A Deep Learning-Based Hybrid Feature Selection Approach for Cancer Diagnosis

Haoran Wu*
1 Pennsylvania State University, State College, PA, 16802, United States
*Corresponding author’s e-mail: wu960721@gmail.com

Abstract. Feature selection plays an important role in machine learning-based classification tasks, especially in high dimensional data, such as biological omics datasets. Recent research has begun to explore the use of deep learning to accomplish this task as a step in feature representation. In this research, we developed a deep learning-based hybrid feature selection approach combining Sparse Autoencoder (SAE) and Logistic Regression-Recursive Features Elimination (LR-RFE) and evaluated our method on TCGA miRNA datasets. The results show that our proposed hybrid method achieves a better performance compared to other comparison methods.

1. Introduction
The emergence of omics technology marks the beginning of exploring complex biological systems from a systematically perspective. Due to the ability of analysing a broad range of molecules simultaneously, omics technologies are able to monitor families of cellular molecules in a high throughput manner [1].

Oomics technologies have been applied to investigate different biological molecular levels, such as genomics, proteomics, metabolomics, transcriptomics, lipidomics, and glycomics. Due to high level of complexity of omics data, using computational tools like machine learning and deep learning to analyse the deep level architecture of these intricated data has become increasingly popular. In machine learning and deep learning, the quality of features decides the upper limit of a model's performance. Feature selection is one of the methods to avoid overfitting and improve final model performance. It can delete redundant features to reduce data dimension, enhance the precision of the model, and decrease the model's running time.

Thus, a variety of feature selection methods have been proposed. Chuang et al. proposed a hybrid filter-wrapper method to select feature genes in microarrays dataset, information gain was used as the filtering step to calculate each feature's importance with respect to each class, and finally the traditional BPSO (Binary Particle Swarm Optimization) was used as a wrapper step to further select informative features [2]. In another study, Maldonado et al. used Support Vector Machine (SVM) with kernel functions as a novel wrapper feature selection method to decide which feature to remove in each iteration [3]. Nevertheless, due to the restriction of traditional machine learning setup in solving data with complicated structures, deep learning methods started to more attentions to accomplish the task of feature selection. Current deep learning-based feature selection method usually involves two steps. In the first step, deep learning is used to extract a high-level feature representation. In the second step, a traditional feature selection method is used to select important features. Some studies have proposed in using deep learning as a feature selection method in biology field. Nezhad et al. developed a two-step feature selection method to deal with a medicine problem which focus on discovering risk factors for hypertension (HTN) in a specific subgroup (African-American), Nezhad used stacked-autoencoder as a
deep architecture for feature representation at higher levels in the first step, and then used random forest (RF) to select important features [4]. Authors in [5] used multi-layered sparse auto-encoder as a deep feature selection method for a system to classify cancer based on the gene expression data. Deep learning as a way to extract features in deep architecture, is also widely used in biomedical image recognition. Biomedical experts always diagnose pneumonia through X-ray images. Togacar et al. used Convolutional Neural Networks (CNN) as a deep feature learning method through extracting discriminative features in lung X-rays images to help shorten the whole diagnosis process [6]. In the study of cancer detection and relevant gene identification, Danae et al. applied Stack Denoising Autoencoder (SDAE) do deeply extract features from high dimensional gene expression datasets [7].

In this paper, we propose a novel hybrid feature selection method based on deep learning. In our method, we apply sparse auto-encoder as a feature abstraction step to represent features in a higher level. In the next step, Recursive feature elimination (RFE) is used as a traditional feature selection method to select most important features.

The remaining part of the paper is organized as follows: Section II will discuss our method in detail, Section III will describe our proposed methods with comparison methods, Section IV presents the report, and Section V makes the conclusion.

2. Method

2.1. Datasets
In order to carry out an comprehensive evaluation, we downloaded four mi-RNA seq gene expression datasets from The Cancer Genome Atlas (TCGA) http://gdac.broadinstitute.org/ which are Stomach adenocarcinoma (STAD), Breast invasive carcinoma (BRCA), Head and Neck squamous cell carcinoma (HNSC) and Prostate adenocarcinoma (PRAD). In our pre-processing steps, we first transposed original dataset and set the Sample Encoding in Hybridization REF barcode as the label as our goal is to predict the status of samples where ‘01’ indicates Primary Solid Tumor and ‘11’ indicates Solid Tissue Normal. Secondly, we deleted the columns whose feature values were all 0. The number of features and samples of these 4 datasets after pre-processing steps are shown in Table 1:

| Datasets | Sample Number | Feature Number |
|----------|---------------|---------------|
| STAD     | 430           | 868           |
| BRCA     | 849           | 897           |
| HNSC     | 532           | 903           |
| PRAD     | 547           | 850           |

2.2. Proposed Approach
We propose a hybrid two-step feature selection method as shown in Figure 1. In our method, we first use Sparse autoencoder (SAE) for higher-level feature representation to discover deep architecture. In the second step, we will apply a Logistic Regression based Recursive Feature Elimination (LR-RFE) to select the 30 most important features as final output. The details of each component of our proposed method is described in Figure 1.

