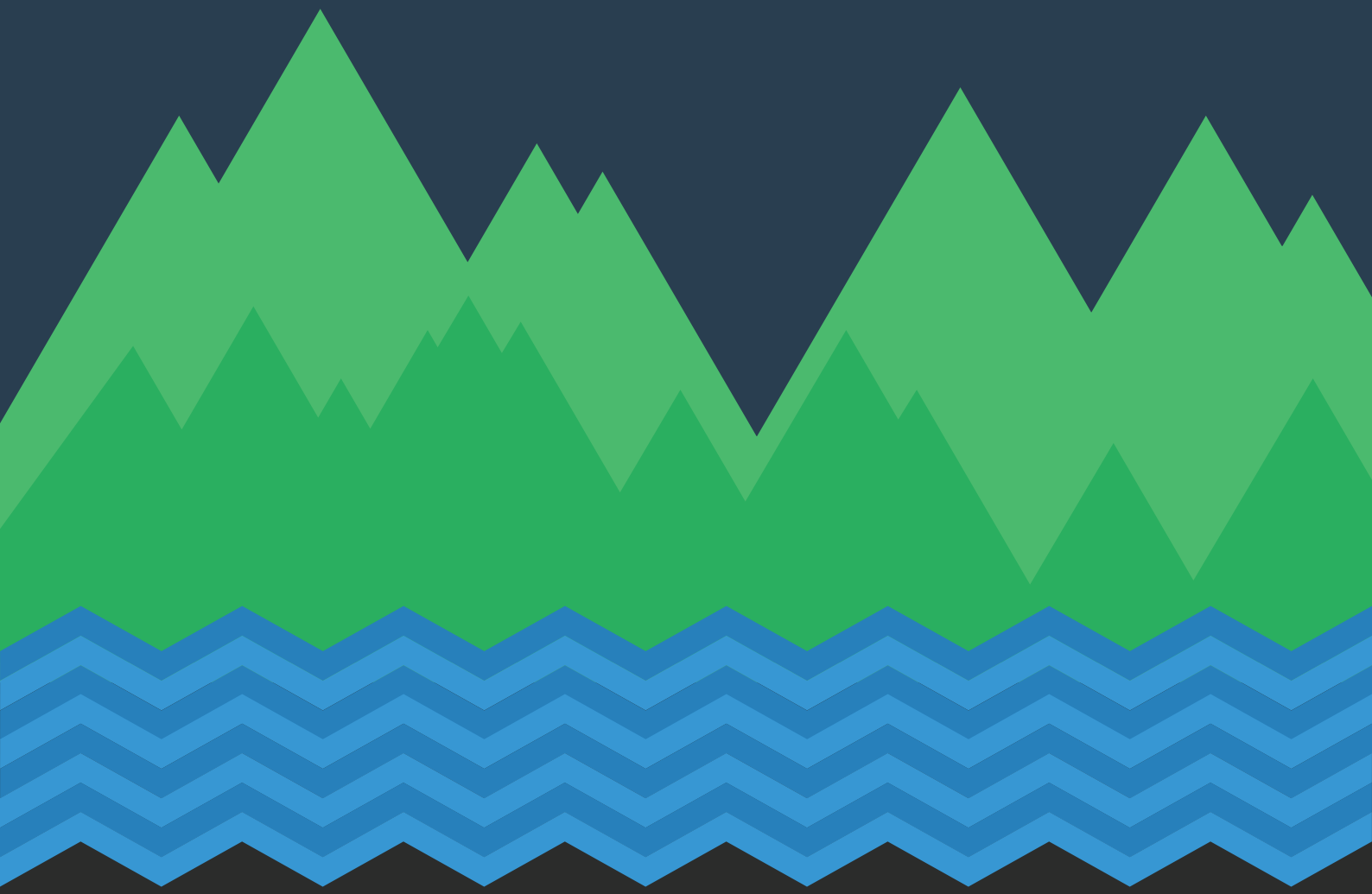


BOOK OF ABSTRACTS

DISCOVERY SCIENCE

08 - 10.10.2014
BLED, SLOVENIA

List of Late Breaking Papers accepted for poster presentation
at the **17th *International Conference on Discovery Science* (DS 2014)**

Tal Galili - *dendextend: an R package for scientific visualization of dendograms and hierarchical clustering*

Klemen Kenda, Luka Stopar and Marko Grobelnik - *Multi-level approach to sensor streams analysis*

Dragi Kocev, Sašo Džeroski, Ivica Dimitrovski, Michelangelo Ceci, Tomislav Šmuc and Joao Gama - *MAESTRA: Learning from Massive, Incompletely annotated, and Structured Data*

Mitja Luštrek and Maja Somrak - *Mining telemonitoring data from congestive-heart-failure patients*

Mariana Oliveira and Luis Torgo - *Ensembles for Time Series Forecasting*

Aljaz Osojnik and Sašo Džeroski - *Modeling Dynamical Systems with Data Stream Mining*

Apurva Pathak, Bidyut Kr. Patra, Ville Ollikainen and Raimo Launonen - *Clustering based approach for balancing accuracy and diversity in collaborative filtering*

Matic Perovšek, Nada Lavrač and Bojan Cestnik - *Bridging term discovery for cross-domain literature mining*

Marko Robnik-Šikonja - *Generator of unsupervised semi-artificial data*

Laszlo Szathmary, Petko Valtchev, Amedeo Napoli, Marton Ispany, and Robert Godin - *CGT: a vertical miner for frequent equivalence classes of itemsets*

Aneta Trajanov, Vladimir Kuzmanovski, Florence Leprince, Benoit Real, Alain Dutertre, Julie Maillet-Mezeray, Sašo Džeroski and Marko Debeljak - *Studying the drainage periods for agricultural fields with data mining: La Jaillière case study*

Takeaki Uno and Yushi Uno - *Mining Graph Structures Preserved Long Period*

Viivi Uurtio, Juho Rousu, Malin Bomberg and Merja Itävaara - *Extracting Sparse Canonical Correlations Between Microbial Communities and Deep Groundwater Geochemistry*

Anita Valmarska and Janez Demšar - *Analysis of citation networks*

Nina Vidmar, Nikola Simidjievski and Sašo Džeroski - *Predictive process-based modeling of aquatic ecosystems*

Denny Verbeeck and Hendrik Blockeel - *iMauve: A Fast Incremental Model Tree Learner*

Vedrana Vidulin, Tomislav Šmuc and Fran Supek - *Speed and Accuracy Benchmarks of Large-Scale Microbial Gene Function Prediction with Supervised Machine Learning*

Martin Žnidaršič, Senja Pollak, Dragana Miljković, Janez Kranjc and Nada Lavrač - *Identifying creative fictional ideas*

dendextend: an R package for scientific visualization of dendograms and hierarchical clustering

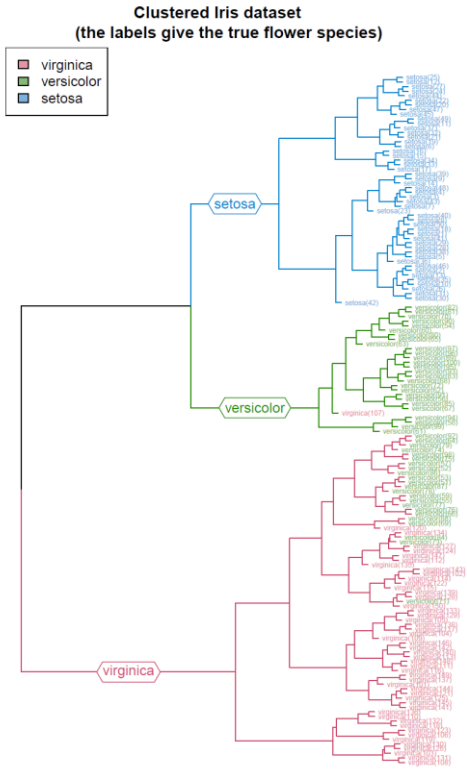
Tal Galili [1,*]

1. Tel Aviv University
*Contact author:Tal.Galili@gmail.com

Keywords: dendrogram, hierarchical clustering, hclust, visualization, tanglegram

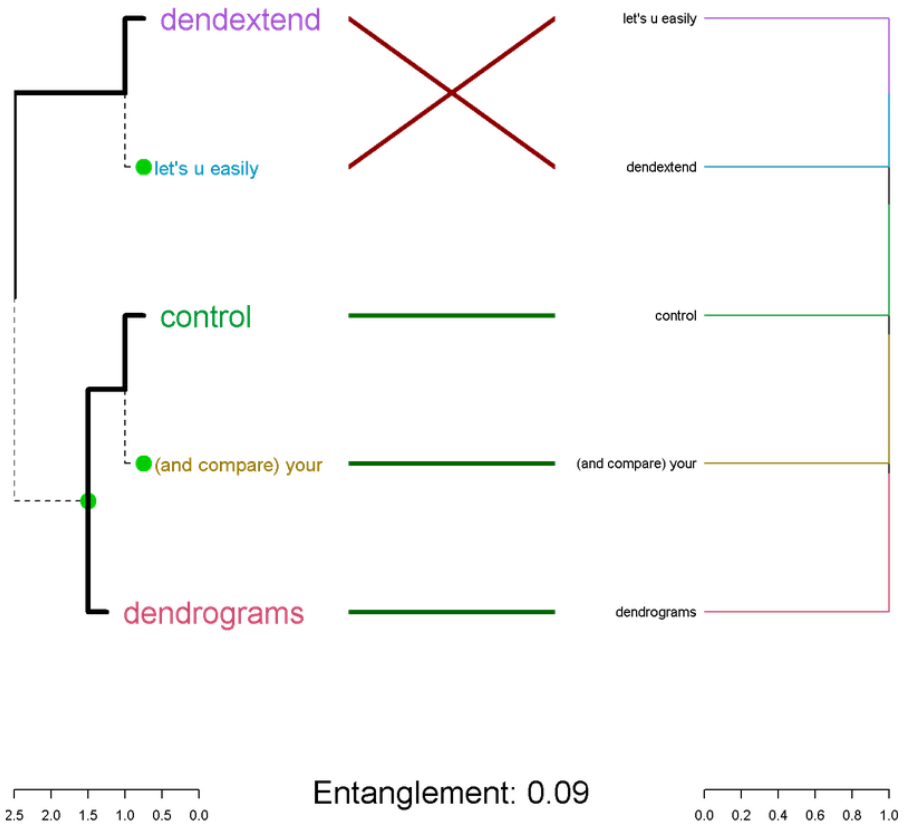
This poster introduces the ***dendextend*** package [1] for extending the palette of functions and methods for the `dendrogram` class in the R statistical environment.

A dendrogram is a tree diagram which is often used to visualize a hierarchical clustering of items. Dendrograms are used in many disciplines, ranging from Phylogenetic Trees in computational biology to Lexomic Trees in text analysis. Hierarchical clustering in *R* is commonly performed using the `hclust` function. When a more sophisticated visualization is desired, the `hclust` object is often coerced into a `dendrogram` object, which in turn is modified and plotted. While **base** *R* comes with several very useful methods for manipulating the `dendrogram` object (namely: `plot`, `print`, `[[`, `labels`, `as.hclust`, `cophenetic`, `reorder`, `cut`, `merge`, `rev`, and `str`), still - the current palette of functions leaves a lot to be desired.



The ***dendextend*** R package offers functions and methods for `dendrogram` class objects in R, allowing for easier manipulation of a dendrogram's shape (via `rotate`, `prune`), color and content (via functions such as `set`, `labels_colors`, `color_branches`, `etc.` function). The package also provides S3 methods for functions such as `labels<-`, `cutree`, and more. ***dendextend*** also provides the tools for comparing the similarity of two dendrograms to one another either graphically using a `tanglegram` plot, or statistically with association measures ranging from `cor_cophenetic` to `Bk_plot`, while enabling bootstrap and permutation tests for comparing the trees.

Since tree structure often requires the use of recursion, which can be slow in *R*, some of the more computationally intensive aspects of the **dendextend** package can be handled with its sister package, **dendextendRcpp** [2], which overrides several basic functions (namely: `cut_lower_fun`, `heights_per_k.dendrogram`, `labels.dendrogram`), with their *C++* implementation.



References

[1] Tal Galili (2014). dendextend: Extending R's dendrogram functionality, http://cran.r-project.org/web/packages/dendextend

[2] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

[3] Tal Galili (2014). dendextendRcpp: Faster dendrogram manipulation using Rcpp, http://cran.r-project.org/web/packages/dendextendRcpp

MULTI-LEVEL APPROACH TO SENSOR STREAMS ANALYSIS

Klemen Kenda, Luka Stopar, Marko Grobelnik
Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: klemen.kenda@ijs.si

1 INTRODUCTION

Monitoring of the systems, which are described with numerous time series, can be a complex task. Too much data is difficult to follow even by an expert human user. In presented work we focused on understanding dynamics of such complex systems and presenting the results in a humanly-comprehensible way.

Instead of following the dynamics of a system through numerous time series, the result of our methodology is a directed state graph (see Figure 1) equipped with corresponding transitional probabilities. To achieve such a result we extract different features (markers) from the time series, aggregate them in a sliding window and pack them into state vectors. We perform clustering on top of a set of such vectors and calculate transitional probabilities between the clusters.

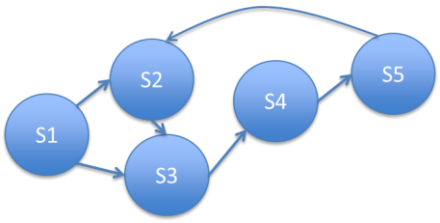


Figure 1: Internal states are clustered, transitional probabilities are calculated.

Typical states (centroids) are identified by domain experts. Such knowledge base can be later used for anomaly detection, root cause analysis and other tasks.

2 METHODOLOGY

Each system is observed by multiple sensors that output regular measurements forming multiple time series. In an energy forecasting setting, for example, such sensors could be sensors measuring weather conditions (temperature, pressure, wind direction, wind speed, cloud cover, humidity), sensors measuring energy demand and consumption and virtual sensors »measuring« day of week, time of day, working hours, holidays, sunrise, sunset and other relevant features

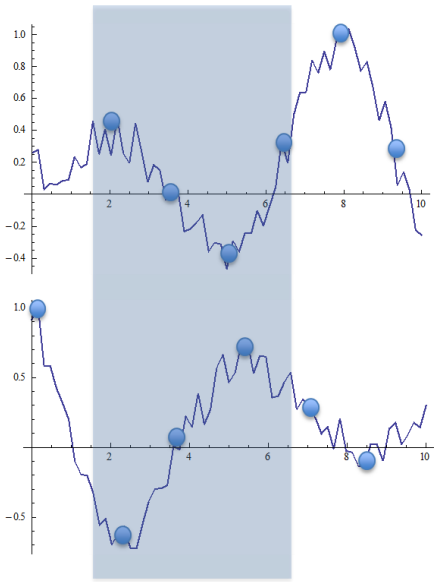


Figure 2: Time series are enriched with the markers, which are aggregated within the sliding window.

