

EDUCATION

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje September, 2011 – May, 2015

Bachelor of Science in Computer Science and Engineering
Thesis: Collaborative Two-level Classification Ensemble by Margin Increase
Graduated November, 2015

GPA: 9.74 / 10.00 (5-10 absolute scale, no curve)

• Relevant Coursework

Machine Learning, Data Mining, Artificial Intelligence, Databases, Advanced Databases, Data Warehouses (and Analytical Data Processing), Distributed Systems, Research Methodologies in ICT, Advanced Computer Architecture, Advanced Algorithms, Algorithms and Data Structures, Linear Algebra, Probability and Statistics, Theory of Programming (Automata and Formal Languages), Embedded Systems, Microprocessor Systems, Discrete Mathematics 1 & 2, Calculus 1 & 2

Josip Broz Tito High School, Skopje

Graduated May, 2011

Mathematics program (GPA: 5.00 / 5.00)

AWARDS & ACHIEVEMENTS

- "Top Information Technology Students in Macedonia" Scholarship Awarded 2012, 2013, 2014
- Best Student Award Awarded 2013, 2014, 2015
- Admitted among the top 10% applicants in the field of Computer Science and Engineering June, 2011
- Graduated from High School in the top 2% in the country May, 2011

SUBMITTED JOURNAL PUBLICATIONS

- Arsov N., Pavlovski M. & Kocarev L. "**Weighted Bagging Predictors**" (in submission to *Machine Learning, Springer*) – preprints available online
- Jovanovski J., Arsov N., Stevanoska E., Siljanoska Simons M. & Velinov G. "**A Genetic Algorithm Approach for RLE Compression in a Column Store Table**" (in submission to *Soft Computing, Springer*) – preprints available online

CONFERENCE POSTERS

- Arsov N., Pavlovski M., Basnarkov L. & Kocarev, L. "**Collaborative Bagging of Boosting Ensembles**", presented at the *SEE Forum on Data Science, Belgrade, June, 2016* – electronic version available online

JOURNAL PUBLICATIONS IN PREPARATION

- Arsov N., Pavlovski M., Basnarkov L. & Kocarev, L. "**Collaborative Bagging of Boosting Ensembles**"

Machine learning research in a collaborative approach to combining boosting (Gentle Boost) ensembles by bootstrap aggregation (bagging) into a complex multi-model. The algorithm introduces a novel collaborative approach. Being theoretically faster than existing boosting methods in machine learning, the algorithm displays an emphatic reduction of the generalization error rate. It presents a new theoretical stability-based approach. Preprint will be available upon submission of the manuscript. I am working on this research since January, 2015 at the Macedonian Academy of Sciences and Arts.

Research supervised by Professors Ljupco Kocarev and Lasko Basnarkov.

- Pavlovski M., Arsov, N. & Kocarev, L. "**Unbiased Ensemble Generation Through Population Distribution Approximation**" – **Working Title**

Statistical assumptions in sample data engender application of various approximation techniques to observed data. Classical ensemble methods, such as Bagging are known to reduce variance, but not bias. We use Gaussian approximations to generate simulated data in order to construct an ensemble less biased than ones generated by other methods. Synthetic samples are "chaotically" generated and filtered through an Apriori-based approach to capture the finest details of feature distributions, based on training data. This artificial "validation" set yields better generalization performance.

The research is taking place in the Macedonian Academy of Sciences and Arts, supervised by Prof. Ljupco Kocarev.

- Velinov G., Arsov N., Dimovski A. & Kon-Popovska M. "**PA2DP: from Predicate Abstraction to Predictive and Adaptive Data Partitioning**"

Research in data warehousing that proposes a theoretical model for predictive and adaptive partitioning of data warehousing systems (mostly OLAP). The approach uses global optimization to adaptively find an optimal partitioned warehouse schema design, where quality is measured with respect to realistic workloads. The method is based on database statistics estimation, using special data structures, making it applicable in real systems.

Research supervised by Prof. Goran Velinov.

