The IADIS European Conference on Data Mining 2007 took place in Lisbon, Portugal, 5-7 July, 2007. This conference was part of the Multi Conference on Computer Science and Information Systems 2007 (MCCSIS), 3 - 8 July 2007, which had a total of 1091 submissions.

This Conference was the First European Conference on Data Mining (ECDM’07). It aimed to gather researchers and application developers from a wide range of data mining related areas such as statistics, computational intelligence, pattern recognition, databases and visualization. ECDM’07 had the goal to advance the state of the art in data mining field and its various real world applications. It provided opportunities for technical collaboration among data mining and machine learning researchers around the globe.

The IADIS European Conference on Data Mining 2007 received 81 submissions from more than 24 countries. Each submission had been anonymously reviewed by an average of four independent reviewers, to ensure that the final accepted submissions were of a high standard. Therefore only 15 full papers were approved which meant an acceptance rate below 19 %. A few more papers were accepted as short papers, reflection papers and posters. The best paper authors were invited to publish an extended version of their paper in the IADIS International Journal on Computer Science and Information Systems (ISSN: 1646-3692) and also in the following Journals:
- International Journal of Computational Intelligence Research (special issue on advances in data mining);
- Journal of Information Assurance and Security (special issue on data mining and information security).

The submissions were accepted under the following areas of interest:

**Core Data Mining Topics**
- Parallel and distributed data mining algorithms
- Data streams mining
- Graph mining
- Spatial data mining
- Text video, multimedia data mining
- Web mining
- Pre-processing techniques
- Visualization
- Security and information hiding in data mining

**Data Mining Applications**
- Databases
- Bioinformatics
- Biometrics
- Image analysis
- Financial modeling
- Forecasting
- Classification
- Clustering

In addition to the presentation of full papers, short papers, reflection papers and posters, the conference also included two keynote presentations from internationally distinguished researchers Professor Janusz Kacprzyk, Deputy Director for Scientific Affairs, Polish Academy of Sciences, Poland, and Professor Bogdan Gabrys, Computational Intelligence Research Group, Bournemouth University, UK. The Conference also offered a tutorial by Professor Janusz Kacprzyk.


Overall the Conference offered an opportunity to all their participants to discuss with success the most significant aspects regarding the theme Data Mining. It served as a forum that gathered researchers, practitioners, students and anyone that was working or studying in the field of the Data Mining.

The IADIS DM 2007 Conference was supported by FCT Fundação para a Ciência e a Tecnologia and Luso-American Foundation.
Keynote Presentations:

K.1 - TO COMBINE OR NOT TO COMBINE? MULTIPLE CLASSIFIER AND PREDICTION SYSTEMS by Prof. Bogdan Gabrys, Computational Intelligence Research Group, Bournemouth University, UK

Abstract:

Automatic classification or prediction model building directly from data with a minimal or no human supervision is already absolutely crucial in order to stay competitive and maximally exploit the data in quickly changing business environments. However, the methodology for ensuring that created models (i.e. classifiers, predictors, etc.) are as good as possible should be in place before using them with confidence. No human supervision in model building also implies that one should use powerful enough techniques which can learn the data to any degree of accuracy. There are currently a lot of methods from soft computing, machine learning, and statistics domains which, in principal, satisfy this requirement.

In the pattern recognition domain the examples include the nearest-neighbour classifiers, decision trees, neural networks with sufficient number of hidden nodes, fuzzy if–then rules systems which are built directly from data, various neuro-fuzzy techniques including those based on hyperbox fuzzy sets, Bayesian networks or logical rule bases. From the statistical point of view most of these methods could be classified as non-parametric models. The main challenge in such cases is to design a model building strategy which would guard against overfitting of the training data or, in other words, would lead to a good generalisation performance.