Each time series could have different marker extractors attached. These marker extractors give us certain aggregated information of the time series in a specified time window. Examples of typical markers are: local minimum and maximum, average value, maximum derivative, value threshold markers (is value in a certain range) etc.

Next, we count appearance of the markers within a certain time window (which is implemented as a sliding window) as depicted in Figure 2. Values and/or counts of the markers are collected in a vector

of features and those are collected for different times (see Figure 3). Clustering is performed on a minimal set of such feature vectors and initial transitional probabilities are calculated. Updating of the transitional probabilities and clusters is later done in an on-line manner.

Time	F_1	F_2	F_3	F_4	F_5	F_6
t_0	*	*	*	*		*
t_1	*	*		*		
...						
t_n	...	...	...	...	...	...

Figure 3: For each sliding window aggregated features represent internal state of the system.

Graph of the system offers human-managable view on the system and gives us a deeper understanding of the dynamics of the system. Many different upgrades are possible on top of such an approach.

3 POSSIBLE APPLICATIONS

The most obvious application could be **anomaly detection.** In case a new state of the system does not fit in any of the clusters, this might be a signal for an anomaly. Anomaly could represent another yet undiscovered typical state of the system and therefore the model would need updating or it could be a real anomaly.

Next logical step – as we do have a directed graph and history of transitions – would include **root cause analysis.** When an anomaly occurs we can trace back and identify the path which lead to the anomaly. Furthermore we could detect typical paths in the system that lead to such anomalies and perform **proactive anomaly detection.** This means that we could detect anomalies before they would happen, based on a subsequence path matching.

Another application might be in usage of the clustering for modelling, where we would use different models for different clusters. Thermal plant demand for example differs significantly between heating season and summer.

4 EARLY RESULTS, CONCLUSIONS AND FUTURE WORK

The idea was tested on a dataset of a thermal plant in Reggio nell'Emilia, Italy. Features were extracted from daily thermal production profiles. Clustering identified typical states of the system, annotated by domain experts, such as: early heating season, transitional season (spring, autumn) or high heating season with a variety of substates.

Early results are promising and can be viewed as a proof of concept. More complex systems need to be analyzed. On the other hand other clustering algorithms need to be tested and more markers (extractors) need to be introduced and tested.

Future work includes also development of interactive GUI, which would make it easy for a domain expert to explore and identify clusters at different levels. There is also a vast field of possible applications described in Section 3 that need to be addressed.

This work was supported by the ICT Programme of the EC under NRG4Cast (FP7-ICT-600074) and Proasense (FP7-ICT-612329).

MAESTRA: Learning from Massive, Incompletely annotated, and Structured Data

Dragi Kocev[1], Sašo Džeroski[1], Ivica Dimitrovski[2], Michelangelo Ceci[3], Tomislav Šmuc[4] and Joao Gama[5]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

[2] Faculty of Computer Science and Engineering, Universitz Ss Cyril and Methodius, Skopje, Macedonia

[3] Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, Bari, Italy

[4] Division of Electronics, Ruđer Bošković Institute, Zagreb, Croatia

[5] INESC Technology and Science – INESC TEC, Porto, Portugal

The need for machine learning (ML) and data mining (DM) is ever growing due to the increased pervasiveness of data analysis tasks in almost every area of life, including business, science and technology. Not only is the pervasiveness of data analysis tasks increasing, but so is their complexity. We are increasingly often facing predictive modelling tasks involving one or several of the following complexity aspects: (a) structured data as input or output of the prediction process, (b) very large/massive datasets, with many examples and/or many input/output dimensions, where data may be streaming at high rates, (c) incompletely/partially labelled data, and (d)data placed in a spatio-temporal or network context. Each of these is a major challenge to current ML/DM approaches and is the central topic of active research in areas such as structured-output prediction, mining data streams, semi-supervised learning, and mining network data. The simultaneous presence of several of them is a much harder, currently insurmountable, challenge and severely limits the applicability of ML/DM approaches.

The project will develop predictive modelling methods capable of simultaneously addressing several (ultimately all) of the above complexity aspects. In the most complex case, the methods would be able to address massive sets of network data incompletely labelled with structured outputs. We will develop the foundations (basic concepts and notions) for and the methodology (design and implementation of algorithms) of such approaches. We will demonstrate the potential and utility of the methods on showcase problems from a diverse set of application areas (molecular biology, sensor networks, mutimedia, and social networks). Some of these applications, such as relating the composition of microbiota to human health and the design of social media aggregators, have the potential of transformational impact on important aspects of society, such as personalized medicine and social media.

MINING TELEMONITORING DATA FROM CONGESTIVE-HEART-FAILURE PATIENTS

Mitja Luštrek[1,2], Maja Somrak[1,2]
[1] Jožef Stefan Institute, Department of Intelligent Systems
[2] Jožef Stefan International Postgraduate School
{mitja.lustrek, maja.somrak}@ijs.si

Abstract. The technology is providing us with increasingly more possibilities for telemonitoring patients, yet it is not clearly obvious how to utilize the obtained data. This paper describes the mining of telemonitoring data of congestive heart failure (CHF) patients, with the purpose to reveal potentially unknown relations between the monitored parameters and patient's overall feeling. The resulting models correlate monitored parameters with feeling of good or bad health, consistent with current medical knowledge, as well as confirming the opinion of some cardiologists, for which there has been less evidence.

1 Study overview

Many studies [1] researched the benefit of telemonitoring of patients, but the results of these studies seem to be contradictory. However, since the methods used in these trials were not particularly advanced, the use of intelligent computer methods on the gathered monitored data could help to reveal previously unknown relations in the data.

This paper describes the mining of telemonitoring data of congestive heart failure (CHF) patients. The data was collected in the Chiron[1] project with the purpose to reveal relations between the monitored parameters and patient's overall feeling of health. The research included 25 CHF patients, who produced a total of 1068 recording days. The 49 static parameters were measured at the beginning of the study (e.g. age, BMI, cholesterol, etc.). The 15 dynamic parameters [2] include environmental variables and measurements of vital signs, which were extensively monitored with on-body sensors (ECG and accelerometers). All dynamic parameters were further averaged (*avg*) for each day and changes (*chg*, relative change since the previous day) of values were calculated. The calculations were made for each type of activity in the day ('lying', 'sitting', 'moving' and 'all' activities) and the health risk of the patient was estimated [3]. A mobile application was used for daily reporting of patient's feeling (e.g. 'feeling better' than yesterday), the parameter which we aim to predict with generated models.

[1] http://www.chiron-project.eu/

2 Results

Feature and algorithm selection: We compared subsets of dynamic all_act features, finding that avg and avg + chg performed better than the rest. While the combination with per _act features reduced accuracy, the extension of the subset with static features performed best of all. We also tested various automatic feature selection methods from Weka[2]. The J48 algorithm was selected for its comprehensibility and at accuracy of 76.9% it was outperformed by the best algorithm for less than 5 percentage points.
Interesting models: Two examples are presented in Figure 1. They show that a high heart rate (HR_avg_all_activities), short QRS interval (QRS_avg_all_activities), high systolic blood pressure (SBP) and low diastolic blood pressure (DBP) are all associated with the feeling of good health, which coincides with existing medical knowledge. Increased weight (DRWChg) – signifying excessive fluid retention – and oxygen saturation below 97 % are associated with feeling of poor health as expected. Association of low humidity (HumA) and decrease in humidity (HumAChg) with good health confirms cardiologists' opinion that CHF patients poorly tolerate humid weather.

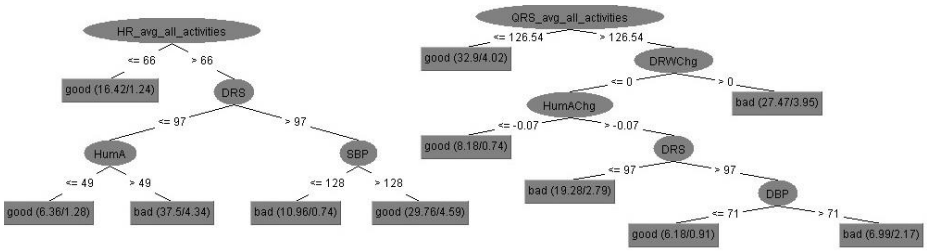


Fig. 1. J48 classification tree on the avg subset of all_act dynamic features (left) and J48 classification tree on the avg + chg subset of all_act dynamic features (right).

Conclusion: This paper presented analysis of the telemonitoring data of CHF patients. The derived models appear consistent with current medical knowledge. However, the models that contain new relations might be even more significant, since they might represent important discoveries, and must be further examined by cardiologists.

References

1. C. Sousa, S. Leite, et al. Telemonitoring in heart failure: A state-of-the-art review. *Revista Portuguesa de Cardiologia* 33 (4), pp. 229–239
2. E. Mazomenos, J. M. Rodríguez, C. Cavero, G. Tartarisco, G. Pioggia, B. Cvetković, S. Kozina, H Gjoreski, M. Luštrek, H. Solar, D. Marinčič, J. Lampe, S. Bonfiglio, K. Maharatna. Case Studies. In *System Design for Remote Healthcare*, 2014, pp. 277–332
3. M. Luštrek, B. Cvetković, M. Bordone, E. Soudah, C. Cavero, J. M. Rodríguez, A. Moreno, A. Brasaola, P. E. Puddu. Supporting clinical professionals in decision-making for patients with chronic diseases. *Proc. IS 2013*, pp. 126–129

[2] www.cs.waikato.ac.nz/ml/weka/

Ensembles for Time Series Forecasting

Mariana Oliveira and Luís Torgo

LIAAD-INESC TEC / Faculdade de Ciências - Universidade do Porto
mrfo@inesctec.pt, ltorgo@dcc.fc.up.pt

Abstract. We describe a new type of ensembles that aims at improving the predictive performance of these approaches in time series forecasting. Previous theoretical studies of ensembles have shown that one of the key reasons for this performance is diversity among ensemble members. The key idea of the work we are presenting is to propose a new form of diversity generation that explores some specific properties of time series prediction tasks. Our hypothesis is that the resulting ensemble members will be better at addressing different dynamic regimes of time series data.

1 Our Proposal

Most existing approaches to time series forecasting use the most recent observed values of the series as predictors for the future values (usually known as an embed of the time series). These approaches require setting a key parameter - how many past values to include. Setting this parameter is not trivial most of the times as there may not exist *one* single correct answer. In effect, non-stationary series and the occurrence of different regime shifts along time may lead to the best value being clearly time-dependent. The key idea of our proposal[1] is that of using different sets of predictors (e.g. different embed sizes) within the members of an ensemble to inject some diversity that is related with specific properties of time series tasks. More specifically, given a maximum embed size k_{max}, in this paper we will consider:

- E a baseline standard bagging approach using the previous k_{max} values of the target variable as predictors
- E+S an extension of standard bagging by adding two extra predictors that try to convey extra information on the dynamics of the series, namely μ_Y and σ_Y^2, calculated using the values within the maximum embed
- DE an ensemble where one third of the models use the maximum embed, another third uses an embed of $k_{max}/2$ and the last third uses $k_{max}/4$.
- DE+S an ensemble similar to DE but all models will have have the μ_Y and σ_Y^2 extra features, although calculated with the respective embed.
- DE±S a variant of DE+S where for each third, half of the models will use the extra statistics, whilst the other half will only use the respective embed.

[1] Further details and an implementation at www.dcc.fc.up.pt/~ltorgo/DS2014

2 Experimental Evaluation

The main goal of our experimental evaluation is to check whether the new variants of bagging are able to outperform standard bagging on time series forecasting tasks. Our baseline benchmark is standard bagging using the approach tagged as E in the list above. All five variants were compared using the same base data (the k_{max} past values) as training set, but some use it in a different way, e.g. by using only part of it or by using it to generate extra features.

All five alternative forms of bagging were tested on fourteen real world time series (details at the paper associated Web site). Mean Squared Error (MSE) was used as evaluation metric to compare the different approaches. In order to obtain reliable estimates of this metric we have used a Monte Carlo simulation. In our Monte Carlo experiments we have randomly selected 10 points in time within the available time intervals of each task. For each of these 10 random points we have used as training set the previous 50% observations and the following 25% cases as test set. All approaches were trained and tested using the same exact data to allow for paired comparisons. Wilcoxon signed rank tests were carried out to test the statistical significance (with $p-$value < 0.05) of the observed paired differences in MSE of the proposed approaches against the bagging baseline.

We have repeated our experimental comparisons using 4 different setups: (i) number of models in the ensemble (M); and (ii) value of the maximum embed used by the ensembles (k_{max}). Table 1 presents the overall results of the paired comparisons. The numbers in column "Wins/Losses" are the wins and losses of each variant against the baseline, on the fourteen problems. Between parentheses we have the number of statistically significant (95% confidence) differences.

M	k_{max}	Variant	Wins/Losses	M	k_{max}	Variant	Wins/Losses
1020	20	E+S	13 (11) / 1 (1)	1500	20	E+S	13 (10) / 1 (1)
		DE	7 (7) / 7 (3)			DE	8 (6) / 6 (3)
		DE+S	13 (10) / 1 (0)			DE+S	13 (10) / 1 (0)
		DE±S	14 (12) / 0 (0)			DE±S	14 (12) / 0 (0)
	30	E+S	11 (9) / 3 (2)		30	E+S	11 (9) / 3 (2)
		DE	10 (6) / 4 (3)			DE	9 (7) / 5 (3)
		DE+S	10 (5) / 4 (2)			DE+S	10 (7) / 4 (2)
		DE±S	10 (9) / 4 (2)			DE±S	10 (9) / 4 (2)

Table 1: Paired comparisons results.

These results clearly show a positive overall balance of our proposed method for adding time series-specific diversity to bagging. In particular, the DE±S variant achieves remarkable results when $k_{max} = 20$, as it always outperforms standard bagging. This is the variant that introduces more variability within the members of the ensemble, which somehow provides further evidence of the advantage of our proposal. Overall, these results are encouraging and provide clear indications of the added value of this research direction even though many more possibilities exist to increase the level of diversity.

Modeling Dynamical Systems with Data Stream Mining

Aljaž Osojnik[1,], Sašo Džeroski[1,]

[a]*Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia*

Abstract

We address the task of modeling dynamical systems in discrete time using regression trees, model trees and option trees for on-line regression. Some challenges that modeling dynamical systems pose to data mining approaches are described: these motivate the use of methods for mining data streams. The algorithm FIMT-DD for mining data streams with regression or model trees is described, as well as the FIMT-DD based algorithm ORTO, which learns option trees for regression. These methods are then compared on several case studies, i.e., tasks of learning models of dynamical systems from observed data. The experimental setup, including the datasets, and the experimental results are presented in detail. These demonstrate that option trees for regression work best among the considered approaches for learning models of dynamical systems from streaming data.

