Fuzzy Systems and Unsupervised Computing

Application of the Paradigms in Biology

Fuyu Cai

蔡甫雨

Fuzzy Systems and Unsupervised Computing

Exploration of Applications in Biology

Fuyu CAI

蔡甫雨

Fuzzy Systems and Unsupervised Computing

Exploration of Applications in Biology

Proefschrift

ter verkrijging van
de grad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnifucus Prof. Mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 12 december 2017
klokke 12:30 uur

door

Fuyu Cai

Geboren te Zhanjiang, China, in 1988

Promotiecommissie

Promotor: Prof. Dr. Ir. F.J. Verbeek

Overige leden: Prof. Dr. B. ter. Haar. Romeny (TU-Eindhoven)

Prof. Dr. V. van Noort (KU Leuven)

Prof. Dr. P. ten. Dijke

Dr. M. Emmerich

Colophon

The research contents described in this thesis were performed at the Leiden Institute of Advanced Computer Science (LIACS) in the Imaging & Bioinformatics group (IB), Leiden University, The Netherlands.

The research was financially supported by Chinese Scholar Council, China (CSC) and partially supported by Cancer Genomics Center, The Netherlands (CGC).

Printed by:

Published by:

ISBN:

Table of Contents

Chapter 1

Introduction on Computation in Biology

As humans we have a good understanding of subjective concepts. For a computer, or simply a computational approach, it is much harder to deal with these subjective concepts as the boundaries in a subjective range are not well defined. In data analysis for life-sciences, we often encounter such problems and are challenged to find a solution that can deal with ranges of measurements in a robust manner.

In this thesis we particularly focus on finding solutions for data-analysis in the life-sciences. The life-sciences cover a broad field of research and approaches to deal with data-analysis which are typically multi-disciplinary. First one has to understand the particular field in of the life-sciences that the data-analysis is applied to and then, often, a number of techniques from statistics, mathematics, physics and computer science are employed to develop a solution. This multi-disciplinary approach for computational problems in the life-sciences is often captured under the umbrella of bio-informatics. One can state that bio-informatics is concerned with the analysis of data. As a consequence it is important to realize that the development of computational tools bio-informatics is therefore an implicit characteristic of this field. The consequence of working with experimental data and results from analysis is that these data need to be organized. These areas pretty much cover the field of bioinformatics.

Data from experimental set-ups in biological research are not always ideal for a straightforward analysis. Experimental conditions and biological variation both contribute to ambiguity. For analysis, the volume of data is not always sufficient, while the distribution of the data is uneven. Moreover, the measurement device itself, due to its electronic components, adds noise to the raw data. All of these issues have to be taken into account for an analysis. In order to further explore solutions for data analysis and typical for data sets without well-defined boundaries between its constituents, we investigate how the use of fuzzy systems theory can be used to enhance computations for such data sets.

Therefore, in this thesis we will explore the use of fuzzy systems theory for applications in bioinformatics. The theory of fuzzy systems is concerned with formulating decision problems in data sets that are ill-defined. It supports the transfer from a subjective human classification to a numerical scale. In this manner it affords the testing of hypothesis and separation of the classes in the data.

The fuzzy systems theory is part of the paradigm of soft computing, a collection of mathematical techniques that supports computing in dealing with uncertainty, inaccuracy, vagueness and incompleteness in data sets.

In the research presented in this thesis, we first formulate problems in terms of a fuzzy systems and then develop and test algorithms in terms of their performance with data from the domain of the life-sciences. From the results and the performance,

we will learn about the usefulness of fuzzy systems for the field, as well as the applicability to the kind of problems and practicality for the computation itself.

As mentioned, computing in bioinformatics is quite interdisciplinary; therefore we will use this introduction to present some of the major concepts from bioinformatics that are important to this thesis as well as provide scope of that field. Next, we will address soft-computing and in particular fuzzy systems and how this links with the analysis of data from the domain of biology. Further to this background information, we will explain our development of a heuristic-based pipe-line for data analysis, applied to generic data analysis as well as to bio-imaging.

1.1. The Scope of Bioinformatics: A Brief Introduction

The research field of Bioinformatics has matured over the past ten years and to date there is consensus on a definition. In general, bioinformatics is considered as the application of computational techniques of analyzing, managing and interpreting biological information [1]. The rationale is to create added value from the data for the field of biology [14]. The research field of bioinformatics encompasses a wide range of subjects, typically referred to as "omics" data, i.e. structural biology, genomics, proteomics, metabolomics, transcriptomics, cytomics, and image-based high-throughput studies.

In trying to understand biology, computational approaches have been probed. These were, in some cases essential for the understanding of phenomena, the discovery of inheritance by Mendel [2] in 1865 stands as a paradigm for computation in biology.

Modern approaches to computation in biology go hand in hand with the development and availability of computers. The notion of bioinformatics is developed in the late 1960's when molecular biologists started the compile their sequencing results of DNA and proteins in databases [3]. Initially, the field of bioinformatics was claimed by the research on the human genome but this progressed into the perception that bioinformatics had a much broader base.

The term bioinformatics is attributed to Hogeweg and Hesper [4], who coined to term as: "the study of information processing in biotic systems". Over the past five decades, however, the field of bioinformatics has evolved in that it now involves various tasks, focusing on the analysis and understanding [14] biological data. Understanding refers to the creation of added value to the data.

Within the scope of bioinformatics different questions on biology are addressed. A common ground of all questions is that the starting point is a large amount of data from which understanding is developed in finding patterns in these data. This means that from the data, a systematic analysis is performed – these studies can be on

various levels; i.e. cellular behavior, molecular design and docking, metabolic networks, RNA/DNA and protein sequence alignment, RNA, DNA and protein structure prediction, analysis of gene expression data, just to name a few important fields.

Taking the data in a bioinformatics study as a starting point, the major emphases of bioinformatics become evident. Having large amounts of data requires these data to be organized in a structured form. In the early days of bioinformatics the important activity was constructing databases for the different data and making these databases available via the web. To date, this data management is still a major activity in the field of bioinformatics. Thus, developing databases and tools to archive and retrieve data is an important activity. The data are, of course, analyzed. A next major activity therefore is analyzing data and developing tools to achieve the analysis. From the analysis a higher level aggregation can be accomplished, combining results from analysis and finding patterns which contribute to the further understanding of biology. This activity, concerns working with statistics and machine learning approaches; it builds upon the other activities, however, its focus is to create added value from the large amounts of data in a manner that is meaningful to biology.

An important part of the field of bioinformatics is therefore the development of computational tools. Here there is common ground with computational biology where the emphasis is on theoretical models and simulations [14]. Nevertheless, computation and tooling is important to both fields that join forces in the quest of understanding the complexity in biology.

The crux for bioinformatics is to have adequate tools for analysis and interpretation available. There are ample computational approaches that have been successfully probed in bioinformatics studies and that, to data, are part of the algorithmic repertoire in bioinformatics. With different datatypes, different computational paradigms have been used. For analysis of sequences, i.e. RNA, DNA or protein, different alphabets are used in string matching procedures. The concept of dynamics programming has been very instrumental in being able to match strings in terms of their similarity. The Basic Local Alignment Search Tool (BLAST), to that respect is a major milestone for bioinformatics as a whole. Finding patterns from data is resolved using machine learning techniques; to a certain extent these techniques are inspired by biology, i.e. neural networks, genetic algorithms. In itself, machine learning techniques, for clustering and classification are deeply rooted in mathematics. Employing these techniques requires therefore, some understanding of the mathematics, e.g. choosing a fitting function in a classification problem.

Classification and clustering allows establishing relations in the data and to reason on behavior of biological entities. Techniques like Bayesian Clustering, Support

Vector Machines (SVM), neural networks and genetic algorithms represent forms of machine learning with and without control. It supports finding groups in the data as well as making predictions from new data based on prior classifications. Nowadays, a new revival of the concept of neural networks is embodied into so called convolutional neural nets, also known as deep learning. This is a very powerful machine learning approach that will further boost the understanding of biology from large amounts of data.

An important concept in computational approaches is that of heuristics, i.e. specific rules, so that a problem can be confined and computational pipelines for a specific bioinformatics study can be designed and implemented.

Our efforts for data analysis extend on existing approaches used in the field of Bioinformatics. Data from biological experiments do contain noise; and this noise complicates the analysis. Techniques that are based on heuristics are capable of confining the computation to solutions that are more probable. Heuristic-based techniques are therefore sometimes preferred in doing computations in large data volumes. Examples of such techniques are Bayesian nets, Neural networks, Fuzzy logic and evolutionary algorithms. In our effort to extend and improve data analysis in biology, we will focus on the so called fuzzy systems to see if we can reinforce solutions for datasets that are otherwise difficult to separate. Fuzzy systems are part of the soft-computing paradigm. Further explanation of this concept will be given in the next section.

1.2. Soft Computing and Fuzzy Systems

Soft computing [5], sometimes referred to as computational intelligence (CI), is a collection of methodologies that has become an area of formal study in computer science in the early 1990s. Soft computing differs from conventional computing in that it, specifically, exploits imprecision and tolerance in order to achieve tractability, robustness and low-cost solutions. The soft computing paradigm represents a number of techniques, its major constituents are the Fuzzy Systems (FS), the Rough Set (RS), Wavelets, Simulated Annealing (SA), the Support Vector Machine (SVM), the Artificial Neural Network (ANN), Evolutionary Algorithms (EAs) and Swarm Intelligence (SI).

The quality of the Fuzzy Systems is that it is rather easy to implement in a system ranging from very small and simple to embedding in large networked systems [17] [18]. As part of an analysis, the fuzzy systems will be part of a specific pipeline for data analysis. This is exactly how we intend to employ the application of the fuzzy systems in analysis of biological problems – and thus the data.

The concepts of **Fuzzy systems** have been applied to a range of different fields, from control theory to artificial intelligence, as well as to computational biology. The fuzzy systems are derived from fuzzy logic and it was first introduced by Lotfi Zadeh [6] in a monograph on fuzzy set theory in 1965.

In traditional computing, a system yields output(s) from the input(s), where conclusion is accepted to be either true or false. However, in real life situations, propositions are given with variable answers; for instance, degree of color between "yellow" and "red", concept of "empty" and "full" in a water-filled bottle, sensation temperature of "cold", "warm" and "hot" in a room, etc. In other words, we consider it natural to reason over a range of subjective concepts. If a certain concept cannot be defined exactly, an amount of quick and ambiguous definitions would develop. This typically happens in a group of people discussing a concept in a certain context. In Figure 1-1, this is exemplified, the fuzzy systems is the control methodology that mimics how a decision (description) is made by humans. Additionally, this decision-making process can be achieved and speeded up on the basis of prior experiences/recognitions of the individual.

Example:

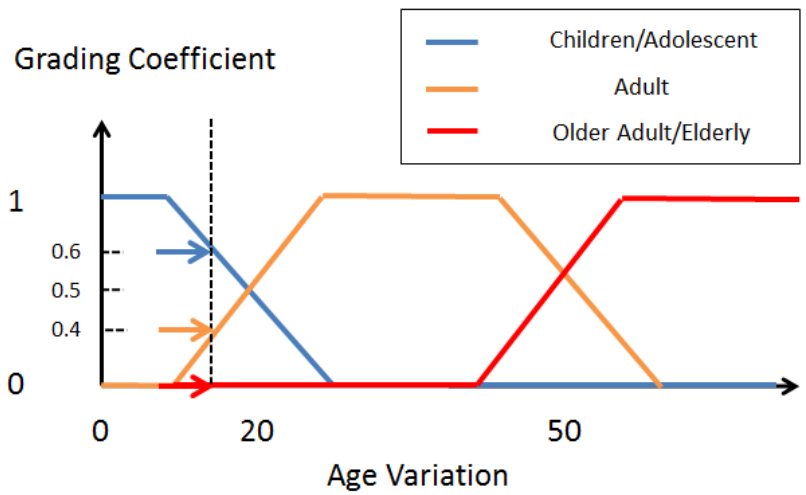


Figure 1-1. The application of Fuzzy Systems in the grading of human age. The grades of age are expressed as "Children/Adolescent", "Adult" and "Older Adult/Elderly"; these are mapped by a pre-defined grading function onto an age scale. The age itself is ambiguous for human perspective inspection and/or sensible feeling. However, this sensation becomes obvious once a human-like metric is defined for the decision, e.g. when the grading coefficient is larger than 0.5, meaning a positive effect, vice versa. Since the red arrow points to zero, people within this age group (grading coefficient = 0, defined on Older Adult/Elder function) can be interpreted as "not old"; while the meanings of the age at orange (coefficient < 0.5, defined on Adult function) and blue (coefficient > 0.5, defined on Children/Adolescent function) arrow can be recognized as "fairly matured" but "still young", simultaneously. Instead of concluding in either young (coefficient = 0) or old (coefficient = 1), fuzzy systems allows for decisions being made by users' knowledge and experience accordingly.

The example of Figure 1-1 demonstrates two key ideas of the fuzzy system: first, the fuzzy system is able to model problems from concepts to mathematical paradigms that are based on the understanding and experience of the decision maker; second, the logic in fuzzy systems accepts the uncertainties that are inherited as realistic inputs, and thereby it is able to cope with these uncertainties (imprecisions) in such a way that their effects are negligible and henceforth, the system will result in a "precise", human-like, output.

As mentioned before, fuzzy systems have been successfully applied to several areas, and also in bioinformatics. It helps in recognizing the hidden essentials in data by a degree of "truth" given by the fuzzy membership [19]. The fuzzy membership is a function that describes the weights for the contribution of the different levels in the system.

In this manner, using the fuzzy membership function, biological information is analyzed and interpreted based upon previous experiences [20] so that knowledge-based systems in biosciences are constructed by vagueness and uncertainty [21]. In this thesis, therefore, fuzzy systems based approaches are proposed and integrated into a dedicated data analysis pipeline(s).

Images form a particular class of data in life-sciences research. The data, i.e. the images, result from an imaging device and are sampled to a digital image. This digital image is input for a first data analysis in order to get measurements out of the image. The measurements themselves are input for a second analysis to find patterns over a collection of images. In general, this collection of images comprises an experiment. The data in the digital image are intensity values and these are ordered in a regular rectangular grid, directly related to the sensor in the digital camera. The data analysis is therefore completely adapted to this organization. Once we obtained the measurements, other approaches need to be probed. In this thesis, the image space based approaches, as well as for the feature space based approaches are

addressed; while the fuzzy systems approach has the capability to augment both the analysis. In order to get an idea of the scope, we briefly discuss data-analysis in bio-imaging in the next section.

1.3. Data Analysis in Bio-Imaging

The use of imaging techniques in biology is currently undergoing a revolution [7] with the availability of all new kinds of biological imaging techniques; i.e. fluorescence lifetime imaging, molecular imaging, diffusion-based imaging, X-ray based imaging, ultrasound-based imaging, magnetic resonance (MR) based imaging, etc. The in vitro and in vivo visualization of organisms, tissues, cells, proteins and macromolecular structures is enabled.

In the laboratory, bright-field and fluorescence imaging are routinely used; while the data from experiments is obtained by bio-imaging components. Analysis of data from these experiments can be performed rapidly by data-science oriented scientists, i.e. computer scientists and bioinformaticists. The data allows for data scientists to test and validate hypothesis related to a range of phenomena, e.g. cellular and molecular behavior, and the data can acquired in different dimensions; i.e. 2D, 3D and time-lapse. Imaging techniques are part of the standard repertoire of a large range of experimentation with a visual control. However, the observations in the data, i.e. images, need to be verified and to that end computational tools are needed. Currently, large volumes of data are more and more the norm, therefore automation of the analysis is absolutely necessary. Such automation will result in data that support the interpretation of the experiment and are an indispensable extension to the imaging system (cf. Figure 1-2). Designing robust and accurate methods are being investigated thoroughly in biosciences and bioinformatics (cf. Section 1.1, [8]). With advances in technology leading to high-throughput systems for imaging, methodology design becomes increasingly important.

In this thesis, we consider data analysis as an extension of the imaging system aiming for understanding and recognizing information from biological image data. We distinguish a sequence of three major steps: (1) **image acquisition**, (2) **image processing**, and (3) **data analysis**. The acquisition is embedded in the imaging device which makes the image data available in digital form on a data repository. The image processing accomplishes a transformation of the raw image data to images from which reliable information can be extracted. In the data analysis step, observations are transformed to numbers and statistical representations so that analysis and validation can be applied. The final step in the analysis is to infer an interpretation from the measurements.

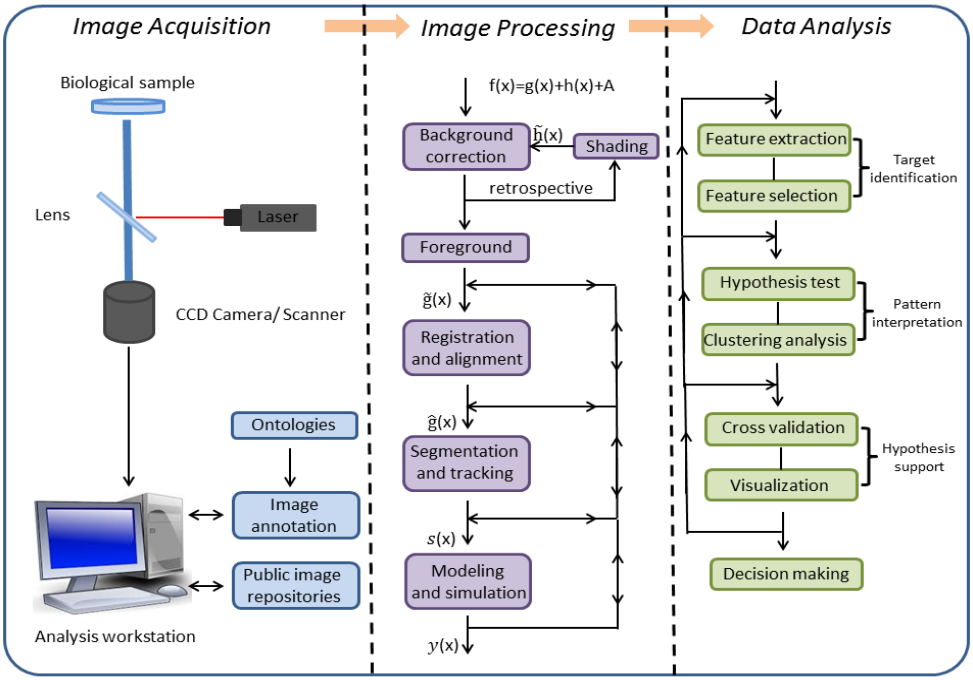


Figure 1-2. Workflow of data analysis in bright-field and fluorescence imaging system

Image acquisition. In bright-field and fluorescence imaging systems, image acquisition procedure is accomplished with a sensor array, often a CCD chip, which has a rectangular layout (CCD camera) or a line layout (flatbed scanner). A CCD camera is mounted on the optical system [9] whereas a line scanner is used to scan larger surfaces such as gels. In this thesis, we use bright-field microscopy images, i.e. cultured cartilage cells (cf. Chapter 2), fluorescence microscopy images, i.e. cardiomyocytes fluorescent images (cf. Chapter 2) and scanner images, gels of protein compositions from a range of cell lines (cf. Chapter 5).

