Final Report

of the

“BioMining: Data Mining for Biomedical Problems”

(National Research, Development and Innovation Office, NKFIH PD 111710)

project

Krisztian Buza, PhD
principal investigator

8th September 2017

1 Executive Summary

The BioMining project focused on data mining for biomedical tasks. In particular, we envisioned to develop new hubness-aware machine learning techniques for challenging biomedical tasks including the classification of gene expression data and biomedical signals as well as drug-target interaction prediction.

All tasks of the research plan were implemented appropriately. Regarding the outcome of the project, we point out the followings:

(i) In terms of scientific output, the project clearly exceeded the expectations: while 3 journal submissions and the publication of 3 conference papers were initially anticipated, by the end of the project, **9 articles have been published in journals indexed by Web of Science**, including Knowledge-Based Systems, Neurocomputing and Frontiers in Neuroscience. Additionally, **8 conference or workshop papers** were presented at international conferences or workshops with proceedings published by Springer, IEEE or within the CEUR Workshop Proceedings series.

(ii) We recorded **16 short videos** with a total length of approx. 100 minutes. These short videos are organized into two playlists: one of them is an online lecture about hubness-aware machine learning[1], while the other one summarizes key achievements[2] of the project.

(iii) We developed the **PyHubs software library** and published further software codes on the BioIntelligence web page (http://www.biointelligence.hu) together with other results of the project.

[1] http://www.biointelligence.hu/course.html

[2] https://www.youtube.com/playlist?list=PLNWnqkAEYZk1ENQAcydQMHdgQ_KV36-oU

2 Summary of Scientific Contributions

In the BioMining project, we envisioned to address biomedical problems by hubness-aware machine learning techniques. In this section, we describe our key contributions and refer to the publications that contain the details. In Section 3, we describe how the aforementioned scientific contributions address the goals and tasks defined in the research plan.

2.1 General contributions to hubness-aware machine learning

At the beginning of the project, with Nenad Tomasev, we studied performance of hubness-aware classifiers in case of label noise [1]. As reliable class labels may be difficult to obtain, this is a particularly relevant situation for the classification of biomedical data. Furthermore, we studied how to combine hubness-aware classifiers with instance selection in order to speed up classification and maintain high accuracy simultaneously [2].

Hubness-aware regressors, such as nearest neighbor regression with error-based weighting, error correction and the combination of these techniques, EW*k*NN, EC*k*NN and EWC*k*NN for short, have been proposed in the principal investigator's first-authored publication [3]. These regressors have been evaluated on real-world datasets from various domains. Additionally, their robustness to label noise has been examined in detail. As the presence of hubs is known to be related to the *intrinsic* dimensionality of the data, we discussed our results in the light of the intrinsic dimensionality of the datasets used in the experiments.

2.2 Link Prediction in Biomedical Networks and its Applications to Drug–target interaction prediction

As drug-target interaction prediction is one of the most prominent applications of link prediction in the pharmaceutical domain, we proposed to use our hubness-aware regressor, EC*k*NN as local model in bipartite local models (BLMs) [4,5]. Additionally, we proposed an enhanced representation of drugs and targets in a multi-modal similarity space. The modalities include external similarities and the similarities calculated based on the known drug–target interactions. Furthermore, we built an ensemble by selecting subsets of features from the enhanced similarity-based representation of drugs and targets. We performed experiments on publicly available real-world drug-target interaction datasets according to standard evaluation protocols. The results show that our approach outperforms state-of-the-art drug-target prediction techniques. We examined the effect of hubness-aware error correction and the ensemble size, and found that both of these techniques are essential for high accuracy. Furthermore, we showed that our approach is able to predict chemically validated new interactions and therefore we hope that it will contribute to design, development and repositioning of drugs.

