# Rough Sets: Theory and Applications

## Workshop at the Joint Rough Set Symposium

July 9th, 2014

Granada, Spain




# Program

11h30 Fuzzy Rough Support Vector Machine for Data Classification
Arindam Chaudhuri

11h50 The Extension of Dominance-Based Rough Set Approach for Solving MCDM Problems
Kao-Yi Shen and Gwo-Hshiung Tzeng

12h10 Professional Profiles and Personality Traits for effective Team Building Processes
Orestes Evangelinos, Christos Zigkolis and Athena Vakali

12h30 Expert-based system for decision making in the ICT industry
Andrés Cid-López, Miguel J. Hornos, Ramón A. Carrasco Gónzalez and Enrique Herrera-Viedma

12h50 Short break

13h00 Instance Selection for Imbalanced Data
Sarah Vluymans

13h20 Using similarity relations in generation of prototypes for problems of classification
Yumilka Fernández, Rafael Bello, Mabel Frias, Yaima Filiberto and Yaile Caballero

13h40 On Exactness, Definability and Vagueness in Partial Approximation Spaces
Davide Ciucci, Tamás Mihálydeák and Zoltán Ernõ Csajbók

14h00 Churn Prediction in Telecommunication Industry Using Rough Set Approach
Adnan Amin, Changez Khan, Imtiaz Ali and Sajid Anwar

16h00 Automated Algorithm Configuration: Beyond Parameter Tuning
Thomas Stützle

17h00 A Concept Analysis View of Rough Sets
Yiyu Yao and JingTao Yao

# Abstracts

## Fuzzy Rough Support Vector Machine for Data Classification

Arindam Chaudhuri[1]
[1]Faculty of Post Graduate Studies and Research, Computer Engineering and Technology, Marwadi Education Foundation's Group of Institutions, Rajkot, India
arindam_chau@yahoo.co.in

Classification of data has been actively used for most effective and efficient means of conveying knowledge and information to users. The prima face has always been upon techniques for making correct decision and extracting useful knowledge from data such that returns are maximized. The challenge lies in analyzing and understanding characteristics of datasets by retrieving useful geometric and statistical patterns. In this direction, support vector machine (SVM) and its variants have emerged as promising classification tools based on the idea of structural risk minimization. The classification task is taken care of by fuzzy rough support vector machine (FRSVM) with hyperbolic tangent kernel. It is a variant of fuzzy support vector machine (FSVM) and modified fuzzy support vector machine (MFSVM). The fuzzy rough set model takes care of the sensitiveness of noisy samples and handles impreciseness in training samples bringing robustness to classification results. The membership function is developed as function of center and radius of each class in feature space and representing it with kernel. It plays an important role towards sampling the decision surface. The success of FRSVM is governed by the choosing appropriate parameter values. The training samples are either linear or nonlinear separable. In nonlinear training samples, input space is mapped into high dimensional feature space to compute separating surface using linear separating method. The different input points make unique contributions to decision surface. The performance of the classifier is also assessed in terms of the number of support vectors. The effect of variability in prediction and generalization of FRSVM is examined with respect to several values of parameter ??. FRSVM effectively handles the nonlinear classification problem. Experimental results on both synthetic and real datasets support the fact that FRSVM achieves superior performance in reducing outliers' effects than existing SVMs.

# The Extension of Dominance-Based Rough Set Approach for Solving MCDM Problems

Kao-Yi Shen[1] and Gwo-Hshiung Tzeng[2]
[1]Department of Banking and Finance, Chinese Culture University (SCE), Taipei, Taiwan
[2]Graduate Institute of Urban Planning, College of Public Affairs, New Taipei City, Taiwan
kyshen@sce.pccu.edu.tw
ghtzeng@mail.ntpu.edu.tw

The use of rule-based model for solving MCDM problems has been widely applied in various fields, such as engineering, marketing, and finance. And the adoption of dominance-based rough set approach (DRSA) has ad-vantages in generating understandable decision rules with reduced attributes, which is useful while facing complex decision problems. DRSA considers the preferential characteristic of attributes in making classification, which makes it suitable to solve multiple criteria decision problems in essence. However, the conventional DRSA method has several limitations in solving certain MCDM problems, such as: (1) define suitable interval ranges for the discretization of at-tributes by natural language; (2) make ranking or selection within the same decision class or un-defined decision class; (3) extend the ranking problem into improvement planning. Therefore, there is a rising trend (need) in integrating or extending DRSA with the other soft computing techniques or MCDM methods to resolve practical problems. Based on this observation, this study attempts to suggest (propose) several approaches to infuse or transform DRSA to address the aforementioned three limitations. First, fuzzy artificial neural network (FANN) is suitable to learn the fuzzy intervals of each attribute in rules with minimized modeling errors, and the authors used this approach in predicting the financial performance of commercial banks with positive outcome in 2013. Second, a rules-based probabilistic weighted decision approach (RPDA) has been proposed in this conference (JRS2014) to make ranking within the same decision class or undefined decision class. Third, the compromise outranking method VIKOR has also been proposed to integrate the probabilistic weights of criteria by the proposed RPDA method to make improvement planning with an empirical case (in JRS2014). The proposed approaches hope to extend DRSA applications into the field of MCDM for solving practical problems with fewer limitations.

