A Comprehensive Comparison of Neural Network-Based Feature Selection Methods in Biological Omics Datasets

Fu Tong
Beijing University of Technology
tong.fu@ucdconnect.ie

ABSTRACT

Omics allows researchers to apply systems biology approaches to identify the novel pathophysiological mechanism of diseases. It yields an unprecedented view of the cellular inner workings and is now often incorporated into the everyday methodology of biological researchers. However, the curse of dimensionality caused by the lack of enough samples and a large number of features is a major impediment to using Omics. In order to improve performance and decrease impediments of dimensionality in machine learning using Omics data, feature selection techniques are adopted. There are different feature selection methods and deep learning-based methods that have been attracting increasing attention in the field. In this paper, I applied the neural network-based feature selection methods to extract the Stomach and Esophageal carcinoma (STES) gene and compared the results. Finally, through comprehensive comparison, I found that the use of neural network-based feature selection methods did not always help improve the performance.

CCS CONCEPTS

• **Computing methodologies** → Machine learning; Learning paradigms; Supervised learning.

KEYWORDS

Omics, Feature selection, Machine Learning

ACM Reference Format:
Fu Tong. 2021. A Comprehensive Comparison of Neural Network-Based Feature Selection Methods in Biological Omics Datasets. In *2021 4th International Conference on Signal Processing and Machine Learning (SPML 2021), August 18–20, 2021, Beijing, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3483207.3483220

1 INTRODUCTION

Omics technology is the method aiming to a comprehensive characterization, quantitation, and quantification of many molecules, grouped according to fundamental structural or functional biological similarities. It includes universal detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) [1-2].

SPML 2021, August 18–20, 2021, Beijing, China

ACM ISBN 978-1-4503-9017-0/21/08.
https://doi.org/10.1145/3483207.3483220

MicroRNAs (miRNAs) are noncoding RNAs which regulate gene expression through targeted binding to mRNA. [01] During biogenesis, a mature miRNA that can participate in posttranscriptional modification once bound to the RNA induced silencing complex (RISC) [3-4]. Many single miRNAs are usually required to adequately inhibit a single mRNA depending on the degree of base pairing [5-6]. MiRNA plays an important role in post-transcriptional regulation. Many miRNAs have been identified as potential biomarkers and targets for the diagnosis of many diseases, such as cancer [3]. The Cancer Genome Atlas (TCGA) researchers have sequenced and analyzed the tumors of more than 30 individuals through large-scale genome sequencing, providing a publicly available data set in which miRAN expression can be used as a signature for computational analysis and prediction of cancer [7].

Cancer is a complex genetic, proteomic, and cellular disease caused by multiple factors, including genetic mutations (hereditary or somatic) and environmental factors [8]. Since early diagnosis may improve the survival rate and overall medical care of cancer patients, the emergence of omics provides a new opportunity in the detection and treatment of cancer.

However, there are some limits to the utilization of omics techniques. Because of a significant amount of time-consuming data acquisition process and high cost, a generalized lack of samples is a challenge in the field. Machine learning methods have been widely used to analyze large data sets, like the ones that derive from multi-omics measurements, and can lead to algorithms with predictive value [9].

The selection of appropriate features in the machine learning method is the key to the accuracy of the prediction results. Every omics dataset can provide thousands of features, so only increasing sample size might not avoid negative impacts, such as overfitting and redundancy. So, I prefer another available method which is selecting some more valuable features for cancer diagnosis.

In this paper, I used the deep belief network method, the autoencoder method, and so on to extract the feature to obtain a higher-level representation of feature abstraction. After that, the abstract features are selected, and multi-component air training is conducted. By analyzing the results that methods without deep learning surpass methods with deep learning, I give suggestions on the diagnosis process using the machine learning data analysis method with Omics technology.

This paper is organized as follows: The second section will introduce the proposed methods in detail; then my results will be present and compared; finally, in the fourth section, I summarize the whole paper and give suggestions.