For the acquisition of images, one should wish for the highest possible quality. However, there is a trade-off between image resolution and acquisition speed. A high-resolution detector allows imaging of objects whereas at lower resolution a significantly higher acquisition speed can be accomplished which is necessary to capture dynamics events. To further accelerate the dynamic acquisition numerous amount of efforts [10] [11] [12] have been made to attempt to acquire high-resolution images at high speed.

Acquired images will be stored in a repository. The increase of the data volume and complexity of biological experiments has made manual-workbook or generic databases unsuitable for keeping track of the images/data produced in experiments.

Therefore, image annotation, i.e. associating images with metadata, such as size, acquisition date, contents, is absolutely necessary for data provenance [22]. These metadata are often required for the further analysis to be able to understand an observation in context; or accomplish that an automated system can "understand" the context.

Image processing. In imaging images are the carrier of information and the content, at some point, should be transformed to a quantitative data presentation. Images as obtained from imaging systems are, however, far from "clean" (noisy) and need to be "polished". The practical approach is to apply restoration through image processing before the data are being analyzed [15]. The term "image processing", means to apply basic level operations on images. This is regarded to a pre-processing step such as enhancement, alignment and segmentation.

A well-performed image processing strategy, to some extent, can minimize data variation. It is, therefore, important to utilize empirical and problem-driven image processing solutions. Referring to Figure 1-2, a specimen is imaged and modeled by the input image $f(x)=g(x)+h(x)+A$, where true information $g(x)$ is masked by background noise $h(x)$ and all absolute multiplicative noise A. This raw image is then restored using an approach that employs additional images obtained at the time of image capturing, or through retrospective shading correction. For the resulting image $\tilde{g}(x)$, also known as the foreground image, now various other processing options are at hand: 1) a registration/alignment process producing $\hat{g}(x)$, 2) segmentation/tracking process, resulting in an image $s(x)$; or 3) an image modeling/simulating block with the output $y(x)$. After these image processing steps, information carried within experimental raw images is now enhanced and can be further explored.

Data analysis. From image processing we have obtained a restored image. Next, we extract features from the image. This is essentially a data reduction; we reduce the image elements to a set of measurements that sustain our observation. The analysis is in the heart of the bioinformatics methodology as it presents contextual approaches for data analysis, representation, and visualization. Image analysis deals with quantification of the amount and localization of signal, and measuring changes in structure over time. Data analysis can help to ensure that resulting measurements are accurate, objective and reproducible. Moreover, data analysis supports the further interpretation of the data by finding patterns in the data of an experiment or relating results to other experiments. In the context of biomedical research, this is typically the domain of bioinformatics. Commonly employed approaches involve target feature extraction and selection, data hypothesis test and data clustering, performance validation and visualization, as well as decision making. The decomposition and comparison of temporal biological data is not yet fully

understood [13]. In general, the data analysis in bioinformatics further augments the impact of knowledge discovery and making good predictions from the measurements.

1.4. Research Scope and Thesis Structure

In this thesis the capability of fuzzy systems is investigated with respect to systems in which simplifying an otherwise complex decision is augmented by allowing more hypotheses on the data. In the introduction (cf. Section1.3), we have provided a dedicated bioinformatics data analysis pipeline, from which subjective and tedious image interpretations are alleviated. This thesis will further address the fuzzy systems in a number of different approaches on data measurement and develop an understanding on how the fuzzy systems can be employed in the concepts with pattern recognition.

The research described in this thesis is structured into 6 chapters. We have provided an introduction in **Chapter 1**. **Chapter 2, "Biological Image Background Correction**", presents a strategy employing a combination of fuzzy logic and rough set theory to constrain a morphological image processing path during the process of image background correction.

Chapter 3, "Feature Selection Strategy in Region of Interest Mask", illustrates a schema of feature selection via a fuzzy criterion in a multi-objective optimization algorithm. From this approach, sets of candidate solutions are provided to the researchers so that they can make decisions based on their own experiences/requirements;

Chapter 4, "Unsupervised Information Classification and Analysis", elaborates on an unsupervised classification technique that hybridizes fuzzy uncertainty-based clustering method with swarm intelligence in order to find a good solution.
In these three chapters (cf. 2,3,4), the performance and efficiency of these algorithms are comprehensively assessed using various benchmark datasets that cover multiple facets of real-life situations. The results are compared with several commonly applied approaches; most are considered state-of-the-art methodologies. The evaluation and validation of these algorithms are used as a theoretical foundation for the design of image analysis workflows for experimental data.

Chapter 5, "A Systematic Study on One Dimensional Gel Electrophoresis Image Analysis", employs the methodologies proposed in Chapter 2, Chapter 3 and Chapter 4 to interpret more complex biological problems in a multi-faceted manner. This case study presents a fuzzy-system based data analysis aiming to investigate and understand the identity of characteristics of proteins that are distinct/ shared

between different subgroups of cancer cell-lines. It further demonstrates the practical implication of image and data analysis workflows following the fuzzy-system designs, as promising.

Finally, in **Chapter 6** conclusions are presented and from a discussion an outlook to the further application of the fuzzy systems is further described.

1.5. Reference

[1] Baldi, P., & Brunak, S. (2001). Bioinformatics: the machine learning approach. MIT press.

[2] Ford, E. B. (1931). Mendelism and evolution. *Mendelism and evolution*.

[3] Dayhoff, M.O. (1966) Atlas of protein sequence and structure. National Biomedical Research Foundation, 215 pp.

[4] Hesper B, Hogeweg P (1970). "Bioinformatica: een werkconcept". 1 (6). Kameleon: 28–29.

[5] Zadeh, L. A. (1994). Fuzzy logic, neural networks, and soft computing. Communications of the ACM, 37(3), 77-85.

[6] Zadeh, L. A. (1965). Fuzzy sets. Information and control, 8(3), 338-353.

[7] Rittscher, J., Machiraju, R., & Wong, S. T. (2008). Microscopic image analysis for life science applications. Artech House.

[8] Eils, R., & Athale, C. (2003). Computational imaging in cell biology. The Journal of cell biology, 161(3), 477-481.

[9] Rochow, E. (2012). An introduction to microscopy by means of light, electrons, x-rays, or ultrasound. Springer Science & Business Media.

[10] Picco, L. M., Bozec, L., Ulcinas, A., Engledew, D. J., Antognozzi, M., Horton, M. A., & Miles, M. J. (2006). Breaking the speed limit with atomic force microscopy. Nanotechnology, 18(4), 044030.

[11] Oheim, M. (2007). High‐throughput microscopy must re‐invent the microscope rather than speed up its functions. British journal of pharmacology, 152(1), 1-4.

[12] Bouchard, M. B., Voleti, V., Mendes, C. S., Lacefield, C., Grueber, W. B., Mann, R. S., ... & Hillman, E. M. (2015). Swept confocally-aligned planar excitation (SCAPE) microscopy for high-speed volumetric imaging of behaving organisms. Nature photonics, 9(2), 113-119.

[13] Yan, K. (2013). Image analysis and platform development for automated phenotyping in cytomics (Doctoral dissertation, Department of Imaging and Bioinformatics, Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University).

[14] F. J. Verbeek. (2016) Lecture Notes, “Bio Modelling and Petri Net”.

[15] F. J. Verbeek. (2017) Lecture Notes, “Image Analysis”.

[16] Jena, R. K., Aqel, M. M., Srivastava, P., & Mahanti, P. K. (2009). Soft computing methodologies in bioinformatics. European Journal of Scientific Research, 26(2), 189-203.

[17] Liao, S. H. (2005). Expert system methodologies and applications—a decade review from 1995 to 2004. Expert systems with applications, 28(1), 93-103.

[18] Castillo, O., & Melin, P. (2014). A review on interval type-2 fuzzy logic applications in intelligent control. Information Sciences, 279, 615-631.

[19] Cai F. & Verbeek F.J. (2015), Dam-based Rolling Ball with Fuzzy-Rough Constraints, a New Background Subtraction Algorithm for Image Analysis in Microscopy.. In: Proceedings International Conference on Image Processing Theory, Tools and Applications (IPTA 2015). 298-303.

[20] Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification. PloS one, 10(3), e0120364.

[21] Jiang, Y., Chung, F. L., Ishibuchi, H., Deng, Z., & Wang, S. (2015). Multitask TSK fuzzy system modeling by mining intertask common hidden structure. IEEE transactions on cybernetics, 45(3), 534-547.

[22] Vargas, E. L., Xia, Z., Slob, J., & Verbeek, F. J. (2016). A Semantic-Based Metadata Validation for an Automated High-Throughput Screening Workflow: Case Study in CytomicsDB. In International Conference on Bioinformatics and Biomedical Engineering (pp. 557-572). Springer, Cham.

Chapter 2

Biological Image Background Correction

This chapter is based on the following publication:

F. Cai, and F.J. Verbeek. "Dam-based rolling ball with fuzzy-rough constraints, a new background subtraction algorithm for image analysis in microscopy." Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on. IEEE, 2015.

F. Cai, and F.J. Verbeek. "Dam-based rolling ball with fuzzy-rough constraints, a new background subtraction algorithm for image analysis in microscopy." Submitted.

Chapter summary

Light microscopy is one of the important techniques in bio-imaging data analysis. The measurements obtained from images provide information on the concentration of molecules in cells, tissues, as well as whole mount. Microscopy images as they are acquired are not ideal; images, specifically shapes and textures often suffer from an uneven background due to flaws in the illumination. The presence of inhomogeneous noise in images, which mostly attribute to factors related to the light path between camera and microscope, can significantly impact the accuracy of downstream measurements.

In this chapter, we seek to contribute to quantitative improvement on the quality of light microscope readouts based on the proposed Dam-Constrained Background Correction (DCBC) method. The strategy we present employs a combination of fuzzy logic and rough set theory to constrain a morphological path at the moment background correction process takes place. To illustrate the competence of this method, a state-of-the-art shading correction based on entropy minimization (EMI) and the frequently used morphological rolling ball method (RBA) are compared via applications on three typical datasets. The reported results and extensive numerical analysis indicate an applausive performance on the proposed method.

2.1. Introduction

The study of biological phenomena using light or fluorescent imaging has transformed optoelectronic signal from a qualitative localization test to quantitative tools for functional analysis. These give rise to, for instance, features such as distance, area, velocity and intensity. A digital image of a specimen is created by the detector in an optical system. Inevitably, errors occur in the process of image acquisition via the specimen, the microscope or the detector itself [1]. An important source of error is the inhomogeneity of the intensity in the background. Sometimes the fluorescence has a shorter exposure time which will result in capturing less emitted photons compared to a procedure of longer exposure time [2]. Collection of fewer photons means that the relative contribution of the noise increases. Additionally, background intensity will accumulate through the surrounding fluorescence/ light sources, which will confound experiment's goals. This result leads to an undesirable contribution of the background with respect to the signal of interests.

In order to apply quantitative measurements in microscopy, a notion of the background must be known and, if possible, it must be removed before measurements in the images. Henceforth, it is always desirable to correct the inhomogeneous background beforehand. Several techniques have become available to mitigate abovementioned problems, also known as intensity inhomogeneity, uneven background and shading (vignetting). The common approaches serve to reduce the amount of inhomogeneity in microscope images, are noticed as Background Correction Processing (BCP).

The work in this chapter introduces a novel retrospective approach, on the basis of mathematical morphology to achieve better performance in a more general way. The proposed method is unsupervised and mostly parameter free. It employs the concepts of fuzzy membership, and approximation of an assumption from the rough set theory which is used to constrain an objective function. Subsequently, inspired and improved by conventional RBA [5], a morphological "dam" is constructed to avoid introducing a topological distortion (artefact), and to eliminate the noisy background while producing an enhanced foreground.

The remainder of this chapter is organized as follows: Section 2.2 introduces research related to background correction methods. In Section 2.3, the notion of contextual knowledge is reviewed. In Section 2.4, description of the objective function and dam-building methodology is presented. Section 2.5 provides the experimental results of the evaluation, followed by discussion and conclusions in Section 2.6.

2.2. Related Work

Existing methods in microscopy for illumination-based background correction can be applied while acquiring images (priori) or after acquisition (posteriori). The main difference between these approaches is a priori correction employs additional images obtained at the time of image capturing; while in posteriori correction, the controlled images are not available and therefore an ideal (hypothetical) background model has to be assumed.

The methods of acquiring prior-knowledge usually employ the background (illumination) images. An additional image is often made by defocusing or removing the specimen from the field of view [3], just capturing the specimen in bright field or dark field mode [15]. Consecutive adjusting the settings of camera and microscope [16] is also of importance to help with resulting images. Yielding results by linear image calculator (transmittance as the ration of transmitted light through specimen), however, these methods cannot cope with objective shading, e.g. shading caused by variation in specimen thickness at transmission imaging or by a non-planar surface in reflecting imaging. More specifically, images acquired from standard or automated microscopes, even with white (dark) referencing, are generally adequate for visual inspection but not completely for quantitative image analysis [17]. Practically, when conditions are not carefully controlled, differences can be more substantial and introduced to downstream evaluations.

Various approaches, namely retrospective (posteriori) correction, have been reported to extract the characteristics of background from a single image that depend on nothing but the actual images acquired during experiments. These methods mostly manipulate the data in both time-domain and frequency-domain using different sorts of filters, e.g. low (high)-pass-filter, linear-filter, compensated Gaussian blurring, etc. [18]. The drawback of these methods, however, is the limitation of the object size and the comparative background scale. The background is assumed to be either darker or brighter than the foreground. Moreover, the overlap of objects with the background is kind of forbidden in restoration, otherwise the mixture of foreground signal will be eliminated while applying the corrections in the frequency domain. Meanwhile, a mathematical morphology structuring element based on the image landscape has been introduced [4] [21] [19], i.e. the rolling ball algorithm. With a pre-defined radius, a virtual ball rolls over the ground of the topographical pixel-landscape. Each pixel that contacts with the surface of the ball will be selected for further processing. These methods, however, have limitations in that they are imprecise in the estimation of a solution and portray uncertainty in the control of the path in the application of microscopy image sets. Another technique [2] [20] which assumes that the image background is more homogenous relative to foreground, and estimate a correction function over the background regions from original images.

Ideally, it is better to correct the images with a priori method as all retrospective approaches make assumptions over the image characteristics that are unlikely to be strictly satisfied in any arbitrary image. However, one should be aware that the reproducibility of acquiring sample images for priori correction requires laborious manipulation, while errors can be even introduced. In this manner, a way more efficient and feasible retrospective DCBC method is proposed in this chapter.

2.3. Primary Notation

A. Fuzzy theory bound

The notion of fuzzy set was firstly introduced in 1965, and pioneered by Zadeh [7], fuzzy logic-based system have been successfully utilized into various application areas. A fuzzy set is the class of objects that contains consecutive grades of membership, with which value ranged from zero to one. This index assigned a "fuzziness" characteristic to the set, meaning a level of belonging. Particularly, a conventional set, referred to as the crisp set (commonly used in k- and/or c-means) will have either a value of zero or one; i.e. a Boolean value. The fuzzy membership function can be written as:

$$\mu_A: U \rightarrow [0,1] \qquad \textbf{Equation 2-1}$$

Where *A* denotes the fuzzy set, and the mapping function μ_A, is the membership function of *A*, while *U* is the universe.

B. Rough theory bound

Proposed by Pawlak [8], rough set theory is an approach to assess imprecision and uncertainty. Objects in the universe characterized by the same information, or knowledge are indiscernible (similar) in the view of available information about them. The concept of the indiscernibility relation is the mathematical basis of rough set theory.

An information system is an aggregation $S= \langle U, A, V, f \rangle$, where *U* is a non-empty finite set of *N* objects $\{x_1, x_2, \dots, x_N\}$ called the universe, and *A* is also a non-empty set of attributes. *V* is a value set such that $a: U \rightarrow V_a$ for every $a \in V$. With every subset of attributes *B* from *A*, we have $B \subseteq A$. We define an equivalence relation on *U* as:

$$I(B) = \{(x, y) \in U \times U: f_a(x) = f_a(y), \forall\, a \in B\} \qquad \textbf{Equation 2-2}$$

Elements belonging to *U* that can satisfy this equation (relation) *I(B)* are objects with the same value for attributes *B* and therefore, these objects are indiscernible with

respect to *B*. Moreover, an equivalence class containing the element *x* will be defined as *I(B)(x)*, in short *B(x)*. The classes of the equivalence sets are the basic concept of *B*. Given any subset of attributes *B*, with concept of $X \in U$ can be approximately defined by employing two exact sets respectively referred to as the lower and the upper approximation sets:

$$BD_*(X) = \{x \in U : B(x) \subseteq X\}$$
$$BD^*(X) = \{x \in U : B(x) \cap X \neq \emptyset\} \qquad \textbf{Equation 2-3}$$

Assigning to every subset *X* of universe *U*, a subset $BD_*(X)$ is referred to as the B-lower approximation of *X*, which can be classified as elements of *X* in the concept of *B*. While $BD^*(X)$ is the upper approximation which elements most probably belong to *X* given the knowledge *B*. The exactness based on the approximation set can be expressed by:

$$\alpha_B(X) = \frac{|BD_*(X)|}{|BD^*(X)|}, \qquad \text{for } X \neq \emptyset \qquad \textbf{Equation 2-4}$$

This equation is referred to as the accuracy of the approximation, where |.| denotes the cardinality of the sets. The accuracy measure captures the degree of completeness of the knowledge about the set *X*. According to the extended report in [9], we obtain a measurement of the roughness index by rewriting the Equation 2-4 as:

$$\rho_r = 1 - \alpha_B(X) \qquad \textbf{Equation 2-5}$$

From this normalized definition it holds that for every *B* and $X \subseteq U$, if $\rho_r = 0$, then the boundary region set *X* is empty. From this moment on, *X* is notated as *B*-definable, e.g. *X* is a crisp set with respect to the knowledge *B*. Otherwise, if $\rho_r > 0$, then this means *X* is *B*-undefinable, e.g. *X* is rough or uncertain with respect to the knowledge *B*.