Email addresses: aljaz.osojnik@ijs.si (Aljaž Osojnik), saso.dzeroski@ijs.si (Sašo Džeroski)

Clustering based approach for balancing accuracy and diversity in collaborative filtering

Apurva Pathak[1], Bidyut Kr. Patra[2], Ville Ollikainen[2], and Raimo Launonen[2]

[1] Microsoft Research and Development Pvt. Ltd., Hyderabad, India.
[2] VTT Technical Research Centre of Finland, Finland.
E-mail: appatha@microsoft.com, {ext-bidyut.patra, ville.ollikainen, raimo.launonen}@vtt.fi

Abstract. In today's world of e-commerce, recommender system plays an important role for both individual user and business. Traditional recommender systems mainly emphasize on individual user satisfaction by recommending high quality items. However, recent study shows that diversity in recommendations is equally important factor from both user and business view points. Diversity in recommended items (individual diversity) provides a user to discover new (novel) items. On the other hand, recommending diverse range of items (aggregate diversity) prevent any item from becoming obscure in a large item space. While there are numerous works done on improving the recommendation accuracy. However, diversity in recommendation is often overlooked. In this paper, we explore clustering technique to generate diverse recommendations at both individual user and aggregate level while maintaining comparable level of accuracy. The results show high diversity gains while applying proposed approach on a real-world rating dataset (MovieLens).
Keywords: Collaborative Filtering, Accuracy, Individual diversity, Aggregate diversity, Clustering.

1 Introduction

Recommender system (RS) is an important tool to cope with information overload problem in the present internet era. Primary task of RS is to provide personalized suggestions of products or items to individual user so that user can select desired products or items directly without surfing over large products space. Collaborative Filtering (CF) is the most successful and widely used recommendation system [1, 2]. In CF, item recommendations to an user are performed by analyzing rating information of the other users and other items in the system. Traditional CF algorithms emphasize on predicting accurate ratings of the items suggested to individual user. Recent studies show that diversity in recommendations can be beneficial to not only users but also to business models. Diversity in recommendation can be viewed from two different perspectives *i.e, user perspective (individual diversity)* and *business perspective (aggregate diversity)* [3]. Increasing in individual diversity helps an user obtain more idiosyncratic items in her recommended list while aggregate diversity of an RS increases selling of obscure items. However, both types of diversity in recommendation are achieved at the expense of accuracy.

In this paper, we propose an approach which explores hierarchical clustering to increase diversity (both individual and aggregate) of recommender systems. Proposed

Table 1. Results obtained from MovieLens dataset with density index 4.66%.

	RMSE	Aggregate Diversity (AD)	Individual Diversity (ID)	Increase in RMSE	Gain in AD	Gain in ID
RSVD	**0.7546**	32	0.0310	NA	NA	NA
Ranking Approach	0.7734	654	0.0562	2.49 %	**1943.75%**	80.82%
Proposed Approach	0.7571	643	0.0747	**0.39%**	1909.09%	**140.47%**

approach is tested with popular MovieLens data. Experimental results show that proposed approach gains significant diversity which is very close to the diversity obtained by ranking based approach, however, our approach suffers little accuracy loss compared to that of the ranking based approach [3].

2 Methodology

The proposed approach has two steps. In the first step, a standard matrix factorization technique such as regularized SVD (RSVD) is employed to obtain predictions of unknown ratings of items and we sort them in descending order in their predicted rating for each user in a system. It can be noted that this step is same as the step performed by many traditional RS which results in low diversity in the recommendation. Therefore, in the next step, we use hierarchical average-link clustering method to recommend top N items to individual user as follows.

We select a set of items S_u with predicted ratings more than a predefined threshold T for an active user u obtained in the first step. Subsequently, we apply hierarchical clustering technique to the set of selected items S_u. We reuse the feature vectors of items obtained from regularized SVD for computing similarity between a pair of items in S_u. We choose N clusters from the dendogram formed by the hierarchical average-link clustering method and select one item with minimum predicted rating in each cluster to recommend N items to the active user u. We choose a threshold T to ensure minimal accuracy loss in our approach. Selecting items from each of the N clusters provide high individual diversity and reuse of global information (feature vectors) ensure high aggregate diversity of the recommender system.

3 Experimental Results

We used top 3000 users (rated maximum number of movies) and top 2000 movies (received ratings from maximum number of users) from MovieLens 1 M dataset for experimental evaluation of the proposed approach. We compared our approach with the ranking based approach [3] and results are reported in Table 1. Experimental results show that our approach outperforms rank based approach in individual diversity (140.47%) and accuracy loss (0.39%) while it produces aggregate diversity very close to the diversity produced by that of the rank based approach (Table 1).

4 Conclusion and Future Work

We proposed an approach which provide significant individual and aggregate diversity maintaining high accuracy in a collaborative filtering system. Proposed approach can be used with traditional CF approach inoder to provide high diversity. The work can be extended in utilizing state-of-art clustering technique in the second step of the approach.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering **17**(6) (2005) 734–749
2. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Advances in Artificial Intelligence **2009** (2009) 4:2–4:2
3. Adomavicius, G., Kwon, Y.: Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. IEEE Transactions on Knowledge and Data Engineering **24**(5) (2012) 896–911

Bridging term discovery for cross-domain literature mining

Matic Perovšek[1,2], Nada Lavrač[1,2,3], and Bojan Cestnik[4,1]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[2] International Postgraduate School Jožef Stefan, Ljubljana, Slovenia
[3] University of Nova Gorica, Nova Gorica, Slovenia
[4] Temida d.o.o., Ljubljana, Slovenia

Understanding complex phenomena and solving difficult problems often requires knowledge from different domains to be combined and cross-domain associations to be taken into account. This kind of context crossing associations, named bisociations [1], are often needed for creative, innovative discoveries. Following Koestler's ideas [1] and based on computational approaches to bisociative knowledge discovery [2], the goal of this work is to develop a computational system able to discover links between two previously unrelated domains, represented by two different document corpora. The work upgrades the CrossBee methodology [3], aimed to detect the bridging terms that represent bisociative links between different domains. The CrossBee methodology employs an ensemble of specially tailored text mining heuristics that assign to each discovered candidate bridging term (B-term) a score, which should reflect their bisociation potential. The resulting ranked list of potential B-terms enables the user to inspect the top-ranked B-terms, which may result in higher probability of finding observations that lead to the discovery of new bridges between the two domains.

We have extended the CrossBee system by integrating a new complementary technique to B-term ranking. The proposed technique uses banded matrices [4] to discover structures which reveal the relations between the rows (representing documents) and columns (representing terms) of a given data matrix representing a set of documents. We use this information in computing new heuristics that evaluate terms according to their potential for B-term discovery. The proposed approach thus encodes the documents from the two domains into the standard Bag-Of-Words (BOW) vector representation and then transforms the binary matrix of BOW vectors into a banded structure.

The proposed banded matrix methodology is based on the assumption that similar documents, as well as the terms that appear in the same document, will appear closer to each other in the matrix and will therefore form "clusters" along the main diagonal of the matrix in its banded form. Our work is based on the intuition that terms that connect different domains will be positioned at the edges of clusters from different domains, and the developed heuristics should be able to identify these B-terms by ranking them high in the ranked list of terms with high potential for cross-domain link discovery.

The methodology we propose works as follows: first, we preprocess the documents from the two domains using standard text mining techniques. This is performed through a number of steps: stop-word removal, stemming or lemmatization, usage of synonym dictionaries, construction of n-grams of words and, finally, transformation to a Bag-Of-Words representation. Next, the result of the preprocessing step, i.e., the binary matrix of "Bag-Of-Words" vectors (the BOW matrix), is transformed into the banded matrix structure. In the next step we permute columns and rows of the binary matrix using

a bidirectional MBA algorithm [4] in order to retrieve a banded structure, followed by using the proposed heuristics to calculate the B-term potential for every term. One of the proposed heuristics calculates this score for term t as the ratio of documents characterised by term t in a document cluster grouped around the diagonal, multiplied by the number of documents from the other domain including term t. The intuition behind this heuristic is that for term t, the more the term represents a domain (has a large proportion of document on the diagonal cluster of the banded matrix) and also the more documents from the other domain that contain t exist, the higher the potential of term t to be a bridging term between the two domains. After completing the step of term score computation, we sort the terms according to the values of the heuristics and present the top-ranked terms (hopefully representing the most interesting B-term candidates) to the expert. The designed heuristics should favor B-terms over non-B-terms by pushing interesting B-term candidates to the top of the ranked term list.

We are mostly interested in the quality of heuristics from the end-user's perspective. Note that the standard ROC curves and AUC statistics do not provide the most significant evidence of the quality of individual heuristics, even though—in general—a better ROC curve reflects a better heuristic. Usually the user is interested in questions like: how many B-terms are likely to be found among the first n terms in a ranked list (where n is a selected number of terms the expert is willing to inspect, e.g., 5, 20 or 100).

In the experiments we used the well-researched migraine-magnesium domain pair as introduced by Swanson [5]. Swanson managed to find more than 60 pairs of articles connecting the migraine domain with the magnesium deficiency via 43 bridging concepts (B-terms). Using the developed methodology we tried to rank these 43 B-terms as high as possible among other terms that are not marked as B-terms. It proves that banded matrices help us discover the structures that indeed reveal the relation between terms and documents, which allows for faster cross-domain discovery than with the original CrossBee tool. Furthermore, we show that by using a predefined vocabulary we can increase the heuristic's capacities to rank the B-terms at the beginning of the term list. Indeed, by applying this approach in the migraine-magnesium domain we got a higher concentration of Swanson's B-terms among the best ranked terms. Consequently, the user is presented with a simpler exploration task, potentially leading to new discoveries.

Acknowledgment. *This work was supported by the Slovenian Research Agency grant as well as the FP7 European Commission projects MUSE (grant agreement no: 296703) and ConCreTe (grant agreement no: 611733).*

References

1. Koestler, A.: The Act of Creation. Volume 13. (1964)
2. Berthold, M., ed.: Bisociative Knowledge Discovery. Springer (2012)
3. Juršič, M., Cestnik, B., Urbančič, T., Lavrač, N.: Cross-domain literature mining: Finding bridging concepts with CrossBee. In: Proceedings of the 3rd International Conference on Computational Creativity. (2012) 33–40
4. Garriga, G., Junttila, E., Mannila, H.: Banded structure in binary matrices. Knowledge and Information Systems **28**(1) (2011) 197–226
5. Swanson, D.R.: Migraine and magnesium: Eleven neglected connections. Perspectives in Biology and Medicine **78**(1) (1988) 526–557

Generator of unsupervised semi-artificial data

Marko Robnik-Šikonja

University of Ljubljana, Faculty of Computer and Information Science,
Večna pot 113, 1000 Ljubljana, Slovenia Marko.Robnik@fri.uni-lj.si

In many important application areas addressed by discovery science, there just isn't enough data available. There are several reasons for this, the data may be inherently scarce, difficult to obtain, expensive, or the distribution of the events of interests is highly imbalanced. This causes problems in model selection, reliable performance estimation, development of specialized algorithms, and tuning of learning model parameters.

Recently we presented a semi-artificial data generator limited to classification problems [3]. This generator first constructs a RBF network prediction model which consists of Gaussian kernels. The kernels estimate the probability density function from the training instances. Due to properties of Gaussian kernels, the learned kernels are used in a generative mode to produce new data. This approach was successfully used for a variety of data sets described with different number of attributes of both, numerical and categorical type. The generator was successfully tested in development of big data tools [1] and is freely available as R package semiArtificial. We expect such a tool to be useful in the data mining algorithm development and adaptation of data analytics algorithms to specifics of data sets. Possible other uses are data randomization to ensure privacy, simulations requiring large amounts of data, testing of big data tools, benchmarking, and scenarios with huge amounts of data.

The problem as well as the strength of the RBF-based generator is that the learned model tries to discriminate between instances with different class values, therefore the approach generates many kernels for each class. The widths of the kernels are estimated from the training instances which activate the particular kernel.

A simple demonstration of the generated data on a simple data set is presented in Fig. 1 The data set forms a two dimensional grid where attributes A_1 and A_2 are generated with three Gaussian kernels with centers at $(-5, -5)$, $(0, 0)$, and $(5, 5)$. Each group of 500 instances is assigned a unique class value (red, blue, and green, respectively) as illustrated in Fig 1a. The RBF generator based on this data consists of eight Gaussian kernels (two for red and blue class each, and four for green class). We illustrate 1500 instances generated with this generator in Fig 1b. As the RBF learner did not find the exact locations of the original centers it approximated the data with several kernels, so there is some difference between the original and generated data.

In this work we propose a different approach using density trees [2]. The density trees are similar to decision trees with a difference that there is no designated class variable. The split criteria used therefore try to separate areas with different density. We experimented with different types of density trees varying the split selection criteria, stopping criteria, and also the density estimation method used in the leaves of the trees. Our preliminary experiments show that in general the density tree approach produces semi-artificial data with better properties compared to the RBF generator (improved similarity to the original, better clustering performance).

The idea of proposed density trees data generator is to construct a forest of density trees using bootstrap sampling and random selection of a subset of attributes in each node (similarly to random forests) to assure tree diversity. When generating a new instance we start at the root of a randomly chosen tree. The splitting attribute in each interior node is used to generate the value of the new instance randomly but following the selected attribute's empirical cumulative distribution function (ecdf). Based on the generated value we recursively repeat the generation of values for yet unobserved attributes by following the left- or right-hand branch of the tree. Arriving to the leaf of the tree we assume that dependencies between attributes were resolved on the path from the the root to the leaf and we generate the remaining values of the attributes using univariate methods (kernel density estimation, log splines, or ecdf).

The working of the generator are demonstrated in Fig. 1. The generator using density trees was not using the class information, but was able to better capture locality information of the instances (Fig. 1c). If class information was provided the density tree based generator was using it as any other nominal attribute (Fig. 1d). The results on this and many other data sets show that the generated data is more similar to the original data set.

In further work we will investigate exact conditions when each of the proposed methods (RBF generator, density tree generator) produces favorable results.

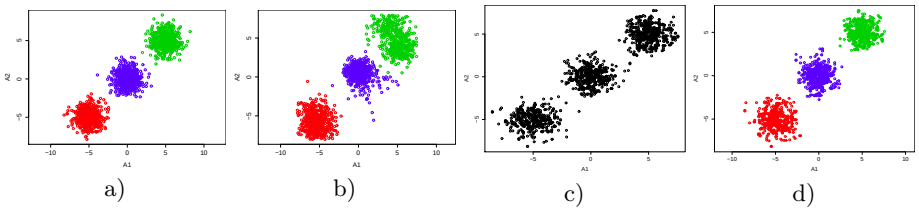


Fig. 1. An illustration of the generated data on the two dimensional dataset: a) the original data, b) data generated with RBF generator, c) data generated with density trees without class information, d) data generated with density trees with class information as a nominal attribute.

Acknowledgments

The author was supported by the Slovenian Research Agency (ARRS) through research programme P2-0209 and European Commission through the Human Brain Project (grant number 604102).

Bibliography

[1] J. Kranjc, R. Orač, V. Podpečan, M. Robnik-Šikonja, and N. Lavrač. ClowdFlows: Workows for big data on the cloud. Technical report, Jožef Stefan Institute, Ljubljana, Slovenia, 2014.

[2] P. Ram and A. G. Gray. Density estimation trees. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD*, pages 627–635. ACM, 2011.

[3] M. Robnik-Šikonja. Data generator based on RBF network. Technical report, University of Ljubljana, Faculty of Computer and Information Science, 2014. URL http://arxiv.org/abs/1403.7308v1.

