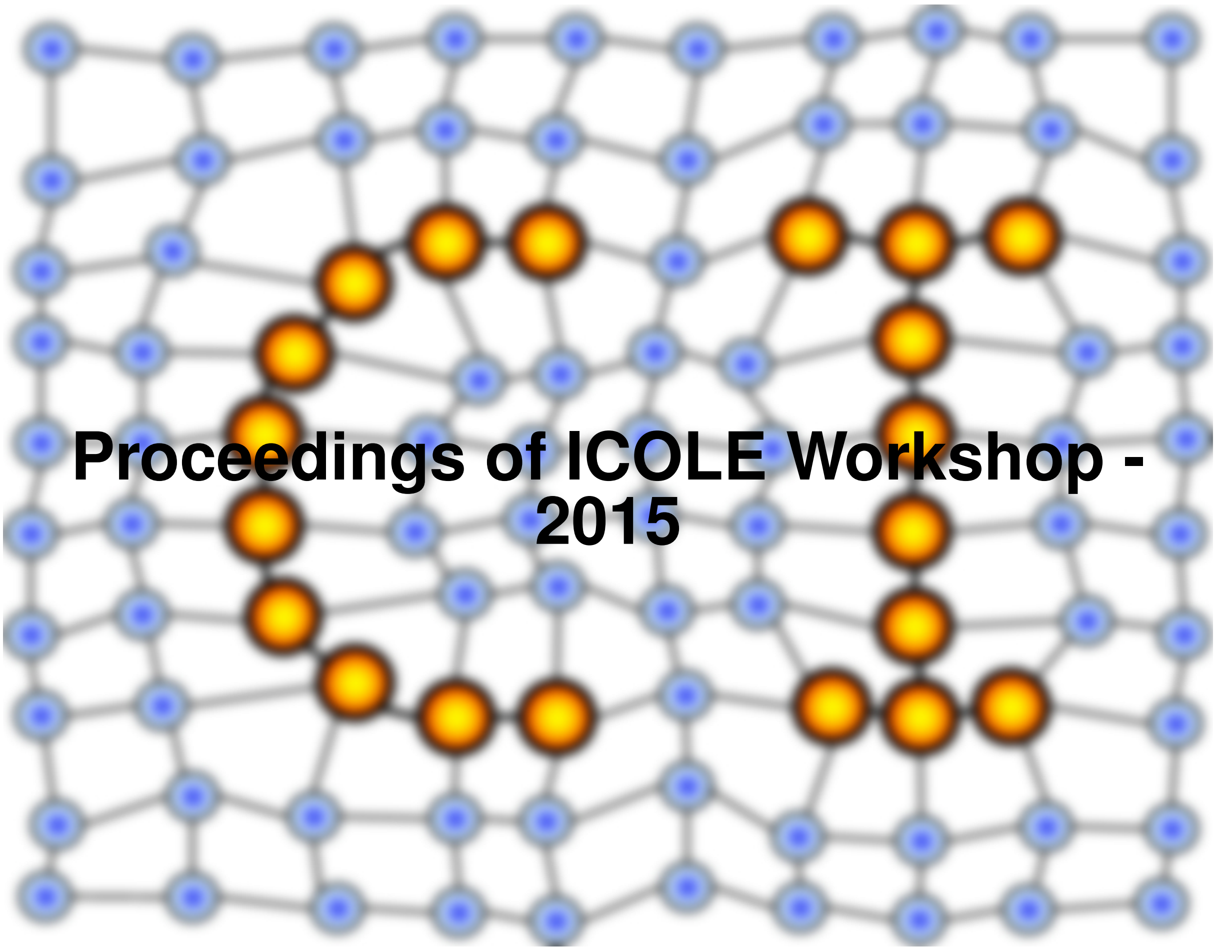
MACHINE LEARNING REPORTS



Proceedings of ICOLE Workshop - 2015

Machine Learning Report 04/2015

Jacek Blazewicz, Klaus Ecker, and Thomas Villmann

Machine Learning Reports
http://www.techfak.uni-bielefeld.de/∼fschleif/mlr/mlr.html

Preface

The annual Polish-German workshop on Computational Biology, Scheduling, and Machine Learning (ICOLE 2015) gathering together nineteen scientists from different universities including Poznań University of Technology, University of Applied Sciences Mittweida, and Bielefeld University. The workshop took place in Lessach, Austria, from 28.9. - 3.10.2010 and continued the tradition of scientific presentations, vivid discussions, and exchange of novel ideas at the cutting edge of research connected to diverse topics in bio-informatics, scheduling, and machine learning.

This report is a collection of abstracts and short contributions about the given presentations and discussions, which cover theoretical aspects, applications, as well as strategic developments in the fields.

Marika Kaden (University of Applied Sciences Mittweida)

Contents

Connections between Sequencing by Hybridization and Next-Generation Sequencing

Jacek Blazewicz[1,2,3], Aleksandra Swiercz[1,2,3,*]

[1]Institute of Computing Science, Poznan University of Technology, Poland
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poland
[3]European Center of Bioinformatics and Genomics, Poznan, Poland
[*]corresponding author: aleksandra.swiercz@cs.put.poznan.pl

Recognition of the genome sequence has been a challenge for many years. Historically, one of the first methods reconstructing a DNA sequence from shorter fragments was proposed in 1988 by Lysov et al. [1]. The method is related to the sequencing by hybridization (SBH), the approach which did not stand the test of time but initiated the industry of microarrays. The goal of the method is to reconstruct the original sequence on the basis of a set of shorter fragments (called *k*-mers, where *k* stands for their length) composing the sequence. Lysov and co-authors modeled the problem as the Hamiltonian path (which is strongly NP-hard), where *k*-mers corresponded to vertices of a directed graph. A computationally easier solution of the same problem was provided by Pevzner in 1989 [2], who proposed a graph construction suitable for the searching for an Eulerian path, the problem solvable in polynomial time. There, *k*-mers corresponded to arcs in a directed graph. The equivalence of these two models is true for all labeled digraphs, with de Bruijn graphs being one of the examples. De Bruijn graphs are “complete” in the sense of labeling – their vertices are associated with all possible labels of a given length over a given alphabet. Lysov graphs, which are called DNA graphs, are vertex-induced subgraphs of de Bruijn graphs labeled over four-letter alphabet.

In the last decade we could observe the evolution of next generation sequencing (NGS) which rapidly increased the cognition of novel genomes. The sequencers are capable of producing massively parallel billions of short DNA sequences, called reads, in no longer than a few days and at far lower costs than Sanger sequencing. These reads come from random positions on the target genome sequence. The computational challenge is to reconstruct the target sequence using only the reads (this process is called de novo assembly). The assembly problem can be seen as a more complicated version of SBH, however it has to tackle also additional problems like sequencing errors, double-stranded DNA and repetitions of fragments in the target. The graph approaches developed for SBH have been exploited in many algorithms for de novo assembly, the ones using the idea of decomposition reads into series of *k*-mers and creating a Pevzner-like graph where reads are represented by paths [3]. Such approach is called in the literature the de Bruijn graph approach. However, in the presence of sequencing errors the reconstruction problem is no longer polynomially solvable, unlike the Pevzner's algorithm working for the errorless input. The problem becomes strongly NP-hard and needs time- and memory-efficient heuristics capable of processing billions of reads within hours. The same concerns the other approach "overlap-layout-consensus" (OLC), where the overlap graph is used, based on the concept of the DNA graph [4,5]. Its vertices represent the input reads and the arcs represent overlaps of the reads. This time the overlaps are allowed to be non-exact and the resulting path(s) may not pass through all the vertices. As the result of the assembly we get a collection of contigs, i.e. contiguous fragments of the genome sequence.

