvement in the performance compared with existing methods in the literature. The results reveal some interesting conclusion with respect to the use of the proposed approach to help the medical practitioners in a better decision making in the diagnosis and treatment with the least number of symptoms in the MLD.

Key words: Multi-Label (ML), Differential Diagnosis (DD), Feature Selection (FS), Deep Learning (DL), Ensemble Method (EM)

INTRODUCTION

Diagnosis is a challenging task because many signs and symptoms are nonspecific. The differential diagnosis plays a vital role in the field of medical diagnosis that distinguish a particular disease from other diseases which all have the similar symptoms by comparing and contrasting all possible explanations of patient data (Shortliffe and Barnett, 2014). Clinical decisions are often made based on the doctor's 'intuition and experience rather than the knowledge-rich data hidden in the database (Palaniappan and Awang, 2008). Machine learning algorithms are used as a tool to extract hidden interesting pattern from the medical database.

Interesting patterns will be used to assist the physicians to improve the diagnosis speed, accuracy and/or reliability (Kononenko, 2001). Decision Support Systems (DSS) are defined as a computer based system developed to assist decision makers in the effective decision making (Rupnik and Kukar, 2007). It reduces the diagnosis time and increase the diagnosis accuracy in complicated diagnosis, decision process as well as the cost of care. Clinical Decision Support Systems (CDSS) are technology based computer systems designed to improve clinical decision-making about individual patients (Berner and Lande, 2007).

In the recent past, the significant growth of research and development in medical industries generates massive data includes clinical examination, vital parameters, investigation reports and drug decision, etc. for diagnosis (Shortliffe and Barnett, 2014). These data are high, ML and multi-dimensional in nature. The representation of MLD is a set of symptoms associated with a set of diseases with binary labels. This MLD degrades the performance of the classifiers and reduces the diagnostic accuracy and processing this data is too complex by traditional methods and need a systematic approach (Senthilkumar and Paulraj, 2013). Therefore, mining the MLD is a challenging task among the recent medical data mining researchers in order to speed up the diagnostic process, reduce overuse of medical tests, save costs and to improve the accuracy of diagnosis (Senthilkumar and Paulraj, 2013).

In clinical practice, diagnosis of disease has complex and nonlinear relationship between symptom and diseases. For diagnosis the disease most of the algorithm

Corresponding Author: D. Senthilkumar, Department of Computer Science and Engineering, University College of Engineering, Anna University, Tiruchirappalli, Tamil Nadu, India

does not consider the hierarchical nature and its cause misunderstanding and bias. Also, it is limited to human for proper diagnosis with more accuracy. Deep Learning (DL) is an artificial neural network and emerging technique in the area of machine learning research. It mimics the nature of the research of the brain for learning to analyze patient data, make diagnoses and suggest treatments. It supports complex nonlinear relationship between the symptoms and diseases. It is so popular among machine learning researchers that it has produced high classification performances in many areas such as image, video, speech and text (Zhao and He, 2014).

In order to overcome this MLC problem, a structured framework with the ensemble DL algorithm for mining ML medical data with the aim of building intelligent decision support system for ML problems are proposed.

Deep neural networks and related works: This study briefly reviews the basics of Deep Learning Network (DLN) and its importance in machine learning as well as related work in the literature on ML. DL can automatically discover clinically-relevant features by first architecting a hierarchy of patterns and then rapidly updating those patterns upon observing examples. It learns very well, very fast and use optimal set of features (Zhao and He, 2014). The beauty of DLN is to compute hierarchical features or representations of the observational data where high dimensional features are expressed as low-dimensional features. DLN learning automatically learns multiple levels of representations of the underlying distribution of the data to be modeled. It automatically extracts the low and high level features necessary for classification. High level features means features that hierarchically depend. It finds the non-linear connections between a given input and output.

DLN for syndrome diagnosis is proposed with a single hidden layer. In this 13 node number value for the hidden layer is used to choose the best node value. Predefined set of hyper parameter is not good procedure to identify the best parameter because it leads to bias in the classification process (Guo and Letourneau, 2013). Restricted Boltzmann machine model is used to solve the multi-label learning with incomplete labels (Xin *et al*., 2015). Classification of MLD using DL is proposed with the set of hyper parameter. Several parameters are fine-tuned to select the best set of parameter from the fixed set of parameter using three fold cross validation (Liu *et al*., 2014). Multi-label classification is proposed in the optical remote sensing application using DL with sparse auto encoders and a single hidden layer. Critical Hyper Parameters (CHP) of the neural network used in (Guo and Letourneau, 2013; Xin *et al*., 2015; Liu *et al*., 2014) are predefined and uses a single hidden layer with less number of cross validation. The proposed structure of DLN for MLD is illustrated in Fig. 1. It is a multilayer network which includes input, output and with many hidden layers.

