Project Report

Machine Learning Applying on Gene Chip Data
VE490, Introduction to Artificial Intelligence

Changxu Luo 515370910009
camelboat@sjtu.edu.cn

***Abstract*—Gene chip analysis uses microarray technology to represent the original chromosome with its annotation, analyzing that representation with mathematical methods such as linear models [1]. With the exponentially growing of gene chip data volume, gene chip analysis becomes more challenging and requires appropriate mathematics tools. In this project, machine learning algorithms, including both classical approaches and deep learning methods, have been applied to solve this problem [2] [3]. On database E-TABM-185, the resulting test accuracy of normal-abnormal prediction has achieved 93.7% with classical approaches and 95.8% with a fully-connected neural network of four layers by dimension reduction to 95% variance.**

I. Methods

Analyzing procedure on gene chip database can be divided into four parts, which are the preparation of data(problem definition included), model construction, training and evaluation [4].

A. Data Preparation

1) Dimension Reduction: According to the original data, there are 22283 genes and 5896 observations, which means it is a large p small n problems. Through the observation of original data, we can find that a large percentage of gene values have no significant change among different observations, contributing little to the observed sample's health condition. Hence, to prevent the potential overfitting possibility and filter those non-significant genes, data needs to be dimensionally reduced. In this project, principle component analysis(PCA) is used to prepare the data and reduce dimension [6].

2) Feature Extraction: Literature description in label file should be transformed into digits value in order to train the models. What we are interested in is the relation between people's gene chip data and their health condition, hence current disease state should be used as the principle label feature. Among all these disease states, we first count number of different types of states, then treat them as normal-abnormal binary classification problem, using 0 for normal and 1 for abnormal as the mathematical representation.

B. Model construction

1) Traditional Approaches: Four traditional machine learning approaches are used in this project, which are supported vector machine(SVM), logistic regression, k-nearest neighbors algorithm(KNN) and linear discriminant analysis.

2) Deep Learning: We represent two fully-connected neural network(NN) of 2-layer and 4-layer in this project as the deep learning models. With input-layer's nodes number same as the dimension of reduced data and output layer's nodes number as 2, we added hidden layer(s) in order to increase the fitting capability. The performance comparison between these two models and comparison between deep learning methods and traditional approaches would be implemented in this project.

C. Training and Evaluation

For traditional approaches, we use cross-validation with folds number of 5 to validate and test the trained model. L1 normalization is applied for each traditional approaches in order to control the model complexity [5]. As for the deep learning part, database is divided into three parts, which are the training set, validation set and test set. Training set contains 70% of data(4126 observations), while validation set and test set either contains 15% of data(885 observations each).

II. Results

A. Data Preparation Results

For original training data, we first use PCA to reduce dimension. The resulting relationship between variance percentage and dimension numbers is shown in table I.

TABLE I
PCA RESULTS

variance percentage(%)	dimension number
95	1271
90	453
85	187

To decrease the biases in training and maintain generality, we choose data with 1271 and 453 dimensions as the training data for traditional approaches, and 1271 dimensions for deep learning part.

For original label data, there are 192 different disease states in total, while several states are not simple single disease states that can be treated as a certain model output. For example, {*acute lymphoblastic leukeia, chemotherapy response*} is the combination of two different disease states that have occurred in other observations, and for one disease state such as *pituitary adenoma*, it has different stage itself like *ACTH-secreting* or *GH-secreting*. Consequently, it is hard to transform these literature labels into digit values without plenty of medical knowledge, hence for this project, labels are parsed into normal(0, correspond to 'normal' and 'healthy') and abnormal(1, correspond to other disease state). The result of label data preparation is shown in table II

TABLE II
LABEL PARSING RESULT

	normal	abnormal
observation count	1953	3943

B. Classical Approaches

SVM is first applied to the prepared data. Figure II-B shows graphical relation between the first three important data columns.