[1] Y. Yu.P. Lysov, V. Florentiev, A. Khorlin, K. Khrapko, V. Shik, and A. Mirzabekov, “Determination of the nucleotide sequence of DNA using hybridization with oligonucleotides. A new method” *Doklady Akademii Nauk SSSR*, vol. 303, pp. 1508-1511,1988.

[2] P. Pevzner, “l-tuple dna sequencing: computer analysis” *Journal of Biomolecular Structure and Dynamics*, vol. 7, pp. 63-73, 1989.

[3] P.A. Pevzner, H. Tang, and M.S. Waterman. A new approach to fragment assembly in DNA sequencing. In Proc. 5th Ann. Inter. Conf. Res. Comput. Molecular Biology (RECOMB), pages 256–267, Montreal, 2001. ACM Press.

[4] T. Jiang and M. Li. DNA sequencing and string learning. *Mathematical Systems Theory*, 29:387–405, 1996.

[5] J. Blazewicz, M. Figlerowicz, P. Gawron, M. Kasprzak, E. Kirton, D. Platt, A. Swiercz, L. Szajkowski, "Whole genome assembly from 454 sequencing output via modified DNA graph concept" *Computational Biology and Chemistry* 33 no.3 , 2009, pp. 224-230.

Model for the audience forecasting in TV advertisement

Maciej Drozdowski, Krzysztof Odasz, Grzegorz Pawlak Małgorzata Sterna

Institute of Computing Science, Poznan University of Technology, Poland

The aim of the research was developing an algorithm for optimizing advertising campaigns according to the predicted TV viewership during the commercial breaks. Results will be used to maximize the personalization and impact of the ads and reduce the costs for advertisers.

In order to provide effective viewership prediction, historical data was analyzed. There are many factors influencing the viewership. The study was based on the data provided by AGB Nielsen Media Research and included:

- telemetry data – the demographic information on panelists, as well as the information on time periods in which particular panelist watched particular TV channels,
- descriptions of TV programs and commercial breaks – the daily TV schedule, the information covered all programs broadcast on all channels,
- descriptions of spots within commercial breaks – very detailed historical data on the aired spots.

The analysis of the viewership requires integration of these sources into one data set.The research also encompassed the following points:

- analysis of the daily viewership of a channel,
- identifying factors that determine probability of watching some program at certain time by a person in a sample,
- identifying factors that determine volatility of watching some program at certain time by a person in a sample,
- analysis of the prediction errors for various methods of viewership estimation.

A model forecasting the audience in TV advertisement periods was based on a factorial design which is a kind of experimental design. In this method certain types of experiment scenarios are designed in such a way that a mathematical model of the analyzed object response can be easily derived from the collected data. Examples of the accuracy of the audience prediction according to the factorial model have been provided.

Scope of the optimization phase, which is the next step of the research, is as follows:

- developing data structures of the genetic optimization code,
- proposing algorithms for converting the code into broadcasting parameters and advertisement broadcasting plan,
- working out algorithms for choosing a cost-effective list of advertisement expositions in commercial breaks,
- establishing and tuning the goal function which should integrate the daily reach of the advertising campaign, given a distribution of strength of the campaign reach in the exposition frequency and gaps between viewer-advertisement contacts.

Application of chemical reaction simulation methods in order to verify RNA World hypothesis

Jaroslaw Synak[1*], Natalia Szostak[1,3], Szymon Wasik[1,3], Jacek Blazewicz[1,2,3]

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland
[3]European Center for Bioinformatics and Genomics, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland

**presenting author, e-mail: j_synak@wp.pl*

The RNA World hypothesis [1] is the most popular and well substantiated theory which tries to explain origins of life on Earth. It is a collaborative effect of work of many scientists who try to guess possible reaction paths and mechanisms which resulted in life in the form that we know today. Taking into account an enormous space of possible reactions and an uncertainty of primordial conditions, these attempts require effective methods of verification and estimation of various possibilities. Such methods should be fast, based on existing knowledge and, at the same time, as accurate as possible.

This presentation demonstrates a variety of algorithms which can be used to simulate interactions between molecules on prebiotic Earth [2], based on our three years long experience on simulating such systems. We underline the accurateness and efficiency of each method. Moreover we present a new approach that treats RNA chains as regular chemical molecules. RNAs diffuse, react with other RNAs, replicate and are the subject of mutations. Some variants of the model built on this assumption have been implemented as computer simulations. Based on these simulations we were able to infer first, biologically significant conclusions regarding evolutionary dynamics of first living systems and verify already formulated hypothesis. [3]

[1] T. R. Cech, “The RNA Worlds in Context,” *Cold Spring Harb. Perspect. Biol. 2012,* vol. 4, no. 7, p. a006742.

[2] K. Takahashi, S. N. V. Arjunan, and M. Tomita, “Space in systems biology of signaling pathways--towards intracellular molecular crowding in silico,” *FEBS Lett. 2005*, vol. 579, no. 8, pp. 1783–1788.

[3] N. Takeuchi and P. Hogeweg, “Evolutionary dynamics of RNA-like replicator systems: A bioinformatic approach to the origin of life,” *Phys. Life Rev. 2012,* vol. 9, no. 3, pp. 219–263.

This work has been partially supported by the DS grant of the Institute of Computing Science, PUT.

Reject Options for Learning Vector Quantization

Lydia Fischer

Abstract

Due to intuitive training algorithms and model representation, prototype-based models are popular in settings where on-line learning and model interpretability play a major role. In such cases, a crucial property of a classifier is not only which class to predict, but also if a reliable decision is possible in the first place, or whether it is better to reject a decision. While strong theoretical results for optimal reject options in the case of known probability distributions or estimations thereof are available, there do not exist well-accepted reject strategies for deterministic prototype-based classifiers. In this contribution, we present simple and efficient distance-based reject options for prototype-based classification, and we evaluate their performance on artificial and benchmark data sets using the example of learning vector quantization. We demonstrate that the proposed reject options improve the accuracy in most cases, and their performance is comparable to an optimal reject option of the Bayes classifier in cases where the latter is available. Further, we show that the results are comparable to a well established reject option for support vector machines in cases where learning vector quantization classifiers are suitable for the given classification task, even providing better results in some cases. Extending this so called global reject option towards local rejection enables the method to include knowledge about local characteristics of the input space. Therefore, we analyse optimal reject strategies for classifiers with input space partitioning, e. g. prototype-based classifiers. We compare reject schemes with global thresholds, and local thresholds for the partitions of the space induced by the classifiers. For the latter, we develop a polynomial-time algorithm to compute optimal thresholds based on a dynamic programming scheme, and we propose an intuitive linear time, memory efficient approximation thereof with competitive accuracy. Evaluating the performance in various benchmarks, we conclude that local rejection is beneficial in particular for simple classifiers, while the improvement is less pronounced for advanced models. An evaluation for biomedical data highlights the benefit of local thresholds. To connect the distance-based measures with probabilistic counterparts obtained from probabilistic models, e. g. robust soft learning vector quantization or Gaussian mixture models, we conducted experiments on artificial and benchmark data to compare the reject option of probabilistic measures against the rejection based on distance-based measures. It turned out that both rejection variants leads to similar results which allows to simply take the distance-based measures since they are easier to calculate.

From High-level Mathematical Equations to Large-scale Distributed Hybrid Computing: Introduction To CaKernel/Chemora

Marek Blazewicz[1,2,*]

1 Introduction

Starting from a high-level problem description in terms of partial differential equations using abstract tensor notation, the *Chemora* framework discretizes, optimizes, and generates complete high performance codes for a wide range of compute architectures. Chemora extends the capabilities of Cactus, facilitating the usage of large-scale CPU/GPU systems in an efficient manner for complex applications, without low-level code tuning. Chemora achieves parallelism through MPI and multi-threading, combining OpenMP and CUDA. Optimizations include high-level code transformations, efficient loop traversal strategies, dynamically selected data and instruction cache usage strategies, and JIT compilation of GPU code tailored to the problem characteristics. Chemora's capabilities have been demonstrated by simulations of black hole collisions.

