

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/57983> holds various files of this Leiden University dissertation.

Author: Cai, F.

Title: Fuzzy systems and unsupervised computing: exploration of applications in biology

Issue Date: 2017/12/12

Summary

In this thesis the application of the fuzzy systems in the domain of bioinformatics is investigated. In the past decades, bioinformatics has gained an increasing interest by both biologists and computer scientists. Biologists because they produce vast amounts of data that need be analyzed in an accurate and robust manner. Computer scientists because they embark on the challenge of creating added value to these large data volumes by developing and designing methodologies through which such accurate and robust analysis can be applied.

Data from life-sciences research in their native, raw form are known to be vague, ambiguous, imprecise, and sometimes points within the data are missing or unknown. In order to cope with the notion that data are not perfect, several new approaches need to be studied and implemented. The Fuzzy systems is a relatively new heuristic technique that has the capability of simplifying an otherwise complex decision by allowing more hypotheses in the analysis of the data. In other words, the logic in fuzzy systems acknowledges the fact that significance is the most important factor in modelling, while in other systems precision is acknowledged as such.

The Fuzzy systems is regarded to have potential for data analysis; therefore, the research described in this thesis intends to focus on designing efficient and reliable heuristic solutions for analysis to uncover the hidden information from biological experiments. Our research aims to build dedicated analysis pipelines based on the fuzzy systems; in the research chapters of this thesis, three different perspectives are elaborated. Finally, we integrate these three different uses of the fuzzy systems into an analysis pipeline illustrated with a case study.

(1) We have used the Fuzzy systems in background illumination correction. An image resulting from microscopy contains noise and other effects that do not contribute to the real signal that is needed to be measured. Apart from random shot noise caused by the electronics of the device, uneven illumination affects the analysis of the image. In order to diminish the impact of this illumination in the background, an appropriate correction must be performed.

In Chapter 2 of this thesis, a background correction method based on mathematical morphology, hybridized with fuzzy and rough constraints is proposed to eliminate shading effects. Compared with most generally used EMI method and the, in commercial software, often employed Rolling Ball algorithm, the DCBC that we developed has demonstrated a robust performance for typical images from biomedical microscopy.

(2) We have studied the fuzzy systems in relation with feature selection and redundancy removal in a data set. The procedure for feature selection plays an

important role in converting phenotypic data to a statistical representation in a matrix form. With respect to this problem, there are two aspects that should be considered: 1) selection of the smallest amount features to best describe dataset and from which predictions can be inferred; 2) reduction of the dataset dimensionality by removal of tedious information.

In Chapter 3 of this thesis we proposed a filter-wrapped feature selection approach in which fuzzy criteria are employed as one of the cost functions. Alongside, a multi-objective evolutionary algorithm is used to produce optimal solutions at the Pareto-front. This approach is different from other feature selection strategies in that the method introduced in our work provides a set of candidates feature combinations. From these combinations, the decision maker can benefit in choosing the most valuable ones for their cases.

(3) We have explored the use of fuzzy systems with information clustering analysis. Among pattern extracting methods, i.e. summarization, association and prediction, information clustering is of the great importance; it is popular in both in research and daily practice. This is especially the case for a dataset which has little or no labels.

In Chapter 4 of this thesis we have accomplished a novel clustering methodology that combines the fuzzy rough c-means approach and particle swarm optimization (PSO) algorithm. This combination integrates into sensible global results. The concept of fuzzy logic can cope with uncertainty, vagueness and overlapping partitions in the dataset, while the PSO algorithm helps with searching for near-optimum solutions.

(4) We have investigated fuzzy systems for a dedicated data analysis pipeline which is a concatenation of approaches presented in previous chapters. The applicability is demonstrated by a case study.

In Chapter 5 of this thesis we describe a case study which focuses on the overall analysis of data from one dimensional gel electrophoresis. The proposed fuzzy-system based data analysis pipeline has shown to be capable of precise extraction of features from gel images of proteins that are typical to cancer cells. Gel images themselves are far from ideal; therefore a number of corrections need to be applied so that accurate measurements can be extracted from these images. These measurements can be employed to distinguish and characterize the significance differences between cancer cell lines and their corresponding group and sub-group. This contributes to the further understanding of cancer development. Additionally, the resulting pipeline can contribute to a better understanding of protein/DNA migrations and expressions with respect to their characteristic features.

In conclusion, the research described in this thesis aims at investigating the use of fuzzy systems, in combination with pattern recognition in the field on bioinformatics. By studying the existing methodologies, an ensemble of fuzzy logic and unsupervised based algorithms are designed and integrated together as an analysis pipeline to understand and address biology-oriented questions and pattern-matching problems