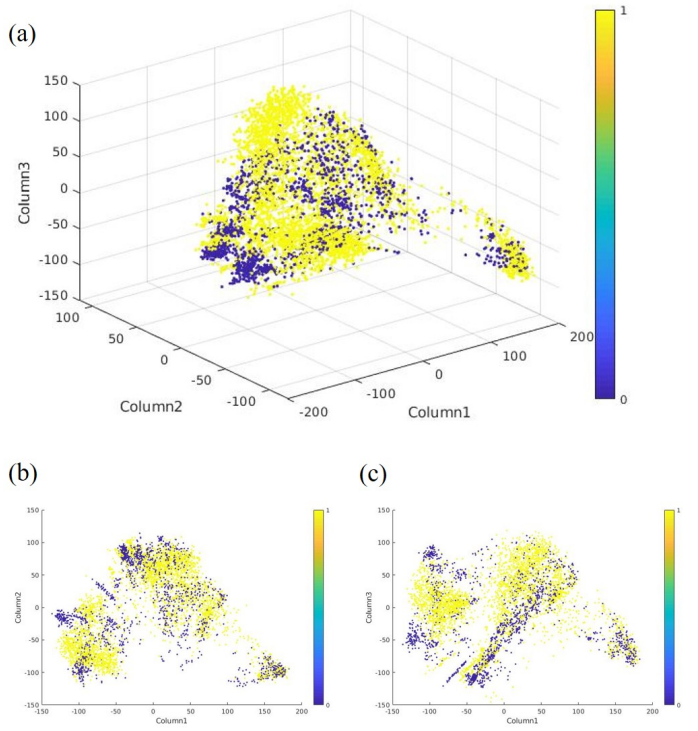

Fig. 1. (a) Scatter figure of first three important data column; (b) column1 $vs.$ column2; (c) column1 $vs.$ column3; in all these three figures, blue(0) represents normal and yellow(1) means abnormal.

According to figure II-B, it can be found that these three columns are not obviously linearly separable, hence to find out the model with highest performance, kernel functions are used in training. The kernel functions include linear kernel, quadratic kernel, cubic kernel and Gaussian kernel [8]. Table III shows SVM models performance with different data dimensions and different kernel functions.

TABLE III
ACCURACY OF DIFFERENT SVM MODELS

	linear(%)	quadratic(%)	cubic(%)	Gaussian(%)
95%	93.7	93.7	92.8	91.1
90%	93.7	94.7	94.8	94.5

Table IV shows prediction speed and training time caused by each SVM models.

TABLE IV
PREDICTION SPEED AND TRAINING TIME OF DIFFERENT SVM MODELS

		linear	quadratic	cubic	Gaussian
prediction speed(obs/s)	95%	940	870	750	620
	85%	5100	4800	3800	2300
training time(s)	95%	75.08	68.42	75.87	88.94
	85%	22.53	17.58	19.95	28.40

From these two tables, we can find that both linear kernel and quadratic kernel SVM have accessed highest accuracy of 93.7% for dimension of 1271 while cubic SVM achieves highest accuracy of 94.8% for dimension of 453. With the dimension reduction from 1271 to 453, prediction speed has been greatly enhanced and the training time has also been drastically decreased, while the best accuracy doesn't change obviously, which shows great importance of dimension reduction in application of machine learning where efficiency and computational speed are highly required. Here we choose linear and quadratic kernel SVM for dimension 1271 and cubic kernel SVM for dimension 453 as the examples. Their confusion matrices and ROC curves are shown in figure 2 and figure 3.