It is interesting to note that the formal definition of the drug-target interaction prediction task is similar to that of tasks considered in the recommender systems community. Therefore, we considered one of the most prominent recommender algorithms, Bayesian Personalized Ranking and extended it to allow for drug-centric predictions that are tailored towards drug-repositioning scenarios [6].

2.3 Semi-supervised classification of gene expression data

In cases, when it is difficult to obtain labelled data (e.g. in case of rare diseases), semi-supervised models may be desired as they aim to learn both from labelled and unlabelled data. Therefore, the principal investigator proposed a semi-supervised extension of a hubness-aware classifier [7,8]. This approach is called Semi-Supervised Naive Hubness-Bayesian *k*-Nearest Neighbor or SSNHBNN for short. SSNHBNN was evaluated on gene expression data related to the diagnosis of various cancer types, such as breast cancer, colon cancer and lung cancer, it was also compared with other methods from the literature: the experiments show that SSNHBNN outperforms state-of-the-art classifiers.

2.4 Classification of biomedical time series

As electroencephalography is one of the most wide-spread approaches to capture brain activity, we examined the performance of hubness-aware classifiers in case of electroencephalograph signals [9]. Furthermore, we developed a projection-based classification technique and applied it to the classification of electroencephalograph signals [10]. We tested the performance of our classifier in various situations, such as the recognition of a disease and the recognition of the stimulus. We also examined the robustness of the approach: in order to do that, we trained the classifier on the signals that were obtained from healthy subjects, and tested the classifier's performance on signals of persons affected by a disease. Our projection-based approach worked well in this setting as well.

Both in case of hubness-aware classifiers, as well as in case of the projection-based approach, we used dynamic time warping (DTW) as distance measure between time series. We also examined how well DTW-based models work on functional magnetic resonance imaging (fMRI) data describing brain activity, and it turned out that DTW may be a proper measure of functional connectivity between brain regions, i.e., DTW may outperform conventional correlation-based functional connectivity in many applications, such as classification of the data according to diseases or other conditions. The resulting models are interpretable by domain experts, they are in accordance with biological knowledge, and they may give us new insights about how the brain works [11, 12, 13].

One of the most wide-spread brain diseases is the Parkinson’s disease. The UPDRS score can be used to assess how sever is the disease in case of a particular patient. There were previous attempts to estimate UPDRS score from speech data, and, if the accuracy was sufficiently high, the estimation of UPDRS scores could be integrated into smartphones and tables: in principle, while the patient makes telephone or skype calls, his or her UPDRS score could be estimated and this would allow to monitor the patient's UPDRS score regularly. As neural networks are known to be universal function approximators and hubness-aware models haven't been used for the task of UPDRS score estimation previously, we decided to examine the performance of neural networks coupled with hubness-aware weighting, and we obtained promising results [14].

Another type of simply observable biometric time-series describe the dynamics of typing which is known to be characteristic to persons, and has been proposed as a biometric for person identification. According to our results, hubness-aware regressors developed in the first year of the project [3] may be used to person identification as well [15].

2.5 Further resources generated by the project: videos, software prototypes, demo

Besides task-specific scripts, we published the PyHubs software package[3] that contains a Python-based implementation of hubness-aware classifiers. We note that PyHubs is complementary to the HubMiner software package both in terms of technology (Python vs. Java) and the implemented algorithms (according to our best knowledge, hubness-aware regression techniques are only available in our PyHubs package, not yet in the HubMiner

[3] http://www.biointelligence.hu/pyhubs/

package). With PyHubs we also aim to extend the hubness-aware machine learning community towards Python-users. For the sake of completeness, PyHubs contains implementations of hubness-aware classifiers and basic methods used to perform machine learning experiments.

We recorded an online lecture[4] about hubness-aware machine learning and prepared a self-check quiz. The lecture consists of 12 short videos focusing on various subtopics. These videos are organized into a youtube playlist, the total length of which is roughly an hour.

We demonstrated user authentication based on our approach for user identification using keystroke dynamics at the forum of the 28th International Conference on Advanced Information Systems Engineering where dozens of conference participants tried out our method.

