Scientific Report

First name / Family name	Indrajit Saha
Nationality	Indian
Name of the *Host Organisation*	University of Wroclaw
First Name / family name of the *Scientific Coordinator*	Leszek Pacholski
Period of the fellowship	01/04/2014 to 31/03/2015

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

During this fellowship, I have done the scientific activities like research, publication in peer-referred journals and conferences, participation in conference and workshop, visiting four other institutes for one week each and serving as a member of the board of reviewers and program committee of various international journals and conferences. The details can be found below.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

Published/Accepted

[1] S. Dey, **I. Saha**, U. Maullik and S. Bhattacharyya, "Multi-level Thresholding using Quantum Inspired Meta-heuristics", Knowledge-Based Systems, Vol. 67, pp. 373-400, 2014. DOI: doi:10.1016/j.knosys.2014.04.006 [Impact Factor: 3.06]

Abstract:
Image thresholding is well accepted and one of the most imperative practices to accomplish image segmentation. This has been widely studied over the past few decades. However, as the multi-level thresholding computationally takes more time when level increases, hence, in this article, quantum mechanism is used to propose six different quantum inspired meta-heuristic methods for performing multi-level thresholding faster. The proposed methods are Quantum Inspired Genetic Algorithm, Quantum Inspired Particle Swarm Optimization, Quantum Inspired Differential Evolution, Quantum Inspired Ant Colony Optimization, Quantum Inspired Simulated Annealing and Quantum Inspired Tabu Search. As a sequel to the proposed methods, we have also conducted experiments with the two-Stage multithreshold Otsu method, maximum tsallis entropy thresholding, the modified bacterial foraging algorithm, the classical particle swarm optimization and the classical genetic algorithm. The effectiveness of the proposed methods is demonstrated on fifteen images at the different level of thresholds quantitatively and visually. Thereafter, the results of six quantum meta-heuristic methods are considered to create consensus results. Finally, statistical test, called Friedman test, is conducted to judge the superiority of a method among them. Quantum Inspired Particle Swarm Optimization is found to be superior among the proposed six quantum meta-heuristic methods and the other five methods are used for comparison. A Friedman test again conducted between the Quantum Inspired Particle Swarm Optimization and all the other methods to justify the statistical superiority. Finally, the computational complexities of the proposed methods have been elucidated for the sake of finding out the time efficiency of the proposed methods.

[2] **I. Saha**, J. P. Sarkar and U. Maulik, "Ensemble based Rough Fuzzy Clustering for Categorical Data", Knowledge-Based Systems, Vol. 77, pp. 114-127, 2015. DOI:10.1016/j.knosys.2015.01.008 [Impact Factor: 3.06]

Abstract:
Categorical data is different from continuous data, where the values of attribute do not follow any natural ordering. Moreover, inherent complexities like uncertainty, vagueness and overlapping among clusters make the analysis of real life categorical data set more difficult. Recent literature review shows that the well-known categorical data clustering techniques are using different similarity/dissimilarity measures to tackle the inherent complexities of the categorical attribute values. Generally, it is hard to find single method and cluster validity measure that can be used as perfect or standard for all kinds of categorical data sets. Hence, in this paper first, a clustering method for categorical data is proposed by fusing rough set and fuzzy set theories. Subsequently, an ensemble based framework is designed with the recently proposed similarity/dissimilarity measures in order to have better clustering results for different types of categorical data sets. For this purpose, the proposed rough fuzzy clustering method is used sequentially with the integration of different measures to evolve the clustering solutions. Using consensus of these solutions, pure classified, semi rough and pure rough points are identified. Thereafter, machine learning method, called Random

Forest, is used in incremental way to classify the semi and pure rough points using pure classified points to yield better clustering results. The performance of the proposed method has been demonstrated in comparison with several other recently developed clustering methods. Additionally, the selection of Random Forest in the proposed framework is justified by comparing its performance with other well-known machine learning methods like K-Nearest Neighbor and Support Vector Machine. Ten categorical data sets are used for the experimental purpose. Finally, statistical significance test has been conducted to judge the superiority of the results.

[3] I. Saha, B. Rak, S. S. Bhowmick, U. Maulik, D. Bhattacharjee, U. Koch, M. Lazniewski, D. Plewczynski, "Binding Activity Prediction of Cyclin-Dependent Inhibitors", Journal of Chemical Information and Modeling, pp. 1469 - 1482, 2015. DOI: 10.1021/ci500633c [Impact Factor: 4.07]

Abstract:

The Cyclin-Dependent Kinases (CDKs) are the core components coordinating eukaryotic cell division cycle. Generally the crystal structure of CDKs provides information on possible molecular mechanisms of ligand binding. However, reliable and robust estimation of ligand binding activity has been a challenging task in drug design. In this regard, various machine learning techniques, such as Support Vector Machine, Naive Bayesian classifier, Decision Tree, and K-Nearest Neighbor classifier, have been used. The performance of these heterogeneous classification techniques depends on proper selection of features from the data set. This fact motivated us to propose an integrated classification technique using Genetic Algorithm (GA), Rotational Feature Selection (RFS) scheme, and Ensemble of Machine Learning methods, named as the Genetic Algorithm integrated Rotational Ensemble based classification technique, for the prediction of ligand binding activity of CDKs. This technique can automatically find the important features and the ensemble size. For this purpose, GA encodes the features and ensemble size in a chromosome as a binary string. Such encoded features are then used to create diverse sets of training points using RFS in order to train the machine learning method multiple times. The RFS scheme works on Principal Component Analysis (PCA) to preserve the variability information of the rotational nonoverlapping subsets of original data. Thereafter, the testing points are fed to the different instances of trained machine learning method in order to produce the ensemble result. Here accuracy is computed as a final result after 10-fold cross validation, which also used as an objective function for GA to maximize. The effectiveness of the proposed classification technique has been demonstrated quantitatively and visually in comparison with different machine learning methods for 16 ligand binding CDK docking and rescoring data sets. In addition, the best possible features have been reported for CDK docking and rescoring data sets separately. Finally, the Friedman test has been conducted to judge the statistical significance of the results produced by the proposed technique. The results indicate that the integrated classification technique has high relevance in predicting of protein−ligand binding activity

[4] J. P. Sarkar, **I. Saha** and U. Maulik, "A new SVM integrated Rough Type-II Fuzzy Clustering Technique", 9th IEEE International Conference ICIIS-2014, Gwalior, India, December 2014. DOI:10.1109/ICIINFS.2014.7036555

Abstract:

Clustering algorithms based on type-I fuzzy set theory have been used for handling overlapping partitioning area over the last few decades. However, these fail to deal with additional degree of fuzziness within the real life datasets, because the membership values of type-I fuzzy set are crisp real numbers. Therefore, since inception, the type-II fuzzy set theory has been studied to address the weakness of type-I fuzzy set theory as the membership value itself is fuzzy in type-II fuzzy set theory. On the other hand, rough set based clustering method helps in great extend to handle the inherent uncertainty and vagueness of the data with the concept of lower and upper approximation. However, in rough clustering, rough points are not so certain to a particular cluster. In that case, machine learning technique such as Support Vector Machine can be used to assign the rough points into proper clusters in order to get the better clustering result. Hence, in this article, Support Vector Machine integrated Rough type-II Fuzzy C-Means clustering technique using both the rough set and type-II fuzzy set theories, is proposed. The effectiveness of the proposed clustering technique has been demonstrated quantitatively and visually on several synthetic and real life datasets in comparison with other well-known clustering techniques. Finally, the superiority of the results produced by theproposed technique has been shown using statistical significance test.

[5] A. Lancucki, **I. Saha** and P. Lipinski, "Evolutionary Gene Selection using Stochastic Embeddings", International IEEE Conference on Evolutionary Computing, Sendai, Japan, pp. 1612-1619, May 2015.

Abstract:
Microarray technology allows to investigate gene expression levels by analyzing high dimensional datasets of few samples. Selection of discriminative, differentially expressed genes from such datasets is important to differentiate, prognose and understand the underlying biological processes. In this regard, the paper presents a new evolutionary gene selection method based on Student-t Stochastic Neighbor Embedding (t-SNE), Differential Evolution (DE) and Support Vector Machine (SVM). Here the underlying classification task of SVM is used as an optimization problem of DE, while t-SNE provides better ordering of genes for selection purpose. Generally, t-SNE is used to reorder the genes in such a way so that similar genes are grouped together and dissimilar genes are kept further apart. These reordered genes are then fragmented into fixed-length partitions. Thereafter, from each partition, a gene is selected randomly to encode the initial population of DE along with the combination of its weight and threshold values in order to participate in fitness computation. In the final generation of DE, a subset of genes is selected based on higher classification accuracy. The proposed technique is tested on six publicly available microarray datasets concerning various cancerous tissues of Homo sapiens and yields a potential set of genes by providing prefect or nearly perfect classification accuracy. Moreover, the superiority of the proposed technique has been demonstrated in comparison with other widely used techniques. Finally, the achieved results have also been justified by a statistical test and allowed us to draw biological conclusions through the identification of Gene Ontologies.