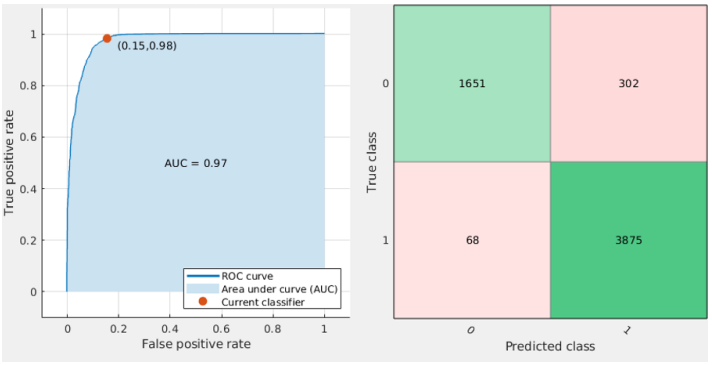

Fig. 2. ROC curve and confusion matrix of linear SVM for dimension of 1271; for ROC curve, positive class means abnormal, AUC=0.97 and this SVM classifier is at (0.15, 0.98); for confusion matrix, TP=1651, FP=302, FN=68, TN=3875.

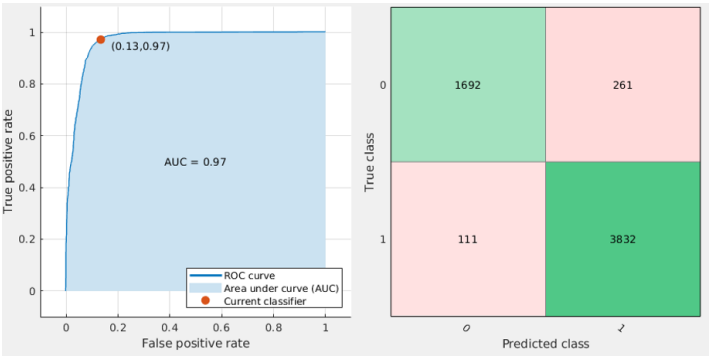

Fig. 3. ROC curve and confusion matrix of quadratic SVM for dimension of 1271; for ROC curve, positive class means abnormal, AUC=0.97 and this SVM classifier is at (0.13, 0.97); for confusion matrix, TP=1692, FP=261, FN=111, TN=3832.

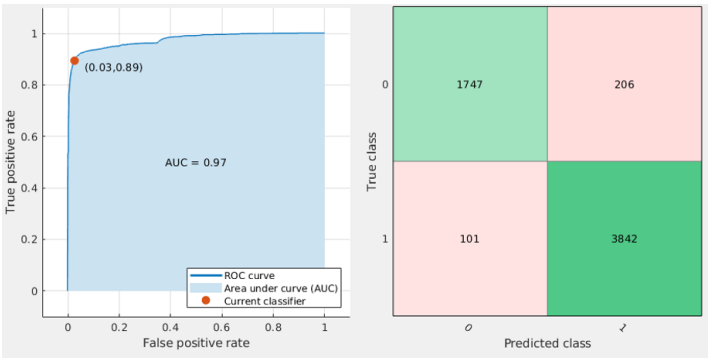

Fig. 4. ROC curve and confusion matrix of cubic SVM for dimension of 453; for ROC curve, positive class means abnormal, AUC=0.97 and this SVM classifier is at (0.03, 0.89); for confusion matrix, TP=1747, FP=206, FN=101, TN=3842.

From these ROC curves and confusion matrices, the robust correctness for our SVM models' prediction abilities can be strongly guaranteed [9].

Use 90% variance percentage data to do the SVM optimization with Bayesian Optimization, the resulting objective function model and relation between min observed objective, estimated min objective and function evaluations are shown in figure 5.

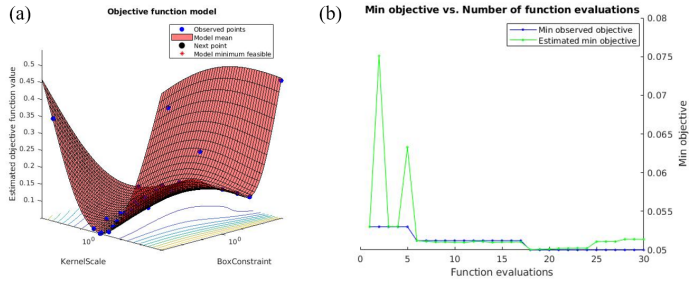

Fig. 5. (a) objective function model for linear SVM model on 90% variance data;(b) Min objective vs. Number of function evaluations, green line is the estimated min objective while blue line is the min observed objective.

