German Medical Data Sciences: Shaping Change – Creative Solutions for Innovative Medicine
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI190824

Utilizing Molecular Network Information via Graph Convolutional Neural Networks to Predict Metastatic Event in Breast Cancer

Hryhorii CHEREDA[a], Annalen BLECKMANN[a,b,c], Frank KRAMER[d], Andreas LEHA[a] and Tim BEISSBARTH[a,1]
[a] *Medical Bioinformatics, University Medical Center Göttingen*
[b] *Hematology & Medical Oncology, University Medical Center Göttingen*
[c] *Internal Medicine-A (Hematology, Oncology, Hemostaseology and Pulmonology), University Hospital Muenster*
[d] *IT Infrastructure for Translational Medical Research, University of Augsburg*

Abstract. Gene expression data is commonly available in cancer research and provides a snapshot of the molecular status of a specific tumor tissue. This high-dimensional data can be analyzed for diagnoses, prognoses, and to suggest treatment options. Machine learning based methods are widely used for such analysis. Recently, a set of deep learning techniques was successfully applied in different domains including bioinformatics. One of these prominent techniques are convolutional neural networks (CNN). Currently, CNNs are extending to non-Euclidean domains like graphs. Molecular networks are commonly represented as graphs detailing interactions between molecules. Gene expression data can be assigned to the vertices of these graphs, and the edges can depict interactions, regulations and signal flow. In other words, gene expression data can be structured by utilizing molecular network information as prior knowledge. Here, we applied graph CNN to gene expression data of breast cancer patients to predict the occurrence of metastatic events. To structure the data we utilized a protein-protein interaction network. We show that the graph CNN exploiting the prior knowledge is able to provide classification improvements for the prediction of metastatic events compared to existing methods.

Keywords. Gene expression data, classification, CNN, prior knowledge, molecular network.

1. Introduction

Technologies as microarray gene-expression profiling and next-generation sequencing are becoming more and more available and play a significant role in cancer prognosis, for example in discovering individual biomarkers [1]. Furthermore, high-throughput technologies produce huge amounts of data that can be used for assessment of metastatic events. At the moment, deep learning techniques are well known to show prominent results in many research fields with big and complex data.

In recent years deep learning was applied to a wide range of problems in various areas. Deep learning methods are aimed at the automatic learning of data representa-

[1] Corresponding Author, Tim Beißbarth, University Medical Center Göttingen, Medical Bioinformatics, Goldschmidtstr. 1, 37077 Göttingen, Germany; E-mail: tim.beissbarth@ams.med.uni-goettingen.de.

tions (features) needed for machine learning task. These methods demonstrated state-of-the-art performance in visual object recognition, object detection, speech recognition as well as other domains such as drug discovery and genomics [2]. One of the most popular methods of deep learning are Convolutional Neural Networks (CNN). They show cutting edge results for data that are spatially structured. Different classes of such data have different spatial dimensionality: 2D for images, 3D for video and 1D for signals and sequences. The main property of CNNs is a capability of capturing local spatial patterns in natural signals and merging them into high-level abstractions.

The usual CNN architecture consists of three types of layers: convolutional layers, pooling layers, and fully connected layers. The first two layers utilize the Euclidean structure of the data preparing informative features for the fully connected neural network layers. For grid-structured data as images, the convolution layer performs filtering operation to extract highly correlated local groups of pixels forming the same pattern in different parts of the image. A nonlinear function is applied to each output of filtering. As a result, the feature map is created per each filter, consisting of the feature values based on the same pattern. As for the pooling layer, since the slightly shifted position by 1 row or 1 column can give slightly different feature values for the same pattern it merges the feature values into one [2]. Usually this operation is performed by computing the maximum of a local patch of features. In such a way, the dimensionality reduction and the gain of invariance to small shifts are performed.

Deep learning and CNNs are already used in the field of bioinformatics [3]. As an example, CNNs were applied to gene expression data for tumor type classification [4]. One should notice that in Lyu and Haque [4] the gene expression data were transformed into images and then CNNs were applied to them. In general, gene expression data do not have any spatial structure, and the number of genes is much higher than the number of patients that might lead to poor classification performance on the test set. Thus to deal with this problem, still approaches are needed that utilize prior knowledge based on known interactions in molecular networks. Here we demonstrate that the classification performance can be improved by a combination of deep learning and prior biological knowledge.