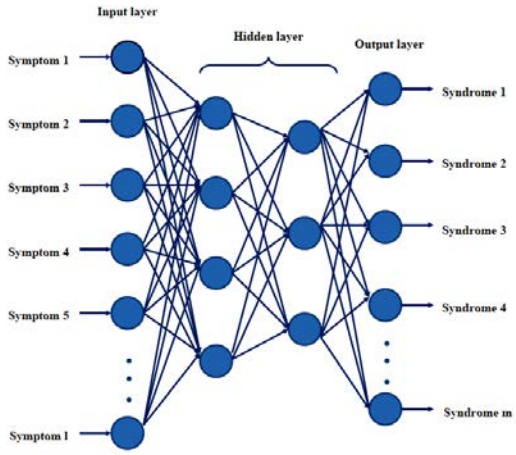

Fig. 1: Proposed structure of DLN for MLD

MATERIALS AND METHODS

Proposed Multi-label Ensemble Deep learning (MED): In this Study DL algorithm is proposed to build the classifier in the MLD. A Proposed MED algorithm for MLD is shown as Algorithm 1. Reduced feature subset (using MFSS) was used directly in the proposed MED (Senthilkumar and Paulraj, 2013). DLN does not deal with the MLL directly; therefore the MLD is transformed into Single Label Dataset (SLD) using any one of the problem transformation methods namely Binary Relevance (BR) and Label Powerset (LP). In this study, MLD is transformed into SLD using BR method. The process of each SLD was shown in Fig. 2.

Algorithm 1: Multi-label Ensemble Deep learning (MED):
Input: Multi-label Reduced Dataset MLD- D'
Output: Multi-Label Ensemble Deep Learning Model with higher accuracy
Transform the Multi-Label Dataset D' into multiple Single Label Dataset (SLD)
i.e., SLDi = {SLD1, SLD2, SLD3 … SLD_m} =; I = 1 ... m
For each single label dataset SLD_i; I = 1 ...m
{
Divide the data set into two parts, namely test data and training data
Train the DLN to identify the optimal hyper parameter using training data
Build the EDLM draft model with training data using the optimal hyper parameters
Validate the EDLM draft model using the test data
}
Once the validation is completed build the final EDLM for each SLD
Build the MED classifier for the prediction by combining the multiple EDLM
End

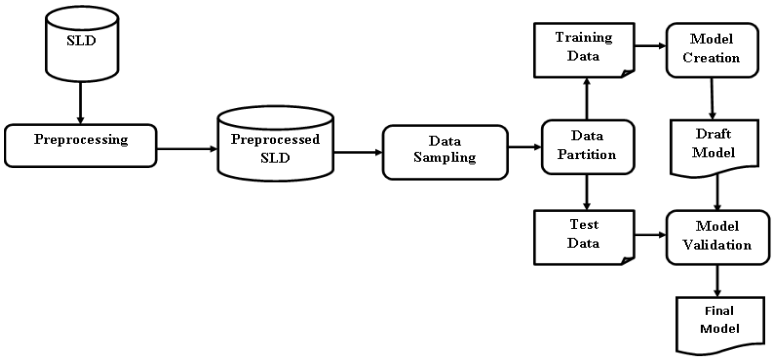

Fig. 2: Process of MED algorithm for each SLD

The proposed framework is implemented in R-language with H_2O. H_2O is fast, scalable online machine learning plat which models, data very fast and easy to make better decisions faster by running advanced data mining algorithms for big-data analysis. Construction of good ensembles of classifiers is the most important research area for the researcher (Rokach, 2010). Sensitivity of the DL model based on the various CHP. In order to improve the performance of the model, there are various CHP needs to be define to train the DLN. Therefore, the ensemble DL model is proposed with optimal hyper parameters to improve the performance of the classifier. For each SLD, select the best combination of CHP by training the DLN with various combinations of parameters using the two different approaches CV and GM then use the hyper parameters to build the draft Ensemble DL Model (EDLM). Finally validate the draft EDLM with the test set. Once the validation is completed combine the result of multiple EDLM to build the MED for the MLC.