2 RELATED WORK

In 2010, Christoph et al. attempted to use principal component analysis (PCA) to map high-dimensional data of DNA microarray data to low-dimension data, and the result showed that with very few dimensions the most significant information in the data can be captured better than the original data [10]. After this, people made active exploration and attempt the method of applying feature selection on biological data. Xie et. al proposed a clustering learning of the feature space named unsupervised deep embedding for clustering analysis, they define a parameterized non-linear mapping from the data space X to a lower-dimensional feature space Z, then use stochastic gradient descent (SGD) via backpropagation on a clustering objective to learn the mapping to achieve the purpose of unsupervised clustering feature selection [11]. Nezhad et al. a new deep feature selection method based on deep architecture. They used stacked auto-encoders for feature representation in higher level abstraction. This novel feature selection approach focused on assessing and prioritizing risk factors for hypertension (HTN) in a vulnerable demographic subgroup (African-American) and has been proved the feature learning approach leads to better results in comparison with others [12].

In recent years, researchers have begun to use feature extraction methods on a wider range of organisms. Gupta et al. apply Gray Level Co-Occurrence Matrix (GLCM) as a feature extraction technique to extract the features from CT scan image, which would be given to classifier to train and detect head and neck cancer [13].

Autoencoder inspired unsupervised feature selection method proposed by Han et al. is a novel AutoEncoder Feature Selector (AEFS) for unsupervised feature selection which combines autoencoder regression and group lasso tasks. AEFS could select the most important features by excavating linear and nonlinear information among various features, which is more flexible than the conventional self-representation method for unsupervised feature selection with only linear assumptions [14].

3 METHODS

In this section, I will introduce the method I proposed in detail, including the deep learning methods involved, the algorithms adopted for feature selection, the data set used, the experimental setting, and the evaluation criteria.

3.1 Methods overview

In this work, I analyzed miRNA-seq expression data downloaded from the cancer genome atlas (TCGA) database which characterized over 20,000 patients and matched normal samples spanning 33 cancer types [15]. I used data sets for Stomach and Esophageal carcinoma (STES).

I deleted genes whose expression is zero in all samples, and all samples with zero genetic characteristics. Then I standardized all the values in the dataset. The cleaned dataset was divided into 70% for training and the rest of 30% for testing.

For the purpose of comparison, I designed three setups, namely, autoencoder, deep belief network (DBN), and no high-level representation. In the first setup, I used a deep belief network to extract a high-level representation of features from gene expression data. Then the high-level represented features are used for feature selection to acquire further selected features (see Figure. 1. A) In the second setup, I replaced the deep belief network with the autoencoder as the method to obtain the high-level representation of gene features. This setup is presented in Figure. 1. B. For the third setup, I did not use the deep learning approach to achieve high-level representation shown in Figure. 1. C.

After the above high-level representation extraction and feature selection, three groups of selected features can be obtained. In the first two setups I obtained 200 high-level representations of features through the extraction step, and then selected the top 50 of these ranked features as the final training features with a classical feature selection approach (including random forest, decision tree, and chi-square).

3.2 Autoencoder

Autoencoder is a type of artificial neural network applied to learn a high-level representation of original data in an unsupervised manner. There are several variants of self-encoders, including Sparse Autoencoder (SAE), Variational Lossy Autoencoder, and Denoising Autoencoder (DAE). In this study, I use a traditional auto-encoder for data representation learning.

In an autoencoder, a neural network architecture is developed with a bottleneck which forces a compressed knowledge representation of the original input. The autoencoder then learns how to reconstruct the data from a reduced dimension representation to a representation as close to the original input as possible. Its structure is divided into two parts: encoder and decoder.