Nowadays, deep learning is extending to Non-Euclidean domains. This extension is based on generalization of CNNs [5] to graphs and manifolds. We applied graph CNN [6] to gene expression data structured by a molecular network representing the connection between genes. In other words, since each vertex of a molecular network is assigned a gene expression value, we are performing a graph-signal classification task. Recently, quite similar methodology was applied to classify breast cancer subtypes utilizing gene expression data structured by protein-protein interaction network [7]. Breast cancer is one of the three most common cancers in industrialized countries [8]. Patients often develop metastases that limit survival, as there has not been any curative therapy for them [9]. We show that graph CNN outperforms more classical machine learning methods at the prediction of metastatic events in breast cancer.

2. Materials and Methods

2.1. Breast Cancer Data

We used the breast cancer patient data previously studied and preprocessed in research [10]. The data consist of 10 public microarray datasets measured on Affymetrix Human Genome HG-U133 Plus 2.0 and HG-U133A arrays. The datasets have accession numbers GSE25066, GSE20685, GSE19615, GSE17907, GSE16446, GSE17705, GSE2603, GSE11121, GSE7390, GSE6532 and are available from the Gene Expression Omnibus (GEO) [11] data repository. The RMA probe-summary algorithm [12] was used to process each dataset after which they were combined together on the basis of HG-U133A array probe names and quantile normalization was applied over all datasets. In the case of few probes mapped to one gene the probe with the highest average value was taken. In the end, we ended up with 12179 genes per each patient. Further, patients with and without metastatic events were selected to formulate two classes for the prediction task: 393 patients with metastasis within the first 5 years, 576 patients without metastasis having the last follow up between 5 and 10 years.

2.2. Protein-Protein Interaction Network

We used the Human Protein Reference Database (HPRD) protein-protein interaction (PPI) network [13] to structure the gene expression data. This PPI network consists of binary interactions between pairs of proteins and can be represented as an undirected graph. One should notice that this graph is not connected. The genes from gene expression data can be mapped to the vertices of the PPI network. In such a way, the resulting PPI graph has 7168 vertices (genes) matched, and it has 207 connected components. The main connected component has 6888 vertices, and each of the 206 other components has from 1 to 4 vertices. The graph CNN requires graph to be connected so all the machine learning methods had 6888 genes as an input.

2.3. Problem formulation

Initially, the problem is formulated as a binary classification of gene expression data $X \in R^{m \times n}$ to target variable $Y \in R^{m}$ representing the occurrence of metastatic event. m is a number of samples (patients) and n is a number of features (genes). Additionally, we incorporate the information of the molecular network which is represented as a undirected graph $G = (V, E, A)$, where V and Ecorrespond to the sets of vertices and edges respectively. A is an adjacency matrix. The number of vertices is equal to the number of genes n. A row x of gene expression matrix X contains data from one patient and can be mapped to the vertices of the graph G. The values of x are interpreted as a graph signal.

2.4. Graph Convolutional Neural Network and Multilayer Perceptron

The graph CNN [6] captures localized patterns of a graph signal via convolution and pooling operations performed on a graph. The convolution operation bases on the spectral graph theory utilizing the convolution theorem and graph Fourier transform. The graph convolutional filter can be approximated by a parameterized expansion of Chebyshev polynomials of graph frequencies [6]. Such filter of polynomial degree k localizes the signal pattern in K-hop neighboring nodes. For the pooling operation, the graph is coarsened exploiting a graph clustering technique. We applied the graph CNN with following hyperparameters for learning. Two convolutional layers were used with 32 convolutional filters and polynomial degree 8 per each layer. Maximum pooling of size 2 applies to both of the convolutional layers. Two fully connected layers have 512 and 128 units consequently. ReLU (rectified linear unit) activation function was used and cross entropy loss was minimized. Application of usual CNN is not straightforward for gene expression data since it is not spatially ordered. Therefore, we applied deep Multilayer Perceptron implemented in Keras [14], on the same set of genes but without prior knowledge structuring the data. The hyperparameters of our deep neural network are the following: 4 hidden layers and each of them consist of 1024 units with ELU (exponential linear unit) activation function. Cross entropy loss was minimized.