[6] G. Mazzocco, S. S. Bhowmick, **I. Saha**, U. Maulik, D. Bhattacharjee, and D. Plewczynski, "MaER: A New Multiclass Classifier for Binding Activity Prediction of HLA Class II Proteins", 6th International Conference on Pattern Recognition and Machine Intelligence (PReMI), Warsaw, Poland, pp. 462-471, June-July, 2015. DOI:10.1007/978-3-319-19941-2_44

Abstract:
Human Leukocyte Antigen class II (HLA II) proteins are crucial for the activation of adaptive immune response. In HLA class II molecules, high rate of polymorphisms has been observed. Hence, the accurate prediction of HLA peptide interactions II is a challenging task that can both improve the understanding of immunological processes and facilitate decision-making in vaccine design. In this regard, during the last decade various computational tools have been developed, which were mainly focused on the binding activity prediction of different HLA II isotypes (such as DP, DQ and DR) separately. This fact motivated us to make a humble contribution towards the prediction of isotypes binding propensity as a multiclass classification task. We analysed a binding affinity dataset, which contains the interactions of 27 HLA II proteins with 636 variable length peptides, in order to prepare new multiclass datasets for strong and weak binding peptides. Thereafter, a new multiclass classifier, called MetaEnsembleR (MaER) is proposed to predict the activity of weak/unknown binding peptides, by integrating the results of various heterogeneous classifiers. It pre-processes the training and testing datasets by making feature subsets, bootstrap samples and creates diverse datasets using principle component analysis, which are then used to train and test the MaER. The performance of MaER with respect to other existing state-of-the-art classifiers is estimated using validity measures, ROC curves and gain value analysis. Finally, a statistical test called Friedman test has been conducted to judge the statistical significance of the results produced by MaER.

Communicated

[1] **I. Saha**, J. P. Sarkar and U. Maulik, "Towards Improving Rough-Fuzzy Clustering for Categorical Data", Knowledge-Based Systems, 2014. (Current status: communicated) [Impact Factor: 3.06]

Abstract:
Clustering categorical data has drawn much attention in recent years with the advancement of data mining techniques. The fact that the categorical data do not have natural ordering in their attribute values, makes it more complex in terms of vagueness, uncertainty and overlapping. For this purpose, integration of rough and fuzzy set theories can be used for better handling of categorical data. In this regard, recently Rough Fuzzy K-Modes clustering technique has been developed. However, it may suffer from the problem of local optima as the choice of initial cluster modes are done randomly. Hence, in order to overcome this problem, here first we have used two different types of well-known metaheuristic methods like Simulated Annealing and Genetic Algorithm in order to perform the clustering better as an underlying optimization problem. The proposed methods are named as Simulated Annealing based Rough Fuzzy K-Modes and Genetic Algorithm based Rough Fuzzy K-Modes. These methods are able to produce crisp and rough points separately. Thereafter, for each case, in order to get the final clustering results, rough points are assigned to a particular crisp cluster using well-known machine learning technique, called Random Forest. For this purpose, crisp points are used as a training set of Random Forest to classify the rough points produced by individual clustering methods. Second, the varying cardinality of the training and test sets for each clustering method motivated us to propose a generalized technique called Integrated Rough Fuzzy Clustering using Random Forest. For this purpose, results of three aforementioned clustering techniques are used to compute the roughness measure. Based on this roughness measure, the set of \textit{best-crisp points} is selected from three sets of crisp points. Subsequently, sets of \textit{semi-crisp points} which are not belonging to the best-crisp points, but are the member of either one or two other sets of crisp points and \textit{pure-rough points} which are neither the member of best-crisp points nor the member of semi-crisp points, are determined. Thereafter, using multi-phase learning, best-crisp points are used to classify the semi-crisp points, and then using both of them, pure-rough points are classified by Random Forest. Experimental results are demonstrating the effectiveness of the proposed methods in comparison with well-known techniques for six synthetic and four real life datasets quantitatively and visually. Finally, statistical significance tests have been conducted to establish the superiority of the results produced by the proposed methods.

.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

[1] Participated in 9th IEEE International Conference on Industrial and Information Systems (ICIIS), Indian Institute of Information Technology, Gwalior, India. Duration: 15-17 December 2014

[2] Participated in ABCDE seminar of ERCIM and their 25th Anniversary meeting, CNR-IIT, Pisa, Italy. Duration: 23-24 October 2014.

[3] Participated in workshop on “Bioinformatiha 3”, National Research Council (CNR) - Institute of Informatics and Telematics (IIT), Pisa, Italy. Duration: 20-21 October 2014.

[4] Participated in annual workshop of ICM, University of Warsaw, Poland. Duration: 26-30 May 2014.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

[1] Visiting Research Scientist, Centrum Wiskunde and Informatica (CWI), Amsterdam, Netherlands, from 9th to 13th March 2015.

[2] Visiting Research Scientist, Institut National de Recherche en Informatique et en Automatique (INRIA), Lozere, Saclay-Paris, France, from 23rd to 27th February, 2015.

[3] Visiting Research Scientist, National Research Council (CNR) - Institute of Informatics and Telematics (IIT), Pisa, Italy, from 14th to 24th November, 2014.

[4] Visiting Research Scientist, University of Warsaw, Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), Warsaw, Poland, from 26th to 30th May 2014.