2 Chemora Framework

Chemora [1] takes a physics model described in a high level and generates highly optimized code suitable for parallel execution on heterogeneous systems. There are three major components in Chemora:

- **Cactus-Carpet computational infrastructur** – is an open-source, modular, highly-portable programming environment for collaborative research using high-performance computing. formats [2].

- **CaKernel programming abstractions** – a set of high level programming abstractions, and the corresponding implementations [3]. CaKernel allows easly to implement code that, depending on the changeable metadata, can be executed on different architectures.

- **Kranc code generator** – Translates equations from a high-level mathematical notation into C or Fortran and discretizes the domain [4]

3 Results

We used part of the binary black hole simulation as a weak-scaling performance benchmark. We chose a local problem size that fitted into the GPU memory, corresponding to 100^3 evolved points plus boundary and ghost zones. The simulation was ran on 40 nodes, we observed 50% scaling efficiency due to synchronization overhead, and achieved a total performance of 500 GFlop/s, what was two times faster than simulation performed in similar configuration on nodes with twelve-core AMD processors. The code was fully generated out of high level mathematical equations.

Acknowledgments

The presented work was a joined research with: Ian Hinder[3], David M. Koppelman[4], Steven R. Brandt[4], Milosz Ciznicki[1], Michal Kierzynka[1,2], Frank Löffler[4], Erik Schnetter[5,4] and Jian Tao[4].

References

[1] Marek Blazewicz, Ian Hinder, David M. Koppelman, Steven R. Brandt, Milosz Ciznicki, Michal Kierzynka, Frank Löffler, Erik Schnetter, and Jian Tao. From physics model to results: An optimizing framework for cross-architecture code generation. *Scientific Programming*, 21:1–16, 2013.

[2] T. Goodale, G. Allen, G. Lanfermann, J. Massó, T. Radke, E. Seidel, and J. Shalf. The Cactus framework and toolkit: Design and applications. In *Vector and Parallel Processing – VECPAR'2002, 5th International Conference, Lecture Notes in Computer Science*, Berlin, 2003. Springer.

[3] Marek Blazewicz, Steven R. Brandt, Michal Kierzynka, Krzysztof Kurowski, Bogdan Ludwiczak, Jian Tao, and Jan Weglarz. Cakernel - a parallel application programming framework for heterogenous computing architectures. *Scientific Programming*, 19(4):185–197, 2011.

[4] Sascha Husa, Ian Hinder, and Christiane Lechner. Kranc: a Mathematica application to generate numerical codes for tensorial evolution equations. *Comput. Phys. Comm.*, 174:983–1004, 2006.

[*1] Applications Department, Poznań Supercomputing & Networking Center, Poznań, Poland

[†2] Poznań University of Technology, Poznań, Poland

[‡*] Email: marek.blazewicz@cs.put.poznan.pl

[3] Max-Planck-Institut für Gravitationsphysik, Albert-Einstein-Institut, Postdam, Germany

[4] Center for Computation & Technology, Louisiana State University, Baton Rouge, LA, USA

[5] Perimeter Institute for Theoretical Physics, Waterloo, ON, Canada

EMAES: Order Completion and Delivery Problem

Mateusz Cichenski, Mateusz Jarus, Michal Miszkiewicz, Malgorzata Sterna, Jaroslaw Szymczak

Keywords computer applications · genetic algorithms · scheduling · order completion and delivery problem

Many real world problems inspire people working in the field of scheduling theory to create mathematical models that solve them efficiently (Blazewicz et al., 2007). However, using existing solvers for linear programming problems is not always providing the solutions quickly enough. There are problems for which one needs good enough solution rather than optimal one.

One such problem has been identified in the charity organizations supply process (Cichenski et al., 2015), in which the organization buys soon to be expired products at much lower prices and delivers them to places in need. An example of such goods could be vaccines, which are going to expire soon or food that will go bad if not eaten. Therefore, one has to consider the distance travelled by the delivery vehicle, but also the prices of the goods that are collected. This can be formulated as order completion and delivery problem or as a variant of traveling purchaser problem (Ramesh, 1981).

To provide an easy way for the organizations to solve their problem an on-line tool has been developed, namely EMAES. It is a website that allows the organization to keep their inventory and request goods from other depots or places that are registered in the system. It also allows to put items that are going to be expired ”on the market”, informing others that they can bought them at lower prices. Registered users can create lists of items they want to buy and the system will try to find the best solution to that particular instance of order completion and delivery problem, minimizing the traveling costs and the total price for items.

This on-line tool has been implemented with the idea that the user might just need good enough solution and want to view possible solutions as soon as possible. Therefore, a genetic algorithm (Holland, 1975) has been employed to solve the problem, which allows the system to display the best solution found so far by that method in the user interface. The search procedure can be stopped at any time, and the user can decide to use the solution found by the algorithm and reserve the items in the depots from the solution.

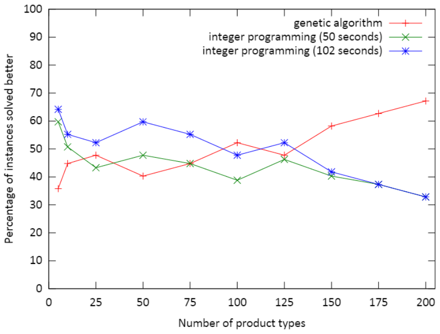


Fig. 1 Comparison of the methods given by the ratio of the number of instances solved better by a given method to the total number of all available instances

Obviously, one could ask if the linear programming solvers cannot perform the same task better, finding the optimal solution. A computational study has been conducted in which a mathematical model for the problem at hand has been solved by Gurobi solver with a time limit set to 102 seconds, which was the average time the genetic algorithm needs to find improved solutions. The percentage of instances solved better by either genetic algorithm or by the solver has been presented in Figure 1. The chart also shows results for linear programming model after 50 seconds, which is much more relevant to the on-line website service. The genetic algorithm is providing solutions almost immediately after it starts and they are constantly being improved, while the solver finds one particular solution and becomes stuck on proving its optimality.

In this paper a real world problem has been modelled and solved using genetic algorithm and linear programming methods. It has been integrated into on-line website that leveraged the responsiveness of genetic algorithm, to support charitable organizations in fulfilling their mission. Both methods have been compared and given the short amount of time that is given to solve the problem, the genetic algorithm did perform better.

References

Blazewicz, J., Ecker, K. H., Pesch, E., Schmidt, G., Weglarz, J., 2007. Handbook on Scheduling: From Theory to Applications.

Cichenski, M., Jarus, M., Miszkiewicz, M., Sterna, M., Szymczak, J., 2015. Supporting supply process in charitable organizations by genetic algorithm. Computers & Industrial Engineering, Volume 88, 39–48

Holland, J. H., 1975. Adaptation in Natural and Artificial Systems.

Ramesh, T., 1981. Traveling purchaser problem. OPSEARCH 18, 87–91.

A new measure of gene semantic similarity in the context of Gene Ontology

Marcin Jaroszewski

Institute of Computing Science, Poznan University of Technology
Poznan, Poland

Gene Ontology (GO) [1] is a species independent biological database dedicated to functional annotation of genes and their products in the context of a cell. GO is hierarchical in structure and organised in the form of three disjoint and directed acyclic graphs. Each graph captures a distinct domain of gene product functions: biological process, cellular component, molecular function. Nodes of a graph are biological terms which carry some information related to the specific domain, covered by the graph. Each gene is associated with the set of terms that characterize its activity within a cell. This set is known as the gene's annotation.

Semantic similarity of genes is a measure of their relationship, calculated based on some ontology (e.g. GO) and relevant annotations. More than a few such measures have been developed in the context of GO, e.g. [2, 3, 4]. All of them can be roughly divided into two categories [5]: edge based and information content based. The second group is further divided into annotation based and topology based measures.