Individual classification/prediction models have recently been challenged by combined pattern recognition or prediction systems, which often show better performance. This talk will be concerned with pattern classification and prediction problems and in particular the motivations and challenges behind using and designing multiple classifier systems. Starting with an attempt to answer the question of why would one like to combine classifiers we will move to an overview of how to combine them. This will be followed by discussing the issues of majority voting limits and an illustration of the potential enhancement for theoretical error limits when using hierarchical multistage organisations for majority vote based combination systems. The talk will also cover the issues of classifier diversity, classifier selection, various practical evolutionary algorithm based implementations of multistage, multilevel and multidimensional selection-fusion models.

Alternatives to combining decisions of the predictors in the context of neuro-fuzzy classifiers together with a number of algorithm independent learning approaches attempting to answer the main question: To combine or not to combine? will also be discussed. The results of extensive testing for many different benchmark problems will be shown throughout the presentation.

K2 - LINGUISTIC DATA SUMMARIES VIA FUZZY LOGIC AS SIMPLE, EFFICIENT AND HUMAN CONSISTENT DATA MINING MEANS by Prof. Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences Ul, Warsaw, Poland

Abstract:

We view data mining as the process of extraction of potentially useful, previously unknown or not explicitly known, information from data meant as large data sets or databases. We assume that, first, at the data mining process is to serve a human being who needs that information for some purposeful activities exemplified by making decisions. Second, we assume that the sheer size of these data sets makes them too large for human comprehension and therefore the process requires a considerable data compression (summarization). These two assumptions of a central role of a human being and a need for a massive summarization imply that the process should be motivated by a choice of tools that are the most human consistent. From this perspective, since for a human being the only fully natural means of communication
and articulation is natural language, one can argue that natural language is the ultimate tool on which such human consistent data mining tools should preferably be based. This philosophy is advocated in this talk.

First, we briefly review some natural language oriented approaches to data mining, notably those involving the natural language generation (NLG) tools and techniques, and point out that although they provide powerful tools, they cannot express in a simple and efficient way imprecision of natural language that is crucial in our context.

We advocate the use of Zadeh’s computing with words and perceptions, based on fuzzy logic, as a viable solution that is simple and efficient.

We present first the concept of a linguistic data(base) summary in the sense of Yager (1982), notaby in its Kacprzyk and Yager’s (2001) and Kacprzyk, Yager and Zadrozny’s (2001) extensions. Basically, its essence is that if we have a (large) set of numerical data (base), e.g. of employees, and we specify an attribute of interest, e.g. age, then such a linguistic summary can be “most employees are young” with some degree of validity (truth). This simple form provides a highly summarized and granulated view of the very essence of data.

We present some extensions of this concept, notably Kacprzyk and Zadrozny’s (1998 – 2007) view of linguistic summaries in terms of protoforms, their relation to other types of summarizations, notably association rules and if – then rules, and the role of user interaction, fuzzy querying and GUIs for the derivation of such summaries. Then, we present Kacprzyk and Zadrozny’s implementations for mining sales data of a small to medium computer retailer.

We extend the concept of a linguistic data summary as mentioned above to the summarization of dynamic sets of data, specifically time series. First, we mention a natural language generation based solution adopted in the SumTime project (EPSRC Funded Project for Generating Summaries of Time Series Data - www.csd.abdn.ac.uk/research/sumtime/) that is close in spirit. We indicate that it does not address the problem of imprecision and varying granularity of our intended natural language expressions, and we present a new approach initiated in a series of papers by Kacprzyk, Wilbik and Zadrozny (2006 - 2007) who provided tools for extracting such linguistic summaries of time series as: "Most of the trends are short", "Most long trends are increasing", "Increasing trends taking most of the time are of a low variability", etc. We show the role of a proper linguistic quantifier driven aggregation of partial scores (trends) via the original Zadeh’s calculus of linguistically quantified proposition and some other methods like the OWA operators, the Sugeno and Choquet integrals.

We present an application of the new technique to the summarization of time series data on daily quotations of a mutual fund and show some linguistic summaries that can be useful for making investment decisions both by an individual and institutional level.

We conclude with some remarks on possible and promising future directions, notably along the philosophy of a human centric computing.