Results for other traditional approaches are shown in table V and table VI.

TABLE V
ACCURACY OF OTHER TRADITIONAL METHODS

	logistic regression	KNN(cosine)	Linear Discriminant
95%	88.3	92.6	93.7
90%	90.9	92.3	93.4

TABLE VI
PREDICTION TIME AND TRAINING TIME OF OTHER TRADITIONAL METHODS

		logistic regression	KNN(cosine)	Linear Discriminant
prediction speed(obs/s)	95%	5300	240	5700
	85%	20000	700	24000
training time(s)	95%	372.86	79.91	14.15
	85%	17.76	27.709	2.33

The result for other traditional approaches shows that besides SVM, linear discriminant analysis has the best performance which gets 93.7% accuracy with data of 95% variance percentage. Logistic regression has the best prediction speed among all of the traditional methods, which explains why it is still popular in classification tasks that highly require algorithm efficiency. Although KNN algorithm has comparably short training time, it doesn't perform in this task especially for fine KNN and Medium KNN, which only achieve 87.0% and 61.7% accuracy separately. This is mainly because our data has strongly overlapping boundary, so KNN algorithm without suitable kernel is not an ideal model.

C. Deep Learning Methods

In this part, two fully-connected neural networks with different structures, including two layers(one hidden layer and one output layer) and four layers(three hidden layers and one output layer), have been implemented on dimension-reduced dataset with 95% variance. The diagrams of their structures are shown in figure 6.

The algorithm applies gradient backpropagation method and keeps training for more iterations until the model has reached its best performances on validation set. Two-layer model uses 26 seconds to train for 32 iterations while four-layer model use 43 seconds for training 41 iterations. Both of the models have reached excellent results for total accuracy over 95%, which

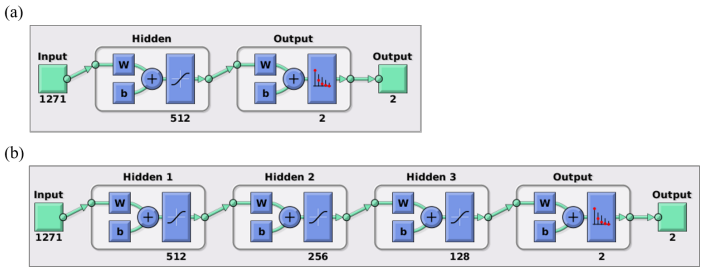

Fig. 6. (a) Structure of a two-layer fully-connected NN, with one hidden layer of 512 nodes and one output layer;(b)Structure of a four-layer fully-connected NN, with three hidden layer of 512, 256, and 128 nodes separately and one output layer

are both better than the best result from the traditional machine learning methods.

The ROC curves and confusion matrices for both models are shown in figure 7 and 8.

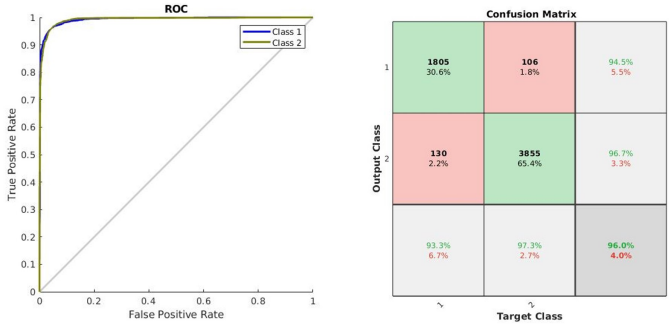

Fig. 7. ROC curve and confusion matrix for two-layer NN, with total accuracy of 96.0%. For confusion matrix, class 1 denotes for normal, while class 2 denotes for abnormal. TP=1805, FP=106, FN=130, TN=3855