CGT: a vertical miner for frequent equivalence classes of itemsets (extended abstract)

Laszlo Szathmary[1], Petko Valtchev[2], Amedeo Napoli[3], Marton Ispany[1], and Robert Godin[2]

[1] University of Debrecen, Faculty of Informatics, Department of IT, H-4010 Debrecen, Pf. 12, Hungary
{szathmary.laszlo, ispany.marton}@inf.unideb.hu
[2] Dépt. d'Informatique UQAM, C.P. 8888, Succ. Centre-Ville, Montréal H3C 3P8, Canada
{valtchev.petko, godin.robert}@uqam.ca
[3] LORIA (CNRS - Inria NGE - Université de Lorraine) BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France
napoli@loria.fr

Abstract. In this extended abstract we present a vertical, depth-first algorithm that outputs frequent generators (FGs) and their associated frequent closed itemsets (FCIs). The proposed algorithm –called *CGT*– is a single-pass algorithm and it explores frequent equivalence classes in a dataset.

Introduction. In data mining, frequent itemsets (FIs) and association rules play an important role. Due to the high number of patterns, various concise representations of FIs have been proposed, of which the most well known representations are the FGs and the FCIs. There are a number of methods in the literature that target both FCIs and FGs, but most of these algorithms are levelwise methods. It is known that depth-first algorithms usually outperform their levelwise competitors. Here we present briefly a single-pass, depth-first, vertical FG+FCI miner.[1,2]

The CGT algorithm. *CGT* is a vertical itemset mining algorithm for finding frequent equivalence classes. *CGT* is based on *Talky* [1], where *Talky* is a modified version of *Eclat* [2]. *Eclat* and *Talky* produce the same output, i.e. they find all FIs in a dataset. However, *Talky* uses a different traversal called reverse pre-order strategy. This traversal goes from right-to-left and it provides a special feature: when we reach an itemset X, all subsets of X were already discovered. As a result, this traversal can be used to filter FGs among FIs. During the traversal procedure *CGT* also filters FCIs and assigns them to the corresponding FGs, thus *CGT* outputs at the end the frequent equivalence classes. In order to filter the FGs, it must rely on the reverse pre-order strategy.

Consider the following 4×6 sample dataset: $\mathcal{D} = \{(1,\ ACDE),\ (2,\ ABCDE),\ (3,\ ABE),\ (4,\ BEF)\}$. The execution of *Talky* on dataset $\mathcal{D}$ with $min\_supp = 2$ is

[1] Due to lack of space, please refer to our paper [1] for the definitions of basic concepts.

[2] This work was partially supported by the TAMOP-4.2.2.C-11/1/KONV-2012-0001 project, itself partially funded by the European Union, co-funded by the European Social Fund. Canadian co-authors acknowledge the support of the National Science and Engineering Research Council through their respective Discovery grants.

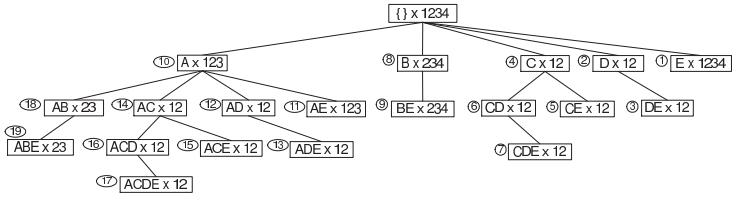


Fig. 1. Execution of *Talky* on dataset $\mathcal{D}$ with $min\_supp = 2$.

Table 1. CGT builds this table, which is actually a hash table. Key of the hash: a tidset. Value of the hash: a row of the table.

tidset	generators	eq. class members (optional)	closure	support
1234	$\emptyset$	E	E	4
234	B	BE	BE	3
123	A	AE	AE	3
23	AB	ABE	ABE	2
12	D, C	DE, CE, CD, CDE, AD, ADE, AC, ACE, ACD, $ACDE$	$ACDE$	2

shown in Figure 1. The processing order of nodes is indicated in circles. For instance, the node $C \times 12$ means that the itemset C is present in the 1st and 2nd row of the dataset, thus its support is 2.

Example. CGT builds a hash table, as depicted in Table 1. A row object represents an equivalence class. In our dataset $\mathcal{D}$, the column E is full, meaning that E is not a generator because it has a proper subset with the same support namely the empty set. Thus, the empty hash table is initialized with the line "1234;$\emptyset$; ;$\emptyset$;4". (That is, the tidset is 1234, the generators and the closure are the empty set, and the support value is 4. Eq. class members is left empty.) Then, the algorithm starts enumerating the 19 FIs of $\mathcal{D}$ using the traversal strategy of *Talky* (as seen in Figure 1). The first node is $E \times 1234$. The tidset 1234 is an existing key in the hash. E has a proper subset with the same support (the empty set), thus E is added to the "eq. class members" and "closure" fields. The "closure" column is the union of the generators and the itemsets that belong to the same equivalence class. The next FI is $D \times 12$. Since 12 is not yet in the hash, a new row is added in the hash table. The next node is $DE \times 12$. The tidset 12 is in the hash, thus DE belongs to an existing equivalence class. It has a proper generator subset, D, thus DE is added to the "eq. class members" and "closure" fields. After adding E, D, and DE, the current state of the hash table looks like this: 1st row is "1234;$\emptyset$;E;E;4"; 2nd row is "12;D;DE;DE;2". The end result is shown in Table 1.

References

1. Szathmary, L., Valtchev, P., Napoli, A., Godin, R.: Efficient Vertical Mining of Frequent Closures and Generators. In: Proc. of the 8th Intl. Symposium on Intelligent Data Analysis (IDA '09). Volume 5772 of LNCS., Lyon, France, Springer (2009) 393–404
2. Zaki, M.J., Parthasarathy, S., Ogihara, M., Li, W.: New Algorithms for Fast Discovery of Association Rules. In: Proc. of the 3rd Intl. Conf. on Knowledge Discovery in Databases. (August 1997) 283–286

Studying the drainage periods for agricultural fields with data mining: La Jaillière case study

Aneta Trajanov[a], Vladimir Kuzmanovski[a], Florence Leprince[b], Benoit Real[c], Alain Dutertre[d], Julie Maillet-Mezeray[e], Sašo Džeroski[a,f,g], Marko Debeljak[a,f]

[a] Department of Knowledge Technologies, Jozef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

Emails: aneta.trajanov@ijs.si, vladimir.kuzmanovski@ijs.si, saso.dzeroski@ijs.si, marko.debeljak@ijs.si

[b] ARVALIS – Institut du végétal, 21, chemin de Pau, 64121 Montardon, France

Email: f.leprince@arvalisinstitutduvegetal.fr

[c] ARVALIS - Institut du végétal, 2, chaussée Brunehaut, CS 30200, 80208 Peronne Cedex, France

Email: b.real@arvalisinstitutduvegetal.fr

[d] ARVALIS - Institut du végétal, Station expérimentale de La Jaillière, 44370 La Chapelle Saint Sauveur, France

Email: a.dutertre@arvalisinstitutduvegetal.fr

[e] ARVALIS - Institut du végétal, Station expérimentale, 91720 Boigneville, France

Email: julie.mailletmezeray@bayer.com

[f] Jozef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

[g] Center of Excellence for Integrated Approaches in Chemistry and Biology of Proteins, Jamova cesta 39, 1000 Ljubljana, Slovenia

The identification of intensive drainage periods is important for determining mitigation strategies for protecting water against pollution with plant protection products. Most attempts to estimate the start, duration and the end of a drainage period are based either on mechanistic modeling approach or on empirical knowledge about tile drainage. The mechanistic modeling requires many parameters, while the empirical approach does not allow for making simulations and predictions needed for proposing reliable mitigation measures. In order to complement these two approaches, we have used a data mining approach on data from 25 (1987-2011) agricultural seasons (campaigns) from the experimental station La Jaillière, France. The models for estimating the start and the end of the intensive drainage periods for a particular campaign have the form of decision trees and tell us which factors influence these dates the most. The start of a drainage period depends mostly on the cumulative drainage and the cumulative rainfall since the beginning of the campaign and the average air temperature of the last seven days. For estimating the end of a drainage period, the most important variables are the cumulative rainfall of the last seven days and the average air temperature of the following seven days.

The obtained models for estimating the start and the end of a drainage period could be used not just to estimate the daily status of the drainage regime on a particular field (e.g., presence or absence of drainage), but they can also be used to predict the drainage status of the field for a time period covered with reliable weather forecasts. Using information from weather forecasts to run simulations on models for the beginning of a drainage period (or the end, depending on the decision at hand) would make it easier for farmers and policy makers to take into account the drainage period when deciding to apply plant protection products in the field. Thus, our data mining models, built from measured data, bring decision making flexibility to their users, because they can be used either for ex-ante or ex-post analysis. The combination of both types of analysis presents a very simple decision support system, which significantly increases the certainty and flexibility of management decisions taken by advisors and farmers in the La Jaillière area (ARVALIS) or in other places with the same field and crop management properties.

Mining Graph Structures Preserved Long Period

Takeaki Uno[1] and Yushi Uno[2]

[1] National Institute of Informatics, Japan. uno@nii.jp
[2] Graduate School of Science, Osaka Prefecture University, Japan.
uno@mi.s.osakafu-u.ac.jp

Abstract. Data mining from sequences of graphs is now increasing its importance in practice. In this research, we newly focus on "preserving structures" while other studies have focused on "changes". We particularly address connected induced subgraphs and cliques that do not change in a long period in the given graph sequence. We propose new polynomial delay algorithms for the problems. The delay is $O(|V||E|^3)$ for the former, and and $O(\min\{\Delta^5, |E|^2\Delta\})$ for the latter, where the input graph is $G = (V, E)$ with maximum degree Δ.

Introduction

In a recent and practical situation, graph structures may change over time (i.e., "dynamic"), and such data is collected periodically along a time series (and thus the data becomes bigger). In this setting, not only information acquired separately from single graphs but also from graph patterns appearing sequentially could be important. Along this direction, there are some research topics of interest so far. Finding graph patterns that appear periodically in a graph sequence is studied. Graph patterns frequently appear during a certain period such as burst patterns are also studied. On the other hand, some research address the change patterns that appear frequently in a graph sequence composed of graphs with edge insertions/deletions, such as changes between two time periods and changes of subsequences. Several studies focus on clustering of vertices. Surprisingly, all these researches look at "changes", but no research does "stability".

We propose a new concept of graph mining; finding graph/subgraph structures belonging to a graph structure class, that are appearing in a long period in a series of dynamically changing graphs. We call such structures *preserving structures* in a graph sequence, and the problem for enumerating all such structures *preserving structure mining* in general. As for such properties, we consider maximal connected induced subgraphs and maximal cliques. For example, a topic on the Web that is controversial for a long time may correspond to a clique that exists in a consecutive sequence of webgraphs during a certain period. As another example, a group of a species in a wildlife environment may constitute a consecutive sequence of connected vertex subsets in a sequence of graphs that are constructed from its trajectory data. To the best of our knowledge, this study is the first case in which a "long-lasting" or "unchanging" structure is regarded as the target structure to be captured.

Related works

(1) Pattern mining in graph sequences. This is already explained in the Introduction, and it tends to capture changes.

(2) Dynamic flow. On a dynamic network defined by a graph with capacities and transit times along its edges, the dynamic flow problem asks the maximum flow from a specified source to a sink within a given time bound. As explained

later, our model for a graph sequence can be naturally generalized so that it implies dynamic flows.

(3) Dynamic graph algorithms. Dynamic graph problems of constructing data structures that enables to answer a given graph property quickly, with small update cost for edge insertions/deletions. Typical properties of concern include connectivity, transitive closures, cliques, bipartiteness, shortest path distance, and so on. Dynamic graph algorithms could also find a period during which a property is satisfied. However, since they are not well designed so that they can extract local structures efficiently, they are not suitable for this purpose.

Contributions

In this paper, we first propose a new concept, that is, a preserving structure in a graph sequence. Then by adopting this notion, as an onset, we pose two problems of mining preserving structures: one for maximal cliques and the other for maximal connected induced subgraphs. As we have seen so far, both structures or properties will have significant meanings in a sequence of graphs that appear in practical situations.

Let $G_1, \ldots, G_T$ be the graph sequence we are given. For a vertex set S, the *active time set* is the set of indexes of the graphs of the graph sequence in which S induces the connected graph. When there is an interval of the active time set is of length at least τ, we say that S is *preserving*. Our first problem is to enumerate all maximal preserving connected induced subgraphs. Our enumeration algorithm for the problem is based on a recursive graph partition. Consider the intersection of the partitions, given by the connected components of graphs $G_{t_1}, \ldots, G_{t_2}$, such that the result is composed of disjoint maximal vertex sets that are not subdivided by the connected components of any graph. Look at a component S of the resulted partition. If S is connected in any graph in the graphs, it is a solution to the problem. Otherwise, its subsets are solutions. Thus, we recursively do this for the set of graphs $G_i[S], i = t_1, \ldots, t_2$, recursively, until the components will be connected in any graph. This idea motivates us to find all solutions to short period, and update them with increase the time span. In this way, we can enumerate all solutions to the problem. When for each edge, the graphs having the edge form an interval, we have the following theorem.

Theorem 1. *All maximal preserving connected induced subgraphs can be enumerated in $O(|V||E|^3)$ time for each, where the input graph is $G = (V, E)$.*

Our algorithm for enumerating maximal preserving cliques is based on the reverse search, which is a framework for designing efficient enumeration algorithms. The idea is to introduce a parent-child relation.

While a straightforward application of maximal clique enumeration to our problem may require a long delay per output, our algorithm exploits properties of the time intervals of edges so that the algorithm will be polynomial delay. Compared to a naive algorithm, this reduces the time complexity with a factor of the number of edges of an input graph. Although these algorithms may seem to be relatively simple, our problem setting is quite fundamental and new. Therefore, it gives a new perspectives for graphs that change over time, together with a way of data representations and analysis of algorithms, and it would be a first step to pioneer a new research field.

Theorem 2. *All maximal preserving cliques can be enumerated in $O(\min\{\Delta^5, |E|^2\Delta\})$ time for each where the input graph is $G = (V, E)$ with maximum degree Δ.*

Extracting Sparse Canonical Correlations Between Microbial Communities and Deep Groundwater Geochemistry

Viivi Uurtio[1], Juho Rousu[1], Malin Bomberg[2], and Merja Itävaara[2]

[1] Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science, Aalto University,
Konemiehentie 2, FI-00076 Aalto, Finland
[2] VTT Technical Research Centre of Finland, Espoo, Finland

Abstract. Microorganisms are found in deep subsurface groundwater, upto kilometers deep. Microbial populations interact in communities with the geochemical resources and conditions of their habitat [1], in ways that researchers are only beginning to understand. Here we describe a study on deep subsurface microbial communities in Fennoscandian crystalline bedrock [2],[3]. The motivation for the particular study is risk assessment for long term disposal of nuclear waste [4], where the geobiochemical stability of the site and the potential chemical and physical effects of microbial activity need to be understood.