Last, but not least, we mention that, in order to ensure visibility of our results, we prepared a youtube playlist[5] summarizing the core contributions of the project.

3 Report about the Implementation of the Tasks of the Research Plan

All tasks of the research plan have been implemented and the results have been made publicly available. Next, we describe in detail, how each task was implemented.

In order to address Task 1.1, we developed the SSNHBNN semi-supervised hubness-aware classification technique and applied it to gene expression data. We prepared a software prototype implementing the algorithm and evaluated its performance. According to Task 1.2, we reported the results at the 9th International Conference on Computer Recognition Systems [7]. We extended the work by more detailed performance analysis and published it in Computer Methods and Programs in Biomedicine [2] which fulfils Task 1.3.

The study of biomedically relevant aspects of hubness-aware machine learning techniques [1,2] is also in accordance with the goals of Task 1.1 and Task 1.3. This was performed in collaboration with researchers outside of the project as it was envisioned in WP4.

According to Task 4.1, we monitored the work of other researchers and observed that a Java-based implementation of hubness-aware machine learning algorithms has been made publicly available meanwhile (in particular, the HubMiner software package). Therefore, we decided to prepare a Python-based implementation as Python is one of the most popular and powerful programming languages of data science. In order to fulfil Task 1.4, Task 2.4 and Task 3.4, we made PyHubs publicly available[6] and we extended it throughout the lifetime of the project. Additionally, we published task-specific scripts related to our publications, such as the codes[7] implementing cross-validated LASSO which can be applied for the classification of high-dimensional data such as the brain imaging data we worked with.

[4] http://www.biointelligence.hu/course.html
[5] https://www.youtube.com/playlist?list=PLNWnqkAEYZk1ENQAcydQMHdgQ_KV36-oU
[6] http://www.biointelligence.hu/pyhubs/
[7] https://github.com/MRegina/crossvalidated_lasso

According to Task 1.5, we prepared an online lecture about hubness-aware machine learning and made it publicly available[8].

As envisioned in Task 2.1 and Task 2.2, we developed a projection-based classification technique for biomedical time-series and presented it at the 14th International Conference on Artificial Intelligence and Soft Computing [10]. For person identification based on biometric time-series (such as keystroke dynamics), we proposed a "classification-via-regression"-approach and presented it at the 11th IEEE International Symposium on Applied Computational Intelligence and Informatics [15].

In accordance with Task 2.3, we performed a detailed analysis of hubness-aware classification techniques in context of EEG, and published the results in a journal paper [9]. Furthermore, as speech can be seen as simply-observable biometric time-series based on which the severity of Parkinsons Disease (the UPDRS score) can be estimated, we proposed hubness-aware neural networks for the estimation of UPDRS score and published the results in a journal paper [14] in accordance with Task 2.3.

In accordance with the goals of WP2 and WP4, we worked together with researchers of the Brain Imaging Center of the Center for Natural Sciences of the Hungarian Academy of Sciences on classification of brain imaging data which can be considered as high-dimensional time series [11, 12, 13].

We developed a hubness-aware approach for link prediction in biomedical networks and applied it to drug–target interaction prediction according to Task 3.1, published the initial results at the 11th IEEE International Symposium on Applied Computational Intelligence and Informatics [4] in order to fulfil Task 3.2, and published the enhanced version of the method in Neurocomputing [5] according to Task 3.3. The publication of our Bayesian link prediction approach [6] and its application to drug–target interaction prediction is also in accordance with the Task 3.3.

Last, but not least, we mention that the playlist summarizing the key contributions of the project fulfils Task 4.2.