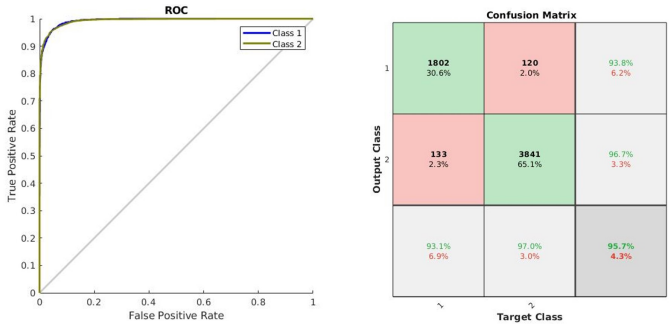

Fig. 8. ROC curve and confusion matrix for two-layer NN, with total accuracy of 95.7%. For confusion matrix, class 1 denotes for nomral, while class 2 denotes for abnormal. TP=1802, FP=120, FN=133, TN=3841

Two-layer model has reached the total accuracy of 96.0%, while this number for 4-layer model is 95.7%. Since we don't see great improvement on the accuracy with the improvement of the complexity of models, the first two-layer model may have already been enough for this task, and the 4-layer model may meet problem of overfitting. To test if the model has met the overfitting problem, validation performance curves are plotted for both models, which are shown in figure 9.

For the validation performance vs. Epochs graphs, if there exist the phenomenon that the test curve significantly increases before

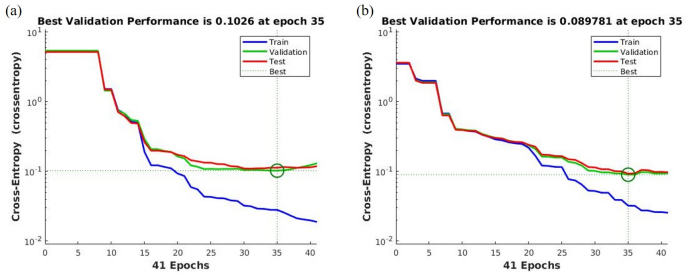

Fig. 9. (a) Validation performance vs. Epochs for 2-layer model, with best validation performance to be 0.1026 at epoch35;(b) Validation performance vs. Epochs for 4-layer model, with best validation performance to be 0.089781 at epoch 35

the validation curve increases, then overfitting might occurred. In our two models, however, both two test curves perform similarly with the corresponding validation curves, hence there is no obvious overfitting problem. Also, our 4-layer model has a little better performance on the validation performance than the 2-layer model, which shows its higher complexity capacity, which however, doesn't guarantee its higher capability of generalization on test dataset.

III. Discussions

A. Comparison between Classical Approaches and Deep Learning Methods

1) Performance comparison: In our results, both classical approaches and deep learning methods have shown great prediction performance with well-developed ROC curves of AUC>0.95. Deep learning method with multi-layer neural network has shown a little better prediction performance(96%) than traditional methods whose best performance is 93.7% for quadratic kernel SVM. This can be explained by the higher complexity capacity and capability of generalization of deep learning methods. Also, since our data have quite a large number of overlapping boundaries between normal-abnormal classes, it can be hard for traditional methods to learn without a suitable kernel function, while NN can use its multiply hidden layer to recombine the features and generate new features that are more important for this classification task.

About the training efficiency, since this is a binary classification task, all of the methods we applies in this project have given us reasonable training times. For traditional methods, linear discriminant analysis algorithms have the best training efficiency while remaining a reliable prediction accuracy. Since the scale for neural network here is not that large compared to other fields like computer vision or data mining, our 2-layer and 4-layer NN can reach a stable validation accuracy in comparably quite little number of iterations(35 in both of our models). If the structure of an NN has been determined, then the training time will just increase linearly with the increasing of the dataset scale, however, for traditional methods such as SVM with non-linear kernel function, their worst time complexity is approximately $O(N^2)$ or even more with N to be the dataset scale, which increases much more rapid than the deep learning method. But for dataset with comparably little scale, for example, with dataset that has observations less than 1000, any multi-level NN with hidden nodes number greater than 10 will have great possibility to face a strong overfitting problem, while traditional methods will continue to show good results, hence it can be concluded that traditional approaches are more suitable for tasks on small datasets than the deep learning methods, while the structure of deep learning models should be determined by both the scale and structure of the datasets.

