Domen Mohorčič
Tbilisijska ulica 30, 1000 Ljubljana, Slovenija
Study programme: Data Science Masters in Computer and Information Science
Enrollment number: 63180210

Committee for Student Affairs
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Večna pot 113, 1000 Ljubljana

The master's thesis topic proposal
Candidate: Domen Mohorčič

I, Domen Mohorčič, a student of the 2nd cycle study programme at the Faculty of computer and information science, am submitting a thesis topic proposal to be considered by the Committee for Student Affairs with the following title:

Slovenian: **Učenje s prenosom pri napovedovanju fenotipa na majhnih podatkovnih naborih o izraženosti genov**
English: **Transfer Learning for Phenotype Prediction from small Gene Expression Data Sets**

This topic was already approved last year: ***YES***

I declare that the mentors listed below have approved the submission of the thesis topic proposal described in the remainder of this document.

I would like to write the thesis in English as I am part of the Data Science program, where English is our primary learning language.

I propose the following mentor:

Name and surname, title: Blaž Zupan, prof. dr.
Institution: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
E-mail: blaz.zupan@fri.uni-lj.si

Ljubljana, 4. december 2023.

Proposal of the masters thesis topic

1 The narrow field of the thesis topic

English: bioinformatics, machine learning transfer learning

2 Key-words

English: gene expression data, data integration, transfer learning, data embedding, knowledge-based data analysis

3 Detailed thesis proposal

Past approvements of the proposed thesis topic:

The proposed thesis was submitted and approved in the 2022/2023 academic year.

During our work on this topic, we realized that it would be very difficult to achieve the explicability of individual instances, since we are dealing with many different data sets with different tasks, which would require additional domain knowledge that is not present in the data sets alone. Instead, we decided that the dissertation should focus on analyzing the embedding space generated by the models and evaluating if and to what extent the embedded structure of the data is preserved after encoding.

3.1 Introduction and problem formulation

Understanding biological data is an integral part of medical research. But we are usually dealing with huge amounts of data, more than a single person or even a team can handle. Analyzing it by hand is nearly impossible, which is where machine learning comes in. Machine learning in bioinformatics is used to analyze and extract results from data and can provide helpful insights [1]. Its prominent uses are in genomics, where researchers use it to identify candidate genes behind certain diseases or unique biological properties, and in proteomics for protein folding tasks [2].

In many cases, machine learning is used for prediction purposes [3]. An example would be the likelihood of a patient developing cancer. We have many large databases, suitable for such learning. The Cancer Genome Atlas (TCGA) [4] is a cancer genomics programme that contains more than 20,000 molecularly characterised samples from 33 cancer types. The Gene Expression Omnibus (GEO) database [5] is an international public repository that contains microarray, next-generation sequencing, and other forms of community-submitted genomics data. The GEO database contains approximately 5.4 million samples which are organized into more than 4,000 data sets. But we also have many smaller data sets where we have a few instances and a large number of attributes. The GEO database has about 1800 data sets with less than 10 instances.

The problem we propose to solve in this thesis is learning from small gene expression data sets and developing accurate models that produce meaningful gene encodings for phenotype prediction. Learning on smaller data sets is difficult due to the small number of samples, but by integrating them, the models have a higher chance of learning a good latent representation and meaningfully embedding the data into the same concept space. Using transfer learning to model smaller data sets could result in accurate and more intuitive models based on the encodings.

3.2 Related work

Autoencoders [6] are neural networks that are used to produce efficient embeddings of given data. They consist of two models: an encoder, which encodes the data into an embedding, and a decoder, which can decode the embedding. They have been successfully used for anomaly detection [7], image denoising [8], and dimensionality reduction. If we are only interested in embeddings, a better representation can be achieved by using multi-task learning, which trains the embeddings for direct use in a task prediction [9].

Transfer learning [10] is a machine learning problem where knowledge is gained in one domain and applied to another, most often by using it for feature representation. It is a popular approach to use pre-trained models as a starting point, especially in image classification [11] and natural language processing [12], where there are huge amounts of data.

Rao et al. in 2019 [13] introduced tasks assessing protein embeddings (TAPE), a set of 5 biologically relevant semi-supervised learning tasks used to assess protein embeddings. Their model of learned protein representation was later used by Ekvall et al. in 2022 [14], where they predicted mass spectrometry intensities of different proteins. They used a transformer architecture and achieved better results than deep models, trained on raw data, which is encouraging.

Chen et al. in 2022 [15] used transfer learning to predict cancer drug responses to RNA sequences. They used denoising autoencoders to learn a compact representation of gene expressions in specific cells and to learn the correlation between drug responses and gene expressions. Another autoencoder was trained to extract features from single-cell RNA sequences. Transfer learning was performed using another deep model to adapt the extracted gene features to have similar distributions. The final prediction was made by taking a single-cell RNA sequence along with the adapted features and returning whether a certain cell is sensitive or resistant to a particular drug.