Figure 1. Proposed Method. We first used SAE for feature representation, then used LR-RFE to select important features. SAE, Sparse Autoencoder; LR-RFE, Logistic Regression based Recursive Feature Elimination
2.2.1. Sparse Autoencoder (SAE)

In order to understand what is sparse-autoencoder, we need to get a basic understanding of autoencoder. As shown in Figure 2. Autoencoder is an unsupervised learning algorithm which learns the deep structure of features. Because of this specialty, we always use autoencoder for feature representation. There are multiple different types of autoencoder such as sparse autoencoder, denoising autoencoder, contractive autoencoder and stack autoencoder. As we can see in Figure 2, there are three layers in the autoencoder structure. Layer L1 is an input layer, layer L2 is a hidden layer and layer L3 is an output layer. The process between L1 and L2 is called the encoding step where the L2 hidden layer tries to get the compressed representation of input. And the process between L2 and L3 is call the reconstruct step. The whole neural network tries to get the function

$$h_{(w, b)}(x) = x$$

where $w$ is the weight for connection between unit $j$ in layer $l$ and unit $i$ in layer $l + 1$, $b$ is the bias for unit $i$ in layer $l + 1$ [8].

![Autoencoder](image)

Figure 2. Autoencoder. From Layer L1 to Layer L2 is the process of encoding, from Layer L2 to Layer L3 is the process of decoding

Ng et al. proposed sparse autoencoder because when the number of units in hidden layer is bigger than that in input layer, we cannot get the compressed representation without sparsity. To be specific, a neuron is active when its output is close to 1, and a neuron is inactive when its output is close to 0 [8]. What sparsity actually do is to constrain the most neurons to be inactive in the most time.

We define $a_{j}^{(2)}(x)$ as the activation value of unit $j$ with input $x$ in layer L2 (Hidden Layer). Let

$$\hat{\rho}_j = \frac{1}{m} \sum_{i=1}^{m} [a_{j}^{(2)}(x)(x(i))]$$

be the average be the average activation value of unit $j$ in hidden layer with input $x$. We want to set the constraint $\hat{\rho}_j = \rho$, and $\rho$ is a sparsity parameter which is a really small number close to 0. We need to set one more penalty term to constraint the $\hat{\rho}_j$ deviating too much from $\rho$. The penalty term is as follows:

$$\sum_{j=1}^{s_2} \rho \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \hat{\rho}_j}$$

(2)

where $s_2$ is the number of unit in hidden layer and unit $j$ is summing over the hidden units in our networks. The formula can also be written as (3):

$$\sum_{j=1}^{s_2} KL(\rho || \hat{\rho}_j)$$

(3)

where is Kullback-Leibler (KL) divergence between Bernoulli random variable with mean $\rho$ and a Bernoulli random variable with mean $\hat{\rho}_j$. Our overall cost function of Sparse Autoencoder can be written as:
\[ J_{SAE} = J(W, b) + \beta \sum_{j=1}^{s_2} KL(\rho || \hat{\rho}_j) \]  

where \( \beta \) is the parameter controlling the weight of the sparsity penalty term. Sparse Autoencoder is one of the popular models used for feature representation in omics data. Xu et al. developed a Stack Sparse Autoencoder (SSAE) for nuclei detection on breast cancer histopathology images [9]. In another study, Chang et al., used Sparse Autoencoder to learn characteristics of protein interactions in cancer cell lines integrated with protein expression data [10].

2.2.2. Traditional Feature Selection Method

In this step, we use Recursive Feature Elimination (RFE) as a wrapper feature selection method. RFE can be used based on different algorithms to choose features and determine how many features we want to choose [11]. In another study, Ding et al. proposed an SVM-RFE method which performs feature selection by recursively training a SVM classifier with the current set of features and removing the features indicated by the SVM [12]. In this paper, we will use Logistic Regression based Recursive Feature Elimination (LR-RFE). Features will be scored by Logistic Regression and ranked by their importance. RFE attempts to eliminate the features recursively until the desired number is obtained.

2.2.3. Evaluation

In our study, we used Support Vector Machine (SVM) and Random Forest (RF) as the classifiers to evaluate the features selected by our proposed feature selection method. We used a 10-fold cross validation accuracy as metrics to compare our method with other comparison methods.

3. Results

In our study, we propose a deep learning-based feature selection approach which starts with SAE for the purpose of feature representation, followed with LR-RFE which outputs final feature selection result. In order to evaluate our approach, we carried out a comprehensive evaluation by comparing our method with other related pipeline setups. As shown in Table 2, we included a pipeline with Principal Components Analysis (PCA) which is also a commonly used dimension reduction method for high dimensional data in biological field [13-14]. However, our result indicates that adding a PCA in the proposed pipeline adversely affects the final performance and it may result from the loss of original information. In the HNSC dataset, if we add PCA to our proposed method, the result is 98.99%, which is lower than the 99.22% obtained from our proposed method.