The encoding step of each layer is a forward process and mathematically described as equation (1), (2), and (3), given the input space x∈X, and feature space h∈F, the auto-encoder solves the mapping f and g of the two, so as to minimize the reconstruction error of input features [16]:

$$f : X \rightarrow F \tag{1}$$

$$g : F \rightarrow X \tag{2}$$

$$f, g = \ argmin||x - g\left[f\left(x\right)\right]||^2 \tag{3}$$

After the optimization process is completed, the encoded feature of the hidden layer output by the encoder, i.e., "encoded feature", can be regarded as the representation of the input data. According to the difference of the auto-encoder, its encoding characteristics can be compression (contraction autoencoder), sparse (sparse autoencoder), or implicit variable model (variational auto-encoder) of the input data [17].

In this paper, the autoencoder model adopts the following training parameters:

In the coding layer, there are three layers. In the first layer, units=889, and activation='relu'. In the second layer, units = 512, and activation='relu'. In the third layer, units=256, and activation='relu'. In the encoder output, the units=200 and activation=None.

In the decoding layer, there are three layers. In the first layer, units=256, and activation='relu'. In the second layer, units = 512, and activation='relu'. In the third layer, units=889 and activation='tanh.

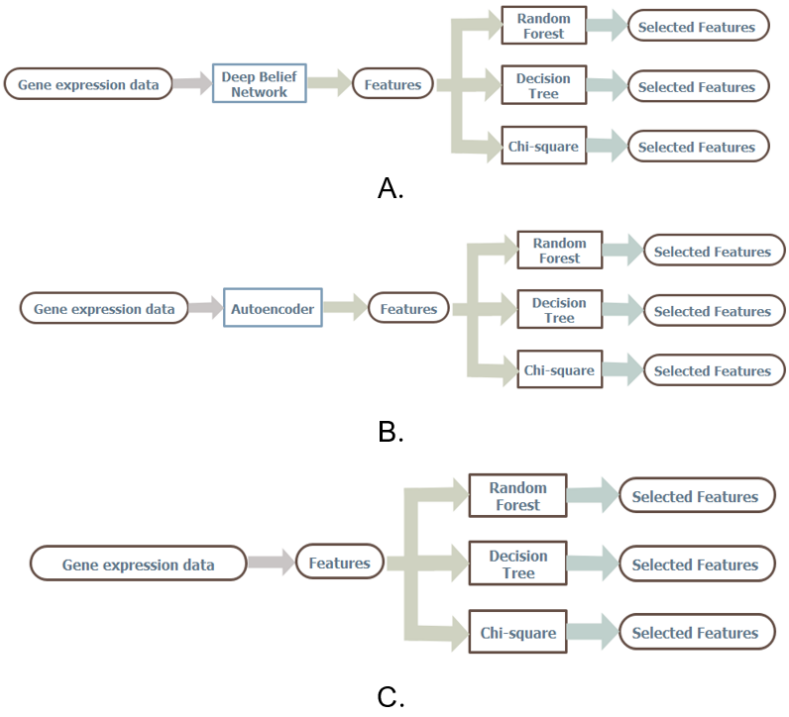

Figure 1: The main process and method of the experiment

3.3 Deep Belief Network

When using traditional neural networks training in deep layered networks, there will be some problems, such as slow learning and being stuck in local minima due to poor parameter selection. To solve these problems, deep belief networks are proposed by Hinton in 2006 [18].

A deep belief network is constructed by multiple Restricted Boltzmann Machine (RBM) stacks. It considers the output of the hidden layer in an RBM as the visible layer's input of another RBM. In this process, I can obtain the extracted feature of samples [16].

As Figure 2, Deep Belief Network, stacked by three RAMs, uses the output of the upper RBM's hidden layer as the input of the lower RBM's visible layer.