Edge based measures associate similarity between terms with the number of edges or nodes separating them. Unfortunately, they ignore a term's position characteristics. Annotation based measures calculate a term's information content, but it depends, e.g., on the set of genes compared. Topology based measures [3, 4] are void of these drawbacks. They take only the underlying graph's topology into account and are thus the preferred solution.

I have designed a new measure of gene semantic similarity based on GO, that would return a score reflecting proximity of biological processes in which genes of interest are involved. This measure relies on topology of the biological process graph and is based on fuzzy sets. First, all genes of interest are annotated, each with its set of terms from GO. To compare any two genes, A and B, their annotation sets are converted into fuzzy sets F_A and F_B respectively. The final score is produced by comparing the size of the union and of the intersection of sets F_A and F_B.

References

[1] Ashburner M., Ball C. A., Blake J. A., et al. (2000) *Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium*, Nature Genetics 25 (1), 25-29.

[2] Couto F. M., Silva M. J. and Coutinho P. M. (2007) *Measuring semantic similarity between Gene Ontology terms*, Data and Knowledge Engineering 61 (1), 137-152.

[3] Mazandu G. K. and Mulder N. J. (2012) *A Topology-Based Metric for Measuring Term Similarity in the Gene Ontology*, Advances in Bioinformatics 2012.

[4] Wang J. Z., Du Z., Payattakool R., et al. (2007) *A new method to measure the semantic similarity of GO terms*, Bioinformatics 23 (10), 1274-1281.

[5] Mazandu G. K. and Mulder N. J. (2013) *Information Content-Based Gene Ontology Semantic Similarity Approaches: Toward a Unified Framework Theory*, BioMed Research International, Article ID 292063.

Smartphone tasks – modelling mobile devices in scheduling theory

K. Ecker[1] and M. Tanas[2]

[1]Clausthal University of Technology, Clausthal, Germany
[2]Adam Mickiewicz University, Physics Faculty, ul. Umultowska 85 61-614 Poznan Poland, phone: +48 660512136, mail: Zaklad Informatyki Stosowanej, Wydzial Fizyki UAM, ul. Umultowska 85, 61-614 Poznan, email: mtanas@amu.edu.pl

Abstract

Theory of scheduling are well researched and widely used domain of computer science. However, in most of existing results classical computer or production systems are considered, for which the desired goal is in general to maximize their processing power and capacity. Nowadays a new class of computer-like devices called mobile devices (smartphones, tablets, etc) whose usage is completely different, becomes more and more polular. The distinguishing features of mobile devices, especially its critical dependency on battery power, combined with specific features of applications executed on these devices, especially requirement for high interactivity and fluency of execution, leads to an entirely new optimization criterion, which is "what is the lowest processing power required to execute the given set of tasks without violation of relative deadlines between them".

The problem of smartphone tasks scheduling has the following distinguishing features. The speed of the processor can be chosen once, but after then it can be changed no more. The chosen processor speed is modelled by machine slowness factor M_{slow}, which describes how much the current speed of processor differs from its basic speed (i.e. speed with $M_{slow} = 1$). Each task contains infinite chain of identical operations and every two subsequent operations in a task are distance constrained, i.e. the subsequent operation must be finished no later than d_j units of time after completion of the preceding operation. The optimization criterion is to minimize M_{slow} in such a way, that the given set of tasks can be processed without violation of the relative deadlines.

There are two hypothesis stated about the smartphone tasks scheduling problem:

1. If relative deadlines are variable the problem is beyond NP or is not even a combinatorial problem (because size of solution can be arbitrary large)
2. However if relative deadlines are constant the problem is NP-hard.

Keywords: scheduling, complexity, distance constraint tasks, smartphone tasks.

New *in silico* approach to assess RNA secondary structures with non-canonical base pairs

Natalia Szostak[1,3], Agnieszka Rybarczyk[1,2,3], Maciej Antczak[1,3], Tomasz Zok[1,3], Mariusz Popenda[2,3], Ryszard Adamiak[1,2,3], Jacek Blazewicz[1,2,3] and Marta Szachniuk[1,2,3]

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland
[3]European Center for Bioinformatics and Genomics, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland

e-mail: Natalia Szóstak <nszostak@cs.put.poznan.pl>

RNA function depends on its structure, therefore an appropriate recognition of the latter is of great importance. One particular concern is the assessment of base-base interactions, described as the secondary structure. It greatly facilitates an interpretation of RNA function and allows for structure analysis on the tertiary level. The RNA secondary structure can be predicted from sequence using in silico methods often adjusted with experimental data, or assessed from 3D structure atom coordinates. Computational approaches consider mostly Watson-Crick and wobble base pairs. Handling of non-canonical interactions, important for a full description of RNA structure, is still a challenge.

Here we present novel two-step in silico approach to asses RNA secondary structures with non-canonical base pairs. Its idea is based on predicting the RNA 3D structure from sequence or secondary structure that describes canonical base pairs only, and next, back-calculating the extended secondary structure from atom coordinates. We have integrate in a computational pipeline the functionality of two fully automated, high fidelity methods: RNAComposer for the 3D RNA structure prediction and RNApdbee for base pair annotation. We have benchmarked our pipeline on 2559 RNAs sequences with the size up to 500 nucleotides obtaining better accuracy in non-canonical base pair assessment than the compared methods that directly predict RNA secondary structure.

This work was supported by grants from National Science Center, Poland [2012/05/B/ST6/03026, 2012/06/A/ST6/00384].

RNAsprite - a concept of tool to calculate RNA structural data

Piotr Ceranek[1]*, Tomasz Zok[1], Maciej Antczak[1], Marta Szachniuk[1,2]

[1]Institute of Computing Science, Poznan University of Technology, Poland
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
*corresponding author: pert.ceranek@gmail.com

1 Introduction

In the recent decades, an increasing interest in molecular biology could be observed in the scientific community. It has grew out of the structural study of proteins, DNA and RNA - three macromolecules essential for all known forms of life. Our project is focused on RNA, the least known of these three building blocks of the living organisms.

RNA is best known from playing a crucial role in the process of protein synthesis by conveying an information encoded in DNA. However, there is much more than that in the RNA story. For example, several types of non-coding RNAs regulate cell metabolism and this is why medicine is more and more often targeting these molecules. Ribonucleic acids are also hypothesized to be the first molecules to replicate and evolve, thus, starting the era of life on Earth. From this, one can agree that a development of tools and technologies to support molecular biology-rooted research is one of the crucial aims in contemporary science.

2 A concept of RNAsprite

There are various ways of representing the 3D structure of RNA (Zok *et al.*, 2014). Each of them gives the other perspective on the molecule three-dimensional shape. Algebraic representation is the most common and the simplest way to describe molecule fold by providing coordinates of its atoms in three dimensional space. Trigonometric representation describes the 3D structure by providing values of torsion angles found along RNA chain, i.e. dihedral angles between two planes that are defined by atom quartets. Finally, geometric representation is given as a matrix of Euclidean distances between particular atoms of the corresponding residues. These representations are used in various computational systems designed to store, compute and analyse structural data (cf. Popenda *et al*., 2010). Each of them is considered in the RNAsprite project.

RNAsprite has been invented as a part of RNApolis (http://rnapolis.pl) - a collection of software tools designed to cover a wide variety of aspects of RNA structure analysis. An idea behind RNAsprite was to develop a tool that could easily translate between different representations of RNA three-dimensional structure. Moreover, the tool is planned to provide its users with the ability to compute non-typical torsional angles (defined by the user), intra-molecular distances between atoms selected by the user, to visualize distance matrices, etc. Since the 3D shape of a molecule is directly connected to its features and functions, we believe, that new perspectives on the structure can enable discovering unrealized structural or functional characteristics of RNAs.