# Professional Profiles and Personality Traits for effective Team Building Processes

Orestes Evangelinos[1], Christos Zigkolis[1], Athena Vakali[1]
[1]Informatics Department Aristotle University of Thessaloniki, Thessaloniki, Greece
{evanores,chzigkol,avakali}@csd.auth.gr

Today's increasing hurriedness in our way of life along with technology advances demand from businesses to react always in a faster mode, while maintaining or, better yet, improving the quality of their products and services. Towards that objective, businesses can gain benefits by employing redefined team-building processes through which teams will be generated in a more cohesive and efficient manner. Aiming at forming productive and effective teams, we present PROTEAS, a framework where professional profiles and personality traits are both taken into account in the team-building process. The professional portrait of users is drawn from their LinkedIn accounts, while hints of their personality are obtained by a well-defined questionnaire based on the Big Five Factor Model. The significance of personality is substantiated by recent research work where it has been proven that personality traits can play a vital role in group dynamics. Motivated by the latter, PROTEAS defines a series of variables where characteristics of team members (e.g. skills, education) are combined with personality traits towards identifying candidate teams. All candidates are evaluated via a weightbased algorithm which calculates pairwise similarities among teams' members and ranks the teams accordingly at the end. To assess the effectiveness of PROTEAS an academic setting has been chosen. A group of students were given two assignments in teams of up to 4 people, while they've also completed the IPIP1 personality questionnaire. Through the experimentation, a number of interesting observations occurred: (a) programming assignments got higher grades when the team members have shown high similarity scores regarding their technical skills, (b) the absence of 'Agree-ableness' personality trait has led to disagreements among members through which new ideas were put on the table, which eventually produced more effective solutions, (c) the presence of 'Conscientiousness' trait among members indicated a more focused team which has completed the assignment in a faster manner. This work has laid down the foundation stone for further development in the future and our intention is to extend our framework in many aspects and in particular, applying PROTEAS in a business setting where employees have more diverse backgrounds.

# Expert-based system for decision making in the ICT industry

Andrés Cid-López[1], Miguel J. Hornos1, Ramón A. Carrasco-Gónzalez[2], and Enrique Herrera-Viedma[1]
[1]Universidad de Granada, Granada, Spain
[2] Universidad Complutense de Madrid
fandrescid,mhornosg@ugr.es; ramoncar@ucm.es; viedma@decsai.ugr.es

In their day to day, most of the companies from the Information and Communication Technologies (ICT) sector face decision problems related to the need for investment in their business, task that requires a significant dedication due to the high economic values these decisions entail. Moreover, the business dynamics driven by the technological

changes and the evolution of the business itself, as well as the thrust of competition create a scenario in which making decisions late or hastily can have a negative impact on these companies' expectations. This will be directly reflected in the income statements of the corresponding economic indicators.
Although these companies allocate significant amounts of resources to feasibility studies, market research, technology trends, etc., there are always qualitative factors influencing the direction to follow that cannot be easily quantified in monetary terms. In this respect, experts play a vital role, as they will be called to clarify the utility offered by new technological developments and their true applicability within an increasingly demanding market, in order to make certain investments or not.
We propose a linguistic multicriteria decision making model based on pertinent information collected through online surveys that allow us to know what the users think regarding the use of ICT. This model, which supports decision making in ICT investments, implements a hybrid system, in the sense that it is made up of a group of previously selected experts (humans) with extensive experience in this type of decision-making problems and a computer model (system) which provides relevant information to be taken into account in the process. In this model, the importance given to each individual of this expert group (according to his/her expertise degree) plays a significant role. This will be determined by the assignment of the corresponding weights, which will lead to a more consistent decision.
Additionally, our proposal is based on a 2-tuple linguistic computational model, which uses fuzzy logic for processing data (i.e. words) without loss of information. Moreover, different input data sources could be used to feed this model.