COMPUTER SKILLS

Programming Languages

- C, C++, Java, Python, R, Lisp, Embedded C, Assembler (MIPS IS, PIC and basic experience in IA-64), basic C#

Computational software

- MATLAB/Octave, Mathematica

Other

- DBMS: OracleDB, PostgreSQL, MySQL, SQL and PL/SQL (Oracle/PostgreSQL dialects), Hadoop, VoltDB, Big Data, Git, MS Office

PROJECTS

Exploiting the power of modern processors and architecture-specific compiling techniques in numerical algorithms

July, 2015

A research project in the Advanced Computer Architecture course focused on analysis of feasible optimizations of numerical algorithms (Gauss-Jordan Elimination and Doolittle's LU Factorization) by the Intel C++ Compiler on an Intel Core i7 CPU. The two not unwittingly naïve C++ implementations were optimized on their own merits by the compiler and manually imported compiler directives, without changing the code. This eventually showed that Gauss-Jordan Elimination, a three times slower algorithm, can be cleverly optimized to run up to ten times faster. This inverts the theoretical algorithm complexity relationship between these two and brings to light the obscurity of algorithm complexity on modern CPUs.

Text retrieval from the Amazon Common Crawl Corpus of crawled web pages

September, 2014 – June, 2015

A research project in the Distributed Systems course, supervised by Professors Boro Jakimovski and Goran Velinov. The project involved the research and implementation of specific Map-Reduce jobs in Hadoop for extraction of large volumes of data from the Amazon Common Crawl corpus, tested on the local CS department cluster. The second, alternative approach is text retrieval from Common Crawl Index of WARC files, organized as a B-tree index over the corpus. The extracted data is intended for a transfer to a PostgreSQL database used in the plagiarism-detection software PlaxHunter, developed by Plannow Technologies, Inc.

Analysis of Spearman's ranked correlation as a measure of similarity in text mining

November-December, 2014

The final research team project in the Research Methodologies in ICT course. The project is a continued idea from my independent research project in the Intelligent Systems course. The paper was graded the best among 66 student research papers. The paper focuses on the analysis of Spearman's ranked correlation coefficient as a measure of text similarity in documents represented in a vector space by the Term Frequency – Inverse Document Frequency (TF-IDF) values. We have discovered that this measure can capture the semantics in a sentence better than other measures, regardless of the order of words.

Data mining in the "Wearable Computing: Classification of Body Postures and Movements" dataset

February – June, 2014

A semester-long project in the Data Mining course. It involved data preprocessing, cleaning and an in-depth statistical analysis of different data mining techniques and algorithms from both supervised and unsupervised learning. I have applied classification as well as clustering algorithms in order to detect additional hidden connections between the five classes of body postures, such as a strong correlation between walking and sitting, presumably concealed by the placement of wearables. These connections were not identifiable from the provided class labels and showed that the placement of wearables had to be changed in order to distinct the five desired body postures. The project was supervised by professor Zaneta Popeska.

Creation and design of an online programming competition system inspired by Top Coder

February – May, 2013

This team project was part of the Databases course. The system was implemented in Oracle Database 11c and Oracle Application Express.

WORKING EXPERIENCE

- Volunteering researcher at the Macedonian Academy of Sciences and Arts (Laboratory for Complex Systems and Networks at the Research Center for Computer Science and Information Technologies) January, 2015 - present

INTERNSHIPS

- DIS DOO July – August, 2014
In collaboration with FCSE, Skopje – Implementation of the TPC-C Benchmark in VoltDB (distributed database system) using Java and setup of a VoltDB distributed experimental environment in FCSE's Nebula Cloud infrastructure.
- FIDKO DOOEL July – September, 2013
In collaboration with FCSE, Skopje – Implementation of a genetic algorithm in C++ for column-store database compression by run-length encoding using heuristic column-reordering approaches. Additionally, byte-aligned variable-length encoding was implemented to optimize compression ratios.
- Tutoring Students 2011-2015
Programming-related courses, Calculus, Probability and Statistics, Discrete Mathematics.

LANGUAGES

- English – fluent
- German – basic proficiency