In order to model the complex network of microbial community interactions with the deep bedrock habitat, we analysed environmental samples obtained from the deep bedrock containing both microbial and geochemical variables. We focused on sulfate reducing bacteria that produce sulphide which may corrode the copper of the nuclear waste capsules. The bacteria were identified by their dissimilatory sulphite reductase marker genes (*dsr*B) that are present in all microorganisms performing dissimilatory sulfate reduction [5]. Computational analysis of this type of data requires a multivariate approach in order to extract correlations among the variables. Since the diverse bacterial interactions with the geochemistry of the habitat are complex and encompassed in high-dimensional data, the visualization and interpretation of the results of multivariate analysis is challenging.

We applied asymmetrical sparse canonical correlation analysis (SCCA) [6] in order to extract subsets of highly correlating sulfate reducing bacterial communities and geochemical measurements from two datasets obtained from deep bedrock drill holes in Finland [5],[7]. SCCA is a multivariate method that seeks semantic projections that use as few relevant features as possible to explain as much correlation as possible [6] by imposing *L1*-norm penalization on variable weights [8].

We imposed sparsity on either of the data views by penalizing dual variables related to the latent variables of the other view. In order to select an optimal level of sparsity for a data view, we performed 3-fold cross validation in which we computed the optimal feature weights at a range of levels of sparsity for a training set and assigned these to an unseen test set in order to obtain a predictive canonical correlation coefficient. The optimal level of sparsity resulted in highest predictive correlation coefficient of the projections. The optimal observed projection correlations were tested statistically by permutation tests in which the optimal level of sparsity was chosen for each permuted dataset.

We analysed the resulting projections by examining the correlation coefficients, i.e. cosine angles of the projections to the feature axes. In this way, we extracted the subsets of features contributing to the highly correlating projection direction. The method of computing linear correlation coefficients among the original and projected measurements was introduced by [9],[10] in the framework of two co-dependent datasets but has not yet been applied to SCCA. The highly correlating subsets of features were visualized by means of correlation plots [9],[10]. We also extended the correlation plot visualization to a clustergram in which the problem of overlapping features was overcome.

When sparsity was imposed on sulfate reducing bacterial data, we discovered a high positive correlation among the *Peptococcaceae* family and the geochemical measurements depth, electrical conductivity, total dissolved salts, total number of cells including the ionic chloride and calcium which points out a possible relation between salinity and sulfate reduction. Another correlation that contributes to this finding was seen when sparsity was imposed on the geochemical measurements, since the *Desulfobulbaceae* family was correlating negatively with ionic chloride. *Peptococcaceae* and *Desulfobacteraceae* families were correlating positively with pH measurements. This could be explained by the fact that sulfate reducers have an impact on it through

consumption of sulfate and production of sulfide which regulate the buffering capacity of the water [11] in their habitat.

Our approach finds biologically relevant correlations that can be used to unravel the complex interactions occurring in a microbial habitat. The optimization of the level of sparsity for each view prior to computation of the feature weights seems to be important in cases where the number of features in the two views differs greatly. Both visualization techniques, correlation plots and clustergrams, enabled biological interpretation of the results.

Asymmetrical SCCA algorithm together with optimization of the level of sparsity and statistical significance testing of the resulting projections provides a means to examine the relations among two sets of co-dependent variables in high-dimensional space. An alternative approach to improve the framework would be to include prior information about the relations among the variables before the computation of the projections.

References

1. Madigan M.T., Martinko J.M., and Parker J. *Brock biology of microorganisms*. Pearson Benjamin Cummings, 2009.
2. M. Itävaara, M. Nyyssönen, A. Kapanen, A. Nousiainen, L. Ahonen, and I. Kukkonen. Characterization of bacterial diversity to a depth of 1500 m in the outokumpu deep borehole, fennoscandian shield. *FEMS microbiology ecology*, 77(2):295–309, 2011.
3. L. Purkamo, M. Bomberg, M. Nyyssönen, I. Kukkonen, L. Ahonen, R. Kietävainen, and M. Itävaara. Dissecting the deep biosphere: retrieving authentic microbial communities from packer-isolated deep crystalline bedrock fracture zones. *FEMS Microbiology Ecology*, 85:324–337, 2013.
4. M. Nyyssönen, M. Bomberg, A. Kapanen, A. Nousiainen, P. Pitkänen, and M. Itävaara. Methanogenic and sulphate-reducing microbial communities in deep groundwater of crystalline rock fractures in olkiluoto, finland. *Geomicrobiology Journal*, 29(10):863–878, 2012.
5. M. Bomberg, M. Nyyssönen, and M. Itävaara. Quantitation and identification of methanogens and sulphate reducers in olkiluoto groundwater. Technical report, Posiva OY and VTT Technical Research Centre of Finland., 2010.
6. D.R. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83, 2011.
7. M. Bomberg, M. Nyyssönen, and M. Itävaara. Characterization of olkiluoto bacterial and archaeal communities by 454 pyrosequencing. Technical report, Posiva OY and VTT Technical Research Centre of Finland., 2012.
8. J. Rousu, D. D. Agranoff, O. Sodeinde, J. Shawe-Taylor, and D. Fernandez-Reyes. Biomarker discovery by sparse canonical correlation analysis of complex clinical phenotypes of tuberculosis and malaria. *PLoS computational biology*, 9(4):e1003018, 2013.
9. B-H. Mevik and R. Wehrens. The pls package: Principal component and partial least squares regression in r. *Journal of Statistical Software*, 18, 2007.
10. I. González, K.-A. Lê Cao, M. J. Davis, and S. Déjean. Visualising associations between paired omics data sets. *BioData Mining*, 5, 2012.
11. G. Arp, V. Thiel, A. Reimer, W. Michaelis, and J. Reitner. Biofilm exopolymers control microbialite formation at thermal springs discharging into the alkaline pyramid lake, nevada, usa. *Sedimentary Geology*, 126(1):159–176, 1999.

ANALYSIS OF CITATION NETWORKS

Anita Valmarska [1,2], Janez Demšar [1]

[1] Faculty of Computer and Information Science, Ljubljana, Slovenia
[2] Jožef Stefan Institute, Ljubljana, Slovenia
anita.valmarska@ijs.si, janez.demsar@fri.uni-lj.si

1 Introduction

Citation networks are directed networks in which one paper cites another. Reasons for citations are various. In most cases the authors cite older publications in order to identify the related body of work, to substantiate claims or establish precedence, or to legitimate their own statements or assumptions. In the scientific world citations are used also to critically analyze or correct earlier work. Intuitively, it can be expected that papers would more often cite other papers from the same research subfield. To confirm this hypothesis, we wanted to find out whether we could detect the research subfields within a single research field, i.e., psychology, using only a citation network of papers published in the given research field. To this end we applied one of the state-of-the-art algorithms for community detection in the hope that we would be able to differentiate among different topics addressed in psychology research.

2 Data collection and network construction

To the best of our knowledge, there is no central repository with publications from psychology. Consequently, we decided to crawl the pages connected with psychology in Wikipedia. From each of the visited pages we collected the references identified by their DOIs in the reference section. This resulted in a collection of 63,826 unique papers.

Next, we queried the Microsoft Academic Research data (MAS) and collected information about the scientific papers citing the initial set of collected papers. This allowed us to construct a citation network whose core contained papers published in the field of psychology. The resulting network consists of 948,791 vertices and 1,539,563 edges.

Due to the nature of our data collection process, we had to perform an initial data pre-processing in order to extract the papers that had a significant impact on the field of psychology. This resulted with a new network of 3,918 vertices connected by 5,732 edges.

3 Community detection and naming the communities

The process of identification of research subfields in the citation network was translated into the problem of community detection. For the purpose of our research, we applied the Louvain method [1]. It is a simple, efficient, and easy to implement method for identifying communities in large networks.

The method is divided into two phases that are repeated iteratively. At the beginning, each of the n vertices of the graph $G(V, E)$ is assigned to a different community. In this initial partition there

are as many communities as there are vertices. Then, for each vertex v we consider the neighbors u of v and we evaluate the gain of modularity that would result by removing v from its community and placing it in the community of u. The vertex is placed in the community for which this gain is maximal, and only if the gain is positive. This process is applied repeatedly and sequentially for all vertices until no further improvement can be achieved. The second phase of the method consists of building a new network whose vertices are now the communities found during the previous phase. Once the second phase is completed, it is possible to iteratively re-apply the two phases on the obtained networks.

Part of the evaluation of the detected communities was to name them and examine their connections. Due to the vast quantity of available data and unfamiliarity with the field of psychology, we named the communities based on the cosine similarity between our initially collected psychological papers and relevant texts for each of the APA (*American Psychological Association*) divisions of psychology.

4 Results

The community detection algorithm implemented in Pajek [4] detected 52 communities. The division into communities can be observed in *Figure 1*. The smallest cluster included 7 papers, while the largest cluster was constructed of 230 psychological publications. In *Figure 1* we can observe the extracted and named communities.

5 Conclusion

Results obtained by the network analysis and community detection are encouraging. The visual representation of the communities reveals sensible relationships between psychology subfields. However, the nature of data collection and the influence of our subjective judgment on community naming offer opportunities for further improvement. This involves improved data collection, developing new and improved methods for community detection, and employing better measures for text similarity. In further work, we would also like to explore the methodology proposed by Grčar et al. [2].

References

1. V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre. "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics-Theory and Experiment*, vol. 10, no. 10, 2008.
2. M. Grčar, N. Trdin, N. Lavrač. "A Methodology for Mining Document-Enriched Heterogeneous Information Networks" , *The Computer Journal*, bxs058, 2012.
3. T. Kamada, S. Kawai. "An algorithm for drawing general undirected graphs", *Information Processing Letters*, vol. 31, no. 1, pp. 7–15, 1989.
4. W. de Nooy, A. Mrvar, V. Batagelj. *Exploratory Social Network Analysis with Pajek*, Cambridge University Press, New York, 2011.

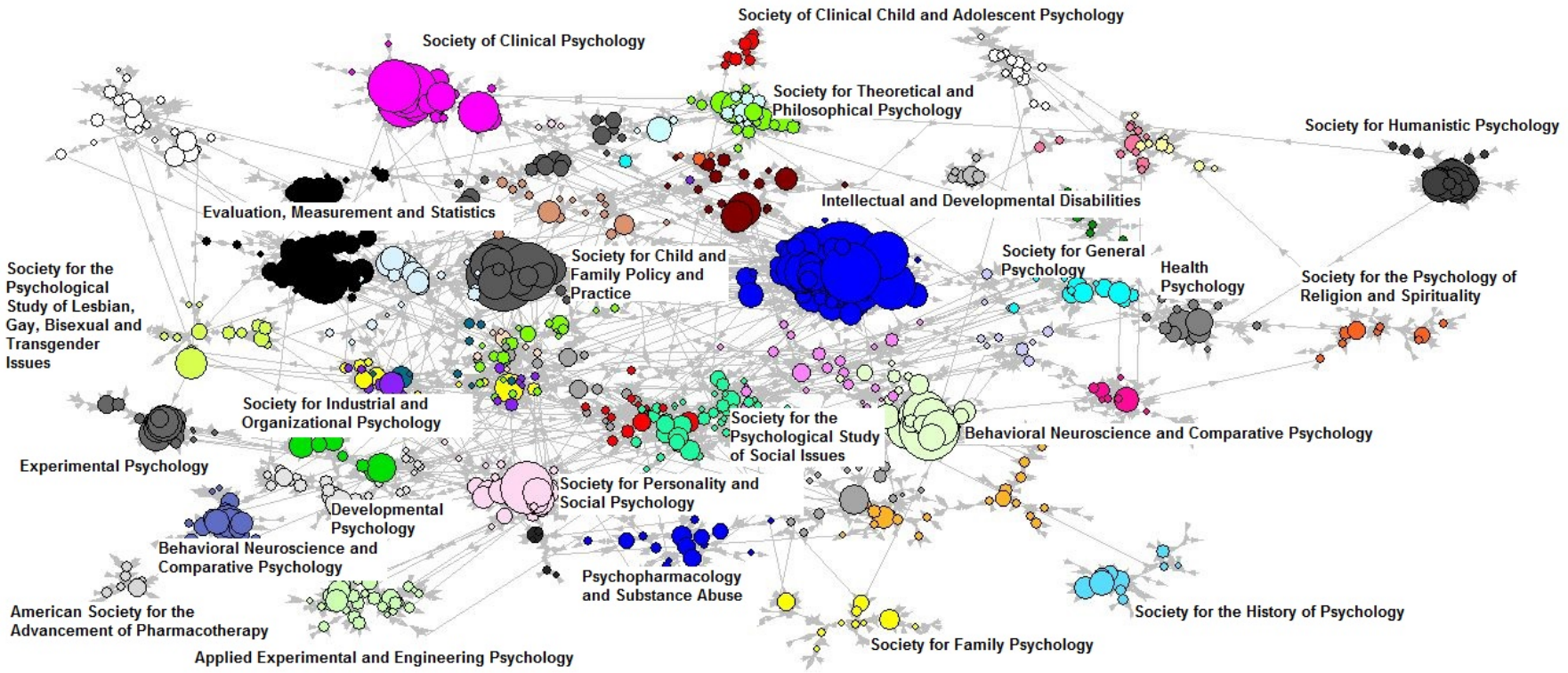


Fig. 1. *Community detection of psychological papers. Louvain method returned a total of 52 communities. The smallest cluster included 7 papers, while the largest cluster was constructed of 230 psychological publications. Vertex size represents the betweenness value of the vertex. Network is visualized with the Kamada-Kawai [3] visualization algorithm. Algorithm was applied twice: firstly for separation of the communities, and then for optimization of the position of vertices within their community. Communities were named based on the measures for cosine similarity between our initial psychological papers and the relevant texts for each of the APA divisions.*

iMauve: A Fast Incremental Model Tree Learner

Denny Verbeeck, Hendrik Blockeel

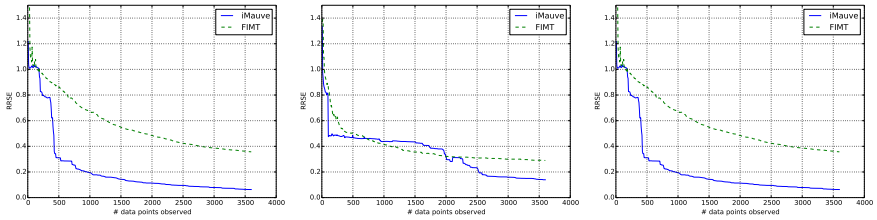
Department of Computer Science, KULeuven,
Celestijnenlaan 200A, 3001 Heverlee, Belgium
{Denny.Verbeeck,Hendrik.Blockeel}@cs.kuleuven.be

We consider the following task: Given a stream of elements of the form $(\mathbf{x}_i, y_i)$, learn a model tree $M : \mathcal{X} \rightarrow \mathcal{Y}$ that predicts y from $\mathbf{x}$. The main difference to the standard learning setting is that data are assumed to arrive sequentially as part of a stream: each data element can be looked at briefly when it arrives, then disappears.