Conference Tutorial:

**- FROM FUZZY LOGIC TO COMPUTING WITH WORDS AND PERCEPTIONS** by Prof. Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences Ul, Warsaw, Poland

Abstract:

First, we present an overview of Zadeh’s fuzzy sets theory, notably the concepts of a fuzzy set, fuzzy relation, etc. We mention basic set theoretic, algebraic, etc. operations. We present an overview of Zadeh’s linguistic approach to the analysis of systems that includes linguistic variables, fuzzy conditional statements, the conditional rule of inference, and fuzzy algorithms. Then, we outline Zadeh’s paradigm of computing with words and perceptions assuming both a logical and systems analytic perspective. We mention relations to some elements of natural language, notably modalities. Our discussion will be illustrated with examples of applications in decision analysis, data analysis and data mining.

Best Paper:

**- MODELING MISSING DATA WITH MARKOV RANDOM FIELDS IN LARGE DATA SETS** by Esa Junttila and Marko Salmenkivi University of Helsinki, Finland
Abstract:

A key issue in data analysis is the treatment of missing data. In spatial domain the assumption of autocorrelation is often employed to make inference about missing data from the observations made in nearby areas. Bayesian data analysis methods provide a well-founded framework for the statistical inference. Due to the computational requirements, they have been used almost exclusively in confirmatory data analysis. Our basic idea is to utilize Bayesian methods as a preprocessing phase in the KDD process. We employ relatively simple but reasonable models, and apply them to large data sets. The resulting posterior distributions can then be analyzed by other data mining methods. In particular, we analyze a large linguistic data set: 17,100 geographic distributions of Finnish dialect words.

Committees:

Program Chair: Ajith P. Abraham, School of Computer Science, Chung-Ang University, South Korea

Conference Co-Chairs:

- Piet Kommers, University of Twente, The Netherlands
- Pedro Isaías, Universidade Aberta (Portuguese Open University), Portugal
- Nian-Shing Chen, National Sun Yat-sen University, Taiwan

Committee Members:

- Alexander Gelbukh, National Polytechnic Institute, Mexico
- Alexandros Nanopoulos, Aristotle University of Thessaloniki, Greece
- Aixun Sun, Nanyang Technological University, Singapore
- Akihiro Inokuchi, Tokyo Research Laboratory, IBM Japan
- Alfredo Cuzzocrea, University of Calabria, Italy
- Andreas König, TU Kaiserslautern, Germany
- Annalisa Appice, Università degli Studi di Bari, Italy
- Arianna Gallo, Dipartimento di Informatica Università degli Studi di Torino, Italy
- Arpad Kelemen, University of Niagara, USA
- Artchil Maysuradze, Moscow University, Russia
- Carlotta Domeniconi, George Mason University, USA
- Chiara Renso, ISTI-CNR, Italy
- Claudio Lucchese, Università Ca’ Foscari di Venezia, Italy
- Cosmin Danut Bocaniala, University of Lancaster, UK
- Daniel Pop, University of the West Timisoara, Romania
- Daniela Zaharie, University of the West Timisoara, Romania
- David Cheung University of Hong Kong, Hong Kong
- Daxin Jiang, Nanyang Technological University, Singapore
- Deepayan Chakrabarti, Yahoo! Research, USA
- Dimitrios Katsaros, Aristotle University of Thessaloniki, Greece
- Ruoming Jin, Kent State University, USA
- Edward Hung, Hong Kong Polytechnic University, Hong Kong
- Enrico Blanzieri, University of Trento, Italy
- Evimaria Terzi, University of Helsinki, Finland
- Frans Coenen, Liverpool University, UK
- Gabriela Kokai, Friedrich Alexander University, Germany
- Gagan Agarwal Ohio State University, USA
- George Pallis, Aristotle University of Thessaloniki, Greece
- Graham Williams, Togaware Data Mining, Australia
- Gregorios Tsoumakas, Aristotle University of Thessaloniki, Greece
- Guoren Wang, Northeastern University, China
- Hidenao Abe, Shimane University, Japan
- Hisao Ishibuchi, Osaka University, Japan
- Hong Gao, Harbin Institute of Technology, China