2.4. Dam-Constraint Background Correction Strategy

A. Classical rolling ball concept

The quantitative measurement of (pixel-) intensity is a mixture of signal and background noise. It can be well estimated by measuring the local background pixels in the region of interest [10]. This procedure can be written as:

$$F_{obj} = \sum_{m=1}^{m-N_{obj}} F_{obj} - N_{obj} \frac{\sum_{n=1}^{n=N_{bkg}} F_{bkg}}{N_{bkg}} \qquad \textbf{Equation 2-6}$$

Where *F* is the fluorescent signal measured at each pixel (*m*, *n*), *obj* is the object, *bkg* is the selected background area or volume, and *N* is the number of pixels in the

selected object or background. This equation describes the framework of the RBA approach by computing the contribution of the background noise per pixel. By taking both advantages, we aggregate the classical RBA algorithm with the concept of fuzzy logic and rough theory to the image domain.

B. Selection of crest and toe for dam

Tedious signals are introduced often attribute to over-processing. The main goal of constructing a "dam" in DCBC method is to constrain the path of the morphological ball rolling into either the valley or summit image landscape, which does not meet our anticipation (selection of proper background). In this manner, a dam with crest and toe are established. This can be formulated into a bimodal threshold modelling on the basis of the composition of objects in the microscope images (cf. Figure 2-1 for instance, which objects foreground are assumed to consist of foreground and sub-foreground). Under the constraint of a dam, original information of a local region of interest will be preserved as much as possible, while over-segmentation will be limited (cf. Figure 2-2).

For a local region of interest, we have to obtain the multi-threshold values in order to construct a dam with a certain crest and toe in the image-landscape. Let x_{mn} be the pixel value with respect to the region size $m \times n$, and will obtain two average grey levels.

$$t_0 = \frac{1}{i}\sum_m \sum_n x_{mn}\ , \quad x_{mn} < t$$

$$t_1 = \frac{1}{j}\sum_m \sum_n x_{mn}\ , \quad x_{mn} \geq t \qquad \textbf{Equation 2-7}$$

where, i and j denote the number of occurrences of x_{mn} according to the threshold intensity-level t, and $i + j = m \times n$. Given by an initial threshold value t, the two average intensity-levels, t_0 and t_1, can be considered as a local background. In this manner two sets X_{t_0}, X_{t_1} are obtained with element x. The relationships of the pixels are $x \in X$, while their corresponding region should be directly depended on the change of the pixel values and the change in the local background. With respect to the membership function, taking the condition of these properties into account, we observed that the smaller the difference of the element pixel and its corresponding local background value, the larger the output of the membership will be. Notice that, it is expected one element should either belongs to set X_{t_0} or X_{t_1}. Consequently, this will result in a membership output value in an interval of 0.5 to 1. It is clear that the membership value equals one only if the element belongs to a crisp set, while it should also be monotonous within the domain of definition. With these notions, we thus obtain a piecewise function $g(x)$ and its convex formulation μ_X:

$$g(x) = \begin{cases} \frac{x_{mn}-t_0}{I_{max}-I_{min}}, & x_{mn} < t \\ \frac{x_{mn}-t_1}{I_{max}-I_{min}}, & x_{mn} \geq t \end{cases}$$ **Equation 2-8**

$$\mu_X(x_{mn}) = \frac{1}{2}[(g(x)-1)^2 + 1]$$ **Equation 2-9**

The I_{max} and I_{min} represent the maximum and minimum image intensity value. For a given value *t*, the membership value μ_X is in the interval [0.5, 1]. Moreover, the membership function μ_X has a convex shape in a line plot; this introduces a slight change of slope near 0.5 with much vagueness near the fuzzy centre (value 0.5), and a dramatic change of slope near 1 with a fast rate of convergence near the crisp point (value 1).

By definition, the fuzzy set is known as an approach to access the quantitative analysis of membership between the elements and sets. To this regard, we can obtain a performance measure of the fuzzy set segmentation result. This provides a better image landscape than the uniformity evaluation method. This is because uniformity is an index of indicating a degree of variance in a segmented region and the mean values belonging to this region. However, a shape evaluation is summary of a generalized gradient value for every pixel by checking the relationship between the determined threshold value and the grey values of its neighbouring pixels. Consequently, the more appropriate the threshold is chosen, the better the representation of resulting image landscape will be accomplished.

We assume that for the proposed bimodal thresholding method, the results of the first segmentation will separate the background and foreground, thereby containing the objects of interest. Then in a second segmentation of foreground, boundary domains (sub-foreground) will be isolated from objects. This procedure, for a better understanding, could take place when referring to cell cytoplasm and cell nucleus in cell microscope images. For our applications, the most likely shape we therefore will obtain is one in which all foreground pixels should contain cytoplasm and nucleus (cf. Figure 2-1). Successful partition of the nucleus from the maximum shape is required to define the boundary between the cytoplasm and the nucleus. Henceforth an upper-approximation set can be made as the pixels belonging to cytoplasm, whereas the nucleus is the lower-approximation set.

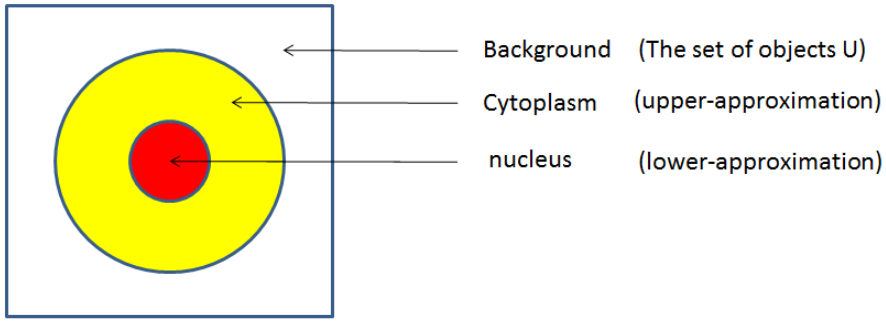


Figure 2-1. Illustration of rough-set reflection in a cell image. Boundary region is a subtraction of cytoplasm (yellow) by nucleus (red) region, while background (white) remains the same.

The upper-estimation of the rough set is estimated by assessing all neighbouring pixels for each pixel in the local region of interest:

$$P_{upper}(x,y) = \frac{2}{\sqrt{n}}\left(\sum_{i=1}^{n} |P_{n-neighboring} - P(x,y)|^2\right)^{\frac{1}{2}} \quad \textbf{Equation 2-10}$$

where n is the number of n-neighbouring pixels for each pixel. Practically, n=8 outperforms the other settings in our examination. The upper-approximation set is a collection of all points, which possibly belongs to one segmented region. In this manner, a correlation of spatial information with respect to those who have same or similar values is set up. The lower-approximation set contains original pixels that definitely belong to a class of known intensity, and therefore the roughness index ρ_r can be formulated as:

$$\rho_r = 1 - \alpha_B(X) = 1 - \frac{P(x,y)}{P_{upper}(x,y)} \quad \textbf{Equation 2-11}$$

The value of roughness index is large when the cardinality of the upper-approximation is larger than the original pixel value in the selected position. This typically occurs when there is large variance of the selected pixel with respect to its surrounding pixels; i.e. the intensity variation dramatically changes if there exists a boundary between two objects or regions. In other cases, the roughness index will be small, e.g. close to zero, as there is no significant change of intensity around a selected pixel.

After the two membership functions have been defined, they are combined by using a decision function, such as a parametric aggregation operator from the fuzzy set theory [11]. To simultaneously satisfy abovementioned criteria, while taken both

advantages from fuzzy logic and rough theory, it is of great importance to aggregate using the product t-Norm operation.

$$\mu_X(x) = A(x) = \rho_r(x) \oplus g(x)$$

$$\text{Subject to} \quad A(x) = \prod_{l=1}^{L} C_l(x) \qquad \textbf{Equation 2-12}$$

where $\oplus$ [1] is the aggregate operator and L=2. C_1 and C_2 correspond to $\rho_r(x)$ and $g(x)$ respectively. In this manner, a bi-objective function is aggregated into single-objective one for satisfaction. Note that, regarding to Equation 2-8, which consists of two partial equations that are depending on the local background, as well as on a temporary threshold *t*. It is essential that a correlation index allowing a membership function μ_X continuous weighting at local domain of definition. Therefore, we introduce a correlation function by taking the minimum and maximum intensity level into account, and weight each component of membership function μ_X as:

$$\wedge = \frac{1}{2} \cdot \frac{I_{max} - I_{min}}{I_{global\_max}} \qquad \textbf{Equation 2-13}$$

In Equation 2-13, I_{max} and I_{min} are the maximum and minimum value of local region respectively, while I_{global_max} is the maximum value in whole image. It is easily seen that the value of $\wedge$ will be in [0, 0.5]. Afterward, the bimodal thresholding cost function can be formulized as:

$$\mu_X(x_{mn}) = \rho_r(x) \oplus (\wedge \cdot g_{x_{mn}<t}(x) + (1-\wedge) \cdot g_{x_{mn}\geq t}(x)) \quad \textbf{Equation 2-14}$$

The appropriate measurement of uncertainty is the key to evaluate the degrees of vagueness whereas an element (pixel) belongs to a certain set (region) or not. Several approaches have been reported in recent decades, but in our case, we propose an evaluation based on Shannon's function to solve abovementioned uncertainty problems. From the information entropy theory [12], the measured entropy of the vagueness can also be experienced within a slightly changed definition:

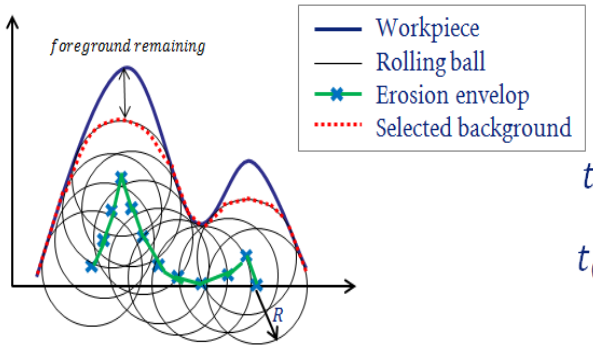
$$E(X) = \frac{1}{MNln2} \sum_m \sum_n S(\mu_X(x_{mn})) \qquad \textbf{Equation 2-15}$$

Where *M* and *N* represent the size of the selected local region. Note that the given Equation 2-15 is monotonically decreasing in the interval [0.5, 1], but monotonically increasing in the interval [0, 0.5]. Hence, it is possible to minimize the lowest energy of the region rolled by the RBA ball path through Equation 2-15, while the definition zone is set to the interval [0.5, 1].

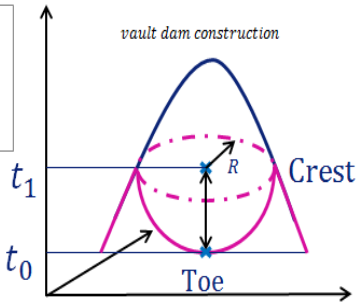
[1] To avoid confusion of the $\oplus$ symbol used in morphological processing as dilation operator.

C. Dam construction strategy

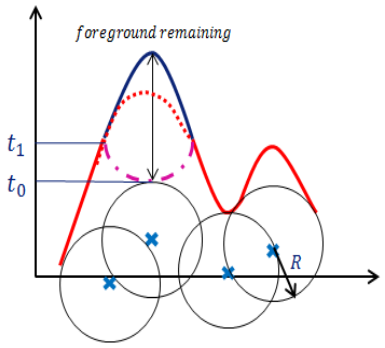
Background illumination correction based on mathematical morphology approach aims to find an estimation of illumination imperfections. For a "ball" rolling under the image landscape, it holds that those pixels the ball cannot touch will be kept in a smoothed foreground; e.g. the local minima and the peak. The smaller the radius of the ball, the deeper the topographical shape can be touched, and vice versa (cf. Figure 2-2). With the existing rolling ball (RBA) method there are some problems that effectuate uncertainty and imprecision in the resulting image. This morphological selection, attributing to over segmentation, will result in a loss of energy and details in the original image. To that end, the proposed method intends to produce a much smoother image and eliminate the artefacts by employing local threshold values t_0 and t_1 conducted from the results of minimization of local entropy, i.e. indicating the crest (t_0) and toe (t_1) of the dam.

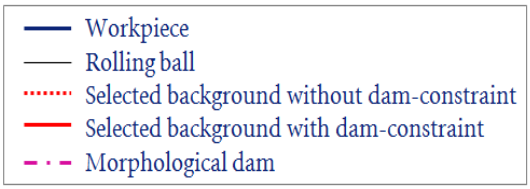


(a) Conventional Morphological BCP



(b) Inllustration of constructing dam in 3D landscape





(c) Dam – constrained BCP

Figure 2-2. Process of constructing morphological dam in a 2D/3D image landscape. The ball path, with a predefine radius *R* (usually set in a range from 50 to 500 pixel), in either red line or dotted line depicts a solution of background selection. (a) classical rolling procedure of RBA in selecting background signal; (b) In a local region of interest, a vault dam is built with regard to local bi-thresholding values. The surface of dam is constructed on the basis of inversed equation (6). However, the extreme values are controlled by both t_0 and t_1; (c) illustration of background selection by means of approaches with and without dam-constraint, respectively.

Intuitively, the morphological ball in the proposed DCBC method, will not roll into the convex area, in which this region of interest is recognized as foreground. In other words, a suppression will occur if the ball is forced to rolling into the region with grey level between t_0 and t_1; and completely forbidden in the area with grey value higher than t_1 (cf. Figure 2-2). Be aware that, the smoothing factor of the foreground during morphological subtraction is relative to the radius of the rolling ball. Unlike the existing algorithm, which requires to be tuned for every step before adapting the ball to the object of interest, the proposed method is more robust as it includes an adaptive radius. The local region of interest is chosen as a sliding window with half the size of the original image, i.e. length and height is *m/2* and *n/2* respectively. Therefore, the radius of the ball is practically in an adaptive way, set as the *minimum (m/2, n/2)*. The procedure of proposed DCBC protocol is illustrated as in Figure 2-3.

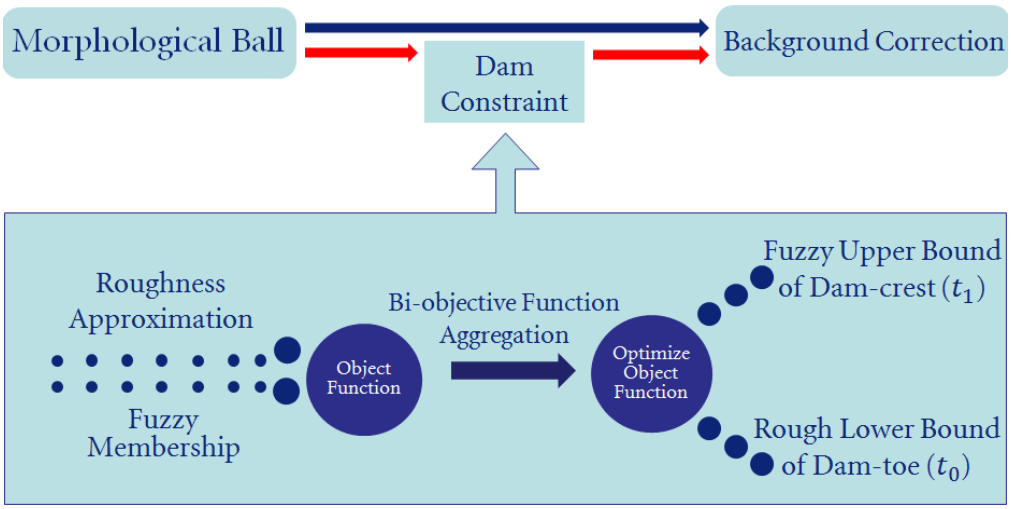


Figure 2-3. Workflow of proposed DCBC approach.

2.5. Experimental Results

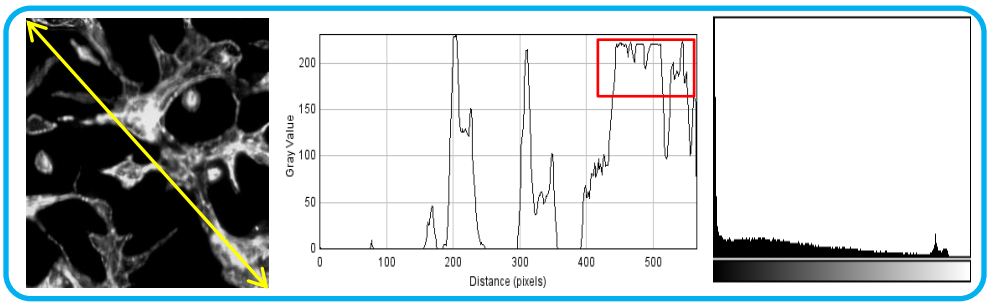
A. Data acquisition

In order to assess the performance of proposed DCBC approach compared with state-of-the-art EMI and typical RBA method, characteristic images are employed. Image Set 1 is typical bright-field imaging, depicting cartilage cell cultures with bright background (256 x 256 pixels, 16 bit); these images were acquired with the standard Zeiss Bright Filed microscope. The other sets (set 2 and set 3) are typical multi-channel fluorescence sets depicting cultured cardiomyocytes with dark background (1024 x 1024 pixels, 8 bit; and 4704 x 3584 pixels, 8 bit); these sets are acquired with the BD-Pathway Imager [6]. Each dataset contains 12 samples in different growth stages. The evaluation methods are then implemented in qualitative and quantitative terms.