The recent work by Ikonomovska (FIMT) [2] is the current state of the art in learning model trees from streams. Using Hoeffding bounds [1], a split is made only if the splitting heuristic of the best split is better than the second best split with a specified confidence δ. The splitting heuristic FIMT uses is *standard deviation reduction*, as in M5 [3]. Compared to regression trees, which predict a constant value in each leaf, model trees store a linear equation in each leaf that wil be used to make a prediction for instances sorted into this leaf. Thus, when learning model trees from streams, one not only needs to learn a tree structure incrementally, but also the linear regression function in the leaves of this tree. FIMT uses a single perceptron without activation function as the linear model in each leaf. The weights of the perceptron are updated using the Widrow-Hoff rule.

In this work, we propose two main adaptations to Ikonomovska's approach. The first change we propose is to modify the splitting heuristic. As pointed out by Vens and Blockeel [4], the standard deviation reduction heuristic used in many tree learners is sub-optimal when used in combination with linear regression in the leafs. In their work, a heuristic based on simple linear regression, called Mauve, is proposed. This heuristic calculates the reduction in standard deviation of the residuals, assuming a linear model with only the split attribute as a regressor. We have extended Mauve to calculate the residual standard deviation using a full linear least squares regression in all attributes, with no need for random access to the data. The second change we propose is the type of linear regressor used in the leafs of the tree. As the tree grows, the number of data points that each leaf observes decreases exponentially. However the perceptron update rule relies on a large amount of observations to converge the weights of the linear model. As an alternative, we propose a linear least squares regression to estimate the weights of the linear model in each leaf. This approach does not need time to converge, but gives the weights which minimize the sum of squared residuals, given the currently observed data. Like in FIMT, a set of statistics is kept up-to-date when each new data point is observed. The same set of statistics can be used for the split heuristic as well as the linear regression model. We call the resulting algorithm iMauve.

As a result of the adaptations described above we hypothesize: (1) iMauve will learn more compact trees without sacrificing accuracy compared to FIMT, (2) iMauve will reach higher accuracy levels after a lesser amount of observations compared to FIMT, (3) iMauve's higher time complexity will not be prohibitive for problem settings with a reasonable amount of attributes. We set up experiments on different data sets using 10-fold cross-validation. Due to space constraints we here only the results for three datasets, containing 4000 instances each. Paraboloid is a 2-attribute problem, while Lexp and Losc are 5-attribute problems. We measure RRSE during the streaming phase (when the algorithms are receiving data observations). The resulting curves are shown in Figure 1. Tree size and run-time are measured at the end of the run, i.e. when the data set is exhausted. These results are averaged over the 10 runs and shown in Table 1.



Paraboloid Lexp Losc

Fig. 1: Comparison of the evolution of the RRSE for different data sets.

	FIMT Runtime (s)	iMauve Runtime (s)	FIMT Tree Size	iMauve Tree Size
Paraboloid	0.394	0.460	22.7	20.8
Lexp	0.713	1.963	23.7	15.2
Losc	0.725	1.431	22.9	16.5

Table 1: Results averaged over 10 runs

Our experiments confirm these hypotheses on almost all data sets used in the comparison. The trade-off made is that the more advanced split heuristic and regression method makes iMauve the slower algorithm of the two in terms of time taken to process each observation. For a small number of attributes, as is the case in the problems shown here, iMauve processes observations at a rate of 2000 to 8000 examples per second. A moderate sized problem, like the Wine quality dataset which has 11 attributes, is still processed at about 1600 observation per second, making iMauve useful for a broad category of applications. iMauve also has a larger memory footprint due to the increased number of statistics that needs to be kept. However iMauve tends to learn smaller and more accurate trees, while needing less observations to accomplish this.

References

1. Domingos, P., Hulten, G.: Mining high-speed data streams. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 71–80. KDD '00, ACM, New York, NY, USA (2000)
2. Ikonomovska, E., Gama, J., Džeroski, S.: Learning model trees from evolving data streams. Data Mining and Knowledge Discovery 23(1), 128–168 (2011)
3. Quinlan, J.R.: Learning with continuous classes. In: Proceedings of the Australian Joint Conference on Artificial Intelligence. pp. 343–348. World Scientific, Singapore (1992)
4. Vens, C., Blockeel, H.: A simple regression based heuristic for learning model trees. Intell. Data Anal. 10(3), 215–236 (May 2006)

PREDICTIVE PROCESS-BASED MODELING OF AQUATIC ECOSYSTEMS

Nina Vidmar[1], Nikola Simidjievski[2,3], Sašo Džeroski[2,3]
Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia[1]
Jožef Sefan Institute, Ljubljana, Slovenia[2]
Jožef Stefan International Postgraduate School, Ljubljana, Slovenia[3]
e-mail: nina.vidmar@student.fmf.uni-lj.si, {nikola.simidjievski, saso.dzeroski}@ijs.si

ABSTRACT

In this paper, we consider the task of learning interpretable process-based models of dynamic systems. While most case studies have focused on the descriptive aspect of such models, we focus on the predictive aspect. We use multi-year data, considering it as a single consecutive dataset or as several one-year datasets. Additionally, we also investigate the effect of interpolation of sparse data on the learning process. We evaluate and then compare the considered approaches on the task of predictive modeling of phytoplankton dynamics in Lake Zürich.

1 INTRODUCTION

Mathematical models play an important role in the task of describing the structure and predicting the behavior of an arbitrary dynamic system. In essence, a model of a dynamic system consists of two components: a structure and a set of parameters. There are two basic approaches to constructing models of dynamic systems, i.e., theoretical (knowledge-driven) modeling and empirical (data-driven) modeling. In the first, the model structure is derived by domain experts of the system at hand, the parameters of which are calibrated using measured data. In contrast, the later uses measured data to find the most adequate structure-parameter combination that best fits the given task of modeling. In both approaches, models often take the form of ordinary differential equations (ODEs), a widely accepted formalism for modeling dynamic systems, allowing the behavior of the system to be simulated over time.

Equation discovery [1, 2] is the area of machine learning dealing with developing methods for automated discovery of quantitative laws, expressed in the form of equations, from collections of measured data. The state-of-the-art equation discovery paradigm, referred to as process-based modeling [3], integrates both theoretical and empirical approaches to modeling dynamics. The result is a process-based model (PBM) – an accurate and understandable representation of a dynamic systems.

The process-based modeling paradigm has already been proven successful for modeling population dynamics in a number of aquatic ecosystems, such as: lake ecosystems [4, 5, 6, and 7] and marine ecosystems [3]. However, these studies focus on obtaining explanatory models of the aquatic ecosystem, i.e., modeling the measured behavior of the system at hand, while modeling future behavior is not considered. In contrast, Whigham and Recknagel [8] discuss the predictive performance of process-based models in a lake ecosystem. However, either they assume a single model structure and focus on the task of parameter identification, or explore different model structures where the explanatory aspect of the model is completely ignored. The method proposed by Bridewell et.al [9] focuses of establishing robust interpretable process-based models, by tackling the over-fitting problem. Even though this method provides estimates of model error on unseen data, these estimates are not related to the predictive performance of the model, i.e., its ability to predict future system behavior beyond the time-period captured in training data. Most recently, the study of Simidjievski et.al [10] focuses on the predictive performance of process-based models by using ensemble methods. However, while their proposed ensemble methods improve the predictive performance of the process-based models, the resulting ensemble model is not interpretable.

In this paper we tackle the task of establishing an interpretable predictive model of a dynamic system. We focus on predicting the concentration of phytoplankton biomass in aquatic ecosystems. Due to the high dynamicity and various seasonal exogenous influences [6, 7], most often process-based models of such systems are learned using short time-periods of observed data (1 year at most). Note however, this short time-periods of data are very sparse, i.e., consist of very few measured values, thus, most often the measurements are interpolated and daily samples are obtained from the interpolation.

The initial experiments to this end, indicate that the predictive performance of such models is poor: While providing dense and accurate description of the observed behavior, they fail at predicting future system behavior. To address this limitation we propose learning more robust process-based models. We conjecture that by increasing the size of the learning data, more general process-base models will be obtained, thus yielding better predictive performance while maintaining their interpretability.

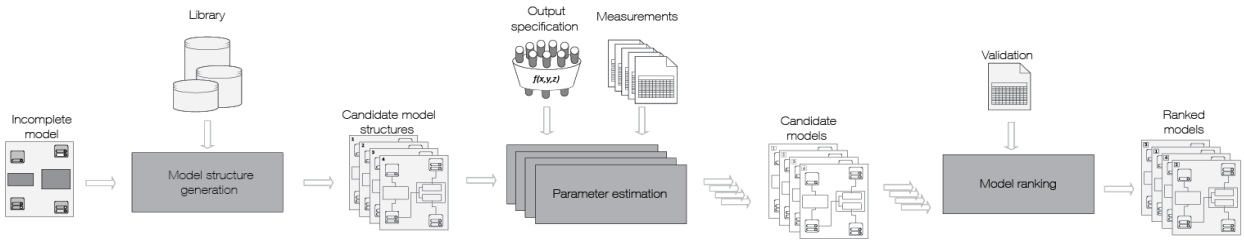


Figure 1: Automated modeling with ProBMoT.

The main contribution of this paper are the approaches to handling the learning data. The intuitive way of increasing the size of the learning data is by sequentially adding predeceasing contiguous datasets, thus creating one long time-period dataset, i.e., learning from sequential data (LSD). In contrast, when learning from parallel data (LPD), the model is learned from all the datasets simultaneously. Figure 2 depicts the both approaches. The two approaches, in terms of learning process-based models, are described in more detail in Section 3.

We test the utility of the two approaches on a series of tasks of modeling phytoplankton concentration in Lake Zürich. We use eight yearly datasets, using six for training, one for validation and one for testing the predictive performance of the obtained models. The aim of this paper is two-fold: besides validating the performance of the two approaches to handling data when learning predictive process-based models, we also test the quality of the training data. For that purpose, we perform additional set of experiments, similar to the previous. However, instead of using the interpolated data for learning the models – we use the original (sparse) measured values, thus examining the influence of the interpolation on the predictive performance of the process-based models.

The next section provides more details of the task of process-based modeling, and introduces a recent contribution to the area of automated process-based modeling, i.e., the ProBMoT [4, 10] platform. Section 3 depicts the task of predictive process-based modeling of aquatic ecosystems. Section 4 describes the data used in the experiments, the design of the experiments, and the task specification. Section 5 presents the results of the experiments. Finally, Section 6 discusses the findings of this paper and suggests directions for future work.

2 PROCESS-BASED MODELING AND PROBMOT

The process-based modeling paradigm, addresses the task of learning process-based models of dynamic systems from two points of view: qualitative and quantitative. The first, provides a conceptual representation of the structure of the modeled system. Still, this depiction does not provide enough details that would allow for simulation of the system's behavior. In contrast, the later, treats the process-based model as a set of differential and/or algebraic equations which allows for simulation.

A process-based model consists of two basic types of elements: entities and processes. Entities correspond to the state of the system. They incorporate the variables and the constants related to the components of the modeled system. Each variable in the entity has its role. The role specifies whether the variable is exogenous or endogenous. Exogenous variables are explanatory/input variables, used as forcing terms of the dynamics of the observed system (and are not modeled within the system). Endogenous variables, are the response/output (system) variables. They represent the internal state of the system and are the ones being modeled. The entities are involved in complex interactions represented by the processes. The processes include specifications of the entities that interact, how those entities interact (equations), and additional sub-processes.

From the qualitative perspective, the unity of entity and processes allows for conceptual interpretation of the modeled system. On the other hand, the entities and the processes provide further modeling details that allow for transformation from conceptual model to equations and therefore simulation of the system, i.e., providing the quantitative abilities of the process-based model. The equations define the interactions represented by the processes including the variables and constants from the entities involved.

The process-based modeling paradigm allows for high-level representation of domain-specific modeling knowledge. Such knowledge is embodied in a library of entity and process templates, which represent generalized modeling blueprints. The entity and process templates are further instantiated in specific entities and processes that correspond to the components and the interactions of the modeled system. These specific model components and interactions define the set of candidate model structures.

The algorithm for inducing models employs knowledge-based methods to enumerate all candidate structures. For each obtained structure, a parameter estimation is performed using the available training data. For this reason each structure is compiled into a system of differential and algebraic equations, which allows for the model to be simulated. In essence, this includes minimizing the discrepancy between the values of the simulated behavior obtained using the model and the observed behavior of the system.

Recent implementations of the PBM approach include Lagrame2.0 [11], HIPM [12] and ProBMoT (Process-Based Modeling Tool) [4, 10], which is next described.

The Process-Based Modeling Tool (ProBMoT), is a software platform for simulating, parameter fitting and

inducing process-based models. Figure 1 illustrates the process of automated modeling with ProBMoT. The first input to ProBMoT is a conceptual model of the modeled system. The conceptual model specifies the expected logical structure of the modeled system in terms of entities and processes that we observe in the system at hand. The second input is the library of domain-specific modeling knowledge. By combining the conceptual model with the library of plausible modeling choices, candidate model structures are obtained.

The model parameters for each structure are estimated using the available training data (third input to ProBMoT). The parameter optimization method is based on meta-heuristic optimization framework jMetal 4.5 [13], in particular, ProBMoT implements the Differential Evolution (DE) [14] optimization algorithm. For the purpose of simulation, each model is transformed to a system of ODEs, which are solved using CVODE ODE solver from the SUNDIALS suite [15].

Finally, the last input, is a separate validation dataset. In both cases (LSD and LPD), the model which has best performance on the validation dataset is the output of automated modeling process.

3 PREDICTIVE PROCESS-BASED MODELING OF AQUATIC ECOSYSTEMS

ProBMoT has been used extensively to model aquatic ecosystems [4, 5, 6]. Most of the case-studies, however, have focused on descriptive modeling – focusing on the content/interpretation of the learned models and not on their accuracy and predictive performance (with the exception of [10]). Predominately, models have been learned from short time-period (one-year) datasets, as considered long time-periods worth of data resulted in models of poor fit. These models, however, had poor predictive power when applied to new (unseen) data.

We use ProBMoT to learn predictive models of aquatic ecosystems from long time-period (multi-year) datasets. ProBMoT supports predictive modeling, as the obtained models can be applied/evaluated on a testing dataset. Taking the input/exogenous variable values from the test dataset, ProBMoT simulates the model at hand, and makes predictions for the values of the output/endogenous (system) variables. Using the output specifications, the values of the output variables of the model are calculated and compared to the output variables from the test set, thus allowing for the predictive performance of the model to be assessed.

Concerning the use of long time-period datasets, ProBMoT supports two different approaches, i.e., learning from sequential data (LSD) and learning from parallel data (LPD). The parameter optimization algorithm uses the available training data from the observed system to estimate the numerical values of the parameters. When learning from sequential data, illustrated in Figure 2a, ProBMoT takes as an input one training dataset. The training dataset is comprised of several contiguous short time-period datasets, thus the parameters are estimated over the whole time-span.

One the other hand, when learning from parallel data, depicted in Figure 2b, ProBMoT takes as an input several short time-period training datasets. The parameter optimization algorithm handles the short time-periods in parallel, i.e., it estimates the optimal model parameters by minimizing the discrepancy between the simulated behavior and each individual training set.