4 Cost-efficient Implementation of the Project

Additionally, we note that the European Union kindly supported the 9th International Conference on Computer Recognition Systems as well as the 3rd European Network Intelligence Conference, and therefore the authors of accepted papers did not need to pay registration fees. Furthermore, the principal investigator was an invited speaker of the 9th Japanese-Hungarian Symposium on Discrete Mathematics and Its Applications, and therefore much of the costs related to the participation at this event were covered by the Japanese organizers, while the travel costs were kindly covered by the project OTKA 108947 K. These favourable circumstances allowed us to increase the number of conference participations compared to what was originally envisioned in the project proposal.

[8] http://www.biointelligence.hu/course.html

References

[1] N. Tomasev, K. Buza (2015): Hubness-aware kNN Classification of High-dimensional Data in Presence of Label Noise, [Preprint] [Audio Slides] in Neurocomputing, Vol. 160, pp. 157-172

[2] N. Tomasev, K. Buza, D. Mladenic (2016): Correcting the Hub Occurrence Prediction Bias in Many Dimensions, Computer Science and Information Systems, Vol. 13, pp. 1-21

[3] K. Buza, A. Nanopoulos, G. Nagy (2015): Nearest Neighbor Regression in the Presence of Bad Hubs [Preprint] [Audio Slides], Knowledge-Based Systems, Vol. 86, pp. 250-260

[4] K. Buza (2016): Drug-Target Interaction Prediction with Hubness-aware Machine Learning, 11th IEEE International Symposium on Applied Computational Intelligence and Informatics, pp. 437-440

[5] K. Buza, L. Peska (2017): Drug–target interaction prediction with Bipartite Local Models and hubness-aware regression [Preprint] [Audio Slides], Neurocomputing, Vol. 260, pp. 284-293

[6] L. Peska, K. Buza, J. Koller: Drug-Target Interaction Prediction: a Bayesian Ranking Approach, to appear in Computer Methods and Programs in Biomedicine

[7] K. Buza (2015): Semi-supervised Naive Hubness-Bayesian *k*-Nearest Neighbor for Gene Expression Data, Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015, pp. 101-110, Springer.

[8] K. Buza (2016): Classification of Gene Expression Data: A Hubness-aware Semi-Supervised Approach [Preprint] [Audio Slides], Computer Methods and Programs in Biomedicine, Vol. 127, pp. 105-113

[9] K. Buza, J. Koller (2016): Classification of Electroencephalograph Data: A Hubness-aware Approach, Acta Polytechnica Hungarica, Vol. 13, pp. 27-46

[10] K. Buza, J. Koller, K. Marussy (2015): PROCESS: Projection-Based Classification of Electroencephalograph Signals [paper] [poster], Proceedings of the 14th International Conference on Artificial Intelligence and Soft Computing, Lecture Notes in Computer Science, Vol. 9120, pp. 91-100, Springer.

[11] Regina J. Meszlényi, Petra Hermann, Krisztian Buza, Viktor Gál, Zoltán Vidnyánszky (2017): Resting State fMRI Functional Connectivity Analysis Using Dynamic Time Warping, Frontiers in Neuroscience, Vol. 11, Article 75

[12] Regina Meszlényi, Ladislav Peska, Viktor Gal, Zoltán Vidnyánszky, Krisztian Buza (2016): A model for classification based on the functional connectivity pattern dynamics of the brain, The Third European Network Intelligence Conference (ENIC 2016), pp. 203-208

[13] Regina Meszlényi, Ladislav Peska, Viktor Gal, Zoltán Vidnyánszky, Krisztian Buza (2016): Classification of fMRI data using Dynamic Time Warping based functional connectivity analysis, 24th European Signal Processing Conference, pp. 245-249

[14] K. Buza, N.Á. Varga (2016): ParkinsoNET: Estimation of UPDRS Score using Hubness-aware Feed-Forward Neural Networks, Applied Artificial Intelligence, Vol. 30, pp. 541-555

[15] K. Buza, D. Neubrandt (2016): How You Type Is Who You Are, 11th IEEE International Symposium on Applied Computational Intelligence and Informatics, pp. 453-456