# Instance Selection for Imbalanced Data

Sarah Vluymans[1]
[1]Ghent University, Ghent, Belgium
Sarah.Vluymans@UGent.be

Class imbalance presents itself in many real-world applications, like anomaly detection and several branches of the medical domain. A dataset is considered to be imbalanced when it displays an unequal distribution of its classes. In particular, when considering two-class imbalanced problems, one class can be denoted as the majority class and the other as the minority. The cardinality of the former can be considerably larger than that of the latter.
A classifier aims to assign newly presented instances to the correct class. It does so by using a classification model, which it constructs based on information contained in a training set of correctly labeled instances. It has been shown that class imbalance within this set can severely hinder the ability of the classifier to recognize any unseen minority instances as belonging to their own class.

To resolve this latter issue and enhance the performance of the classifier, we develop a set of new preprocessing techniques, the $IS_{Imb}$ methods, which are based on existing Instance Selection (IS) methods. Preprocessing by IS selects a subset of the available training instances to use in the construction of the classification model. In general, when executed on the training set, IS methods can improve the classification process. However, the existing methods have proven to be hindered by the data imbalance as well and do not yield the desired results when the training set is imbalanced. Our contribution is the modification of 33 existing methods for the specific application on imbalanced datasets and we show that these algorithms do lead to a significant improvement of the classification.
We conducted an extensive experimental evaluation of the 33 newly proposed $IS_{Imb}$ methods. Our new methods significantly improve the original IS methods, as well as the baseline classification without preprocessing. We also observe that our methods are highly competitive with the state-of-the-art techniques in this domain and we can place $IS_{Imb}$ among the best solutions to enhance the classification of imbalanced data.

# Using similarity relations in generation of prototypes for problems of classification

Yumilka B. Fernández[1], Rafael Bello[2], Yaima Filiberto[1], Mabel Frías[1] Yaile Caballero[1]
[1]Departamento de Computación, Universidad de Camagüey,Camagüey, Cuba
[2]Departamento de Ciencia de la Computación. Universidad Central de Las Villas, Santa Clara, Cuba
{yumilka.fernandez, yaima.filiberto, mabel.frias,yaile.caballero}@reduc.edu.cu
rbellop@uclv.edu.cu

This paper proposes a new method for constructing prototypes, using similarity relations to perform granulation of the universe. The main motivation is based on the existence of the algorithm NP-BASIR (presented by Bello in MICAI 2013) used for function approximation; taking this into account comes up NP-BASIR-Class for building prototypes.
The Prototype Generation aims at getting a training set, as small as possible, that allows classifying with the same or better quality than the original training set. This reduces the space complexity of the method and reduce the computational cost. Besides, its accuracy can be sometimes improved by eliminating noise.
After analyzing various approaches found in the literature, this paper proposes a new method for constructing prototypes in classification problems based on the concepts of granular computing. The granulation of the universe is made from a similarity relation, this generates similarity classes of objects in the universe, and to each kind of similarity a prototype is built. The method proposed by Filiberto in ISDA2010 is used to construct the similarity relation.
The purpose the NP-BASIR- Class is to build a prototype or centroid for a set of similar

objects. A set of prototypes called ProtoSet is obtained as the purpose of the method. This set is used by the classifier to classify new instances.
The classifier from a new x object and the ProtoSet set calculates the similarity between x and each prototype; then selects the most similar k prototypes and calculates the class of x from the class of the selected k prototypes. In the study, the presented algorithm is compared to 12 algorithms, mentioned in the article A Taxonomy and Experimental Study on Prototype Generation for Nearest Neighbor Classification by Triguero, Derrac and Garcia in 2012. The algorithms that according to this article offer the best results and are implemented in the tool KEEL are LVQ3, GENN, DSM, VQ, MSE, ENPC, AVQ, PSCSA, Chen, HYB, PSO, and SGP.
To evaluate the results of NP-BASIR-Class multiple comparison test in order to find the best algorithm were used. Experimental results show that the proposed method had the best ranking, so that they can determine which is statistically higher than the other in terms of classification accuracy. The reduction factor achieved was also analyzed, showing satisfactory results.