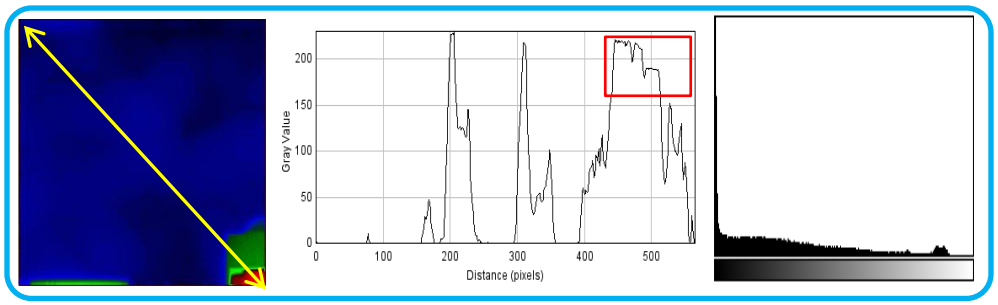
B. Qualitative tests

1) Artefact removal:

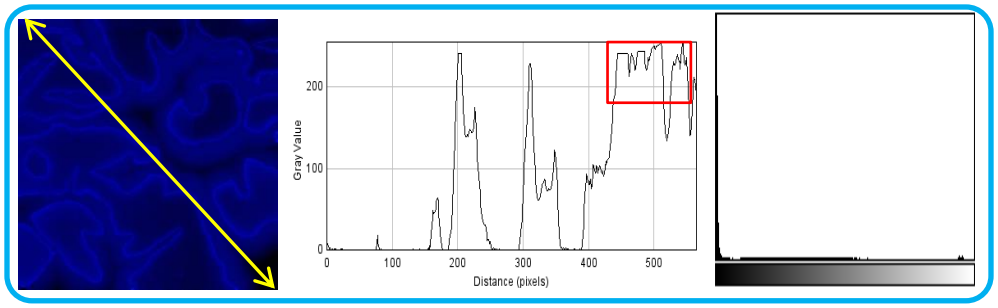
Information utilized for the further processing is actually kept in the background and artefacts are enhanced as well. Artefacts from the connection will, in general, occur between the edge or the corner of an image due to the start centre of conventional RBA and results in an embedded effect of the rolling ball. This is shown in Figure 2-4. (b) as irregular high energy (green and red) in shading image. The rectangular region in image domain, reflects a significant change before and after BCP procedure in image.



(a) Original set



(b) RBA processing



(c) DCBC processing

Figure 2-4. Artefacts removal of cardiomyocytes cell image cultures. (a) Original image, diagonal (along the yellow line) profiling plot and global pixel histogram. (b) and (c) are the strategies of background correction with the resulting shading, diagonal profiling plot and global pixel histogram, RBA and DCBC respectively.

2) Multilevel background subtraction effect:

A well performed BCP procedure would eliminate background signal that present in the image, while preserving the valuable (foreground) signal. Figure 2-5 illustrates the performance of three methods on image sets of cartilage cell cultures (2 weeks, 4 weeks and 7 weeks respectively). The original and the corrected image samples are in the first column, the corresponding surface diagrams (15% Gaussian smoothing processing) are shown in second column, while image global intensity profiles are depicted in the third column. Note that the specified surface plot in Figure 2-5 describes the ability of remaining all information in original foreground and the smoothed and evenly distributed background.

The partitioning of differences in background illumination and foreground is quite difficult in most of the fluorescence and bright-field microscopy images. However in the proposed DCBC method, the information is retained better on the basis of bimodal thresholding strategy, while the distribution of the global intensity is more coherent.

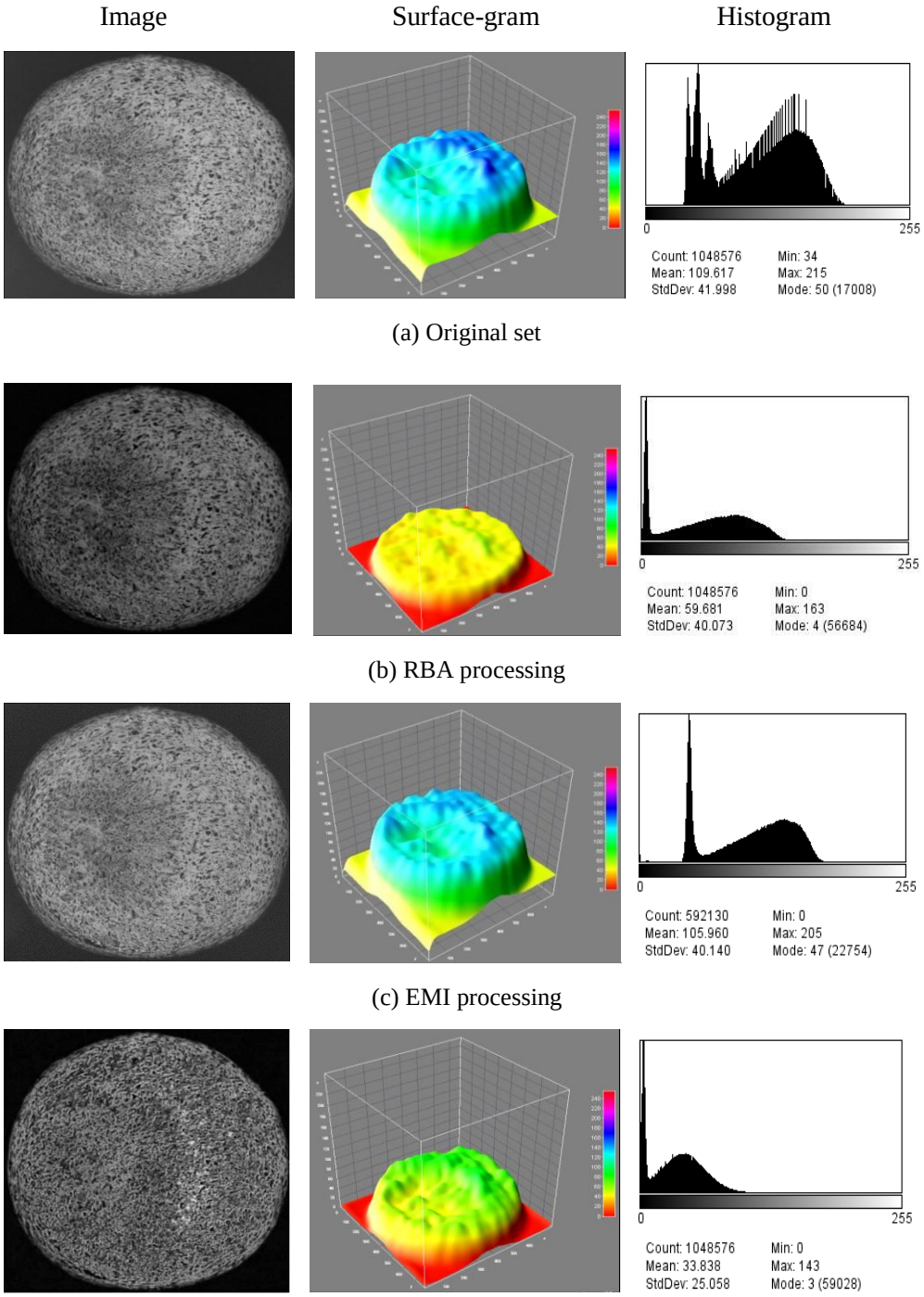


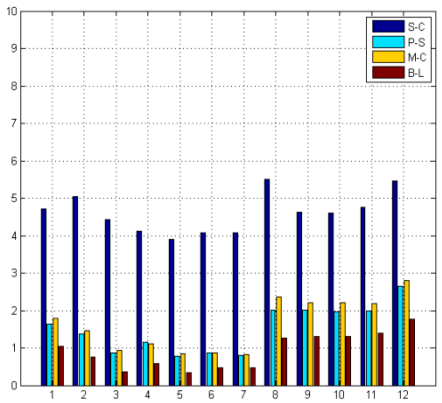
Figure 2-5. Qualitative comparison of different background correction methods in terms of distribution of grey histogram.

C. Quantitative evaluation

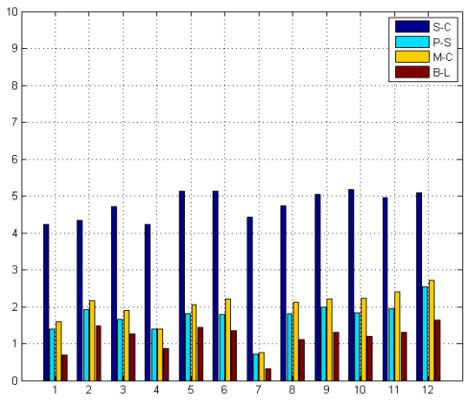
To gain insight in proposed method, a widely acceptable index referred to as Coefficient of Joint Variation (CJV) [13] is utilized. The CJV is characterized by invariance to uniform transformation, i.e., multiplicative and additive illumination. This can be extended to a bimodal formulation:

$$CJV(I_1, I_2) = [\mu(I_1) - \mu(I_2)]^{-1}[\sigma(I_1) + \sigma(I_2)] \qquad \textbf{Equation 2-15}$$

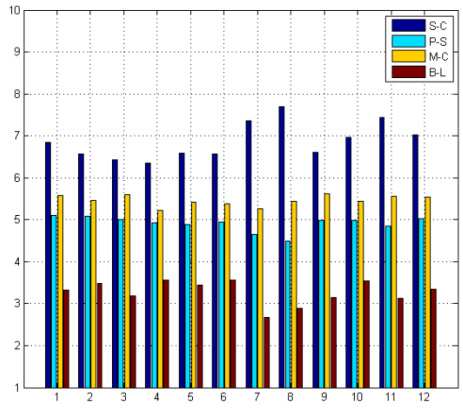
Where, CJV is the sum of the standard deviations of images before I_1 and after I_2 BCP procedure, normalized by the difference of their means. The performances of background subtraction methods are quantitatively evaluated compared with a baseline value, which is computed by the variation of global intensity in the intact (original) image. This can be derived as a limiting form of Equation 2-15 as: $CJV_{Baseline}(I_1) = [\mu(I_1)]^{-1}[\sigma(I_1)]$, where post-processing image is regarded as shading-free (global intensity is zero in fluorescence image type, or max-bit value in bright-field image type; while standard deviation equal to zero as well).



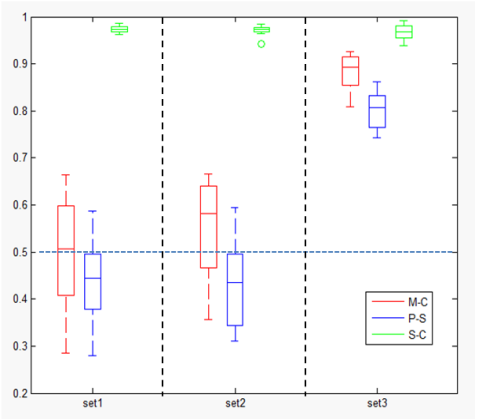
(a) Cartilage weekly #1



(b) Cartilage weekly #2



(c) Cardiomyocyte



(d) Distribution of changes for CJV

Figure 2-6. Performance of three methodologies compared with baseline on three different datasets (a), (b) and (c) respectively. The CJV values are accordingly evaluated and visualized in (d)

The results of background correction are illustrated in Figure 2-6, in which each subfigure stands for one data set. Three representative under-evaluated methodologies, i.e. entropy minimization for shading correction (S-C), morphological correction in RBA (M-C) and proposed DCBC approach (P-S) are compared with baseline value (B-L). In a case of light microscope images after BCP procedure, an ideal distribution of image variation should have smaller global intensity to guarantee background signals are well-eliminated; while global deviation should stand, at least not larger than the intact image to make sure that there is not any tedious signal is introduced. This means that the methodology obtaining the smallest CJV value that outperforms the others. Given by box-whiskers diagram in Figure 2-6 (d), the statistical variation of CJV is then investigated.

From all independent runs, we obtain the average performance of 12 samples for each datasets in terms of CJV value and shown in Table 2-1. The computing time is then illustrated in Table 2-2 for systematically comparison. The best performances (except baseline index) are marked as bold italic in each table.

Table 2-1. Performance of background correction strategies of three dataset in CJV index

	Set1		Set 2		Set 3	
	ΔStd	Δmean	ΔStd	Δmean	ΔStd	Δmean
RBA (M-C)	0.4489	0.6431	0.3391	0.8029	2.9065	23.688
EMI (S-C)	6.5609	11.4812	4.2305	12.4983	51.1308	105.7754
DCBC(P-S)	**0.3547**	**0.5428**	**0.2541**	**0.6144**	**2.1814**	**13.7637**
Baseline (B-L)	0.1390	0.2808	0.1058	0.3388	0.6952	2.7300

Table 2-2. Time complexity in average (seconds)

	Set 1	Set 2	Set 3
EMI (S-C)	58.67	167.23	458.73
RBA (M-C)	1.84	**2.54**	**4.01**
DCBC(P-S)	**1.59**	3.22	6.15

2.6. Discussion and Conclusion

The impacts of the qualitative comparison of BCP procedure performance are shown in Figure 2-4 and Figure 2-5. These results suggest the following evaluation: (i) In Figure 2-4. (c), compared with (b), shows a much better result in terms of the elimination of artefacts signal (removal of square-like high energy region). (ii) In both the red rectangle region in diagonal intensity plot and shown in global pixel histogram, the proposed method kept most information and the intensities across region (slope in the 2D line change) are enhanced for further analysis. (iii) In Figure 2-5, resulting images of visualisation of cartilage cells are shown, where EMI processing eliminates less background than other methods do. (iv) Image after DCBC processing is shading-free, while has clearer and smoother foreground shape that contains all relevant signal. The dam- constraint strategy prevents the over elimination of mixed illumination, and then a more unambiguous and complete cartilage contour can be seen by visual inspection.

The statistics of the three tested approaches can be investigated via Figure 2-6, Table 2-1 and Table 2-2: (i) Evident differences of CJV expression are observed from Figure 2-6 (a) to (c); the likely ranges and interquartile ranges of variation indicated in figure 6 (d) suggests a better performance of proposed method on the three

datasets. (ii) All methods are accomplished in successfully reducing background, while DCBC method always produces better results in terms of smallest CJV value; specifically in set 1 and set 2, a significant improvement can be noticed with a value change of CJV of more than 50%. (iv) Compared with the baseline in Table 2-1, it can be observed that BCP procedure is essential and efficient by removal of noise and redundancy. (v) Proposed method always has a lower CJV value on various datasets, while has a lower time complexity in set 1. (vi) On the basic definition of prevalent EMI method, a retrospectively procedure for estimating shading components consumes a large time budget in application; while convergent parametric components play a role in equalizing original image, meaning a lesser variation compared with intact image and mostly unchanged global intensity. This results in an insufficient elimination of background in fluorescence and bright-field images, and yielding relatively larger CJV value. (vii) In the proposed method, a dam is erected to constrain a path by utilizing both the fuzzy and rough set framework. The membership function of fuzzy logic can handle overlapping partitions; whereas the lower and upper approximations of rough sets can characterize the vagueness and incompleteness in its bimodal class definition. This results in smoothing of the foreground information and a global minimization of the image intensity, while there is not tedious variation introduced.

In this chapter, we propose a Dam-Constraint Background Correction (DCBC) algorithm, which is a novel hybridized approach. The algorithm successfully overcomes the drawback of existing method and includes fully automated data driven parameter tuning. With the innovated morphological concept in image domain, namely “dam-constraint”, the proposed method outperforms the widely and commonly utilized RBA algorithm and the EMI method in both qualitative and quantitative tests. The new method is very promising for application to microscopy images in which further analysis is hampered by undesired effects to background illumination. The subsequent processing steps in proposed data analysis track will be further illustrated in next chapters.

2.7. References

[1]J. C. Waters, “Accuracy and precision in quantitative fluorescence microscopy.” J. Cell Biol., vol. 185, no. 7, pp. 1135–48, Jun. 2009.

[2]B. Likar, J. B. Maintz, M. a Viergever, and F. Pernus, “Retrospective shading correction based on entropy minimization.” J. Microsc., vol. 197, no. Pt 3, pp. 285–95, Mar. 2000.

[3]M. L. Schultz, L. E. Lipkin, M. J. Wade, P. F. Lemkin, and G. M. Carman, “High Resolution Shading Correction” J. Histochem. Cytochem., vol. 22, no. 7, pp. 751–754, Jul. 1974.

[4]A. J. Hanson, “The Rolling Ball” Graphics Gems III, D.Kirk ed., Academic Press, 1992, pp. 51–60.

[5]P. Hall, B. U. Park, and B. A. Turlach, “Rolling-ball method for estimating the boundary of the support of a point-process intensity” no. June, 1998.

[6]F. Zanella, J. B. Lorens, and W. Link, “High content screening: Seeing is believing” Trends in Biotechnology, vol. 28. pp. 237–245, 2010.

[7]U. S. Navy, “Fuzzy Sets *,” vol. 8, no. 3, pp. 338–353, 1965.

[8]Z. Pawlak, “Rough set approach to knowledge-based decision support” Eur. J. Oper. Res., vol. 99, no. 1, pp. 48–57, May 1997.

[9]M. M. Mushrif and A. K. Ray, “Color image segmentation: Rough-set theoretic approach” Pattern Recognition. Lett., vol. 29, no. 4, pp. 483–493, Mar. 2008.

[10]J. C. Waters and J. R. Swedlow, “Techniques Interpreting Fluorescence Microscopy Images and Measurements” 2008.

[11]A. Petrosino and G. Salvi, “Rough fuzzy set based scale space transforms and their use in image analysis” Int. J. Approx. Reason., vol. 41, no. 2, pp. 212–228, Feb. 2006.

[12]A. De Luca and S. Termini, “A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory” Inf. Control, vol. 20, no. 4, pp. 301–312, May 1972.

[13]B. Likar, M. Viergever, and F. Pernus, “Retrospective correction of MR intensity inhomogeneity by information minimization” Med. Imaging, IEEE, 2001.

[14]Z. Ji, Q. Sun, Y. Xia, Q. Chen, D. Xia, and D. Feng, “Generalized rough fuzzy c-means algorithm for brain MR image segmentation.” Comput. Methods Programs Biomed., vol. 108, no. 2, pp. 644–55, Nov. 2012.