ProBMoT offers wide range of objective functions for measuring model performance such as sum of squared errors (SSE) between the simulated values and observed data, mean squared error (MSE), root mean squared error (RMSE), relative root mean squared error (ReRMSE), which is used in all experiments presented here for when learning the models and evaluating their performance. Relative root mean squared error (*ReRMSE*) [16] is defined as:

$$ReRMSE(m) = \sqrt{\frac{\sum_{t=0}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=0}^{n}(\bar{y} - \hat{y}_t)^2}}, \qquad (1)$$

where n denotes the number of measurements in the test data set, $\hat{y}_t$ and y_t correspond to the measured and predicted value (obtained by simulating the model m) of the system variable y at time point t, and $\bar{y}$ denotes the mean value of the system variable y in the test data set.

The data on the aquatic systems are very sparse (e.g. measure on a monthly basis). In the above mentioned studies, often they have been typically interpolated and sampled at a daily interval. Here, to assess the effect of the interpolation to the performance of the models, we also consider using only the original measured values when establishing the predictive process-based model.

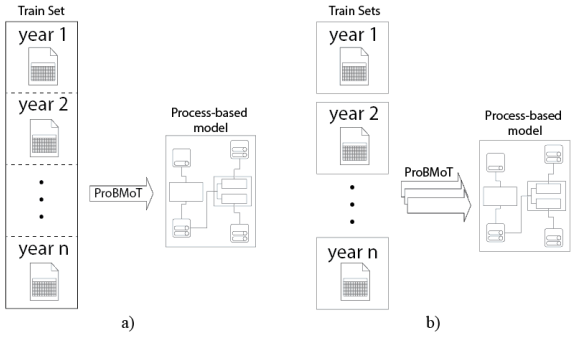


Figure 2: Two approaches to predictive modeling. a) Learning from sequential data (LSD), and b) Learning from parallel data (LPD).

4 EXPERIMENTAL SETUP

In this study, we apply the automated modeling tool ProBMoT to the task of predictive modeling of phytoplankton dynamics in Lake Zürich, Switzerland. We empirically evaluate the two different approaches for learning predictive models, LSD and LPD, on this task. We apply those two on interpolated data (sampled daily) and on the original (sparse) data.

4.1 Data & domain knowledge

The datasets used for our experiments were obtained from the Water Supply Authority of Zürich. Lake Zürich is a lake in Switzerland, extending southeast of the city of Zürich. It has an average depth of 49 m, a volume of 3.9 km^3 and a surface area of 88.66 km^2. The measurements consist of physical, chemical and biological data for the period from 1992 to 1999, taken once a month at 19 different sites, and averaged to the respective epilimnion (upper ten meters) and hypilimnion (bottom ten meters) depths.

The data were interpolated with a cubic spline algorithm and daily samples were taken from the interpolation [17]. Both the original and interpolated data from the first six years were used for training the models (1992-1997), data from year 1998 for validation and data from 1999 to estimate the predictive performance of the learned models.

The population dynamics model considered, consists of one endogenous/output (system) variable and multiple exogenous/input variables structured within a single ODE. The phytoplankton biomass is represented as a system variable, while the exogenous variables include: the concentration of zooplankton, dissolved inorganic nutrients (nitrogen, phosphorus, and silica) and two environmental influences of water temperature and global solar radiation (light).

The library for process-based modeling of aquatic ecosystems used in our experiments, is the one presented by Atanasova [18]. Particularly, to reduce the computational complexity of our experiments, we use a simplified version of the library which results in total of 128 candidate models.

4.2 ProBMoT parameter settings

For the parameter calibration procedure we use Differential Evolution with rand/1/bin strategy, 1000 evaluations over a population space of 50 individuals. For simulating the ODEs we use the CVODE simulator with absolute and relative tolerances set to 10^{-3}. For measuring the model performance we use objective function *ReRMSE*, described in Section 3. To further assess the significance of the differences in performance between the single dataset approach and multiple datasets approach we use Wilcoxon test for statistical significance [19] as presented by Demšar [20].

4.3 Experimental design

In this paper we compare the performance of the two different approaches (LSD and LSP) to learning predictive process-based models. For each approach we learn six process-based models using the available training data of six successive years (1992-1997). For both cases, we start with one short time-period training dataset (year 1997), and continue for five steps adding one preceding year to the training data set. At each step we learn the process-based models accordingly to the two approaches described in the previous section.

First, we apply this two approaches on the interpolated data, or more precisely, daily samples of interpolated data. Second, we apply the two learning approaches to the original (sparse) training data. In all of the experiments the validation dataset (year 1998) and the test dataset (year 1999) remain the same.

5 RESULTS

Table 1 summarizes the performance comparison between models learned from sequential data (LSD) and models learned from parallel data (LPD), using both interpolated (left-hand side) and original (right-hand side) training data. Note that, in both cases, learning from sequential data, yields better predictive performance than learning from parallel data. The results of the Wilcoxon test (in Table 1 below) shows that using LSD is better than using LPD, however, the difference in performance is not substantial nor significant (p-value=0.11).

Table 1: Comparison of the predictive performances (ReRMSE on test data) of models learned from sequential data (LSD) and models learned from parallel data (LPD), from both interpolated and original samples. The numbers in bold indicate the best result for the given years.

Train data (years)	Interpolated		Original	
	LSD	LPD	LSD	LPD
‘97	**1.398**	**1.398**	1.074	1.074
’96-’97	**1.099**	1.391	1.381	1.469
’95-’97	1.006	1.044	**0.984**	1.084
’94-’97	**0.986**	1.094	1.004	1.112
’93-’97	**1.075**	1.109	1.105	1.085
’92-’97	**0.934**	0.998	1.074	0.974
Wilcoxon test	LSD > LPD; p-value = 0.11		LSD > LPD; p-value = 0.11	

Next, as shown in Table 1, using the original measured values when learning the models, did not improve their predictive performance.

Finally, most importantly, from both experiments performed, we can conclude that using large amounts of training data (even interpolated) improves the overall predictive performance of the learned process-based models. Note however, that for one case (’93-97) the performance of the models does not improve. Further investigations are required to determine whether this phenomena is due to the quality of the data of that particular dataset (‘93), or to the dynamics of the system at that particular period significantly differing from the rest.

6 CONCLUSION

In this paper, we tackle the task of learning predictive interpretable process-based models of dynamic systems. In the process of establishing general and robust predictive models, we investigate learning from parallel data (LPD), in contrast to the state-of-the-art approach of learning from sequential data (LSD). We apply the both approaches to the task of modeling phytoplankton dynamics in Lake Zürich, using ProBMoT, a platform for simulating, parameter fitting and inducing process-based models. Additionally, besides validating the performance of the approaches to learning

predictive process-based models, we also test the quality of the training data by learning models from the original measured values, in contrast to learning models from daily samples of interpolated data.

The general conclusion of this paper is that using larger amounts of training data for learning process-based models yields improved predictive performance for tasks of modeling aquatic ecosystems. Both, Atanasova et al [5] and Taškova et al. [6] clearly state that one-year datasets produce models with poor predictive performance. We show that using data from a longer period, considered either consequently (LSD) or parallel (LPD) helps in deriving more general models, and therefore, better predictive models.

Even though the statistical significance comparison shows that the LSD approach has better performance than the LPD approach, the difference in performance is neither substantial nor significant. Nevertheless, when learning from sequential data, due to the mater of simulation and parameter optimization, the available training data considered for learning process-based models should be contiguous. On the other hand, one useful feature of the LPD approach is that can handle missing data (e.g. intermediate period with no measurements) for establishing robust process-based models.

Our empirical evaluation of learning from the original uninterpolated and sampled interpolated data, showed that the interpolation does not affect the performance of the learned process-based models. On the contrary, the models learned using the interpolated values yielded better performance than the ones learned using the original values. We conjecture that this is due to the sparsity of the original measured values (~ 12 time-points per year), which is insufficient to capture the dynamics of such a system. Moreover, considering the relative performance between the two approaches, the LSD approach performed insignificantly better than the LPD approach

Taken all together, some new questions arise for further investigation. How strongly the quality of measurements affects the results? Would the results change significantly in the case of ideal measurements? Considering this, possible directions for further work are as follows. First, performing more experiments using multiple parallel sets of data from different periods and, data from various different lake ecosystems should be used. In order to achieve more controlled experiments, we consider testing the presented approaches on synthetic data, that is, data obtained by simulating a well-established model of an arbitrary aquatic ecosystem. Finally, we would like to extend our approach to different ecosystems and other domains.

Acknowledgements

We would also like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (grant number ICT-2013-612944).

References:

[1] P. W. Langley, H. A. Simon, G. Bradshaw, J. M. Zytkow. Scientific Discovery: Computational Explorations of the Creative Processes. MA: The MIT Press, Cambridge, MA, USA. 1987.

[2] S. Džeroski, L. Todorovski. Learning population dynamics models from data and domain knowledge. *Ecological Modelling 170*, pp. 129–140. 2003.

[3] W. Bridewell, P. W. Langley, L. Todorovski, S. Džeroski. Inductive process modeling. *Machine Learning 71*, pp. 1–32. 2008.

[4] D. Čerepnalkoski, K. Taškova, L. Todorovski, N. Atanasova, S. Džeroski. The influence of parameter fitting methods on model structure selection in automated modeling of aquatic ecosystems. *Ecological Modelling 245* (0), pp. 136–165. 2012.

[5] K. Taškova, J. Šilc, N. Atanasova, S. Džeroski. Parameter estimation in a nonlinear dynamic model of an aquatic ecosystem with meta-heuristic optimization. *Ecological Modelling 226*, pp. 36–61. 2012.

[6] N. Atanasova, F. Recknagel, L. Todorovski, S. Džeroski, B. Kompare. Computational assemblage of Ordinary Differential Equations for Chlorophyll-a using a lake process equation library and measured data of Lake Kasumigaura. In: Recknagel, F.(Ed.), Ecological Informatics. Springer, pp. 409–427. 2006a.

[7] N. Atanasova, L. Todorovski, S. Džeroski, R. Remec, F. Recknagel, B. Kompare. Automated modelling of a food web in Lake Bled using measured data and a library of domain knowledge. *Ecological Modelling 194* (1-3), pp. 37–48. 2006c.

[8] P. Whigham, F. Recknagel, F. Predicting Chlorophyll-a in freshwater lakes by hybridising process-based models and genetic algorithms. *Ecological Modelling 146* (13), pp. 243–251. 2001.

[9] W. Bridewell, N. B. Asadi, P. Langley, L. Todorovski. Reducing overfitting in process model induction. In: Proceedings of the 22nd International Conference on Machine learning. (ICML '05). ACM, pp. 81–88. 2005.

[10] N. Simidjievski, L. Todorovski, S. Džeroski. Learning ensembles of population dynamics models and their application to modelling aquatic ecosystems. *Ecological Modelling (In Press)*. 2014.

[11] L. Todorovski, S. Džeroski. Integrating domain knowledge in equation discovery. In: Džeroski, S., Todorovski, L. (Eds.), Computational Discovery of Scientific Knowledge. Vol. 4660 of Lecture Notes in Computer Science. Springer Berlin, pp. 69–97. 2007.

[12] L. Todorovski, W. Bridewell, O. Shiran, P. W. Langley. Inducing hierarchical process models in dynamic domains. In: Proceedings of the 20th National Conference on Artificial Intelligence. AAAI Press, Pittsburgh, USA, pp. 892–897. 2005.

[13] J. J. Durillo, A. J. Nebro. jMetal: A Java framework for multi-objective optimization. *Advances in Engineering Software 42*, pp. 760–771. 2011.

[14] R. Storn, K. Price. Differential Evolution – A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization 11* (4), pp. 341–359. 1997.

[15] S. D. Cohen, A. C. Hindmarsh. CVODE, a stiff/nonstiff ODE solver in C. Computers in Physics 10 (2), pp. 138–143. Mar. 1996.

[16] L. Breiman. Classification and Regression Trees. Chapman & Hall, London, UK. 1984.

[17] A. Dietzel, J. Mieleitner, S. Kardaetz, P. Reichert. Effects of changes in the driving forces on water quality and plankton dynamics in three swiss lakes long-term simulations with BELAMO. *Freshwater Biology 58* (1), pp. 10–35. 2013.

[18] N. Atanasova., L. Todorovski, S. Džeroski, B. Kompare. Constructing a library of domain knowledge for automated modelling of aquatic ecosystems. *Ecological Modelling 194* (13), pp. 14–36. 2006b.

[19] F. Wilcoxon. Individual comparisons by ranking methods. Biometrics, 1:80–83. 1945.

[20] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research 7*, pp. 1–30. Dec. 2006.

Speed and Accuracy Benchmarks of Large-Scale Microbial Gene Function Prediction with Supervised Machine Learning

Vedrana Vidulin[1], Tomislav Šmuc[1], Fran Supek[1,2]

[1]Division of Electronics, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia

[2]EMBL/CRG Systems Biology Unit, Centre for Genomic Regulation, Dr. Aiguader 88, 08003 Barcelona, Spain

1 Introduction

Machine learning approaches for microbial gene function prediction (MGFP) from genome context data are mostly unsupervised [1] and rely on pairwise distances between individual examples arranged into "functional interaction networks" [2]. When supervised approaches were used, most of them typically predicted a limited set of functions and/or used a single-label approach to classification [3, 4], constructing a separate classifier for each function and ignoring the relationships between the functions. Multi-label approaches may perform better, especially those that can exploit the relations between functions readily available in gene function ontologies [5].

Our aim is to compare predictive accuracy and computational efficiency of single *vs.* multi-label approaches on supervised MGFP. High accuracy is a prerequisite for applying the classifier in real-life tasks, where confidence in predicted functions is of key importance for prioritizing downstream experimental work. Many such predictions have indeed been validated in biological experiments [6, 7]. A lower demand for computational time is of importance when the number of considered functions is high.

2 Data and Experimental Setup

We collected microbial genome data from three databases (NCBI Genome, eggNOG [8] and Gene Ontology – GO) and constructed separate data sets for three types of microbial data representations [1]: (1) the phyletic profiles (PP [9]) data set represents co-occurrences of clusters of orthologous groups (COG) across multiple genomes and contains 6018 examples representing COGs, 1690 attributes representing genomes, and binary values representing COGs' presence/absence in the genomes; (2) the translation efficiency profiles (TEP [10]) data set indicates COGs' predicted expression levels across the genomes and differs from PP in attribute values – for present COGs expression levels are measured with MELP measure [11] and absent COGs are represented with missing values (providing a representation orthogonal to PP); (3) in the conserved gene neighborhood (CGN) data set both attributes and examples represent COGs and attribute values represent relative pairwise distances between COGs (distances are averaged over all genomes where both genes of a pair are present). In all three data sets, the classes are functions taken from the controlled vocabulary of GO. When a COG is labelled with a function at the lower level of the GO, then all superordinate functions are attributed too. In our data sets, we accounted for a subset of 776 GO functions represented with 50 or more examples (COGs).