# On Exactness, Definability and Vagueness in Partial Approximation Spaces

Davide Ciucci[1], Tamás Mihálydeák[2], and Zoltán Ernõ Csajbók[3]
[1]Dipartimento Di Informatica, Sistemistica e Comunicazione, Universitá di Milano-Bicocca, Milano, Italia
[2]Department of Computer Science, Faculty of Informatics, University of Debrecen, Debrecen, Hungary
[3]Department of Health Informatics, Faculty of Health, University of Debrecen, Nyíregyháza, Hungary
ciucci@disco.unimib.it
mihalydeak.tamas@inf.unideb.hu
csajbok.zoltan@foh.unideb.hu

In our generalized approach to rough set theory, lower/upper, boundary, and negative regions/operators of set approximations have been considered as primitive ones. Usually, they are not independent, that is two of them are sufficient to define the others. In this talk, this dependence is relaxed and assumed that they are independent of each other and we study how they can interact.
First, a general basic framework for our investigations is outlined. In the approximation process, the main building blocks are base sets (equivalence classes in the standard case) and definable sets (union of equivalence classes) obtained by some set operations on base sets. In a general setting, the base sets do not necessarily form a covering. Definable sets constitute all the available knowledge about the objects of interest. Consequently, lower/upper approximations, boundaries and negative sets should be all considered as definable sets. Vagueness has a central role in the motivation basis of rough set theory.

Vagueness in Pawlak's information-based proposal was expressed by the boundary regions of sets represented by the difference of upper and lower approximations. Definability, exactness and roughness of sets are defined via these differences as well. However, in general situations, the above notions and their relations can become more tricky. For example, we cannot definitely say neither that definable sets are exact nor that exact sets are definable. In standard, i.e., equivalence based rough set theory, boundaries can be defined in three equivalent ways. In generalized theories, however, these definitions lead different notions of boundaries. Therefore, the mutual relations among lower/upper approximations, boundary and negative region have to be rethought.

# Churn Prediction in Telecommunication Sector using Rough Set Approach

Adnan Amin[1], Changez Khan[1], Imtiaz Ali[1], Sajid Anwar[1]
[1]Institute of Management Sciences, Hayatabad Phase 7, Peshawar Pakistan
{geoamins, sajidanwar.2k}@gmail.com

The Customer churn is a crucial activity in rapidly growing and mature competitive telecommunication sector and is one of the greatest importance for a project manager. Due to the high cost of acquiring new customers, customer churn prediction has emerged as an indispensable part of telecom sectors strategic decision making and planning process. It is important to forecast customer churn behavior in order to retain those customers that will churn or possible may churn. This study is another attempt which makes use of rough set theory, a rulebased decision making technique, to extract rules for churn prediction. Experiments were performed to explore the performance of four different algorithms (Exhaustive, Genetic, Covering, and LEM2). It is observed that rough set classification based on genetic algorithm, rules generation yields most suitable performance out of the four rules generation algorithms. Moreover, by applying the proposed technique on publicly available dataset, the results show that the proposed technique can fully predict all those customers that will churn or possibly may churn and also provides useful information to strategic decision makers as well.

# Invited talks

## Automated Algorithm Configuration: Beyond Parameter Tuning

Thomas Stützle
Université Libre de Bruxelles, Brussels, Belgium
stuetzle@ulb.ac.be

The design and configuration of optimization algorithms for computationally hard problems is a time-consuming and difficult task. This is in large part due to a number of aggravating circumstances such as the NP-hardness of most of the problems to be solved, the difficulty of algorithm analysis due to stochasticity and heuristic biases, and the large number of degrees of freedom in defining and selecting algorithmic components and settings of numerical parameters. Over the recent years, automatic algorithm configuration methods have been developed to effectively search large configuration spaces for superior algorithm designs. These methods have by now proved to be instrumental for developing high-performance algorithms.
In this talk, we will describe the main automatic algorithm configuration techniques, present the main available configuration tools and highlight various successful applications of automatic algorithm configuration to the generation of hybrid stochastic local search algorithms, the design of multi-objective optimizers, and the improvement of algorithm anytime behavior. Finally, I will argue that automatic algorithm configuration will transform the way heuristic algorithms are designed and developed in the future.

## A Concept Analysis View of Rough Sets

Yiyu Yao and JingTao Yao
Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada
jtyao@cs.uregina.ca
yyao@cs.uregina.ca

Rough set theory was proposed by Pawlak for analyzing data and reasoning about data. From a concept analysis point of view, we review and reformulate main results of rough set theory in the context of data processing and analysis. We present a formulation of rough set theory consisting of two models. A semantically sound model for defining rough set approximations and a computationally efficient model for constructing approximations. This enables us to see clearly the motivations for introducing rough set theory, its basic components and appropriate applications, leading to an appreciation for the theory.