2.5. Random Forest and Lasso Logistic Regression

Random Forest and lasso penalized Logistic Regression were used as baseline methods. Random Forest is a tree-based ensemble machine learning technique combining bagging and random subspace method. It is widely used for high-dimensional data analysis, and considered as a standard tool for class prediction and gene selection with microarray data [15]. Logistic regression with lasso regularization is another classical method for classification of high-dimensional data. Lasso penalty allows shrinking of some coefficients to zero so that the variable selection is automatically performed. For the both baseline methods we utilized RandomForestClassifier and LogisticRegression classes implemented in Scikit-learn package [16].

3. Results

3.1. Our approach

Our approach is to structure gene expression data by applying it to prior knowledge on molecular interactions and to feed this structured data as input for the graph CNN deep learning method. The workflow for predicting metastasis events is shown in Figure 1.

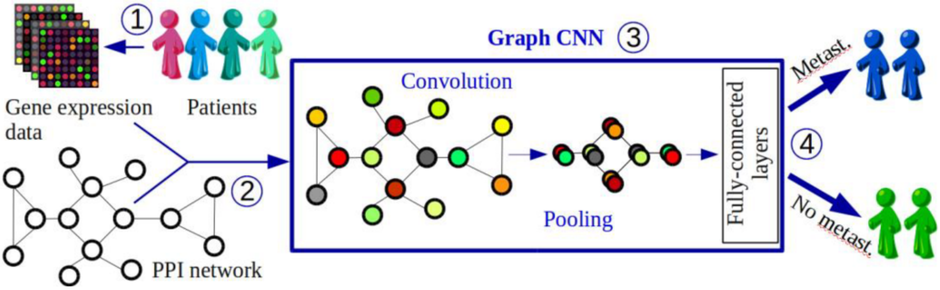

Figure 1. The schema of suggested workflow: 1. Patients' microarray data is preprocessed. 2. Genes are mapped to the vertices of PPI network. 3. The graph CNN processes gene expression data as graph signals. 4. The graph CNN predicts whether the patient is getting metastases during the first 5 years or not.

The endpoint is to predict the occurrence of a metastatic event for a patient. In other words, to classify patients into 2 groups, metastatic and non-metastatic. The first group corresponds to patients with metastasis within the first 5 years and the second concerns patients who are metastasis-free within first 5 years. Graph CNNs were developed recently, and according to the knowledge of the authors the approach described in the paper was not used for metastatic event prediction.

3.2. Comparison of machine learning methods

We compared the graph CNN approach with Multilayer Perceptron, Random Forest and Lasso Logistic Regression. The performance was assessed by 10-fold cross validation. For each of the data splits the model was trained on 9-folds and the classification was evaluated using 10th fold as a validation set. For training, the input was standardized and the validation sets were scaled according to the means and standard deviations on the training set. For each machine learning algorithm the hyperparameters were the same. For each data split the graph CNN and Multilayer Perceptron were trained on the same number of epochs. In this paper we used the most common metrics: area under ROC curve (AUC), accuracy and F1-weighted score. The metrics were averaged over folds and the standard errors of their means were calculated (Table 1).

Table 1. Performance comparison of machine learning methods on metastatic event prediction.

Method	**100*AUC**	**Accuracy, %**	**F1-weighted, %**
Graph CNN	82.16±1.25	76.18±1.36	75.86±1.35
Random Forest	81.40±1.76	74.74±1.67	74.00±1.82
Multilayer perceptron	81.01±1.84	73.92±1.48	73.64±1.54
Lasso Logistic Regression	80.95±1.61	74.74±1.27	74.53±1.27

The graph CNN demonstrates higher values for all three metrics estimating the quality of metastatic event prediction. In such a way we show that utilization of prior knowledge into graph CNN is beneficial in comparison to standard machine learning methods for discriminating classes of patients with or without metastases within 5 years after treatment.

4. Discussion

We demonstrated that the graph CNN applied to graph-structured data predicts metastatic event better than other classical methods that are trained on the same set of features (gene set) and that do not incorporate any prior knowledge. We predicted the occurrence of metastatic events in a breast cancer data set. We have shown that even under the limitations of available data (from deep learning perspective) graph CNN could still outperform other methods. It is well known for breast cancer that molecular subtypes show metastatic differences [10] and thus molecular subtypes influence metastasis-free survival. However, additional confounding factors (e.g. age) may exist that mask the association between input and output variables. The consideration of such confounding factors may be additional future work to consider to evaluate the practical value of such a classifier. Turning event times into a binary endpoint might lead to information loss. One could adapt our method to predict metastasis-free survival.