[15] Piccinini, F., Lucarelli, E., Gherardi, A. and Bevilacqua, A., 2012. Multi‐image based method to correct vignetting effect in light microscopy images. Journal of microscopy, 248(1), pp.6-22.

[16] Zwier, J.M., Van Rooij, G.J., Hofstraat, J.W. and Brakenhoff, G.J., 2004. Image calibration in fluorescence microscopy. Journal of Microscopy, 216(1), pp.15-24.

[17] Singh, S., BRAY, M.A., Jones, T.R. and Carpenter, A.E., 2014. Pipeline for illumination correction of images for high‐throughput microscopy. Journal of microscopy, 256(3), pp.231-236.

[18] Leong, F.W., Brady, M. and McGee, J.O.D., 2003. Correction of uneven illumination (vignetting) in digital microscopy images. Journal of clinical pathology, 56(8), pp.619-621.

[19] Poon, S.S., Wong, J.T., Saunders, D.N., Ma, Q.C., McKinney, S., Fee, J. and Aparicio, S.A., 2008. Intensity calibration and automated cell cycle gating for high‐throughput image‐based siRNA screens of mammalian cells. Cytometry Part A, 73(10), pp.904-917.

[20] LEE, S.C. and Bajcsy, P., 2006. Intensity correction of fluorescent confocal laser scanning microscope images by mean‐weight filtering. Journal of microscopy, 221(2), pp.122-136.

[21] Huang, Adam, Chung-Wei Lee, and Hon-Man Liu. "Rolling ball sifting algorithm for the augmented visual inspection of carotid bruit auscultation." Scientific Reports 6 (2016).

Chapter 3

Feature Selection Strategy in Region of Interest Mask

This chapter is based on the following publication:

Cai, Fuyu, et al. "Fuzzy Criteria in Multi-objective Feature Selection for Unsupervised Learning." Procedia Computer Science 102 (2016): 51-58

Chapter summary

Feature selection, a procedure in which most informative variables are selected for model generation is an important step for pattern recognition. It is also a crucial step that converts information acquired from a bio-imaging experiment to quantitative data representation. In this effort, one often tries to optimize multiple criteria such as discriminating power of the descriptor, performance of model, and cardinality of subset.

Therefore in this chapter, a fuzzy criterion in multi-objective unsupervised feature selection by applying hybridized filter-wrapper approach (FC-MOFS) is proposed. These formulations allow for a way more efficient approach to pick features from a pool; and to avoid misunderstanding of overlapping features via crisp clustered learning in a conventional multi-objective optimization procedure. Moreover, the optimization problem is solved by using non-dominated sorting genetic algorithm, type two (NSGA-II). The performance of the proposed approach is then examined on six benchmark datasets including multiple disciplines, and different number of features. Systematic comparisons of the proposed method and representative non-fuzzified approaches are illustrated in this work. The experimental studies show a superior performance of the proposed approach in terms of accuracy and feasibility.

3.1. Introduction

Feature selection (FS), in some areas also referred to as dimensionality reduction, deals with selection of one or several optimal sets of attributes that are necessary and/or essential for the recognition process. The main idea of FS to choose a subset of available features that are used to predict the entire population is threefold, i.e. to overcome: (i) working with sets of data with high dimensions and scale to practical and computational proportions; (ii) effects of noise, irrelevant and redundant features that otherwise hinder correct and efficient analysis; (iii) feature dimensions exceeding the sample size as it will induce bias in statistical analysis [1].

The challenge of FS is to decide a minimum subset of features with little or no loss of classification/clustering accuracy. This can be formulated as a multi-objective optimization (MOO) problem. The task is the selection of relevant features, elimination of redundant features, and minimization of selected set cardinality. To date, a range of MOO-based FS techniques have been reported [14]. Cross-applications the related FS approaches can be categorized into four groups:

- Filter-supervised, i.e. class-labels known: features are selected based on their discriminating power with respect to the target classes.
- Wrapper-supervised, i.e. class labels known: subsets of features are evaluated from a classification, at the point where comparison of resulting labels and actual labels occurs.
- Filter-unsupervised, i.e. class-labels unknown: features are ranked from the performance histogram of all feature dimension vectors and one or several criteria are chosen for deciding a group of features.
- Wrapper-unsupervised, i.e. class-labels unknown: computation of the subset of features is applied in terms of the performance of a clustering algorithm. In this case, tuning of parameters in clustering process will contribute in obtaining an acceptable subset of features.

The search for proper supervised predictors can usually be regarded as a pursuit for optimization, where the number of wrong-predicted operators for a known dataset should be minimized [2]. However, figuring out a similar criterion for validation in unsupervised schemas is a difficult task [3]. It cannot be relied upon that a new-found pattern obtained by optimizations resulting from an unsupervised algorithm, is able to decide if a given pattern is trustful or not. To some extent, the validity of pattern discovery is depended on a priori knowledge and intentions of decision makers. This brings us to the assumption that one often desires to employ unsupervised learning schemas in order to produce several candidate solutions for users. Additionally, some tasks in FS, cover inherent data groups and thereby omit features which might reveal the nature of hidden patterns. Therefore, the

unsupervised-based multi-objective heuristic optimization algorithm is becoming an attractive approach, that has been given and increasing attention this decade.

There has been reported on development of evolutionary algorithms for multi-objective (MOEA) for unsupervised feature selection [4]. Oliveira, et al. [5] proposed a Pareto-based approach to generate a so-called Pareto-optimal front in a supervised context. Sensitivity analysis and neural networks (NN) enable to representative evaluation of fitness values. About the same time, Kim, et al. [6], used k-means clustering and Expectation Maximization (EM) as embedded unsupervised approach to evaluate a feature subset encoded in chromosomes. The MOEA employed in this case is called evolutionary local search algorithm (ELSA). With these results as a starting point, research of unsupervised learning in feature selection was expanded. Morita, et al. [7] used the k-means clustering algorithm in a wrapper approach, which was encoded with Non-dominated Sorting Genetic Algorithm, type two (NSGA-II). Moreover, two objective functions, i.e. the number of features in a set, and a clustering validation (e.g. Davies-Bouldin (DB) [8]) index are introduced. Handl and Knowles [9] examined different combinations of objective functions and Mierswa [2] investigated different indices, i.e. the normalized DB index. More recent work [10] stated that their multi-objective unsupervised feature selection algorithm (MOUFSA) outperforms several other multi-objective and conventional single-objective methods, by using redundant measurements and negative epsilon-dominance. In addition, three new mutation methods are designed to enhance MOUFSA.

However, the defined criteria in classical objective functions used in unsupervised MOEA, fail to predict the performance of clustering results, i.e. the overlapping information (features) in-between classes which probably highlights the essentials that are shared within these classes. To solve this problem, we employ fuzzy criteria in a hybrid filter-wrapper approach. Pioneered by Zadeh [11], fuzzy logic-based systems have been successfully utilized to various application areas, e.g. control system and pattern classification [12]. The comprehensibility of fuzzy criteria, namely the linguistic interpretability of fuzzy partitions and the simplicity of fuzzy if-then rules [13], makes it a promising method to access qualified optimization in MOEA when employed into unsupervised learning. Although fuzzy criteria are addressed in a supervised manner [14], it rarely has been reported in unsupervised cases, in which the natural patterns are discovered according to fuzzy clustering validity and fuzzy objective functions.

In this chapter, FS procedure is optimized using the generic heuristic search algorithm NSGA-II, and fuzzy criteria are employed in both filter and wrapper approaches. In the unsupervised learning procedure a new fuzzy index is specifically proposed as one of the objective functions. The target functions are: (i) value of

Correlation Membership Measurement (CMM); and (ii) cardinality of feature subset. Here we intend to contribute to the further development of the hybrid methodology, by realizing a sensible integration of fuzzy criteria and MOEA approach in FS area This methodology is applied to a wide set of benchmark datasets and it is compared with commonly used approaches to show its general applicability and competitive advantages.

The remainder of this chapter is organized as follows. In Section 3.2, we introduce the methodology including application of fuzzy criteria and fuzzy model in FS; subsequently, the utilization of NSGA-II in an unsupervised context is presented. In Section 3.3 experimental results are given and Section 3.4 conclusions are presented.

3.2. Methodology

A. Fuzzy entropy in filter-approach

In information theory, entropy is a measure of chaos or uncertainty associated with the variables. The concept of entropy has been defined in various ways and used in different fields; fuzzy logic is becoming commonly used in the estimation of entropies. On this basis, we propose an approach embedding fuzzy c-means (FCM) [15] clustering algorithm to estimate the fuzzy entropy by automatically computing the feature memberships. To depict the level of similarity, the feature membership index assigned with a fuzziness characteristic that can be expressed as:

$$u_{ij} = (\sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{\frac{2}{mf-1}})^{-1} \qquad \textbf{Equation 3-1}$$

Here $mf \in (1, \infty]$ is a scalar that is termed fuzzifier for FCM, and d_{ij} is the product norm distance from object $a_j \in a_1, a_2, \dots, a_n$, to the cluster centroid $v_i \in v_1, v_2, \dots, v_m$. This membership function is subject to the following objective function:

$$J_{fcm} = \sum_{j=1}^{n} \sum_{i=1}^{m} d_{ij} (u_{ij})^{mf} \qquad \textbf{Equation 3-2}$$

In this manner, according to De Luca and Termini [16], the fuzzy entropy can be defined as:

$$H\big(u_j(x)\big) = \frac{1}{n \ln 2} \sum_{j=1}^{n} -u_j(x) \ln u_j(x) - (1 - u_j(x)) \ln(1 - u_j(x))$$

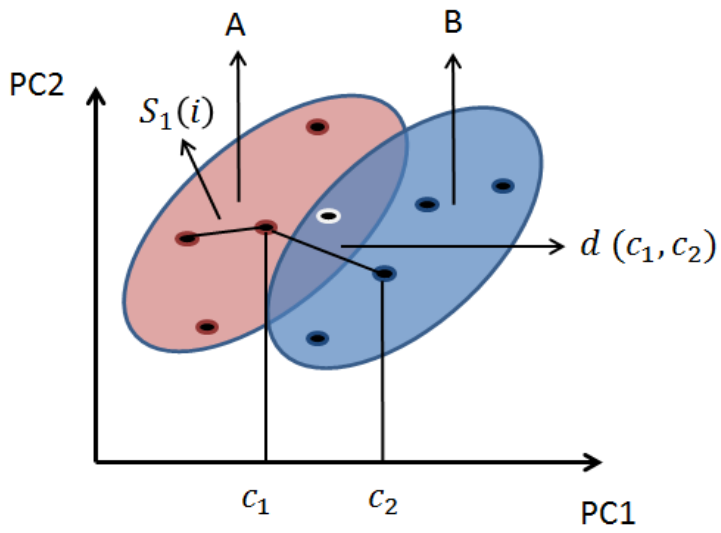
Equation 3-3

In Equation 3-3, $u_j(x)$ denotes the membership index of the j^{th} feature in the feature pattern vector, meaning every individual feature entropy is computed along all the

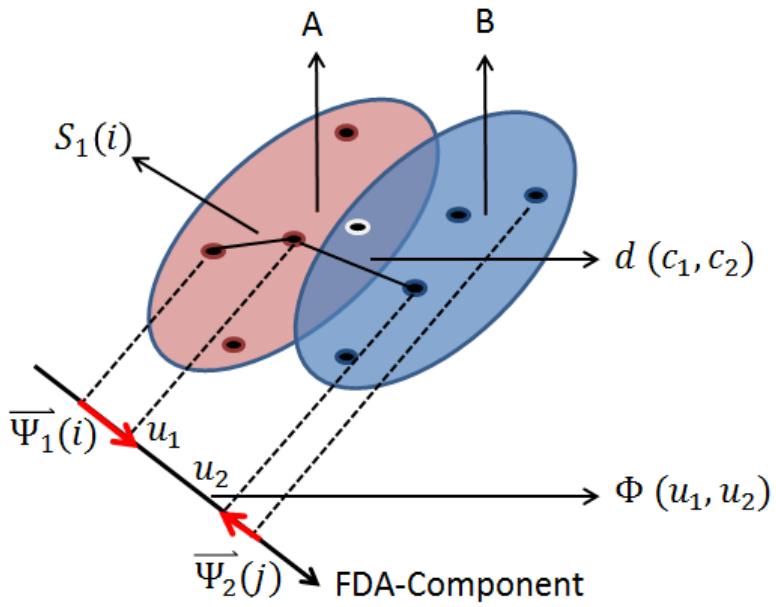
samples *x*. Subsequently, the entire set of features is ranked for guiding the optimization procedure in the wrapping approach, via maximizing their corresponding fuzzy entropies.

B. Fuzzy cost function in wrapping-approach

Multi-objective function optimization, by means of a wrapper technique for unsupervised feature selection, relies on the use of an internal technique of cluster validation. In other words, clustering validation techniques have been designed specifically for the selection of the best clustering solution on the basis of its distance performance. Sometimes the clustering performance is estimated by considering the ratios between intra-class compactness, and inter-class separation. As reported Handl and Knowles [9], this generally suffers from the bias of these measurements with respect to the dimensionality of the feature space. The conflict of this bias can be noticed when dimensionality of a given dataset is enlarged: i.e. the mean of the distribution tends to increase while simultaneously the variance of the distribution decreases. This will cause such a validation technique to be unable to sensitively estimate the difference between all pairs of points, especially in a high dimensional dataset.



(a)



(b)

Figure 3-1. Sketch diagram for CMM. (a) In a high dimensional feature space, two overlapping classes (A and B) with centroids c_1 and c_2 are projected onto two principle component axis PC1 and PC2. *S* is the distance between objects and its belonging center; *d* is the in-between cluster center distance. (b) After Fisher discriminant analysis (FDA) linear projection (project onto the FDA-Component axis), the corresponding components can be rewritten as $d\ (c_1, c_2) \mapsto \Phi(u_1, u_2)$ and $S \longmapsto \vec{\Psi}$ respectively.

To tackle this bias, we propose a fuzzy cost function, the correlation membership measurement (CMM). This function employs both individual clustering information and shared (overlapping) information (cf. Figure 3-1 (a)). We measure the similarity between pairs of vectors using their scalar distance and their directions in high-dimensional attribute space are compared via the projection onto low-dimensional space (cf. Figure 3-1 (b)). This is defined as:

$$\mathrm{CMM} = U_{A \cup B} + U_{A \cap B} \qquad \textbf{Equation 3-4}$$

subject to

$$\begin{cases} U_{A\cup B} = \frac{\frac{1}{N}\sum_{i=1}^{N} S_1(i) + \frac{1}{M}\sum_{j=1}^{M} S_2(j)}{d\,(c_1, c_2)} \\ U_{A\cap B} = \frac{1}{NM}\sum_N \sum_M \{\frac{|\overrightarrow{\Psi_1}(i) - \overrightarrow{\Psi_2}(i)| \cdot |\overrightarrow{\Psi_1}(j) - \overrightarrow{\Psi_2}(j)|}{||\Phi(u_1, u_2)||^2}\} \end{cases} , \quad i \in \text{A and } j \in \text{B}$$

Equation 3-5

Where in the first term of equation 3-4, i.e. the dependent membership $U_{A\cup B}$ of class A and class B are measured, $S_1(i)$ and $S_2(j)$ are the distance of the vector i and j to their corresponding centroid c_1 and c_2; while $d\,(c_1, c_2)$ is the distance between two cluster centroids, and ‖.‖ is distant norm as well. N and M are the numbers of the elements that belong to their classes. The evaluation of performance for the overlapping clusters can be achieved by estimating the positions of every individual vectors in a feature subspace. In a high-dimensional domain, however, the comparison of vectors in terms of directions and angles is not applicable. Therefore, principle projection in FDA [17] is used to find a linear combination of features that characterizes two or more classes. The projection matrix can be defined as:

$$\omega = S_w^{-1}(c_1 - c_2) \qquad \textbf{Equation 3-6}$$

Where $$S_w = (i - c_1)(i - c_1)^T + (j - c_2)(j - c_2)^T$$

Equation 3-7

Subsequently, in the second term of Equation 3-4, i.e. in the correlated membership $U_{A\cap B}$, the projected vector Ψ and Φ can be obtained by multiplying *S* norm and *d* norm with FDA projection matrix *ω* respectively. Moreover, one should realize that, when applied on a real dataset, the S_w, i.e. the with-in class scatter matrix, normally is a singular matrix and thus non-invertible. We have added a tiny perturbation factor to prevent the projection program from being trapped and the projection matrix is rewritten as:

$$\omega = (S_w + \varepsilon I)^{-1}(c_1 - c_2) \qquad \textbf{Equation 3-8}$$

Here *I* is a unit diagonal matrix. The objective is to achieve proper clustering by minimizing the CMM index. With respect to the aim of feature selection, it is more efficient and direct to use the cardinality of feature subsets as a second cost function. However, one can observe (cf. Figure 3-3) that the CMM value decreases with increasing feature numbers. Therefore, a constraint is that at least one feature count in the second objective function should be set.

C. Bi-objective optimization

In the previous section, two objective functions (the cardinality of feature subsets and Equation 3-4) are formulated as quality indicators for the feature extraction procedure. Those two objective functions are conflicting and form a combinatorial bi-objective optimization problem. Therefore, we aim at searching for the Pareto front [20], which represents the non-dominated solutions of the proposed feature selection procedure and which can be used to assess the trade-off. In order to achieve this, Evolutionary Multi-objective Optimization Algorithm (EMOA) is adopted due to its capability of handling combinatorial problems. We specifically utilized the well-known NSGA-II [22] algorithm (Non-dominated Sorting Genetic Algorithm) which is the multi-objective extension to the classical Genetic Algorithm [23]. NSGA-II has the ability to generate well-spread Pareto fronts with relatively low computational overhead and it is proved to be robust in real-world applications through numerous testing and applications. In this chapter, we omit the detailed discussion on the optimization procedure and use NSGA-II as a 'standard' multi-objective optimizer.

As we are dealing with combinatorial optimization problem, discrete Pareto fronts are obtained from NSGA-II, in which each point on the resulting Pareto front represents a candidate feature subset. Each candidate solution will be used for the clustering algorithm and the one giving the best clustering performance (cf. the performance indicators in Section 3.2 B) is chosen. Note that the functionality of the bi-objective optimization is to pre-screen the 'bad' candidate solutions (Pareto dominated feature subsets) from all the possible solutions, leaving the Pareto optimal candidates, the number of which is very small compared to the entire number of solution candidates, to be tested in clustering.