Hanczar et al. in 2022 [16] used transfer learning to predict different types of cancer from two different types of data: gene expressions, compiled by Torrente et al. [17], and RNA sequences from the TCGA. They used both supervised and unsupervised learning techniques. They focused on evaluating performance based on model hyperparameters, which they found to have very little effect on the final results. Their best models performed well on unseen data.

3.3 Expected contributions

We will explore whether transfer learning can be used to improve learning on a collection of small gene expression data sets and whether meaningful encodings can be produced. The contribution of the proposed thesis will be a database of small gene expression data sets, a framework for training and testing models, and a framework for qualitative evaluation of the methods used.

3.4 Methodology

We propose to implement the following work plan:

Collecting relevant data sets. Our target is at least 50 data sets. The number of samples per data set would ideally be around 50 as we have found that it is difficult to improve the results on very small data sets.

Data preprocessing and cleaning. We will inspect the data and check that their attributes are comparable between the data sets. We expect the data sets to have different attributes and we will only focus on 978 landmark genes described in a profiling method called L1000 [18]. We will use these because the authors have shown that they are suitable for inferring the expression levels of 81% of other genes.

Developing a transfer learning framework. The framework will integrate the data sets and train an autoencoder. Training will be done on larger data sets to infer the autoencoder which can capture cross-domain patterns, improve accuracy and produce good embeddings on smaller data sets.

Testing the encodings. We will test the encodings of smaller data sets by training a simple classification model such as logistic regression and using the log loss and the area under the ROC curve metric. We will compare the predictive power of the encodings to the raw data and expect to get comparable results. We plan to investigate how the size of the encoding space affects these results.

Qualitative assessment of encodings. We will examine the encodings produced and check whether they can discriminate between different classes when a clustering algorithm is applied to them and the groups are visualised.

3.5 References

[1] P. Larrañaga, B. Calvo, R. Santana, et. al., Machine learning in bioinformatics, Briefings in Bioinformatics 7 (1) (2006) 86–112.

[2] J. Jumper, R. Evans, A. Pritzel, et. al., Highly accurate protein structure prediction with alphafold, Nature 596 (7873) (2021) 583–589.

[3] R. E. Soria-Guerra, R. Nieto-Gomez, D. O. Govea-Alonso, et. al., An overview of bioinformatics tools for epitope prediction: implications on vaccine development, Journal of biomedical informatics 53 (2015) 405–414.

[4] The cancer genome atlas program (2006).
URL `https://www.cancer.gov/tcga`

[5] R. Edgar, M. Domrachev, A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Research 30 (1) (2002) 207–210.

[6] R. S. Zemel, G. Hinton, Autoencoders, minimum description length and helmholtz free energy, in: Proceedings of the Neural Information Processing Systems, Citeseer, 1994.

[7] C. Zhou, R. C. Paffenroth, Anomaly detection with robust deep autoencoders, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 665–674.

[8] L. Gondara, Medical image denoising using convolutional denoising autoencoders, in: 2016 IEEE 16th international conference on data mining workshops (ICDMW), IEEE, 2016, pp. 241–246.

[9] Z. Huo, D. Shen, H. Huang, Genotype-phenotype association study via new multi-task learning model, in: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium, World Scientific, 2018, pp. 353–364.

[10] S. Bozinovski, A. Fulgosi, The influence of pattern similarity and transfer learning upon training of a base perceptron b2, in: Proceedings of Symposium Informatica, Vol. 3, 1976, pp. 121–126.

[11] M. Hussain, J. J. Bird, D. R. Faria, A study on cnn transfer learning for image classification, in: UK Workshop on computational Intelligence, Springer, 2018, pp. 191–202.

[12] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, arXiv preprint arXiv:1801.06146 (2018).

[13] R. Rao, N. Bhattacharya, N. Thomas, et. al., Evaluating Protein Transfer Learning with TAPE, Curran Associates Inc., Red Hook, NY, USA, 2019.

[14] M. Ekvall, P. Truong, W. Gabriel, et. al., Prosit transformer: A transformer for prediction of ms2 spectrum intensities, Journal of Proteome Research 21 (5) (2022) 1359–1364, pMID: 35413196.

[15] J. Chen, X. Wang, A. Ma, et. al., Deep transfer learning of cancer drug responses by integrating bulk and single-cell rna-seq data, Nature Communications 13 (1) (2022) 6494.

[16] B. Hanczar, V. Bourgeais, F. Zehraoui, Assessment of deep learning and transfer learning for cancer prediction based on gene expression data, BMC bioinformatics 23 (1) (2022) 262.

[17] A. Torrente, M. Lukk, V. Xue, H. Parkinson, J. Rung, A. Brazma, Identification of cancer related genes using a comprehensive map of human gene expression, PloS one 11 (6) (2016) e0157484.

[18] A. Subramanian, R. Narayan, S. M. Corsello, et al., A next generation connectivity map: L1000 platform and the first 1,000,000 profiles, Cell 171 (6) (2017) 1437–1452.e17.