To structure the gene expression data we utilized only the main connected component of the PPI graph. The majority of other vertices are just single nodes of the PPI graph, thus the prior knowledge of the molecular network does not structure them. In future work we consider utilization of the rest of genes that were not mapped to the main connected component as additional input units of fully-connected layer of graph CNN. The authors in Rhee et al [7] applied the graph CNN to RNA-seq gene-expression data structured by the PPI network extracted from STRING database [17] to predict breast cancer subtypes. The STRING PPI network contains weights on pairs of proteins that interact with each other. 4303 genes were selected. It was also shown that graph CNN could outperform the baseline machine learning methods for the specified classification task. In our case, we have 6888 genes and a binary topology which lead to the hypothesis that graph CNN is able to capture meaningful data representation even if edges do not have weights. For future work we are planning to check how the weighted graph of STRING PPI would improve the classification performance and compare the two methods.

5. Conclusion

In this study we showed that graph CNN applied to microarray gene expression data structured by PPI network outperforms other machine learning methods that do not use any prior knowledge.

6. Conflict of Interest

The authors declare no conflict of interest.

7. Acknowledgements

This work was funded by the German Ministry of Education and Research (BMBF) e:Med projects *MyPathSem* (031L0024A), *Her2Low* (031A429C) and *MMML-Demonstrators* (031A428B).

References

[1] J. Perera-Bel, A. Leha, T. Beißbarth, Bioinformatic Methods and Resources for Biomarker Discovery, Validation, Development, and Integration: Applications in Precision Medicine, In S. Badve, G. Kumar, *Predictive Biomarkers in Oncology*, Springer, Switzerland, 2019. doi:10.1007/978-3-319-95228-4_11

[2] Y. LeCun, Y. Bengio, G. Hinton, Deep Learning, *Nature* **521** (2015), 436-444. doi:10.1038/nature14539

[3] S. Min, B. Lee, S. Yoon, Deep learning in bioinformatics, *Briefings in Bioinformatics* **18** (2017), 851–869. doi:10.1093/bib/bbw068

[4] B. Lyu, A. Haque, Deep Learning Based Tumor Type Classification Using Gene Expression Data, *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 89-96. doi:10.1145/3233547.3233588

[5] F. Monti et al, Geometric Deep Learning on Graphs and Manifolds Using Mixture Model CNNs, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5115-5124. doi: 10.1109/CVPR.2017.576

[6] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, *Advances in Neural Information Processing Systems* (2016), 3844–3852.

[7] S. Rhee, S. Seo, S. Kim, Hybrid Approach of Relation Network and Localized Graph Convolutional Filtering for Breast Cancer Subtype Classification. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization* (2018), 3527–3534. doi:10.24963/ijcai.2018/490

[8] J. Ferlay et al, Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012, *International Journal of Cancer* **136** (2015), 359-386. doi: 10.1002/ijc.29210

[9] F. Bray et al, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: A Cancer Journal for Clinicians* **68** (2018), 394-424. doi: 10.3322/caac.21492

[10] Bayerlová et al, Ror2 Signaling and Its Relevance in Breast Cancer Progression, *Frontiers in Oncology* **7** (2017). doi:10.3389/fonc.2017.00135.

[11] T. Barrett et al, NCBI GEO: archive for functional genomics data sets – update, *Nucleic Acids Res* **41** (2013), 991–995. doi:10.1093/nar/gks1193

[12] R. A. Irizarry et al, Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** (2003), 249–264. doi:10.1093/biostatistics/4.2.249

[13] T. S. K. Prasad et al, Human Protein Reference Database - 2009 Update, *Nucleic Acids Research* **37** (2009), 767-772. doi:10.1093/nar/gkn892

[14] F. Chollet, Keras, *GitHub* (2015). https://github.com/fchollet/keras

[15] R. Díaz-Uriarte, S. Alvarez de Andrés, Gene selection and classification of microarray data using random forest, *BMC Bioinformatics* **7** (2006). doi:10.1186/1471-2105-7-3

[16] F. Pedregosa et al, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12** (2011), 2825–2830.

[17] D. Szklarczyk et al, String v10: protein–protein interaction networks, integrated over the tree of life, *Nucleic acids research* **43** (2014), 447–452. doi:10.1093/nar/gku1003