In this paper, DBN model adopts the following training parameters:

n_ins = 20271, hidden_layer_sizes = [200, 200], n_outs = 1, rng= np.random.RandomState(123)

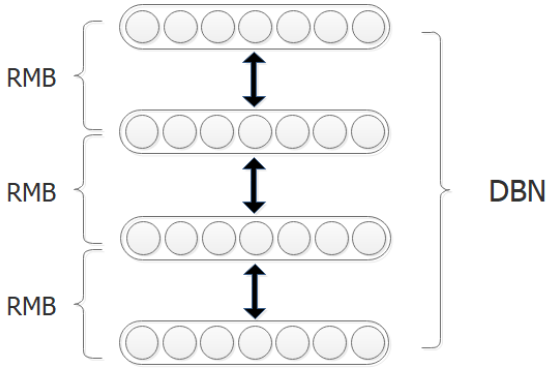

Figure 2: A sample DBN

3.4 Traditional feature selection methods

Selecting the right set of features has been proved helpful for data modeling, reducing computational costs, and solve the curse of dimensionality. Therefore, the feature selection method is justified as an important step towards the model. In this paper, some classical feature selection methods are used, including support vector machines (SVM), random forest (RF), decision tree (DT), and chi-square.

The SVM is already known as a tool that discovers informative patterns [19]. The method is based upon finding those features which minimize bounds on the leave-one-out error. This search can be efficiently performed via gradient descent. Weston et al. has proved SVMs have both quantitative and qualitative advantages in a comparison with several other gene selection methods on Colon cancer data [20].

The DT selects features by using the mean decrease impurity or the mean accuracy decrease criteria. It can include explicit conditions at each branching node, which are based on single features. Then information gain or entropy will be used to compute the significance of variables. The ranking of feature importance is based on the average impurity decrease to achieve effective feature selection [21].

The RF performs as a classifier that is constructed with many DTs. Each DT is a classifier. For each input sample, every node has multiple classified results. To classify an input vector in an RF, the vector is submitted to each tree of the forest. Then the RF integrates all the classification voting results and designates the category with the most votes as the final output [22].

Chi-square is mainly used to assess two kinds of comparing: tests of independence and tests of goodness of fit. In feature selection, the test of independence is assessed by chi-square and estimates whether the class label is independent of a feature. Assesses test of independence Chi-square score with c class and r values. The formula is expressed as equation (4).

$$x^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij} - \mu_{ij}}{\mu_{ij}} \tag{4}$$

μ_{ij} is the amount of sample value with the i^{th} value of the feature. n_{ij} is the amount of samples with the i^{th} value of the feature value in class j [23].

3.5 Feature selection method

I respectively used the random forest, decision tree, and Chi-square methods for feature selection. Through the following comparison, it can be found that the method of support vector machine Performed better on several of these criteria, but the decision tree algorithm has significantly less computation and memory overhead. In terms of feature selection, some current researches use some more advanced methods to select features, such as the neural network. However, based on the good results I have obtained now, I leave this exploration to the future.

3.6 Dataset and experimental environment

This paper uses the miRNA sequencing data on the Stomach and Esophageal carcinoma in TCGA, which encodes the expression of a large number of sequencing sites and statistics the data. The dataset contains more than 20273 samples and covers 646 gene detection sites that could be used as initial features [7]. The samples are classified as normal and abnormal, and the abnormal sample is the one with cancer. I selected the valuable features from these data. After that, the incomplete data is removed and the normalized processing is carried out. As for the training data and test data, I randomly selected 70% data as the training data and the rest as the test data. This reduces the influence of uneven sample distribution on experimental results to a certain extent.

The experiment is conducted using AMD Ryzen 7 3700U 2.30GHz,4GB memory, Windows10 system, pytorch1.4.0, Python 3.7 programming language, CPU Memory 500G, and GPU Memory 8G.

3.7 Evaluation

In this paper, classifier random forest (RF) classifier or support vector machine (SVM) classifier was applied to train the final fifty selected features to obtain the training results. I also used a series of evaluation metrics to evaluate the feature selection performance of three different setups, including classification accuracy, f-score, precision, and recall.