2) Model meaning problems: For traditional methods' results, we usually have a lot of ways to explain what they mean in the real world. For example, the result of a SVM model with kernel functions can be explained as the non-linear relation between the data and labels. However, for the result of deep learning methods, especially for results generated by the fully-connected NN we applied here, we can't practically explain their meanings. Due to this black-box usage, although we here have received deep learning models' high prediction accuracy, we still can't judge whether this result is brought by the high capacity of the NN or we have exactly found the correct model. Unlike partial-connected convolution-based NN applied on computer vision, where new features that come from the decomposition and recombination of original features by the hidden layers can be explained as the image features of high-level abstraction, in this project, our models' hidden layers lack reliable explanation of all its trained weights, hence what our models have actually learned is still required to discuss.

3) Applicable tasks: Due to our analysis of the comparison between traditional approaches and deep learning methods on their training time efficiency and model result explanation, their applicable tasks can be induced here. Traditional methods, especially for SVM with non-linear kernel functions are suitable for dataset with comparably small scale, while traditional methods without non-linear kernel may have a reasonable training time on large-scale dataset, but they are not expected to have a good training result. Deep learning methods, especially neural networks with multiple layers, are too complicated and unnecessary for simple data classification and clustering with small dataset. But for comparable large dataset, only deep learning methods with correctly adjusted structures can guarantee the best results we can achieve now, and with their ability to understand materials in highly abstract level, we can have an important step towards to real artificial intelligence.

B. Possible Improvements

1) Improvement on feature selection: In this project, we use the original numerical value on data of different gene position as the original features, and directly apply data dimension reduction operation to them. This is an incomplete process for a good feature selection, because many other important features may be neglected [10]. For example, some features that are intuitively felt to be related with different disease states, such as a patient's gender, age, current and former living habit, as well as his or her family disease history, are not included in our model training, which may have influence on the generalization capability of our trained models to a large extent. What's more, we don't apply any professional medical concern or knowledge during the whole project, which may tell us that some genes have totally no relation to the body development of a person, or some genes actually have their combined function when treating them as a group. From this point, our models may lack enough real meanings out of this dataset, and have their limitation on medical research.

To overcome these disadvantages, some approaches can be applied in the future research. New features mentioned above should be added into the model, and parsing on these data should

be combined together with medical research and experiments. For example, if our model analyzes that some genes have certain effects on one disease state, it should be proved by the later clinical trials, but not simply applied on disease prediction only based on gene chip data.

2) Improvement on data dimension reduction: In this project, we apply model training on data that has been dimension-reduced for variance of 95% and 90%. However, since initially, there are over 20000 dimensions, whether our dimension reduction operation is reasonable requires to discuss [7]. The plot for relation between variance of dataset and dimension counts is shown in figure 10. From this figure, we can see that the first 200 dimensions have already guaranteed for 85.45% of variance, however, if we want variance for 95%, this number rapidly increases to over 1200, while the remaining 5% variance calls for even more than 4000 remaining dimensions.

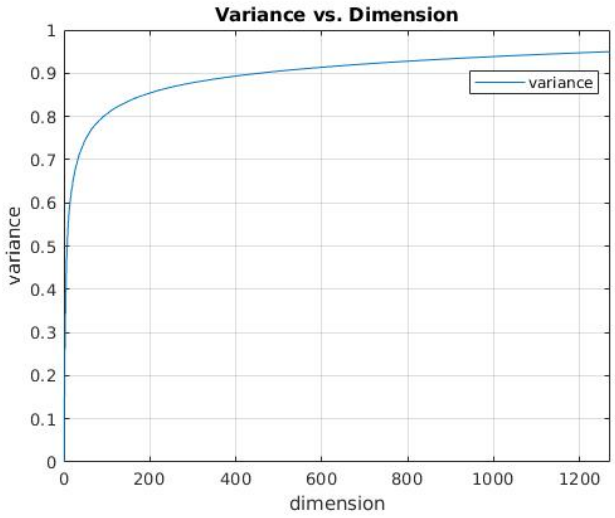

Fig. 10. Relation between variance and dataset dimension. First 200 dimensions account for 85.45%, while 95% calls for 1271 dimensions.