3 Conclusions and future work

RNA three-dimensional fold can be defined by different structural parameters which constitute various structure representations. Here, we present a project of RNAsprite tool that will be capable of providing structure description in various notations and converting between them. RNAsprite is an on-going project and still remains under development. In the nearest future, the following work is planned to be done within the project: providing support for all structure representations, implementing an option of user-defined torsional angle computation, visualization of distance matrices and the structure itself.

References

1. M. Popenda, M. Szachniuk, M. Blazewicz, S. Wasik, E.K. Burke, J. Blazewicz, R.W. Adamiak. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures, *BMC Bioinformatics* 11, 2010, 231.
2. T. Zok, M. Popenda, M. Szachniuk. MCQ4Structures to compute similarity of molecule structures, *Central European Journal of Operations Research* 22(3), 2014, 457-474.

Computational approaches for quality analysis of RNA structures

Piotr Lukasiak, Tomasz Ratajczak, Maciej Antczak

Institute of Computing Science, Poznan University of Technology, 60-965 Poznan, Poland

Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-714, Poznan, Poland

Prediction of 3D structure, recognition of function responsible for selected processes in living organisms or simulation of such processes are between others the most important challenges in medicine and biology. This can lead us to develop new drugs to protect our organisms against dangerous diseases as HIV or cancer. Nowadays, computational methodologies are in the scope of interest in molecular biology area because experiments in so called ‘wet laboratories’ are expensive and time consuming. Within set of mentioned problems, modelling of 3D structure is the first step in the analysis of molecular pipeline. Currently, many computational approaches were developed to support development of 3D RNA structure, and domain experts have available tens of structural models that corresponds to the original sequence. Thus, the quality analysis of models became the crucial point. The largest problem of prediction methods is not the imperfectness of current prediction methods but, it is the lack of estimated error or the quality evaluation of a model. It is important to establish quality estimation methods for predicted models in the way that the model can be used wisely by having knowledge regarding the limitations of the model.

The quality evaluation of models in the context of the reference structure can be performed in various ways using wide range of measures. RNA 3D structures are most often evaluated by numerical measures, namely Root-Mean-Square Deviation (RMSD), Interaction Network Fidelity (INF), Deformation Index (DI), Mean of Circular Quantities (MCQ), CAD–score, P–value. Moreover, visual approaches such as Deformation Profile which depicts the conformation differences between compared 3D structures in local, inter–domain, and intra–domain scales, and visualizations provided by developed by us the quality server RNAssess showing both global and local coherence between a particular model and a reference structure (e.g. 2D map and 3D landscape). All listed measures are available in our server together with new idea called cutoff, which determines the number of local neighborhood areas observed around nucleotide for a fixed sphere radius predicted below a certain threshold. Cutoff measure needs more attention because it gives the information about local precision of prediction from global point of view. Cutoff plot shows how accurate is the prediction of a particular model from a local point of view or, in other words, which part of the model structure is predicted correctly. The calculation is performed based on the fixed precision value (cutoff) for each sphere radius value, defined by the user in spheres radii vector. As a result, the user receives information about the atoms set predicted below cutoff (%). The value of cutoff can be changed interactively by the user. With fixed radius of the sphere such measure can give one number that can correspond to the model quality similarly to GDT.

Validation of protein tertiary structure models is a very important issue in structural biology. The presented methods meets this challenge. Proposed approaches provides the research community with a easy to use methodology for the comprehensive quality inspection conducted between the set of predicted RNA models and the native structure. Our method may be used to evaluate structures from different points of view, to observe prediction difficulty and to identify promising predictions even if they seem to be inappropriate.

Two crowdsourcing approaches for solving optimization problems

Szymon Wąsik[1,2,*]

[1]*Institute of Computing Science, Poznan University of Technology, Poznan, Poland*
[2]*European Centre for Bioinformatics and Genomics, Poznan University of Technology, Poznan, Poland*

** e-mail: szymon.wasik@cs.put.poznan.pl*

First formulation of the crowdsourcing problem, understood as outsourcing work to a large network of people in the form of an open call, comes from the 18th century [1]. Since that time the concept of crowdsourcing have been utilized many times, however, the rapid development of this technique started with the development of the Internet in 1990s. The best examples of its capabilities are services like Wikipedia, GitHub, TripAdvisor or OpenStreetMap. Anonther interesting application of crowdsourcing concept is an implementation of computer games which objective is to solve a scientific problem by employing users to play the game, so called Crowdsourced Serious Game [2]. Such approach has already helped to discover several interesting biological phenomena using games such as Foldit [3], EteRNA [4] or EyeWire [5].

The main objective of the first presented approach was to verify if the crowdsourcing can be successfully applied for finding mathematical equations that explains data gathered from the biological experiments. Moreover, we wanted to compare it with the approach based on artificial intelligence that uses symbolic regression to find such formulas automatically [6]. To achieve this we designed and implemented the game in which players tries to design a spaceship representing an equation that models the observed system. The game was tested by playing almost 10000 games by several hundred players and by conducting users opinion survey. Results prove that the proposed solution has very high potential. The function generated during week long tests was almost as precise as the analytic solution of the system of differential equations and it explained data better than the solution generated automatically by the Eureqa – the leading software implementing symbolic regression [9]. Moreover, we observed benefits from the use of the crowdsourcing technique – the chain of consecutive solutions that leaded to the best solution was obtained by continues collaboration of several players.

The main scientific objective of the second presented approach is to develop a methodology for continuous evaluation of optimization algorithms using crowdsourcing technique along with the methodology for a representative group of problems and their verification in close to real environmental conditions. The goal will be achieved by the design, implementation and deployment of the Internet portal that will allow using a methodology developed and with the help of which it will be possible to verify this methodology in practice. The objective of the presentation is to present main assumption and the current state of the project and to consult it with the audience.

[1] E. Estelles-Arolas, F. Gonzalez-Ladron-de-Guevara, *J INFORM SCI*, **2012**, *38*, 189–200.
[2] U. Tellioglu, GG. Xie, JP. Rohrer, C. Prince, *CGAMES*, **2014**, 1-7.
[3] S. Cooper, et al., Foldit players, *NATURE*, **2010**, 466, 756–760.
[4] J. Lee, et al., *P NATL ACAD SCI USA*, **2014**, 111, 2122–2127.
[5] M. Helmstaedter, KL. Briggman, SC. Turaga, V. Jain, et al., *NATURE*, **2014**, 514, 394–394.
[6] MJ. Willis, *GALESIA*, **1997**, 446, 314-319.
[7] F. Galton, *NATURE*, **1907**, 75, 450–451.
[8] H. Dahari, RM. Ribeiro, AS. Perelson, *HEPATOLOGY*, **2007**, 46, 16–21.
[9] M. Schmidt, H. Lipson, *SCIENCE*, **2009**, 324, 81–85.

Reconstruction *de novo* of genome sequence using GPU computing

Aleksandra Swiercz[1,2,3,*],Wojciech Frohmberg[1,3], Michał Kierzynka[3,4], Paweł Wojciechowski[1,2,3], Piotr Zurkowski[1,3], Marta Kasprzak[1,2,3],Jacek Blazewicz[1,2,3],

[1]Institute of Computing Science, Poznan University of Technology, Poland
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poland
[3]European Center of Bioinformatics and Genomics, Poznan, Poland
[4]Poznan Supercomputing and Networking Center, Polish Academy of Sciences, Poland
[*]corresponding author: aleksandra.swiercz@cs.put.poznan.pl

In the last years we could observe the increase of recognized new genome sequences. Although, many organisms of not yet known genomes can be resequenced with the help of genomes within the same family of species, still many have to be discovered *de novo*, without information of any reference genome. With the rapid growth of the development of next-generation sequencers the process of reading DNA sequence has become cheaper and in this connection more available. Sequencers produce billions of short DNA sequences, called reads, which can be assembled together in order to reconstruct the original genome sequence. The assembly is possible only when reads cover the whole target sequence and neighboring reads sufficiently overlap and can be merged together.