Combining fuzzy entropy in priori evaluation of feature sets in filtering approach and fuzzy criterion in objective function in wrapping approach, the proposed FC-MOFS algorithm manages to assess best candidate feature subsets using NSGA-II. To that end, the detailed procedure of proposed methodology is shown in Figure 3-2.

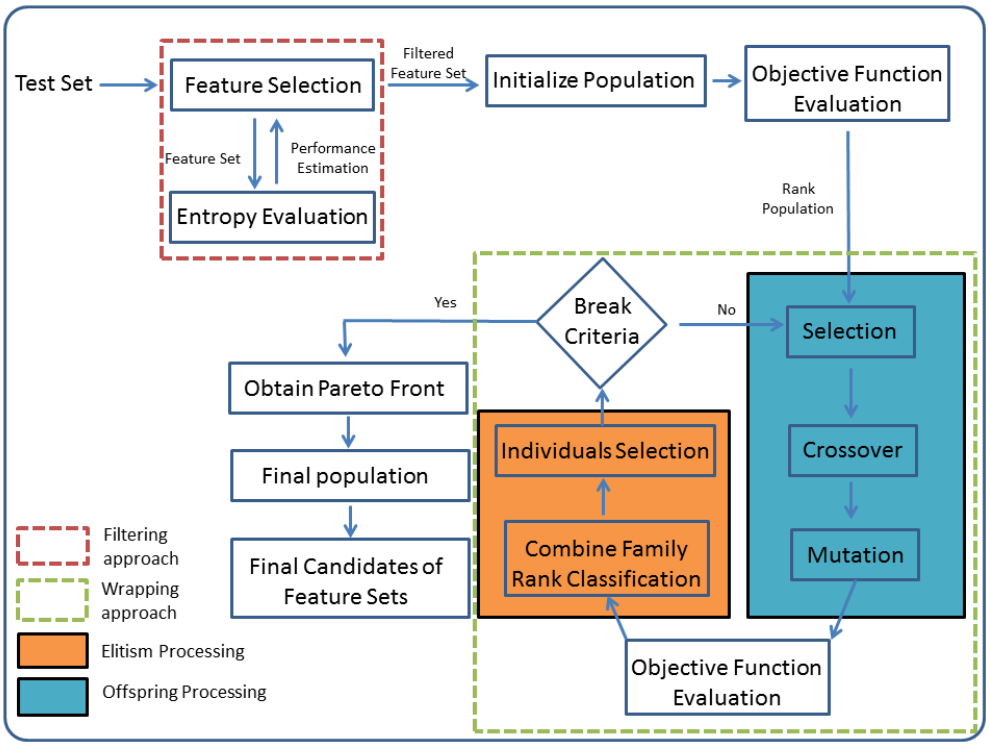


Figure 3-2. The overview of Fuzzy Criteria in Multi Objective Feature Selection (FC-MOFS) process.

3.3. Experimental Results

The objective of this section is to assess the performance of integrating fuzzy criteria into unsupervised multi-objective feature selection procedure. Acceptable results in terms of developing either searching optimization or clustering validation algorithms has been reported in a number of papers. However, for a fair and effective validation of the proposed FC-MOFS method, a commonly used approach without fuzzy constraint [9], referred to as NF-MUFS, is used. Additionally, all datasets are employed in Baseline, using the full feature set. The experiments are conducted on six publicly available datasets, representing multiple disciplines and real life problems (cf. Table 3-1).

Table 3-1. Dataset description

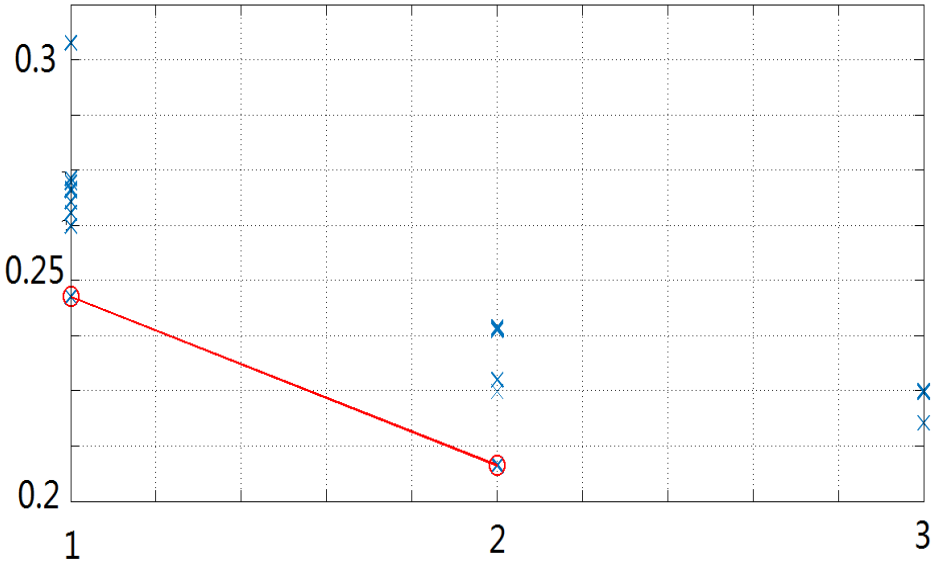
Dataset	Type	Size	Dimension	Class
Glass	Numerical data	214	9	6
Wine		178	13	3
WDBC		569	30	2
Libras		270	90	15
Sonar	Voice	208	60	2
UMIST	Image	575	644	20

A. Parameter setting

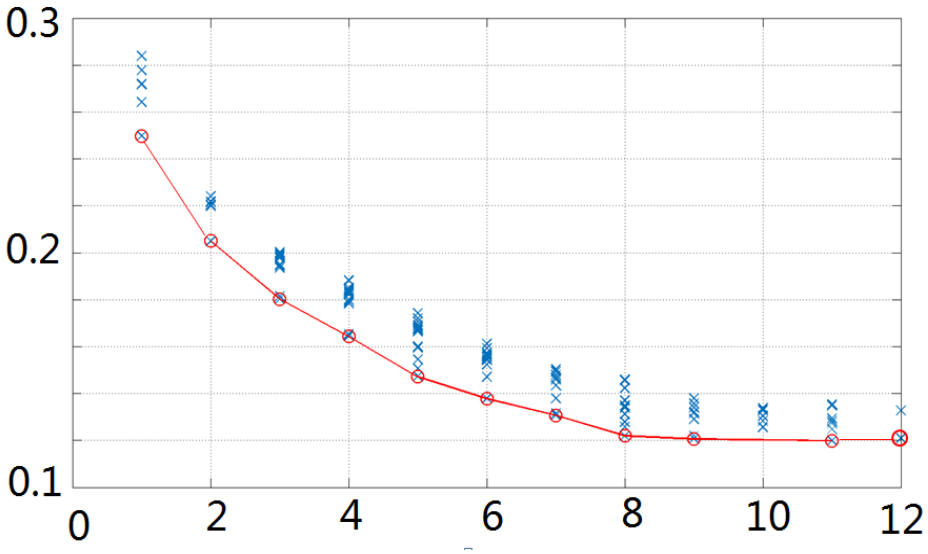
In both FC-MOFS and NF-MUFS, the maximum generation and population size are set as same to 100 and 25 respectively; the crossover percentage is 0.9 and the mutation percentage is 0.4, while the rate of mutation is adaptively selected according to the non-dominated sorting performance and expected number of local optima. The clustering algorithm in unsupervised learning of FC-MOFS is fuzzy c-means, which is substituted by k-means in NF-MUFS.

B. Validation of FS approach

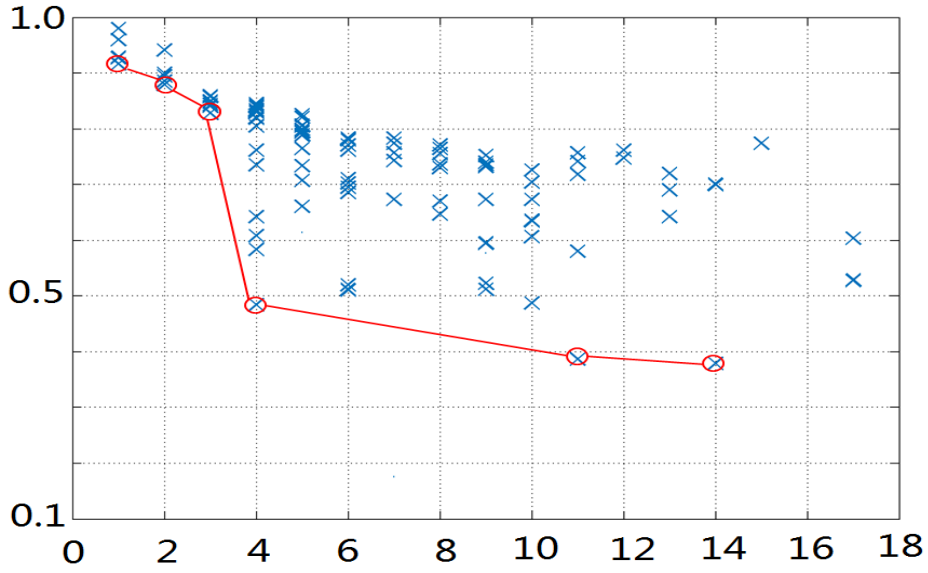
Following previous works, three widely used evaluation metrics, i.e., Accuracy [18] (ACC), Normalized Mutual Information [19] (NMI) and Rand Index [20] (RI) are computed in this chapter. To gain insight in the proposed method, we investigated some aspects that influence clustering performance after feature selection schemes. In the filter approach, the fuzzy entropy feature selection runs once to rank all features for guiding the process in NSGA-II algorithm as initialization; then the results of 20 independent runs of NSGA II to obtain global non-dominated features (cf. Figure 3-3) set are tested on six different benchmarks (cf. Table 3-2 to Table 3-4). Setting three different evaluation strategies, i.e., the application on full sample population (f-s), random sampling (r-s) on the basis of bootstrapping, and uniform distribution sampling (u-s), the accuracy and general capability of FC-MOFSA are measured in overall 50 times.



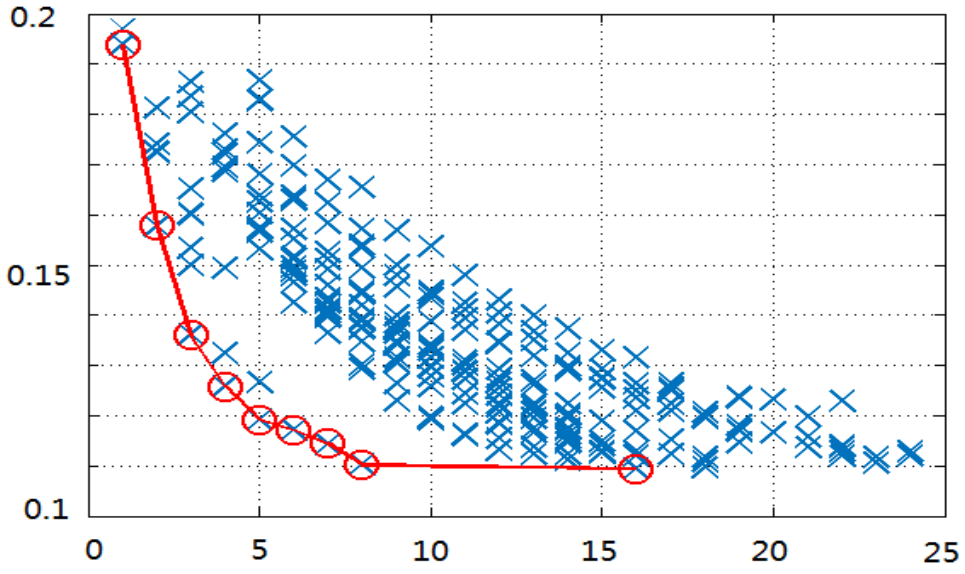
(a) Glass



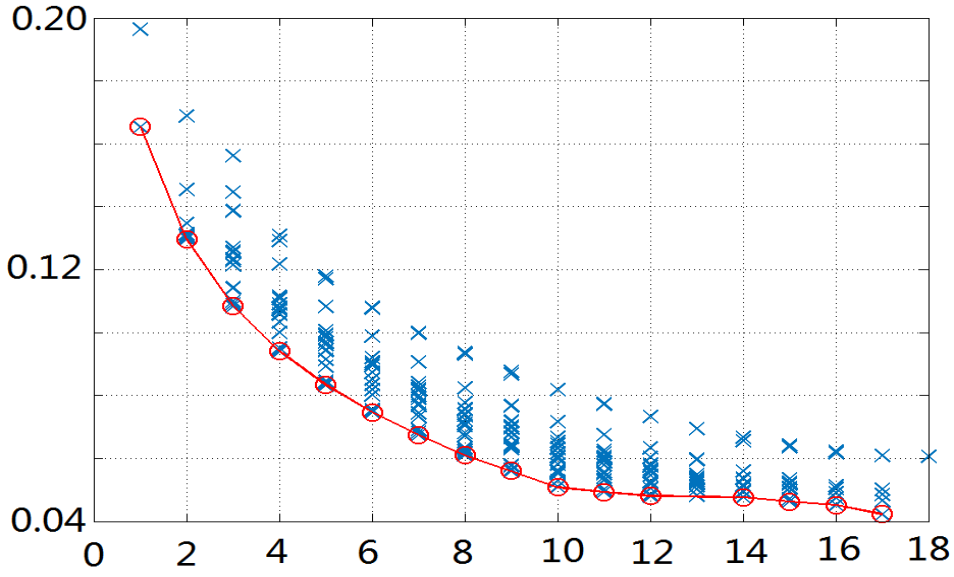
(b) Wine



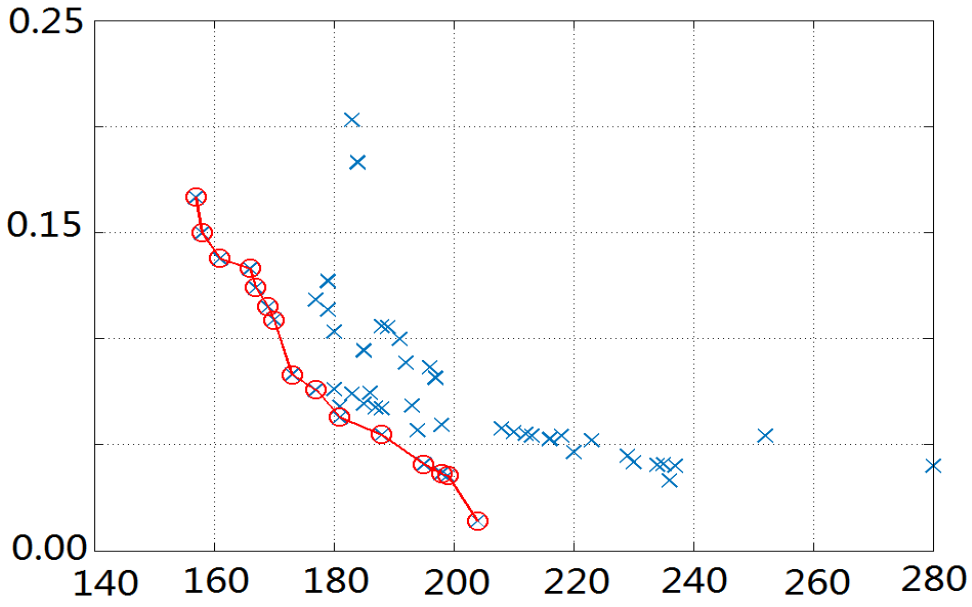
(c) WDBC



(d) Libras



(e) Sonar



(f) UMIST

Figure 3-3. The Pareto fronts for all dataset ((a) to (f)), consisting of 20 independent runs for each database, including 100 generations per run; the global non-dominated sets are selected (red circle) from local non-dominated sets (blue cross). The vertical axis is CMM error w.r.t the number of features on the horizontal axis.