Accuracy is an index used to evaluate classification models. In another word, accuracy refers to the proportion of all the samples that my model predicts correctly. For the dichotomy problem, according to the definition of accuracy, it can be obtained with equation (5):

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

Precision (shown as equation (6)) quantifies the number of positive class predictions that actually belong to the positive class.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

Recall (shown as equation 7)) quantifies the number of positive class predictions made out of all positive examples in the dataset.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TN}} \tag{7}$$

F-Measure provides a single score that balances both the concerns of precision and recall in one number.

4 EXPERIMENTAL RESULTS AND ANALYSIS

To compare results with or without high-level representation, I analyzed the results with the use of the auto-encoder and deep belief network method as the deep learning methods and without deep learning methods. Then, to ensure the universality of the results, six experiments were applied for each setup. I applied three traditional feature selection methods, including RF, DT, and Chi-square. For every feature selection method, I used 2 classifiers, including RF and SVM. The results are evaluated with accuracy, precision, recall, and f-score, respectively. Particularly, I performed 5-fold cross-validation and presenting results in terms of the average evaluation.

From these tables, in the comparison between the deep learning method and the no deep learning method, I can see that methods without deep learning outperform methods with deep learning whether it is an autoencoder or a deep belief network. For example, in the experiments using RF as a feature selection method and SVM as a classifier, the results without deep learning are 0.8 higher than those using DBN and those using auto-encoder in f-score.

I then compared the performance of the two approaches using deep learning. The results of using auto-encoders are generally slightly higher than that of deep belief networks. The only exception is the experiment using chi-square for selecting features and SVM for classification (in Table 1).

5 CONCLUSION AND DISCUSSION

In this paper, I compared performance differences with or without deep learning method in the biological Omics datasets. The deep belief networks or auto-encoders extract a high-level representation of features using all the gene datasets. I selected features from high-level representations of features with traditional feature selections. By evaluating these results, I found that experiments without a deep learning approach perform better than experiments with a deep learning approach. I can see that when solving high-dimension data, using a deep learning method to obtain high-level representations for training will not always help improve training performance. In this paper, I have not invited medical experts to conduct new verification, and I will do so in the next step. DBN and auto-encoder models are always used in the feature selection of Omics data set

Table 1: The results of comparison for different methods

Deep Learning	Feature Selection	Classifier	Accuracy	F-score	Precision	Recall
Auto -encoder	RF	RF	0.93	0.89	0.86	0.93
		SVM	0.92	0.88	0.85	0.92
	DT	RF	0.92	0.88	0.86	0.92
		SVM	0.92	0.88	0.84	0.92
	CS	RF	0.93	0.89	0.87	0.93
		SVM	0.92	0.88	0.85	0.92
DBN	RF	RF	0.90	0.86	0.81	0.90
		SVM	0.92	0.88	0.85	0.92
	DT	RF	0.91	0.87	0.83	0.91
		SVM	0.92	0.87	0.92	0.92
	CS	RF	0.91	0.87	0.83	0.91
		SVM	0.93	0.89	0.86	0.93
No Deep Learning	RF	RF	0.95	0.94	0.95	0.95
		SVM	**0.97**	**0.96**	**0.97**	0.92
	DT	RF	0.96	0.95	**0.97**	**0.96**
		SVM	0.96	0.96	0.96	**0.96**
	CS	RF	0.94	0.93	0.92	0.94
		SVM	0.95	0.95	0.95	0.95

in existing studies, so I chose them in the experiment. My future work will try more neural network models to get more credible conclusions.

This study has potential limitations. At present, I used TCGA data sets on Stomach and Esophageal carcinoma. In the future, I will apply this method to a larger data set to verify the scalability and correctness of this method. In addition, only limited feature selection methods and classification methods are used in this paper, which may affect the correctness of my conclusion.

REFERENCES

[1] Leroy Hood. 2002. A personal view of molecular technology and how it has changed biology. Journal of Proteome Research, 1(5), (pp. 399-409).