Since the variance that is accounted by the last 5000 dimensions are quite small, it is hard to say whether they have relation with the labels' change. The bias-variance dilemma occurred again in the process of dimension reduction, that is, if we want to make sure that our variance percentage is large enough, we may add many so-called noises information into our training process, however, if these dimensions with seemly little-important are totally abandoned, we may lose some information inside these dimensions, especially when these key parts are not represented in a single-column form, but in other forms like linearly combination of many different columns, which consequently can bring our model with high biases. For traditional methods, this dilemma seems hard to solve, however, by deep learning methods, if those concealed features can be found by models, this dilemma may be efficiently relieved.

3) Improvement on deep learning model structure: Currently there is no efficient method to determine the best NN structure before complete model training, and even after first several trial of training, the so-called best structure is still the result based on our experience and existing model performance. Since for this project, the structures of our two deep learning models are both determined by experience dogma that the amount of nodes on hidden layers should be approximately the mean of input nodes and output nodes, there is still a huge space for the improvement of our deep learning model structure [11]. For the possible improvement way, one is to change the fully-connected NN structure into a convolution-based NN structure, applying sparse interactions, sharing parameters as well as adding a pooling layer [4] [12].

C. Improvement on task frame

In this project, we only treat the problem as a binary classification task due to the limited scale of data. Although there are 192 different disease states included in dataset, for each of them there are only several samples which is not enough for model training. Hence if we want to extend project into a multi-class classification task, i.e., using the model not only to predict the normal-abnormal state, but also to predict the specific possible disease state, more data are required. If dataset of similar scale as this one is available for a single disease state, the task frame can be modified in this way.

In addition, improvements on label parsing can be done through the decomposition of disease state. For instance, medical knowledge may tell us whether we can treat a single disease state as a combination of several fundamental ones. If this can be accessed, dataset scale may be able to increase only with the current data.

ACKNOWLEDGMENT

The author thanks the support of course professor Bo Yuan. This project was also supported by TA Tianfan Fu, whose patient and creative answers of the author's questions have brought a lot of helps for the whole project.

REFERENCES

[1] Smyth, Gordon K. "Limma: linear models for microarray data." Bioinformatics and computational biology solutions using R and Bioconductor. Springer, New York, NY, 2005. 397-420.
[2] Furey, Terrence S., et al. "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics 16.10 (2000): 906-914.
[3] Wang, Yu, et al. "Gene selection from microarray data for cancer classificationa machine learning approach." Computational biology and chemistry 29.1 (2005): 37-46.
[4] Goodfellow, Ian, et al. Deep learning. Vol. 1. Cambridge: MIT press, 2016.
[5] Refaeilzadeh, Payam, Lei Tang, and Huan Liu. "Cross-validation." Encyclopedia of database systems. Springer US, 2009. 532-538.
[6] Abdi, Herv, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.
[7] Cands, Emmanuel J., et al. "Robust principal component analysis?." Journal of the ACM (JACM) 58.3 (2011): 11.
[8] Scholkopf, Bernhard, and Alexander J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.
[9] Hanley, James A., and Barbara J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." Radiology 143.1 (1982): 29-36.
[10] Guyon, Isabelle, and Andr Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.
[11] Baum, Eric B., and David Haussler. "What size net gives valid generalization?." Advances in neural information processing systems. 1989.
[12] Egmont-Petersen, Michael, Dick de Ridder, and Heinz Handels. "Image processing with neural networksa review." Pattern recognition 35.10 (2002): 2279-2301.