Table 3-2. Impact of fuzzy and non-fuzzy feature selection algorithms to the clustering results in ACC index (best-row performance is marked as bold italic)

dataset	Sampling strategy	ACC ± std (%)				
		FC-MOFSA	*nf*	NF-MUFS	*nf*	Baseline
Glass	f-s	***53.93*** ± 2.11	2	44.72 ± 2.12	2	44.20 ± 4.15
	r-s	***54.36*** ± 3.77	2	42.61 ± 3.39	3	41.76 ± 4.43
	u-s	***55.50*** ± 4.09	2	43.36 ± 2.47	1	41.85 ± 4.81
Wine	f-s	***80.34*** ± 3.15	2	75.28 ± 3.22	3	70.22 ± 2.28
	r-s	***80.27*** ± 5.15	2	75.20 ± 3.45	10	69.02 ± 5.21
	u-s	***78.31*** ± 6.34	2	74.58 ± 4.97	3	68.00 ± 7.39
WDBC	f-s	***88.40*** ± 2.38	2	83.83 ± 1.85	14	85.41 ± 2.49
	r-s	***88.27*** ± 2.15	2	84.38 ± 1.95	14	84.51 ± 2.11
	u-s	***88.37*** ± 2.54	3	84.83 ± 2.34	6	84.53 ± 2.64
Libras	f-s	***47.79*** ± 4.44	16	44.44 ± 4.29	20	44.81 ± 2.21
	r-s	***28.85*** ± 3.35	16	27.67 ± 4.30	20	17.23 ± 2.01
	u-s	28.46 ± 4.22	16	***28.55*** ± 4.31	29	17.93 ± 2.28
Sonar	f-s	***57.44*** ± 2.99	15	51.44 ± 2.46	4	55.29 ± 3.85
	r-s	***59.67*** ± 2.89	5	54.35 ± 2.30	14	55.20 ± 3.73
	u-s	***60.67*** ± 3.34	4	54.46 ± 2.70	16	56.37 ± 4.03
UMIST	f-s	***47.91*** ± 4.11	167	45.78 ± 2.88	197	43.65 ± 1.48
	r-s	***25.78*** ± 2.39	199	22.50 ± 2.48	197	13.43 ± 1.53
	u-s	***25.56*** ± 3.20	204	23.33 ± 3.05	197	13.78 ± 1.45

Table 3-3. Impact of fuzzy and non-fuzzy feature selection algorithms to the clustering results in NMI index (best-row performance is marked as bold italic).

dataset	Sampling strategy	NMI ± std (%)				
		FC-MOFSA	*nf*	NF-MUFS	*nf*	Baseline
Glass	f-s	**41.25** ± 4.35	2	33.12 ± 2.63	2	39.37 ± 5.42
	r-s	**45.14** ± 4.25	2	35.14 ± 2.93	2	38.60 ± 5.08
	u-s	**47.01** ± 3.33	2	36.62 ± 3.76	2	39.11 ± 5.50
Wine	f-s	**52.37** ± 5.43	2	41.63 ± 5.22	10	42.87 ± 5.19
	r-s	**53.36** ± 0.64	2	44.60 ± 4.86	10	44.95 ± 6.40
	u-s	**52.09** ± 8.40	2	44.09 ± 5.61	7	44.95 ± 7.90
WDBC	f-s	**44.79** ± 5.35	1	38.02 ± 4.28	28	42.20 ± 5.08
	r-s	**41.17** ± 5.01	1	38.56 ± 4.44	4	40.41 ± 4.32
	u-s	39.85 ± 5.45	2	39.90 ± 5.24	6	**41.42** ± 5.25
Libras	f-s	**62.10** ± 2.83	16	56.36 ± 3.33	29	60.84 ± 3.44
	r-s	**25.98** ± 3.00	16	20.80 ± 3.24	16	19.93 ± 3.39
	u-s	**28. 85** ± 2.59	16	22.67 ± 3.39	29	22.01 ± 3.52
Sonar	f-s	**0.91** ± 0.81	14	0.91 ± 1.83	4	0.88 ± 0.87
	r-s	**2.53** ± 0.71	5	1.82 ± 0.79	14	1.21 ± 0.73
	u-s	**2.84** ± 1.11	4	1.95 ± 1.47	16	1.64 ± 0.81
UMIST	f-s	63.84 ± 4.04	167	**64.74** ± 4.83	167	63.82 ± 1.83
	r-s	**25.57** ± 3.95	199	20.17 ± 3.86	197	13.10 ± 2.12
	u-s	**28.39** ± 4.67	204	22.43 ± 4.98	197	14.86 ± 1.63

Table 3-4. Impact of fuzzy and non-fuzzy feature selection algorithms to the clustering results in RI index (best-row performance is marked as bold italic).

dataset	Sampling strategy	RI ± std (%)				
		FC-MOFSA	*nf*	NF-MUFS	*nf*	Baseline
Glass	f-s	***65.49*** ± 2.15	2	58.94 ± 2.88	2	53.63 ± 4.32
	r-s	***65.97*** ± 2.17	2	58.22 ±2.62	2	48.89 ± 3.60
	u-s	***65.59*** ± 2.03	2	58.35 ± 2.93	2	44.35 ± 1.58
Wine	f-s	***77.86*** ± 3.02	1	73.00 ± 2.99	3	71.86 ± 5.58
	r-s	***78.03*** ± 3.90	1	74.53 ± 3.20	3	43.66 ± 5.02
	u-s	***76.48*** ± 4.90	1	74.01 ± 4.10	3	44.91 ± 5.46
WDBC	f-s	73.79 ± 3.58	1	73.08 ± 2.99	5	***75.04*** ± 2.19
	r-s	***74.34*** ± 3.02	1	73.64 ± 2.68	14	50.70 ± 5.51
	u-s	***74.46*** ± 3.89	2	74.27 ± 3.26	6	50.46 ± 5.92
Libras	f-s	***90.40*** ± 4.85	16	90.16 ± 3.25	20	90.37 ± 7.85
	r-s	90.68 ± 4.76	16	***91.30*** ± 4.66	29	83.87 ± 7.87
	u-s	***91.55*** ± 4.95	16	91.29 ± 6.26	18	82.34 ± 2.45
Sonar	f-s	***50.80*** ± 6.55	4	49.70 ± 6.88	4	50.32 ± 4.19
	r-s	***51.11*** ± 5.98	5	50.16 ± 5.08	4	49.97 ± 3.91
	u-s	***51.18*** ± 8.53	4	50.14 ± 8.78	4	49.99 ± 6.42
UMIST	f-s	***95.51*** ± 6.11	198	88.51 ± 4.37	197	92.80 ± 1.48
	r-s	***94.69*** ± 5.24	199	86.95 ± 4.58	167	88.01 ± 1.04
	u-s	***94.40*** ± 6.65	199	89.11 ± 5.12	167	85.72 ± 1.24

The results of bi-objective optimization are illustrated in Figure 3-3, in which each subfigure stands for one data set. The blue crosses in the figure represent different candidate feature subsets after the termination of NSGA-II optimizer. Because of the stochasticity of the NSGA-II optimizer, 20 independent runs are conducted for each data set, resulting in a 'layering structure' of the blue crosses. From all the independent runs, we only selected the non-dominated ones using the non-dominated sorting technique. The Pareto fronts generated from 20 independent runs are marked by red circles in Figure 3-3. Most of the Pareto fronts are convex, except for Figure 3-3(a), in which only 3 features are present and which indicates the existence of trade-off solutions. In addition, the points on the Pareto front are well-spread. In Figure 3-3(c), the distribution of the points is not as good as the rest, which suggests that using more evaluation budget in the multi-objective optimization might improve the quality of the Pareto front on the WDBC dataset. On the basis of our candidate solutions, the resulting Pareto fronts are reliable for using later in the clustering algorithm.

The details of six datasets are shown in Table 3-1. The results of comparisons of clustering performance are listed in Table 3-2, Table 3-3 and Table 3-4. The values indicated in bold are the best results among the algorithms in the same situation and *nf* denotes the number of features used in the clustering. These results suggest the following evaluations: (1) compared with the baseline, it can be observed that the feature selection procedure is necessary and efficient by removal of noise and redundancy. (2) the best solutions of the proposed FC-MOFSA mostly have a higher accuracy, mutual information and RI other than the non-fuzzified feature selection algorithms (NF-MUFS and Baseline employment). In spite of the slightly less performance on WDBC, Libras and UMIST dataset, the u-s and f-s value are still competitive compared with the best results of other methods. (3) The average r-s means that even though with less samples (information) obtained from entire population, still, in most situations, the results of FC-MOFSA are better than those of NF-MUFS and Baseline. (4) The proposed method, in most cases, has the least numbers of features for prediction of the best results. In the second highest cases, FC-MOFSA still obtains the lowest cardinality of feature sets. (5) By expressing the descriptor of similarity in RI and descriptor of redundancy in NMI, our method achieves an accurate clustering performance. This is due to the exploitation of discriminative and overlapping information in an unsupervised context. (6) The accuracy and the similarity grouping capability of the experimental algorithms suffer from a serious degradation when down-sampling is applied on the Libras and UMIST dataset. The sparse distribution of these dataset complicates the unsupervised categorization scheme. However, it is observed that FC-MOFSA is superior to the rest approaches by uncovering the underlying patterns and possibly skewed structure.

3.4. Conclusion

In this chapter, we present a new multi-objective feature selection algorithm utilizing the fuzzy hybrid filter-wrapper approach. We introduce a fuzzy criterion-based manner in multi-objective optimization problems and thereby increase the clustering accuracy in unsupervised feature selection schemas. The proposed method outperforms the commonly used multi-objective feature selection method with non-fuzzified parameters, in terms of accuracy and general capability. In addition to the fuzzy entropy in pre-selection, we also present a new fuzzy index called Correlation Membership Measurement (CMM), which produces superior results, particularly on sparse and skewed datasets. This methodology engages a way that attributes can be promisingly selected from high dimensional yet sparse and skewness dataset. The chosen of the sets of feature candidates provides according means for decision maker to efficiently and precisely draw prediction.

3.5. References

[1] Mierswa I, Wurst M. Information preserving multi-objective feature selection for unsupervised learning. Proceedings of the 8th annual conference on Genetic and evolutionary computation, pp. 1545-1552. ACM, 2006.

[2] Li Z, Lu H. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Transactions on Knowledge and Data Engineering. 2014 Sep; 26(9):2138-50.

[3] Mukhopadhyay A, Coello CA. A survey of multiobjective evolutionary algorithms for data mining: Part I. IEEE Transactions on Evolutionary Computation. 2014 Feb; 18(1):4-19.

[4] Oliveira LS, Suen CY. Feature selection using multi-objective genetic algorithms for handwritten digit recognition. In Pattern Recognition, 2002. Proceedings. 16th International Conference on 2002 Vol. 1, pp. 568-571. IEEE.

[5] Kim Y, Menczer F. Evolutionary model selection in unsupervised learning. Intelligent data analysis. 2002 Jan 1; 6(6):531-56.

[6] Morita ME, Suen CY. Unsupervised Feature Selection Using Multi-Objective Genetic Algorithms for Handwritten Word Recognition. In ICDAR 2003 Aug 3 Vol. 2, pp. 666-670.

[7] Davies DL, Bouldin DW. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence. 1979 Apr; (2):224-7.

[8] Handl J, Knowles J. Feature subset selection in unsupervised learning via multiobjective optimization. International Journal of Computational Intelligence Research. 2006 Jun; 2(3):217-38.

[9] Xia H, Yu D. Multi-objective unsupervised feature selection algorithm utilizing redundancy measure and negative epsilon-dominance for fault diagnosis. Neurocomputing. 2014 Dec 25;146:113-24.

[10] Zadeh LA. Fuzzy sets as a basis for a theory of possibility. Fuzzy sets and systems. 1999 Dec 31;100:9-34.

[11] Lajoie SP, Derry SJ, editors. Computers as cognitive tools. Routledge; 2013 May 13.

[12] Ishibuchi H, Yamamoto T. Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. Fuzzy sets and systems. 2004 Jan 1;141(1):59-88.

[13] Vieira SM, Kaymak U. Fuzzy criteria for feature selection. Fuzzy Sets and Systems. 2012 Feb 16;189(1):1-8.

[14] Trivedi MM, Bezdek JC. Low-level segmentation of aerial images with fuzzy clustering. IEEE Transactions on Systems, Man, and Cybernetics. 1986 Jul;16(4):589-98.

[15] De Luca A, Termini S. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. Information and control. 1972 May 31;20(4):301-12.

[16] Scholkopft B, Mullert KR. Fisher discriminant analysis with kernels. Neural networks for signal processing IX. 1999 Aug;1(1):1.

[17] Papadimitriou CH, Steiglitz K. Combinatorial optimization: algorithms and complexity. Courier Corporation; 1982.

[18] Ghosh J, Acharya A. Cluster ensembles. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011 Jul 1;1(4):305-15..

[19] Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association. 1971 Dec 1;66(336):846-50.

[20] Zitzler E, Thiele L. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. IEEE transactions on Evolutionary Computation. 1999 Nov;3(4):257-71.

[21] Deb K, Meyarivan TA. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation. 2002 Apr;6(2):182-97.

[22] Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. U Michigan Press; 1975.

Chapter 4

Unsupervised Information Classification and Analysis

This chapter is based on the following publication:

Cai, F., and F. J. Verbeek. "Rough fuzzy c-means and particle swarm optimization hybridized method for information clustering problem." J. Commun 11 (2016): 1106-1113.

Chapter summary

This chapter presents a hybrid unsupervised clustering algorithm for biological data analysis, referred to as the rough fuzzy c-means (RFCM) algorithm and particle swarm optimization (PSO). The PSO algorithm features high quality of searching in the near-optimum. At the same time, in RFCM, the concept of lower and upper approximation can deal with uncertainty, vagueness and indiscernibility in cluster relations while the membership function in a fuzzy set can handle overlapping partitions. To illustrate the competence of this method, a number of state-of-the-art hybrid methods (FPSO, Fuzzy-FPSO, RCM-PSO, K-means PSO) are compared through application on datasets obtained from the UC Irvine Machine Learning Repository. The reported results and extensive numerical analysis indicate an excellent performance on the proposed method.

4.1. Introduction

Among pattern finding methods, i.e., summarization, association and prediction etc. [17], information clustering is of the great importance and popularity both in research and implementation. Clustering analysis is a technique aiming at grouping a set of objects, based on the similarities and dissimilarities between the data objects. Clustering can be processed in a supervised, semi-supervised and unsupervised manner and consequently it has received considerable amount of attentions from researchers.

However, the exact number of natural groups in the data is sensitive to outliers and local maxima or minima, algorithmic complexity, and degeneracy [11], etc., are the sorts of issues that cause bottlenecks in the performance of a particular clustering technique [3]. To tackle these problems, nowadays, an amount of approaches and diverse cross-discipline theories are being proposed. Specifically, optimization algorithms are increasingly hybridized with information clustering algorithms.

Particle swarm optimization (PSO) was first introduced in [10]. Particle optimization evolved from swarm intelligence (SI). PSO is one the optimization techniques which has been successfully applied as an approach to a range of clustering quests. It is a population-based metaheuristic algorithm that is inspired by the movement of individuals in a bird flock. PSO consists of a collection of particles, as well as rules to update the status of those particles. The process of updating is based on the history information of the individual and the behavior of its neighbor. Based on these intrinsic properties of PSO, recently hybridized clustering using PSO approaches have been widely and successfully applied in a range of different disciplines [25], i.e., image clustering [18], network clustering [1], clustering analysis [7], and clustering in bioinformatics [23].

Research shows that natural behavior of group animals can be successfully used as an inspiration to solve clustering problems in natural systems [21]. Due to its robust ability to perform a global search, approaches such as K-means, K-Harmonic mean, Fuzzy c-means, etc., can be significantly improved with the help of PSO. Ahmadyfar proposed [2] a new method combining PSO with the K-means clustering algorithm, i.e. PSO-KM. An initial process is set up by randomly choosing k centroids, and PSO operates by searching all dimensions for a global optimization. In [6], a hybrid PSO and K-means algorithm method on document clustering is presented. The initial centroids are constructed via PSO and subsequently, the K-means algorithm continues until the termination conditions are no longer satisfied. Alternatively, a faster convergent result can be produced [24] with a low computational cost, which is based on a K-Harmonic means with a PSO-based data clustering algorithm (KHM-PSO). The hybridization approaches in fuzzy clustering

problems also produce acceptable results. It is stated that [22] the clustering quality is highly correlated with the initialization of centroids in a typical fuzzy c-means (FCM) approach. Such approach is referred to as FPSO and it results in a better performance if centroids are initialized by PSO; traditional FCM can deal well with the fuzzy clustering problem. Additionally, a fuzzier, hybridized FPSO method named FCM-FPSO [12], is proposed to further reduce the minima in the objective function. Based on FPSO, this algorithm initiates an extra FCM approach to research the centroid space in order to reduce the possibilities of being trapped into local minima. In this manner it provides a better convergence. In addition to such an approach, fuzzy c-means algorithm based on Picard iteration hybridized with PSO (PPSO-FCM) is proposed [14] in order to overcome the drawbacks of the typical FCM.

The rough c-means approach has shown successful utilization in feature selection, in addition, in clustering analysis it can also provide good results. Rough set theory is pioneered and introduced by Pawlak [20]. Moreover, a method is proposed to combine Rough c-means with PSO [7], i.e. Rough-PSO. In this method, each cluster is modelled as rough set and PSO is employed to tune the threshold and the relative importance of upper and lower approximation of the rough set.

In this chapter, we propose an efficient approach hybridized with evolutional PSO and RFCM clustering method. We intend to contribute to the further development of hybrid methodology, in which a sensible integration of rough and fuzzy c means approach with particle swarm optimization algorithm is realized. In clustering problems, the principle of the membership in a fuzzy set enables efficient handling of overlapping partitions, the lower and upper sets of rough theory deal with uncertainty, vagueness and incompleteness in the class definition. At the same time, PSO has the characteristic to be reasonably accurate and able to avoid being trapped into local optima.

The remaining of this chapter is organized in the following manner: Section 4.2 gives a primary overview of RFCM and PSO, respectively. The proposed rough fuzzy c-means hybridized with PSO method is illustrated in Section 4.3. Section 4.4 elaborates experimental results, and we conclude the chapter in Section 4.5.

4.2. Primary Theory Bound

A. Rough Fuzzy C-means Algorithm

The idea of dealing with uncertainty information in a dataset has led to a combination of employing both fuzzy set and rough set theory. These hybridized algorithms referred as rough fuzzy c-means (RFCM), [15], [16] and [13], have been

widely and frequently used in real life data clustering problems. In this manner, RFCM algorithm is elaborated as follows.

First, fuzzy c-means is used as an partition-based algorithm that clusters a set of *n* objects $\{x_1,\dots, x_j, \dots, x_n\}$ into *c* fuzzy centroids with $\{v_1,\dots, v_i, \dots, v_c\}$. The membership index assigned "fuzziness" characteristic of a set depicted as level of belonging, can be expressed as u_{ij}.

$$u_{ij} = \left(\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{\frac{2}{mf-1}} \right)^{-1} \qquad \textbf{Equation 4-1}$$

where $mf \in (1, \infty]$ is a scalar referred to as the fuzzifier for FCM algorithm and d_{ij} is the distance from object x_j to the cluster centroid v_i.