| Datasets | PCA | DL | Traditional Feature Selection | SVM | Datasets | PCA | DL | Traditional Feature Selection | SVM |
|----------|-----|----|-------------------------------|-----|----------|-----|----|-------------------------------|-----|
| STAD     | SAE | DBN| LR-RFE                        | 97.11% | PCA     | SAE | DBN| LR-RFE                        | 98.99% |
|          |     |    | Chi-Squared Test              | 92.07% |          |     |    | Chi-Squared Test              | 94.75% |
|          |     |    | LR-RFE                        | 95.39% |          |     |    | LR-RFE                        | 97.01% |
|          |     |    | Chi-Squared Test              | 93.19% |          |     |    | Chi-Squared Test              | 95.78% |
| HNSC     | SAE | DBN| LR-RFE                        | 98.37% | SAE     | DBN|    | LR-RFE                        | 99.22% |
|          |     |    | Chi-Squared Test              | 89.73% |          |     |    | Chi-Squared Test              | 94.58% |
|          |     |    | LR-RFE                        | 94.69% |          |     |    | LR-RFE                        | 96.94% |
|          |     |    | Chi-Squared Test              | 91.37% |          |     |    | Chi-Squared Test              | 95.40% |
| BRCA     | PCA | SAE| LR-RFE                        | 96.79% | PRAD    | PCA | SAE| LR-RFE                        | 94.82% |

Table 2. Comparison our Proposed Method with other Related Methods
As another comparison method, Deep Belief Networks (DBN) is a probabilistic generative deep architecture composed of multiple layers of stochastic and latent variables [15]. As shown in Table 2, SAE has a better feature representation performance compared to DBN. In the STAD dataset, pipeline with SAE achieves an accuracy of 98.37% which is nearly 2% higher than the pipeline with DBN. Also, in the HNSC dataset, our proposed approach with SAE obtains an accuracy of 99.22% which is nearly 3% higher than the approach with DBN.

We also compared Chi-Squared test with our proposed LR-RFE method. As shown in Table 2, the result of BRCA dataset shows that pipelines with LR-RFE is 97.46%, while the pipeline with Chi-Squared test is only 91.81%, which is much lower compared to the one with LR-RFE method. Similarly, in PRAD dataset, LR-RFE is also about 3% higher than Chi-Squared method. In conclusion, our proposed pipeline obtains the best performance in term of accuracy across all of comparison methods in the four TCGA datasets.

4. Conclusion

In our study, we proposed a hybrid deep learning based feature selection approach for cancer diagnosis. We applied Sparse-Autoencoder for feature representation at higher levels and get 100 new features. Following SAE, we used LR-RFE to select 30 most important features for classifiers. The results of four mi-RNA seq gene expression datasets obtained from TCGA show that our approach outperforms other comparison approaches.

References

[1] Debnath, Mousumi, Godavarthi BKS Prasad, and Prakash S. Bisen. Molecular diagnostics: promises and possibilities. Springer Science & Business Media, 2010. 11-31.
[2] Chuang, Li-Yeh, Chao-Hsuan Ke, and Cheng-Hong Yang. "A hybrid both filter and wrapper feature selection method for microarray classification." arXiv preprint arXiv:1612.08669, 2016.
[3] Maldonado, Sebastián, and Richard Weber. “A Wrapper Method for Feature Selection Using Support Vector Machines.” Information Sciences, Elsevier, 5 Mar. 2009.
[4] Nezhad, Milad Zafar, et al. "Safs: A deep feature selection approach for precision medicine." 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2016.
[5] Fakoor, Rasool, et al. "Using deep learning to enhance cancer diagnosis and classification." Proceedings of the international conference on machine learning. Vol. 28. New York, USA: ACM, 2013.
[6] Toğacar, M., et al. "A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models." IRBM 41.4, 2020. 212-222.
[7] Danaee, Padideh, Reza Ghaeini, and David A. Hendrix. "A deep learning approach for cancer detection and relevant gene identification." PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017. 2017.
[8] Ng, Andrew. "Sparse autoencoder." CS294A Lecture notes 72.2011, 2011. 1-19.
[9] Xu, Jun, et al. "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images." *IEEE transactions on medical imaging* 35.1, 2015. 119-130.

[10] Chang, Ji-Wei, et al. "A deep learning model based on sparse auto-encoder for prioritizing cancer-related genes and drug target combinations." *Carcinogenesis* 40.5, 2019. 624-632.

[11] Brownlee, Jason. “Recursive Feature Elimination (RFE) for Feature Selection in Python.” *Machine Learning Mastery*, 27 Aug. 2020, machinelearningmastery.com/rfe-feature-selection-in-python/.

[12] Ding, Yuanyuan, and Dawn Wilkins. "Improving the performance of SVM-RFE to select genes in microarray data." *BMC bioinformatics*. Vol. 7. No. S2. BioMed Central, 2006.

[13] Sanguansat, Parinya, ed. *Principal Component Analysis: Multidisciplinary Applications*. BoD–Books on Demand, 2012.

[14] Zhang, Tonglin, and Baijian Yang. "Big data dimension reduction using PCA." *2016 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2016.

[15] Hinton, Geoffrey E. "Deep belief networks." *Scholarpedia* 4.5, 2009. 5947.