DLN is prone to over fitting, therefore optimal parameter setting is very important in order to improve the performance of the model. Cross Validation (CV) is one of the methods to identify the best HP and then the ensemble model is constructed using the top few models. Check Point (CP) is another method to construct the ensemble model. Therefore, in this study two ensemble models are proposed; first one uses the CV in the pre-training phase of DLN to identify the best HP, then the ensemble model is constructed using the top few models; the second one uses the grid search in the pre-training phase to identify the best HP, then CP of a single model is used to construct the ensemble model.

Results and analysis: This study illustrates the evaluation of proposed MED with two ML Traditional Chinese Medicine (TCM) dataset in terms of the various performance metrics.

Datasets used in this study: In this study, two ML Traditional Chinese Medicine (TCM) dataset namely; chronic fatigue and coronary heart disease are used to evaluate the effectiveness of proposed MFSS and MED (Liu *et al*., 2010; Wang *et al*., 2014).

Performance metrics: One of the important tasks in the data mining is to estimate the performance of the proposed model. Various performance metrics are used to evaluate the quality of the proposed model are accuracy, AUC and F1 (Senthilkumar and Paulraj, 2013; Devaraj and Paulraj, 2010).

RESULTS AND DISCUSSION

This study explores the inferences of the proposed MED model. The physician has to consider many factors to diagnose a disease. Most of the researchers' aim is to identify the significant symptoms which are used for diagnosis and prediction. The most significant symptoms are always enhancing the predictive accuracy of the model (Senthilkumar and Paulraj, 2013). The details of selected symptoms using MFSS to build the MED are shown in Table 1 (Rokach, 2010).

The enhanced performance rate of proposed (MED) is evaluated with different performance metrics. Statistical methods are used to assess the effects of different classification methods on classification accuracy. The performance of various classifiers (i.e., Accuracy, AUC, F-measure ROC) in high dimensional medical data (i.e., with great quantities of features, large number of instances and 'n' number of class labels) is analyzed with Friedman's and Iman and Daveport test with the null hypothesis that the calculated average ranks are significantly different from the mean rank Rj = 2.5 (Garcia and Herrera, 2008). The different models are ranked based on the performance metrics of the model and then average rank of each model is computed. The ranking of

different models for the dataset CHD and CFD are depicted in Table 2. The test results of Friedman's and Iman and Daveport are depicted in Table 3 and 4 for the dataset CHD and CFD, respectively.

From Table 3 for the dataset CHD, Friedman test statistic value of accuracy, AUC and F1 are (χ_F^2 = 3.75, 7.45, 1.95). For F distribution with 3 degrees of freedom, the critical value is 7.81 for α = 0.05 and 11.34 for α = 0.01 and the p-value is 0.29, 0.06 and 0.58. Iman and Davenport test statistic value for accuracy, AUC and F1 are (F_F = 1.32, 3.53, 0.61) with (3,15) degrees of freedom, the critical value is 3.29 for α = 0.05 and 5.42 for α = 0.01 for and the p-value is 0.31, 0.04 and 0.62. In both the tests, the critical values are greater than the respective test statistics, therefore accept the null hypotheses for both the significance level α = 0.05 and α = 0.01.

From Table 3 for the dataset CFD, Friedman test statistic value of accuracy, AUC and F1 are (χ_F^2 = 6.6, 6.38, 6.6). For F distribution with 3 degrees of freedom, the critical value is 7.81 for α = 0.05 and 11.34 for α = 0.01 and the p-value is 0.9, 0.9 and 0.9. Iman and Davenport test statistic value for accuracy, AUC and F1 are (F_F = 3.67, 3.4, 3.67) with (3,9) degrees of freedom, the critical value is 3.86 for α = 0.05 and 6.99 for α = 0.01 for and the p-value is 0.06, 0.07 and 0.06. In both the tests, the critical values are greater than the respective test statistics, therefore, accept the null hypotheses for both the significance level α = 0.05 and α = 0.01.

Based on the analysis of statistical tests it inferred that there is a significant difference between CVM, CVEM, GM and CPEM. Finally, the experimental results conclude that the best model for the dataset CFD is CPEM and the next model is CVEM. But for the dataset CHD the best model is CVEM and the next model is CPEM.