There are mainly two approaches to computational reconstruction of DNA sequences. The first one, called **overlap-layout-consensus** (OLC), builds a graph, with vertices representing all the reads, and arcs representing feasible (can be defined as a parameter) overlapping between reads. The solution of the problem can be seen as a path in the graph, passing through all the vertices. Due to a huge number of reads it is impossible to calculate the overlap between every pair of reads, thus disjoint components of the graph may occur, and in that case several paths would be given as a result instead of one. The other approach uses **so called de Bruijn graphs**. It decomposes every read into a collection of *k*-mers. *K*-mers become vertices in the graph, and reads are represented as paths containing all the subsequences (*k*-mers) in the respective order. There is no need to calculate the overlap between reads, because overlapping reads pass through the same *k*-mers. However, by reads decomposition we lose the information necessary for the reconstruction, especially in case of repetitions in genome sequence. For the comparison of genome reconstruction methods see for example [1,2].

We propose a novel algorithm for *de novo* reconstruction following OLC approach. Our method is composed of a few steps. At the beginning an overlap graph is constructed out of all input reads and their reverse and complementary counterparts. The key point is to select *promising pairs* of reads, which possibly overlap, and for which the alignment is calculated with an exact algorithm implemented on a Graphics Processing Unit (GPU). The overlap graph is next traversed with the detection of the forks in the path. Each path without junctions is changed into contig (a contigous sequence). In the last step we build a scaffold graph, where contigs become vertices and arcs include the information about contigs overlapping and paired-end reads divided into separated contigs.

[1] G. Narzisi, B. Mishra, “Comparing de novo genome assembly: the long and short of it”, *PLoS ONE* 6(4): e19175, 20111

[2] JR Miller, S Koren, G Sutton, “Assembly algorithms for next-generation sequencing data”, *Genomics* 95, 315-327, 2010.

Clustering approach to create libraries of representative RNA conformers

Tomasz Zok Maciej Antczak Martin Riedel David Nebel Thomas Villmann
Piotr Lukasiak Jacek Blazewicz Marta Szachniuk

The structure of RNA plays a key role in the functioning of this molecule. For many years already, computer scientists in cooperation with biochemists have proposed a variety of software solutions in the field of structural biology of RNA. Recently, we have focused on the problem of structure refinement, which can be summarized in the following way. Given coordinates of a modeled RNA, make a number of small alterations which will improve the structure quality. The quality is measured in terms of free energy and general stereo-chemical correctness such as bond lengths in their valid ranges. Several approaches to refine 3D structures exist and they are often based on a generalization of knowledge present in the literature. We decided to take a different course of action and use nucleotide-level templates from a library of reference conformers. Our approach to construct such library has been first presented in [1].

We can look at RNA chains as bio-polymers made of nucleotides. Each of those however, can be divided into smaller fragments and analyzed separately. A phosphate group, a ribose and a base (adenine, guanine, cytosine or uracil) constitute single nucleotide. They adopt certain conformations, i.e. shapes in 3D space, and our assumption is that they are related to each other. For example, a given set of coordinates of phosphate group atoms limits in a way how can the ribose and base bind. Another assumption is that each nucleotide fragment adopts only one of several conformations, with some local variations. This leads to a hierarchical model of conformer library. The library contains nucleotides annotated with metadata about each fragment organized in a tree structure.

We decided to found our approach on the representation of RNAs in torsion angle space [2]. Then, each nucleotide is described by a vector of 12 real values (6 for backbone, 5 for ribose and 1 for base fragment). The dataset consisted of more than 500 RNA structures of high quality (solved by X-Ray with resolution below 2.4 Å) which after filtering resulted in more than 60,000 library entries corresponding to various nucleotides. For each nucleotide type (A, G, C or U), we performed clustering on the backbone fragment data (subvector of 6 values out of 12). We used a median version of neural gas algorithm – a prototype-based clustering method which finds centers of clusters only among input data. This way, we got a set of clusters, each with a representative nucleotide originated from an experimentally solved structure. For each cluster, we reapplied the median neural gas method using the remaining 6 values in the subvector. Finally, among the resulting clusters, we searched for distinguished conformations to become the final representatives. This way, a full hierarchical division of an input was obtained.

The aim of our current work is to propose a software solution to the problem of structure refinement. We plan to replace coordinates of selected atoms in a 3D model of RNA with different ones to increase the quality of the structure. A major component of the pipeline, is a library of reference nucleotide conformations. We have successfully constructed one using clustering approach in a hierarchical manner.

References

[1] Tomasz Zok, Maciej Antczak, Martin Riedel, David Nebel, Thomas Villmann, Piotr Lukasiak, Jacek Blazewicz, and Marta Szachniuk. Building the Library of RNA 3D Nucleotide Conformations Using the Clustering Approach. *Int. J. Appl. Math. Comput. Sci*, 25(3):689–700, 2015.

[2] Tomasz Zok, Mariusz Popenda, and Marta Szachniuk. MCQ4Structures to compute similarity of molecule structures. *Central European Journal of Operations Research*, 22(3):457–473, April 2014.

Data similarities, dissimilarities and types of inner products - a mathematical characterization in the context of machine learning

T. Villmann[1], M. Kaden[1], D. Nebel[1], and A. Bohnsack[2]
1-University of Applied Sciences Mittweida, Computational Intelligence Group
Technikumplatz 17, 09648 Mittweida, Germany
2- Staatliche Berufliche Oberschule Kaufbeuren
87600 Kaufbeuren, Josef-Fischer-Straße 5 , Germany

Abstract

Data dissimilarities and similarities are the key ingredients of machine learning. We give a mathematical characterization and classification of those measures based on structural properties also involving psychological-cognitive aspects of similarity determination, and investigate admissible conversions. Finally, we discuss some consequences of the obtained classification and relations for machine learning algorithms.

1 Introduction

Data in machine learning are usually compared in terms of dissimilarities or similarities. These values maybe obtained either by mathematical calculations or by other judgements like human rater values. Examples are the Euclidean and other distance measures, correlations or divergences but also questionnaire scales, joint probabilities etc. Often, the specific origin of the dissimilarity/similarity values is not known or at least not easy to explore for the data analyst. Thus, respective tasks have to deal just with the given dissimilarity/similarity matrix. Otherwise, algorithms and computational approaches in machine learning frequently require data assumptions to be fulfilled. For example, support vector machines for classification learning suppose a symmetric positive definite similarity matrix [1, 2, 3] whereas online learning vector quantization models assume differentiable dissimilarities [4, 5].

The prevailing methodology in machine learning applications is to convert dissimilarities in similarities and vice versa under mild conditions supposing an almost similar behavior. The respective assumption usually is based on an intuitive understanding of the concept of proximity. This intuitive thinking, however, can be misleading and, hence, result in false interpretation of the approaches regarding their abilities for data analysis. For example, kernel methods like support vector machines are frequently identified as a similarity based approach although, as we will explore later in detail, this is not valid for all kernels. Thus we plead for a more faithful purpose of these aspects. However, a precise categorization of proximity measures is only available from the geometric interpretation of dissimilarity mainly influenced by the mathematical definition of metrics. A finer grained differentiation is provided by PEKALSKA&DUIN in [6] but still starting from the mathematical distance definition. Otherwise, the similarity paradigm is intensively studied in psychology and cognitive science [7, 8, 9]. It is pointed out that several aspects of mathematical distances like symmetry or triangle inequality are apparently not valid in many situations. Starting with the pioneering work from TVERSKY [8, 10] a more conceptual way was suggested based on a feature-theoretical contrast model considering the difference between common and distinguishing features of objects [11]. This approach led to a better understanding of similarities based on weaker assumptions. These assumptions are made according to plausibility arguments and can be related to paradigms in information processing [8, 12]. Yet, it turns out that most attempts return to the geometrical distance interpretation during the detailed considerations [11, 13, 14, 15]. To our best knowledge only two authors remained in the conceptual line for formal description of similarities based on properties figured out to be important object discrimination in perception and tried to characterize it mathematically [16, 12]. Both approaches as well as the early Tversky-approach have in common that they do not differentiate several kinds of similarity.