Classifiers were constructed with four algorithms: CLUS-HMC [12], Fast Random Forest (FRF [13]), k-nearest neighbors (kNN) and Naive Bayes (NB) (from Weka 3.7.11 [14]). Both CLUS and FRF construct a random forest ensemble of decision trees. However, CLUS is hierarchical multi-label classification algorithm, which is aware of the hierarchical relationships between GO functions, while FRF is a single-label algorithm. In both cases, ensembles of 500 trees were constructed. A single-label kNN (with k of 1 and inverse distance weighting) was selected as an algorithm that represents the closest approximation to the unsupervised MGFP approaches. We selected a single-label NB (with default Weka parameters) since it is typically faster on big data in comparison to other algorithms. A wrapper that creates 776 binary-class data sets, one for each function, and computes multiple classifiers was built around the single-label algorithms. Predictive accuracy was evaluated using 10-fold cross-validation and measured as an area under the precision-recall curve (AUPRC).

3 Results and Discussion

The results of comparisons on a subset of 776 GO functions belonging to the "biological process" GO namespace are presented in Figure 1. We have also made comparisons on subsets of functions from the other two GO subhierarchies, molecular function and cellular component, and they exhibit the same general behavior.

Accuracy of the two versions of random forest algorithms, that is, FRF and CLUS, significantly outperforms kNN and NB over all representations and GO functions (Wilcoxon signed-rank test $p<2.2 \times 10^{-16}$). Ensembles, here represented by the random forest classifiers, are often used in biological domains [15] since they capture complex relations in data while being robust to overfitting.

kNN exhibits better performance than NB for PP and CGN. For TEP, in most of the cases, NB performs better. TEP data set has a large proportion of missing values (77%) in comparison to the other two data sets (PP is without missing values, CGN has 18%). While NB ignores missing values while determining the priors, kNN sets the distance to maximum if one of the two attributes being compared has a missing value. With that strategy, kNN decreases classifier's accuracy with the increase of missing values.

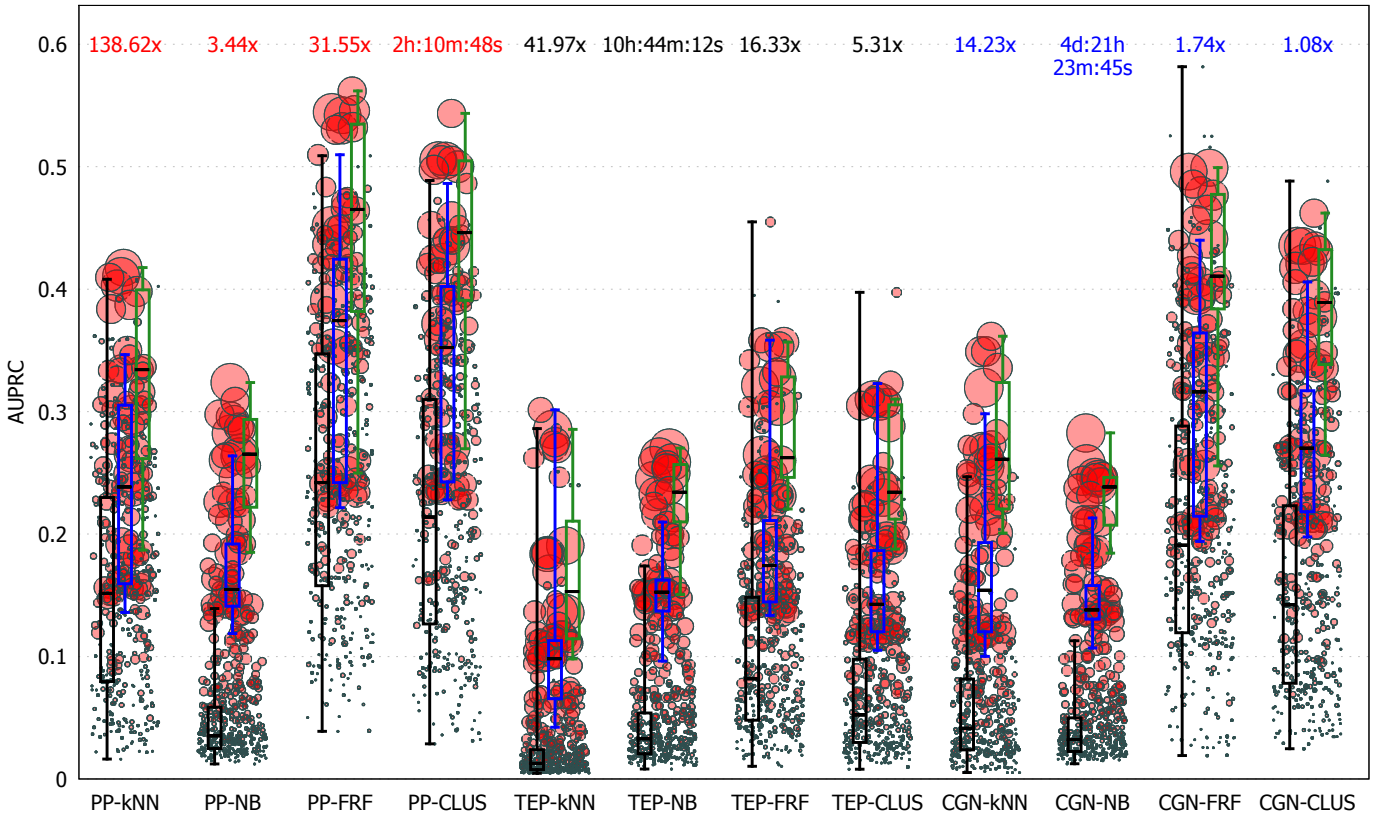


Figure 1: Results of comparisons: The names of classifiers on x-axis are composed of representation and algorithm name abbreviations. Circles represent GO functions and their size denote GO categories' frequencies. Very general functions (frequency $> 30\%$) are omitted. The three box-plots for each classifier represent classifier's performance on small (frequency $<= 10\%$ – black), medium ($10\% <$ frequency $<= 20\%$ – blue) and large ($20\% <$ frequency $<= 30\%$ – green) categories. Box-plots represent minimum, first quartile, median, third quartile and maximum accuracy. Execution times are given above the box-plots, where comparisons for each representation are marked in different color. The fastest execution times are expressed in absolute numbers, while the other executions times are expressed relative to the fastest times.

In single to multi-label comparisons of the two ensemble classifiers, FRF outperformed CLUS (Wilcoxon signed-rank test $p<2.2 \times 10^{-16}$). This result suggests that CLUS does not fully benefit from the hierarchical relationships between the functions in this particular setup, although it has been shown to do so in a non-ensemble (single tree) applications to other datasets [5]. This agrees with research showing that hierarchical and/or multi-label setups are consistently beneficial for predictive accuracy of decision trees across various datasets, but this advantage is less evident with tree ensembles [16]. Consistently, a previous application of CLUS-HMC ensembles to phylogenetic profiling found no effect of varying the weights of GO hierarchy levels, implying that the ensemble did not draw on the information contained in the hierarchy [7].

In computational time requirement comparisons, CLUS has an advantage over FRF: FRF needs 31.55× more time than CLUS to compute the classifier for PP, 3.08× for TEP and 1.61× for CGN). Please note that in our setup FRF computed 10 trees in parallel, while CLUS was computing the trees one after another.

In the case of TEP and CGN, NB has the shortest execution time. The advantage of NB over the random forest approaches is higher for TEP than for CGN. The main cause is in handling missing values: decision trees, when testing on an attribute with a missing value, fragment an instance, while NB ignores the missing values. While TEP has 77% of missing values, in the case of CGN with 18% of missing values CLUS has almost comparable execution time (1.08× slower than NB). In the case of PP, which are without missing values, CLUS performs better that NB (3.44× faster than NB).

CLUS and FRF have the shortest execution time on PP, since, in contrast to TEP and CGN, PP has categorical attribute values. When splitting on a categorical attribute, a decision tree tests one possible split, while when splitting on a numeric attribute several possible splitting points are tested.

In the presented setup, kNN has the longest execution time for all representations because the construction of each classifier was treated as a separate process. The kNN execution time can be reduced by computing the distance matrix once and reusing it in all processes.

In conclusion, the best choice for MGFP is CLUS when a trade-off between predictive accuracy and computational time demands are considered. When the aim is to strictly achieve the highest possible predictive accuracy, FRF may be preferrable. NB is generally the fastest, but constructs classifiers of considerably lower quality that are not suitable candidate for real-life MGFP applications.

Acknowledgement

We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

[1] V.Y. Muley and V. Acharya. *Genome-wide prediction and analysis of protein-protein functional linkages in bacteria*. Springer, 2013.

[2] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl 1):D561–D568, 2011.

[3] F.P.Y. Lin, E. Coiera, R. Lan, and V. Sintchenko. In silico prioritisation of candidate genes for prokaryotic gene function discovery: an application of phylogenetic profiles. *BMC bioinformatics*, 10(1):86, 2009.

[4] P. Hu, S.C. Janga, M. Babu, J.J. Díaz-Mejía, G. Butland, W. Yang, O. Pogoutse, X. Guo, S. Phanse, P. Wong, et al. Global functional atlas of escherichia coli encompassing previously uncharacterized proteins. *PLoS biology*, 7(4):e1000096, 2009.

[5] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.

[6] P.R. Kensche, V. van Noort, B.E. Dutilh, and M.A. Huynen. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *Journal of The Royal Society Interface*, 5(19):151–170, 2008.

[7] N. Škunca, M. Bošnjak, A. Križko, P. Panov, S. Džeroski, T. Šmuc, and F. Supek. Phyletic profiling with cliques of orthologs is enhanced by signatures of paralogy relationships. *PLoS computational biology*, 9(1):e1002852, 2013.

[8] S. Powell, K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, T. Gabaldón, T. Rattei, C. Creevey, M. Kuhn, et al. eggnog v4. 0: nested orthology inference across 3686 organisms. *Nucleic acids research*, page gkt1253, 2013.

[9] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

[10] A. Križko, T. Čopić, T. Gabaldón, B. Lehner, and F. Supek. Inferring gene function from evolutionary change in signatures of translation efficiency. *Genome biology*, 15(3):R44, 2014.

[11] F. Supek and K. Vlahoviček. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC bioinformatics*, 6(1):182, 2005.

[12] Clus. `http://dtai.cs.kuleuven.be/clus`, 2014.

[13] Fast random forest. `https://code.google.com/p/fast-random-forest`, 2014.

[14] Weka 3: Data mining software in Java. `http://www.cs.waikato.ac.nz/ml/weka`, 2014.

[15] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, and S. Džeroski. Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC bioinformatics*, 11(1):2, 2010.

[16] J. Levatić, D. Kocev, and S. Džeroski. The importance of the label hierarchy in hierarchical multi-label classification. *Journal of Intelligent Information Systems*, 2014.

Identifying creative fictional ideas

Martin Žnidaršič[1,2], Senja Pollak[1], Dragana Miljković[1], Janez Kranjc[1,2], Nada Lavrač[1,2]

[1] Jožef Stefan Institute, Jamova ulica 39, SI-1000 Ljubljana, Slovenia.
name.surname@ijs.si,
[2] Jožef Stefan International Postgraduate School, Jamova cesta 39, SI-1000 Ljubljana, Slovenia.

Abstract. Ideas are the main driving force of creative work, so it is important to know how to detect or generate them. People can usually identify a creative idea when they see it. However, the assessment and identification of the best and most valuable creative ideas is a difficult problem that justifies the employment of experts for such tasks in many professional environments. For computers, the evaluation of creativity is even harder, as it requires common knowledge and awareness of the context of each specific use case. Computers currently provide little support in evaluation and detection of creative ideas, and even less in their generation or refinement. However, thanks to the recent developments in terms of computing power, resources and methodologies, first promising results were reported and attracted a wider scientific attention [3, 2].

We are targeting the assessment of fictional *what-if*[1] type of ideas, which can be used in advertising, media, art and other creative industries. The goal of our work is to build a human evaluation model of creativity in this context. The (computational) creativity is usually measured in terms of novelty and usefulness or quality [1]. In our case, the usefulness corresponds to the narrative potential of a fictional idea.

Development of the evaluation model of creative processes and artefacts is tightly connected to discovery of characteristics of the artefacts that are perceived as creative. We are thus focusing on discovery of relations between the characteristics of fictional sentences, evaluators' characteristics and the evaluation scores. The results of this work can help the systems for automatic generation of (fictional) ideas and can contribute to better understanding of the human perception of creativity.

We will address this task with data analysis and pattern discovery methods on databases of human-assessed textual *what-if* ideas, which are produced either by humans or by computers. In the latter case, the features of the idea generation processes can also be considered as features of the evaluation models. This can be particularly useful, as it enables the evaluation models to assess and guide the generation processes.

We are considering syntactic and semantic features of *what-if* ideas. On the syntactic level we are considering features, such as: part-of-speech information, negation, modality markers, number of adjectives, sentence length and rhymes. On the semantic level we are considering features

[1] Example: *What if dragons drove you to school?*

related to semantic resources, for example various distances in semantic networks, such as ConceptNet[2]. Other semantic features that we intend to use are: the bag of words, novelty, actuality, sentiment, ambiguity, presence of fictional characters, named entities and others. While for the calculation of some of these features there exist several proven methodologies (e.g., for sentiment classification), for some others (e.g., ambiguity) we will have to develop and test new techniques for their assessment. We are presenting the first steps of our work: the framework and infrastructure for opinion gathering and the preliminary insights from pilot experiments. The gathered opinions of human evaluators on the *what-if* ideas will be used as the basis for the development of human evaluation model of creativity. We are gathering data in two distinct ways: (I) through an online opinion gathering platform, which targets the opinions of the general public and (II) with targeted questionnaires, which are controlled and focus on the opinions of selected target groups in specific experiments.

The presentation of *what-if*s and the assessment characteristics on the online platform suit the online context and the expected behavior of online evaluators. The evaluation procedure thus favors simplicity of use and clarity of the interface over the thoroughness and completeness of the assessments. The first version of the platform is available at:
http://www.whim-project.eu/whatifmachine2/

It allows the users to set the main input parameters for the current *what-if* sentence generation machine and run it to get the results. The sentences are presented in an environment for evaluation, tailored to motivate the user participation.

Targeted questionnaires are a form of opinion gathering that we use for specific experiments, which demand elaborate questions and a specific amount of opinions gathered in a limited amount of time. These surveys are conducted on groups of people, which we can control regarding the number of participants and their characteristics.

Preliminary results of both kinds of experiments indicate a weak inter-annotator agreement and limited abilities of evaluation modelling with the current set of observed features.

This research was supported through EC funding for the project WHIM 611560 by FP7, the ICT theme, and the Future Emerging Technologies FET programme.

References

1. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17(1):67–99, 2007.
2. Simon Colton, editor. *Proceedings of the 5th Int. Conference on Computational Creativity, ICCC-14*, 2014.
3. Mary Lou Maher, Tony Veale, Rob Saunders, and Oliver Bown, editors. *Proceedings of the 4th Int. Conference on Computational Creativity, ICCC-13*, 2013.

[2] http://conceptnet5.media.mit.edu/

SPONSORS:

Univerza *v Ljubljani*