From Table 4, Liu *et al*., 2010 used ML-kNN with 52 symptoms achieved 66.4% and proposed model MED achieves 79% with the 7 symptoms for the dataset CHD. Wang *et al*. (2014) used CP-RF with 95 symptoms achieved 98.30% and proposed model MED achieves 1005% with 7 symptoms for the dataset CFD. Therefore, proposed model MED performs well with the less number of symptoms compared with existing methods in the literatures.

From Table 6 it is inferred that proposed model CVEM achieves highest rank with the less number of symptoms compared with state-of-the are methods in the literature. From the study it is inferred that CVEM runs for several hours to identify the best HP and to construct the EM. But the grid search with checkpoint CPEM runs with reasonable time compared with the CVEM. Also CV has a drawback which includes the suboptimal models to construct the ensemble model.

In this study also it is concluded that performance of the CVEM is less compared with the CPEM for the data set CFD. Therefore, identifying the optimal parameters using CVEM is not feasible because it leads to wrong decisions. CP has advantages such as supports parallel processing; model tuning with reasonable accuracy and less run time compared with the CVEM.

Table 1: Details of selected symptoms using proposed MFSS for the two datasets

Dataset	Total no of symptoms in the dataset	Selected symptoms using MFSS	No. of symptoms selected MFSS
Chronic Fatigue Disease) (CFD	95	f82, f52, f20, f44, f59, f41, f40	7
Coronary Heart Disease (CHD)	125	f50, f51, f24, f100, f139, f55, f58	7

Table 2: Ranking of different models for CHD and CFD based on Accuracy, AUC and F1

Dataset	CHD				CFD			
Model ranking	CVM	CVEM	GM	CPEM	CVM	EVEM	GM	CPEM
Accuracy	2.08	2.08	3.33	2.50	3.25	3.25	2.250	1.250
AUC	1.92	1.67	3.33	3.08	3.25	3.25	2.250	1.250
F1	2.50	1.92	2.92	2.67	3.50	3.00	2.125	1.375

Table 3: Results of Friedman and (Iman and Daveport) test statistics for accuracy, AUC and F1 for the CHD and CFD

Test (dataset)	Friedman test (CHD, CFD)				Iman and Daveport test (CHD, CFD)			
Measures	Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
Accuracy	3.75	0.29	6.60	0.09	1.32	0.31	3.67	0.06
AUC	7.45	0.06	6.38	0.09	3.53	0.04	3.40	0.07
F1	1.95	0.58	6.60	0.09	0.61	0.62	3.67	0.06

Table 4: Comparison of accuracy metrics with the literature for the two dataset

		Literature				Proposed deep learning with MFSS	
Dataset	Total No. of symptoms in the dataset	Researcher and year	Algorithm	No. of used symptoms	Accuracy (%)	No. of used symptoms	Accuracy (%)
CHD	125	Liu *et al*. (2010)	ML-kNN	52	66.4	7	79
CFD	95	Wang *et al*. (2014)	CP-RF	95	98.30	7	100

Table 5: Comparison of accuracy metrics for each class with the literature for the CHD dataset

	Accuracy of CHD dataset			
	Wang YQ *et al.* (2010, 2011) No. of symptoms used-125		Proposed MED; No. of symptoms used-7	
Pattern of syndrome	OCON-NN	RBF-KFSVM	CVEM	CPMEM
z1 Deficiency of heart qi syndrome	60.67	73.20	1.00	0.68
z2 Deficiency of heart yang syndrome	78.08	81.70	0.80	0.73
z3 Deficiency of heart yin syndrome	65.16	68.63	0.72	0.66
z4 Qi stagnation syndrome	87.07	85.62	0.83	0.83
z5 Turbid phlegm syndrome	60.11	50.33	0.63	0.64
z6 blood stasis syndrome	62.35	76.47	0.74	0.78
Overall accuracy	68.91	72.66	0.79	0.72

Table 6: Ranking for the MED with the literature for the CHD

Model	OCON-NN	RBF-KF-SVM	CVEM	CPEM
Ranking	3.17	2.17	2.08	2.58

CONCLUSION

Multi-Label Classification (MLC) refers to the problem where each instance is associated with more than one class labels. Due to its complex nature classifier built from the MLD are typically more expensive or time-consuming and deteriorates the performance of the classifier in terms of diagnostic accuracy. In this study, MLC with ensemble DL algorithm is proposed. A proposed framework addresses the four important research challenges in the MLC:

- Extraction of relevant subset of features of the large number of features
- Efficient algorithm to deal with complicated and large data to reduce the computational complexity
- To design, efficient online MLC framework that scale to large and sparse domains

The proposed MED is applied to two publicly available ML medical data sets. A systematic study is proposed to evaluate the performance of the proposed model MED with different performance metrics using statistical tests. Proposed framework uses MFSS for feature subset selection. Experiment results show that the time complexity is reduced for further analysis or to build a classifier by using MFSS. Also MFSS performs efficiently with the least number of features (6% of CHD and 7.4% of CFD) without affecting the classification accuracy. Therefore, MFSS is an effective feature selection algorithm in those applications generating MLD.

CPEM allows parallel processing and time complexity is very low in model tuning compared with the CVEM. CVEM has a drawback which includes the suboptimal models to construct the ensemble model. Therefore CVEM deteriorates the performance of the classifier by selecting the top few models. The proposed framework MED yields significant performance improvement when compared with existing methods in the literature. The results reveal some interesting conclusion that the proposed framework supports online MLC for medical practitioners in a better decision making in the diagnosis and treatment effectively.

REFERENCES

Berner, E.S. and T.J. Lande, 2007. Overview of clinical decision support systems. Clinical Decision Support Systems, pp: 3-22.

Devaraj, S. and S. Paulraj, 2015. An efficient feature subset selection algorithm for classification of multidimensional dataset. Sci. World J., 2015: 1-9.

Garcia, S. and F. Herrera, 2008. An extension on statisticalcomparisons of classifiers over multiple data sets for all pairwise comparisons. J. Mach. Learn. Res., 9: 2677-2694.

Guo, H. and S. Letourneau, 2013. Iterative classification for multiple target attributes. J. Intell. Inf. Sys., 40: 283-305.

Kononenko, I., 2001. Machine learning for medical diagnosis: History, state of the art and perspective. Artif. Intell. Med., 23: 89-109.

Liu, G.P., G.Z. Li, Y.L. Wang and Y.Q. Wang, 2010. Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning. BMC. Complementary Altern. Med., 10: 1-12.

Liu, G.P., J.J. Yan, Y.Q. Wang, W. Zheng and T. Zhong *et al.*, 2014. Deep learning based syndrome diagnosis of cronic gastritis. Comput. Math. Methods Med., 2014: 1-8.

Palaniappan, S. and R. Awang, 2008. Intelligent hear disease prediction system using data mining techniques Proceedings of the International Conference on Compute Systems and Applications, March 31-April 4, 2008, Doha, pp: 108-115.

Rokach, L.,2010. Ensemble-based classifiers. Artif. Intell. Rev., 33: 139.

Rupnik, R. ad M. Kukar, 2007. Decision support system to support dcision processes with data mining. J. Inf. Organizational Sci., 31: 217-232.

Senthilkumar, D. and S. Paulraj, 2013. Diabetes disease diagnosis using multivariate adaptive regression splines. International J. Eng. Technol., 5: 3922-3929.

Shortliffe, E.H. and G.O. Barnett, 2014. Biomedical Data: Their Acquisition, Storage and Use. In: Biomedical Informatics. Shortliffe, E.H. and J.J. Cimino (Eds.). Springer, London, England, ISBN: 978-1-4471-4474-8, pp: 39-66.

Wang, H., X. Liu, B. Lv, F. Yang and Y. Hong, 2014. Reliable multi-label learning via conformal predictor and random forest for syndrome differentiation of chronic fatigue in traditional Chinese medicine. PloS One, 9: 1-14.

Wang, Y.Q., H.X. Yan, R. Guo, F.F. Li and C.M. Xia *et al.*, 2011. Study on intelligent syndrome differentiation in Traditional Chinese Medicine based on multiple information fusion methods. Int. J. Data Min. Bioinf., 5: 369-382.

Xin, L. and Z. Feipeng and G. Yuhong, 2015. Conditional restricted boltzmann machines for multi-label learning with incomplete labels. Proc. AISTATS., 38: 635-643.

Zhao, Y. and L. He, 2014. Deep Learning in the EEG Diagnosis of Alzheimer's Disease. In: Asian Conference on Computer Vision. Jawahar, C.V. and S. Shan (Eds.). Springer International Publishing, Berlin, Germany, pp: 340-353.