The aim of the present paper is to characterize similarities based on properties in relation to dissimilarity measures. For this purpose, we start with the cognitive understanding of similarity and the respective mathematical description as already done in the mentioned approaches. Thereafter, we fine tune the similarity categorization scheme according to the thoroughly defined kinds of dissimilarities. Further, we shortly discuss aspects of equivalent measures and show implications for machine learning applications.

2 Mathematical description of similarities based on their properties

In this section we develop a scheme to relate kinds of similarities to adequate dissimilarity types based on mathematical properties. For this purpose, we start with the basic (mathematical) assumptions made in cognitive science for similarities, turn over to dissimilarities and extend both lines in sequel to obtain a mirror-inverted description scheme. Thereby we match our differentiation to categories for dissimilarities suggested by Pekalska&Duin in [6].

As pointed out by Santini&Jain in [16], the mathematical distance axioms seem to be too rigid for a system of similarities. The common sense in cognitive science according to the contrast model is to endow a similarity measure s for the object space X with at least the following properties [8]:

1. Maximum or dominance principle: $s(x,x) \geq s(x,y)$ and $s(x,x) \geq s(y,x)$

2. Non-negativity: $s(x,y) \geq 0$

whereby $s(x,y)$ increases with x and y sharing more properties and decreases with the number and the degree of discriminative features, i.e. the similarity is a function of object commonalities and differences. We denote a map $s : X \times X \to \mathbb{R}$, fulfilling the maximum principle, as a basic similarity. If additionally the non-negativity is given s is said to be a primitive similarity. These basic properties are usually accompanied by a consistency requirement often chosen as $s(x,x) = c_s$ with $c_s = 1\ \forall x$. Here, we propose a more gradual definition with either an arbitrary constant c_s or data dependent $c_s(x)$ to take a local view. Accordingly, we adequately suppose for a basic/primitive dissimilarity d

1. Minimum principle: $d(x,x) \leq d(x,y)$ and $d(x,x) \leq d(y,x)$
2. Non-negativity: $d(x,y) \geq 0$

which can come along with the consistency property $d(x,x) = c_d(x)$ yielding a weakly-consistent dissimilarity. The consistency is often tightened to be c_d independent from x or required as $d(x,x) = 0$ (reflexivity). A reflexive dissimilarity obeying these properties is denoted as hollow metric [6].

Definiteness as well as inequality constraints lead to further dissimilarity variants, which adequately transfer to similarities. We collect the properties in Tab.1 and depict the respective similarity and dissimilarity types in Tab. 2.

With these definitions we are able to classify dis-/similarity measures. We can conclude that inner and semi-inner products belong to similarities, if further requirements are fulfilled. For example, if the data are normalized, the maximum principle is verified by the Cauchy-Schwarz-inequality (CSI) [17], however the consistency properties are generally not fulfilled. Accordingly, the Gaussian kernel $\kappa(x,y) = \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)$ defines a Minkowski-like similarity, because it is an inner product bounded by $c_\kappa = 1$, whereas general kernels fail the maximum principle, as inner products of Hilbert spaces [2], and, hence, are not similarities in general. An example for basic similarity is a negative distance measure $-d$ as it is used for the affinity propagation clustering [18] or negative divergences [19, 5]. Further, for Pseudo-Euclidean spaces, as used in relational methods [20], the respective indefinite inner product is not a similarity because the CSI does not hold such that the maximum principle is violated.

	$\mathbf{d}(\mathbf{x},\mathbf{y})$ $d : X \times X \to \mathbb{R}$	$\mathbf{s}(\mathbf{x},\mathbf{y})$ $s : X \times X \to \mathbb{R}$	
MIN	minimum principle $d(x,x) \leq d(x,y) \wedge$ $d(y,y) \leq d(x,y)$	maximum principle $s(x,x) \geq s(x,y) \wedge$ $s(y,y) \geq s(x,y)$	MAX
NN	non-negativity $d(x,y) \geq 0$	$s(x,y) \geq 0$	NN
wC	weak consistency $d(x,x) = c_d(x)$ $(c_d(x), c_s(x) : X \to \mathbb{R})$	$s(x,x) = c_s(x)$	wC
sC	strong consistency $d(x,x) = c_d$ $c_d, c_s \in \mathbb{R}$	$s(x,x) = c_s$	sC
R	reflexivity $d(x,x) = 0$	normalized sC $d(x,x) = 1$	nC
S	symmetry $d(x,y) = d(y,x)$	$s(x,y) = s(y,x)$	S
D	definiteness (non-degeneration) $d(x,y) = d(x,x) \vee$ $d(x,y) = d(y,y) \Rightarrow x = y$	$s(x,y) = s(x,x) \vee$ $s(x,y) = s(y,y) \Rightarrow x = y$	D
T	triangle inequality $d(x,y) + d(y,z) \geq d(x,z)$	reverse triangle inequality $s(x,y) + s(y,z) \leq s(x,z)$	rT
UM	ultra-metric inequality $\max\{d(x,y), d(y,z)\} \geq d(x,z)$	reverse ultra-sim. inequality $\min\{s(x,y), s(y,z)\} \geq s(x,z)$	rUS

Table 1: General properties for dissimilarity and similarity measures.

dissimilarities/properties	MIN	NN	consistency			S	D	inequalities	
			wC	sC	R			T	UM
basic dis.	x								
primitive dis.	x	x							
weakly-consistent dis.	x	x	x						
strongly-consistent dis.	x	x		x					
general dis. (hollow metric)	x	x			x				
pre dis. (pre-metric)	x	x			x	x			
usual dis. (quasi-metric)	x	x			x	x	x		
semi-metric	x	x			x	x		x	
distance (metric)	x	x			x	x	x	x	
ultra metric	x	x			x	x	x		x

similarities/properties	MAX	NN	consistency			S	D	inequalities	
			wC	sC	nC			rT	rUS
basic sim.	x								
primitive sim.	x	x							
weakly-consistent sim.	x	x	x						
weak sim.	x		x						
strongly-consistent sim.	x	x		x					
general sim. (hollow sim.)	x	x			x				
pre-sim.	x	x			x	x			
usual sim. (quasi-sim.)	x	x			x	x	x		
semi-sim.	x	x			x	x		x	
(Minkowsky-like) similarity	x	x			x	x	x	x	
ultra sim.	x	x			x	x	x		x

Table 2: Kinds of dissimilarity and similarity measures

3 Relations between dissimilarities and similarities

As mentioned in the introduction, in context of machine learning, frequently similarities and dissimilarities are converted into each other or are synonymously used. Otherwise, kernel methods often rely on the implicit non-linear mapping to expect better results. Therefor, it is desirable to compare respective measures in terms of their topological properties.

Definition 3.1 *Two dissimilarities d and $\widehat{d}$ are said to be rank-equivalent, if $\forall x, y, v, w \in X$ the following relations hold for d and $\widehat{d}$*

1. $d(x,y) < d(v,w)$ *iff* $\widehat{d}(x,y) < \widehat{d}(v,w)$
2. $d(x,y) = d(v,w)$ *iff* $\widehat{d}(x,y) = \widehat{d}(v,w)$.

Two similarities s and $\widehat{s}$ are said to be rank-equivalent, if $\forall x, y, v, w \in X$ the following relations hold for s and $\widehat{s}$

3. $s(x,y) < s(v,w)$ *iff* $\widehat{s}(x,y) < \widehat{s}(v,w)$
4. $s(x,y) = s(v,w)$ *iff* $\widehat{s}(x,y) = \widehat{s}(v,w)$.