[2] Bonnie Berger, Jian Peng, and Mona Singh. 2013. Computational solutions for omics data. Nature reviews genetics, 14(5), pp.333-346.

[3] David P. Bartel, and Changzheng Chen. 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat Rev Genet, 5, (pp. 396–400)

[4] John R. Chevillet, Inyoul Lee, Hilary A. Briggs, Yuqing He, and Kai Wang. 2014. Issuesand prospects of microRNA-based biomarkers in blood and other body fluids. Molecules, 19, (pp. 6080–6105).

[5] Eva van Rooij. 2011. The art of microRNA research. Circ Res, 108, (pp. 219–234).

[6] Ramesh S.Pillai, Suvendra N.Bhattacharyya, and Witold Filipowicz. 2007. Repression of protein synthesis by miRNAs: how any mechanisms? Trends Cell Biol,17, (pp. 118–126).

[7] Tomczak, K., Czerwińska, P. and Wiznerowicz, M., 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology, 19(1A), p.A68.

[8] William C. S. Cho. 2010. Omics approaches in cancer research. In An omics perspective on cancer research (pp. 1-9). Springer, Dordrecht.

[9] Tommi Suvitaival, Isabel Bondia-Pons, Laxman Yetukuri, Päivi Pöhö, John J. Nolan, Tuulia Hyötyläinen, Johanna Kuusisto, and Matej Orešič. 2018. Lipidome as a predictive tool in progression to type 2 diabetes in Finnish men. Metabolism, 78, (pp.1-12).

[10] Christoph Bartenhagen, Hans-Ulrich Klein, Christian Ruckert, Xiaoyi Jiang, and Martin Dugas. 2010, Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. BMC bioinformatics, 11(1), (pp. 1-11).

[11] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016, June. Unsupervised deep embedding for clustering analysis. In International conference on machine learning (pp. 478-487).

[12] Milad Zafar Nezhad, Dongxiao Zhu, Xiangrui Li, Kai Yang, and Phillip Levy. 2016, December. Safs: A deep feature selection approach for precision medicine. In 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 501-506). IEEE.

[13] Pooja Gupta, Avleen Kaur Malhi. 2018, July. Using deep learning to enhance head and neck cancer diagnosis and classification. In 2018 IEEE International Conference on System, Computation, Automation, and Networking (ICSCA) (pp. 1-6). IEEE.

[14] Kai Han, Yunhe Wang, Chao Zhang, Chao Li, and Chao Xu. 2018, April. Autoencoder inspired unsupervised feature selection. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2941-2945). IEEE.

[15] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology, 19(1A), p.A68.

[16] Yuming Hua, Junhai Guo, and Hua Zhao. 2015, January. Deep belief networks and deep learning. In Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things (pp. 1-4). IEEE.

[17] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. Deep Learning, vol. 1. Cambridge: MIT press.

[18] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. Neural computation, 18(7), (pp.1527-1554).

[19] Isabelle Guyon, Nada Matic, and Vladimir Vapnik. 1996. Discovering Informative Patterns and Data Cleaning.

[20] Tomaso A. Poggio, Weston Jason, Mukherjee Sayan, Pontil Massimiliano, Chapelle Olivier, Vapnik Vladimir. 2000. Feature selection for SVMs. In Advances in neural information processing systems (pp. 668-674).

[21] Ivo D. Dinov. 2018. Natural Language Processing/Text Mining. In Data Science and Predictive Analytics (pp. 659-695). Springer, Cham.

[22] Pall O. Gislason, Jon A. Benediktsson, and Johannes R. Sveinsson. 2004, "Random Forest classification of multisource remote sensing and geographic data", In International Geoscience and Remote Sensing Symposium (pp. 1049-1052). IEEE.

[23] Rachburee, Nachirat, and Wattana Punlumjeak. 2015. October. A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining. In 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 420-424). IEEE