A dissimilarity d and a similarity s are said to be rank-equivalent, if $\forall x, y, v, w \in X$ the following relations hold for s and d

5. $s(x,y) < s(v,w)$ *iff* $d(x,y) > d(v,w)$
6. $s(x,y) = s(v,w)$ *iff* $d(x,y) = d(v,w)$.

This definition leads to the following important observation:

Remark 3.1 *If two dissimilarities d and $\widehat{d}$ are given and $d(x,y) = f\left(\widehat{d}(x,y)\right)$ with f being a monotonously increasing function, both dissimilarities d and $\widehat{d}$ are rank-equivalent.*

According to his remark we can replace in the famous k-nearest-neighbor classifier (kNN,[21]) respective rank-equivalent distances without a change in the classification behavior of the kNN. For example, if a Gaussian kernel distances is considered, this distance is rank-equivalent to the Euclidean distance. Hence, the respective classification ability of kNN remains unchanged if we replace them by each other. Otherwise, vector quantizer may benefit during learning: The prototype vectors are distributed following the magnification law according to the data density [22], which depends on the underlying dissimilarity measure. After training of those models, we can return to a rank-equivalent dissimilarity with lower computational costs in case of time restricted applications. For example, the Gaussian kernel could be replaced by the simpler Euclidean, because they are rank-equivalent. Similar thought may apply for support vector machines.

It is easy to generate a dissimilarity from a similarity measure. A convenient way is just to take

$$(d(x,y))^2 = s(x,x) + s(y,y) - s(x,y) - s(y,x) \tag{1}$$

whereas the reverse way is more complicate. However, the topological equivalence is not guaranteed. For example, if s is the Euclidean inner product, then d is the squared Euclidean distance. However, unit circles with respect to d do not have an analogue for s, in general. To see this, we consider $X = \mathbb{R}^2$ and a constant Euclidean inner product $s(x,y) = x_1 y_1 + x_2 y_2 = c$ for a fixed $x \in \mathbb{R}^2$. Depending on x, the set S of solutions of this equation can take different values: for $x = \mathbf{0}$ we obtain $S = \emptyset$, whereas for $x = (a, 0)$ with $a \neq 0$ we get $S = \left\{ \left(\frac{1}{a}, y_2 \right) : y_2 \in \mathbb{R} \right\}$ determining a straight line. Thus, comparison of unit (constant) circles around x according to d and s show substantial topological differences. Similarly, topological changes occur if flipping or clipping is applied for relational methods in case of Pseudo-Euclidean data [20].

4 Conclusion

In this contribution we consider and define types of similarities and dissimilarities from a mathematical point of view. These types are based on structural properties and are in agreement with earlier approaches as well as involve psychological-cognitive aspects of similarity theory. One immediate consequence of these investigation is that one has to distinguish carefully between similarities and kernels/inner products although both can be seen as counterparts to dissimilarities.

Further, in the sense of this paper, we remark that RSLVQ [23], originally introduced as a probabilistic variant of LVQ, can be seen as a similarity-based algorithm. The reason for this is that the probabilities are estimated by Gaussians, which are, in fact kernels. Hence, they are similarities in the view of this paper. Thus a natural generalization of RSLVQ would be to apply other similarities than Gaussian kernels but keeping the optimization strategy. However, if other similarities replace the Gaussians we obtain a ratio of summarized similarities, which can still serve as probabilistic assignments of data to classes, i.e. we model the probabilities not longer by Gaussians but by a mixture of similarities. Respective studies are planned for the future.

In summary, a general reassessment of kernels, inner products and dis-/similarities in machine learning is demanded for an adequate application use.

References

[1] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, Cambridge, 2002.

[2] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[3] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer Verlag, Berlin-Heidelberg, 2008.

[4] T. Villmann, S. Haase, and M. Kaden. Kernelized vector quantization in gradient-descent learning. *Neurocomputing*, 147:83–95, 2015.

[5] T. Villmann and S. Haase. Divergence based vector quantization. *Neural Computation*, 23(5):1343–1392, 2011.

[6] E. Pekalska and R.P.W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications.* World Scientific, 2006.

[7] R.N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(11):1317–1323, 1987.

[8] A. Tversky. Features of similarity. *Psychological Review*, 84(4):324–352, 1977.

[9] D.N. Osherson. New axioms for the contrast model of similarity. *Journal of Mathematical Psychology*, 31:93–103, 1987.

[10] A. Tversky and I. Gati. Studies of similarity. In E. Rosch and B.B. Lloyd, editors, *Cognition and Categorization*, pages 79–98. Hillsdale, NJ: Erlbaum, 1978.

[11] A. Tversky and I. Gati. Similarity, separability, and the triangle inequality. *Psychological Review*, 89(2):123–154, 1982.

[12] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.

[13] F. Jäkel, B. Schölkopf, and F.A. Wichmann. Does cognitive science need kernels? *Trends in Cognitive Sciences*, 13(9):381–388, 2009.

[14] F. Jäkel, B. Schölkopf, and F.A. Wichmann. Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, 52:297–303, 2008.

[15] A. Tversky and D.H. Krantz. The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology*,

7:572–590, 1970.

[16] S. Santini and R. Jain. Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

[17] G. Lumer. Semi-inner-product spaces. *Transactions of the American Mathematical Society*, 100:29–43, 1961.

[18] B.J. Frey and D. Dueck. Clustering by message passing between data points. *Science*, 315:972–976, 2007.

[19] A. Cichocki and S.-I. Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12:1532–1568, 2010.

[20] B. Hammer, D. Hofmann, F.-M. Schleif, and X. Zhu. Learning vector quantization for (dis-)similarities. *Neurocomputing*, 131:43–51, 2014.

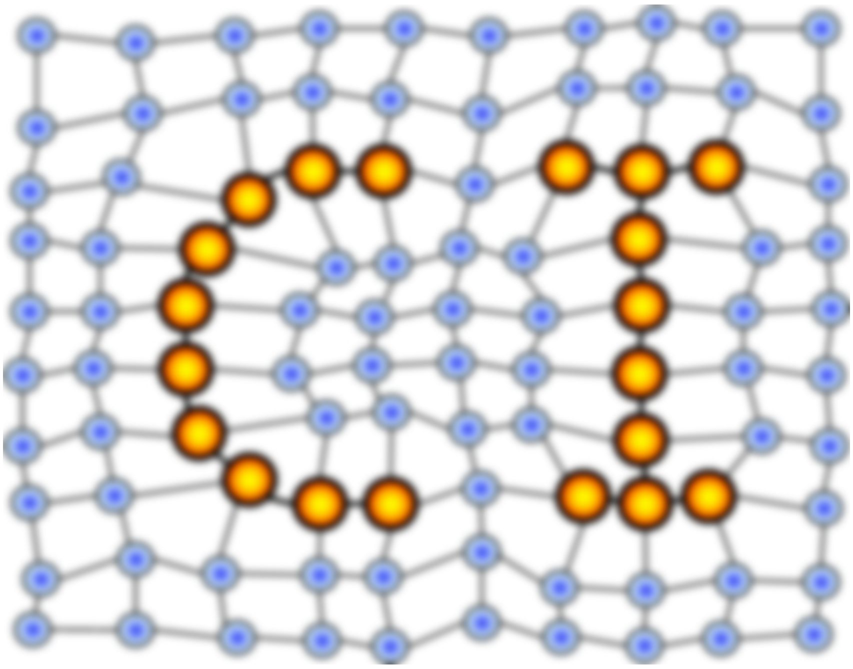
[21] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[22] T. Villmann and J.-C. Claussen. Magnification control in self-organizing maps and neural gas. *Neural Computation*, 18(2):446–469, February 2006.

[23] S. Seo and K. Obermayer. Soft learning vector quantization. *Neural Computation*, 15:1589–1604, 2003.

Machine Learning Reports

Machine Learning Report 04/2015



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.