Taking the advantage of FCM, the boundary domain of a cluster is roughened through incorporation with the approximation sets. The sets are characterized by the lower and upper approximations $\underline{\boldsymbol{R}}(\boldsymbol{X})$ and $\overline{\boldsymbol{R}}(\boldsymbol{X})$, respectively, with the following properties: (i) an object $\boldsymbol{x}$ can be part of at most one lower approximation; (ii) if $\boldsymbol{x}$ is not a part of any lower approximation, then it belongs to two or more upper approximations; and (iii) if $\boldsymbol{x} \subseteq \underline{\boldsymbol{R}}(\boldsymbol{X})$ of class *X*, then simultaneously $\boldsymbol{x} \subseteq \overline{\boldsymbol{R}}(\boldsymbol{X})$. Based on the defined approximations, the R-positive and R-boundary are defined:

$$\begin{cases} \boldsymbol{Positive_R}(\textbf{in short } P_R) = \underline{\boldsymbol{R}}(\boldsymbol{X}) \\ \boldsymbol{Boundary_R}(\textbf{in short } B_R) = \overline{\boldsymbol{R}}(\boldsymbol{X}) - \underline{\boldsymbol{R}}(\boldsymbol{X}) \end{cases} \qquad \textbf{Equation 4-2}$$

Consequently, the objective function of RFCM needs to be minimized and subsequently broken into three conditional equations [15]:

$$J = \begin{cases} \omega \times \left(\sum_{i=1}^{c} \sum_{x_j \in P_\delta(v_i)} d_{ij}^2 \right) + \tilde{\omega} \times \left(\sum_{i=1}^{c} \sum_{x_j \in B_\delta(v_i)} u_{ij}^{mf} d_{ij}^2 \right), & if\ P_\delta(v_i) \neq \emptyset,\ B_\delta(v_i) \neq \emptyset \\ \sum_{i=1}^{c} \sum_{x_j \in P_\delta(v_i)} d_{ij}^2, & if\ P_\delta(v_i) \neq \emptyset,\ B_\delta(v_i) = \emptyset \\ \sum_{i=1}^{c} \sum_{x_j \in B_\delta(v_i)} u_{ij}^{mf} d_{ij}^2, & if\ P_\delta(v_i) = \emptyset,\ B_\delta(v_i) \neq \emptyset \end{cases}$$

Equation 4-3

where the parameters ω and $\tilde{\omega} = (1 - \omega)$ are the weighting factors that are tuned to balance the relative importance between the crisp region and fuzzy boundary. Since objects lying in a lower set denote definite belongings, and will be assigned with a higher weight ω compared with $\tilde{\omega}$ of objects lying in a boundary set. In RFCM algorithm, each cluster is characterized by its own boundary set and lower

approximation, which influences the fuzziness of the final partition. Therefore, the values of the weighting factors are given by [0, 1]. In one cluster, all data are grouped into either lower approximation set or boundary set via a selected attribute δ, which is practically defined as:

$$\delta = \frac{1}{n}\sum_{j=1}^{n}(u_{vh|j} - u_{sh|j}) \qquad \textbf{Equation 4-4}$$

here n is the total number of objects, $u_{vh|j}$ and $u_{sh|j}$ are the highest and second highest membership indexes of object x_j. The meaning of δ is to determine in a degree if one object is "close" enough to the center it belongs to. Therefore, a good clustering procedure should have a value of δ as high as possible. According to the definitions of lower approximation and boundary set, and based on the predefined attribute δ, one object x_j can be characterized as:

$$x_j \begin{cases} \begin{cases} \in P_\delta(v_{vh}) \\ \notin P_\delta(v_{sh}) \\ \notin B_\delta(v_{vh}) \end{cases}, & \delta < u_{vh|j} - u_{sh|j} \\ \begin{cases} \in B_\delta(v_{vh}) \\ \in B_\delta(v_{sh}) \end{cases}, & \delta \geq u_{vh|j} - u_{sh|j} \end{cases} \qquad \textbf{Equation 4-5}$$

When $\delta < u_{vh|j} - u_{sh|j}$, and $x_j \in P_\delta(v_{vh})$, then the impacts of the objects in lower approximation of one cluster should be independent of in-between clusters and centroids, and should have similar influence on with-in cluster and centroid. Otherwise$x_j \in B_\delta(v_{vh})$, the objects belonging to the boundary set in one cluster can also have a different influence on the other clusters and centroids. Therefore, in the RFCM algorithm, the membership index of an object belonging to the lower approximation has to be reset as $u_{ij} = 1$; while the object belonging to its corresponding boundary set will remain u_{ij} (according to Equation 4-1) as in FCM. The new i^{th} centroid is modified using equation 4-6, which also considers the effect of the lower and upper bounds, as well as the fuzzy membership index. In this manner, the extended RFCM algorithm is obtained via:

$$v_i = \begin{cases} \omega \times \left(\frac{1}{|P_\delta(v_i)|}\sum_{x_j \in P_\delta(v_i)} x_j\right) + \widetilde{\omega} \times \left(\frac{\sum_{x_j \in B_\delta(v_i)} u_{ij}^{mf} x_j}{\sum_{x_j \in B_\delta(v_i)} u_{ij}^{mf}}\right), & if\ P_\delta(v_i) \neq \emptyset,\ B_\delta(v_i) \neq \emptyset \\ \frac{1}{|P_\delta(v_i)|}\sum_{x_j \in P_\delta(v_i)} x_j & ,if\ P_\delta(v_i) \neq \emptyset,\ B_\delta(v_i) = \emptyset \\ \frac{\sum_{x_j \in B_\delta(v_i)} u_{ij}^{mf} x_j}{\sum_{x_j \in B_\delta(v_i)} u_{ij}^{mf}} & ,if\ P_\delta(v_i) = \emptyset,\ B_\delta(v_i) \neq \emptyset \end{cases}$$

Equation 4-6

where $|\cdot|$ represents cardinality operator, and the cluster centroid v_i is calculated by the RFCM procedure.

B. Particle Swarm Optimization Prototype

Particle swarm optimization is a population and generation based algorithm modelled after the movements in a "bird flock" and/or a school of fish. Sharing of experience and information of each individual that takes place during stochastic optimization in PSO procedure. Every individual (particle) in the population (swarm) of one generation is assumed to "fly", in order to gain its own best fitness according to its neighboring individuals and prior knowledge of its former history. In this manner, the PSO algorithm maintains a swarm of candidate solutions of the optimization problem, while, each candidate solution is regarded as a particle.

When particles are flying through search space, their positions adjusted that governed by the distance from their own personal best position, as well as the global best position of the swarm. For a swarm of *n* particles with D-dimension vectors, i^{th} particle ($part_i$) contains the following information (notations):

- $pos_i = (pos_{i1}, pos_{i2}, \ \dots, pos_{iD})$, the current position of the i^{th} particle;
- $vel_i = (vel_{i1}, vel_{i2}, \ \dots, vel_{iD})$, the current velocity (change of position) of the i^{th} particle;
- $p_i = (p_{i1}, p_{i2}, \ \dots, p_{iD})$, the best previous position of the i^{th} particle;
- p_g, the best position of a swarm, and $t = (1, 2, \dots, G)$, the current generation.

Every particle in a swarm is manipulated via the following updating equation:

$$vel_{id}(t+1) = \alpha \cdot vel_{id}(t) + \beta_1 r_1[p_{id}(t) - pos_{id}(t)] + \beta_2 r_2\left[p_{gd}(t) - pos_{id}(t)\right]$$

Equation 4-7

$$x_{id}(t+1) = x_{id}(t) + vel_{id}(t+1)$$

Equation 4-8

where $i = 1, 2, \dots, n$ and $d = 1, 2, \dots, D$. In Equation 4-7, α is the positive inertia weight, β_1 and β_2 are the acceleration constants, meaning the correlation between social and individual behavior, and r_1, r_2 are the displacement deviators in the range [0, 1]. $p_{id}(t) - pos_{id}(t)$ is the personal influence, and $p_{gd}(t) - pos_{id}(t)$ is the social influence on the global experience. At present, research on this simple PSO concept is still being performed. Its success is given by the few parameters that are

required for the specification of the problem, i.e. dimensionality of the data space and few weighted factors for control of the convergence.

4.3. Rough Fuzzy C-Means and Particle Swarm Optimization Hybridized Method (RFM-PSO)

Taking both RFCM clustering and the intrinsic properties of PSO into account, we propose an efficient model-combined algorithm, namely RFCM-PSO. In RFCM algorithm, each centroid is considered a vector that updates according to an iterative operation. A representation of the centroid vectors therefore, can refer as elements in particles. In other words, the i^{th} particle $(part_i)$ can be defined as $part_i = (v_1, v_2, \ldots, v_i \ldots, v_c)$, where v_i, $i = 1,2, \ldots, c$ is the cluster center. Consequently, a swarm in PSO represents an amount of candidate solutions of centroids in RFCM algorithm. Thus a fuzzy membership function and roughness definitions are assigned on every single object for its clustering decision making. For each iteration in the RFCM-PSO procedure, the centroids in clusters change and their positions are updated based on the particles. Several extra notations (cf. Section 4-2.B) for RFCM-PSO need to be considered before employing this algorithm:

- n, number of objects;
- c, number of pre-defined centroids;
- v_i, vectors of centroids containing $pos_i(t)$;
- $pos_i(t)$, the current position of the i^{th} particle at generation t; and
- $u_{ij,k}(t)$, the RFCM membership index of the i^{th} object with respect to the j^{th} cluster of the k^{th} particle at generation t it belongs to.

Due to the fast convergence and tenable setup of membership index, we suggest an improvement of the performance of PSO searching algorithm, is to initialize the swarm with FCM. The fit, or in other words the objective function, is then measured and minimized by Equation 4-3.

The approximation optimization of RFCM is based on Picard iteration through Equation 4-1 and Equation 4-6. The process calls the training of the RFCM parameters which starts by randomly choosing centroids and initiating membership in FCM. Subsequently, it progresses in approximation evaluation for modifying u_{ij} parameter. With a pre-set number of particles, the resulting centroids from RFCM are represented by particles that are given as inputs to optimization procedure of PSO. The best solution, i.e., global optimum, is looked for by a stochastic search from solution space of candidates.

In the proposed method, PSO performs as a standard optimizer in FES/per iteration, where FES represents the maximum amount of function evaluations allowed. Thus, time complexity cost of RFCM-PSO tends to be determined by the cost function in RFCM, which is $O(n^2)$. Furthermore, the implementation of the RFCM – PSO method is described in the pseudo-code as:

Schema 1 *Rough fuzzy c-means and PSO hybrid algorithm*:

Input: fuzzifer *mf*, weighting factor ω, cluster number *c*, α, β_1, β_2

Given: integral generation $t \in (1, \infty]$.

Initializing: stochastic centroid v_i, membership matrix$u_{ij,k}$, *vel* velocity, pos position of particles at generation t=1.

for each *t* generation **do**

training RFCM parameter:

Compute the norm distance d_{ij} for each *n* objects and *c* clusters.

if δ check **then**

Reset $u_{ij,k}(t)$.

end if

Update new centroid as v_i (*t+1*) per equation 4-6.

Update $u_{ij,k}$ to (*t+1*) via equation 4-1.

Optimization procedure:

Training the personal best and global best position, $p_i$andp_g.

Update pos_i (*t+1*) and vel_i (*t+1*) for each particle using equation 4-7 and 4-8.

Convergence check; break

end for

4.4. Experimental Results

The main objective of this section is to assess relative performance of clustering technique hybridized with particle swarm optimization algorithm. The algorithms that are compared with proposed method are: k-means PSO (K-PSO) [2], fuzzy c-means PSO (FPSO) [22], fuzzy c-means and fuzzy PSO (FCM-FPSO) [12] and rough c-means PSO (RPSO) [7]. All the methods are coded and implemented in the *Matlab* 2014a environment running on an Intel (R) Core (TM) i7-3770 (CPU 3.4GHz, 16GB RAM) machine. In practice, our input parameters produce with higher performance compared to other settings. We kept the input parameters

constant across all runs (cf. Table 4-1). To analyze the clustering performance of our method, two indices are introduced in the next subsection.

Table 4-1: RFCM-PSO parameter settings

Parameter settings				
Clustering	*mf* =2		ω=0.95	
Optimizing	$\alpha \in [0.1, 0.9]$	$\beta_1 = \beta_2 = \sqrt{2}$	Population=10	Generation=50

A. Quantitative Measurement

The problem of validation in a clustering algorithm is an important consideration since all of its applications have their own sets of partially successful validation scheme. None of any separate index can comprehensively depict the performance of these clustering algorithms [4] of unlabelled data. After conducting a study in several indexes that are used for performance validation, we propose:

Davies-Bouldin Index: Introduced in [8] is:

$$DB = \frac{1}{c}\sum_{j=1}^{c} max_{j, j\neq k}\{\frac{S(v_j)+S(v_k)}{d(v_j, v_k)}\} \qquad \textbf{Equation 4-9}$$

Dunn's Index: Given by [9] is:

$$Dunn = min_j\left\{min_{k,\ k\neq j}\left\{\frac{d(v_j, v_k)}{max_i\{\Delta(v_i)\}}\right\}\right\} \qquad \textbf{Equation 4-10}$$

Validation standard build: the higher the similarities in within-cluster and dissimilarities in between-cluster, the lower the DB value will have; the well-separated the clusters are, the larger the Dunn index will obtain.

B. Validation of Clustering Algorithm

The PSO-combined algorithms have been applied on several bench mark datasets obtained from UCI repository, which cover a range of different type of problems in information science.

Five algorithms are implemented and applied on these datasets (i.e. Table 2), and the quality of each algorithm is investigated. The particular test dataset is Iris, with different pre-set cluster numbers, namely cluster = 2 and cluster = 3. The Iris dataset represents a four-dimensional structure that contains 50 samples in each of the three flower categories. One of the three clusters, *Iris setosa*, is well separated with the

other two, while there are some overlaps within the *Iris sirginica* cluster and the *Iris sersicolor* cluster. We have setup a separate test of the different partition strategies.

Table 4-2: Attribute of selected datasets

Dataset	Feature	Instance	category
Iris	4	150	3
Glass	9	214	7
CMC	9	1473	3
Wine	13	178	3
WBCD	30	569	2

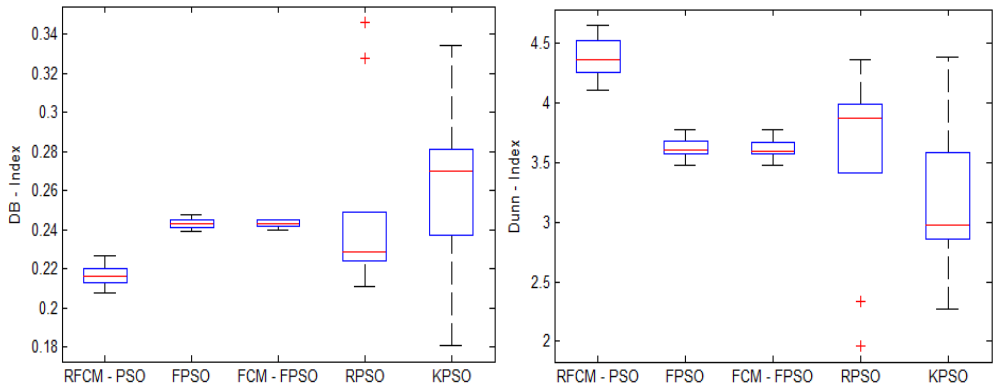
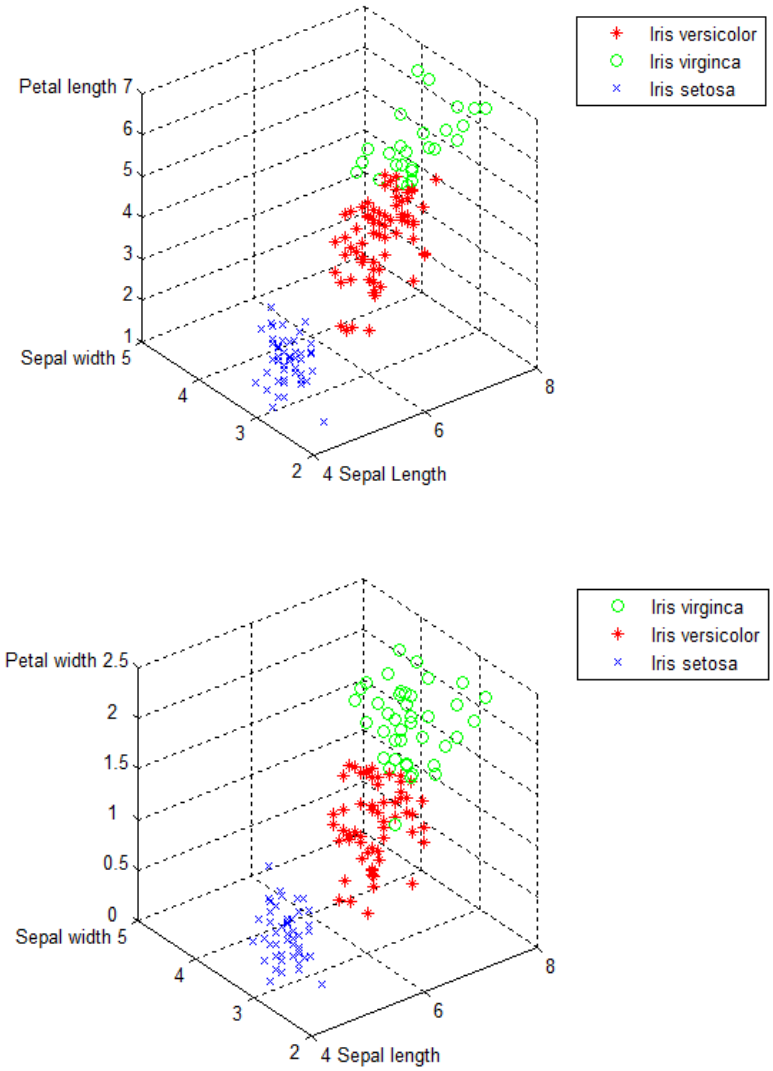


Figure 4-1. Box – plot of investigated algorithm on Iris dataset (left: DB Index and, right: Dunn Index, Cluster = 3.

Performances of different algorithms are depicted in Figure 4-1. This shows that RFCM-PSO has better results by having the lowest DB index and the highest Dunn index in case of Cluster = 2 and Cluster = 3. An evident difference of the Dunn value occurred in case Cluster = 3. This which is a result of the fact that our method outperforms the others while dealing with overlapped clustering problem. The likely range of variation is coherent and acceptable compared to the four clustering methods in our evaluation. Additionally, the interquartile ranges (IQR) of FPSO and FCM-FPSO are smaller in the relative sense compared with RFCM-PSO. This is because in fuzzy c-means, the membership of an individual is inversely related to the relative distance from every centroid, thus tenable results of FCM-/FPSO can be obtained in a dataset of low dimensions. Nonetheless, it is very sensitive to noise

and outliers and it will easily fall into local optima when confronted with dataset of higher dimension.

In the Iris dataset, there are two overlapping clusters of the three clusters in a total. This may sometimes result in a clustering of just two clusters. An efficient classifier (the clustering algorithm in unsupervised learning), however, should be able to identify the boundless and vague features classes. As an example, in Figure 4-2, it is shown as scatterplots depicting the different views of feature. It is observed that the three different *Iris* species, through inspection of the flowers can be well categorized using sepal width, sepal length, petal width and petal length.



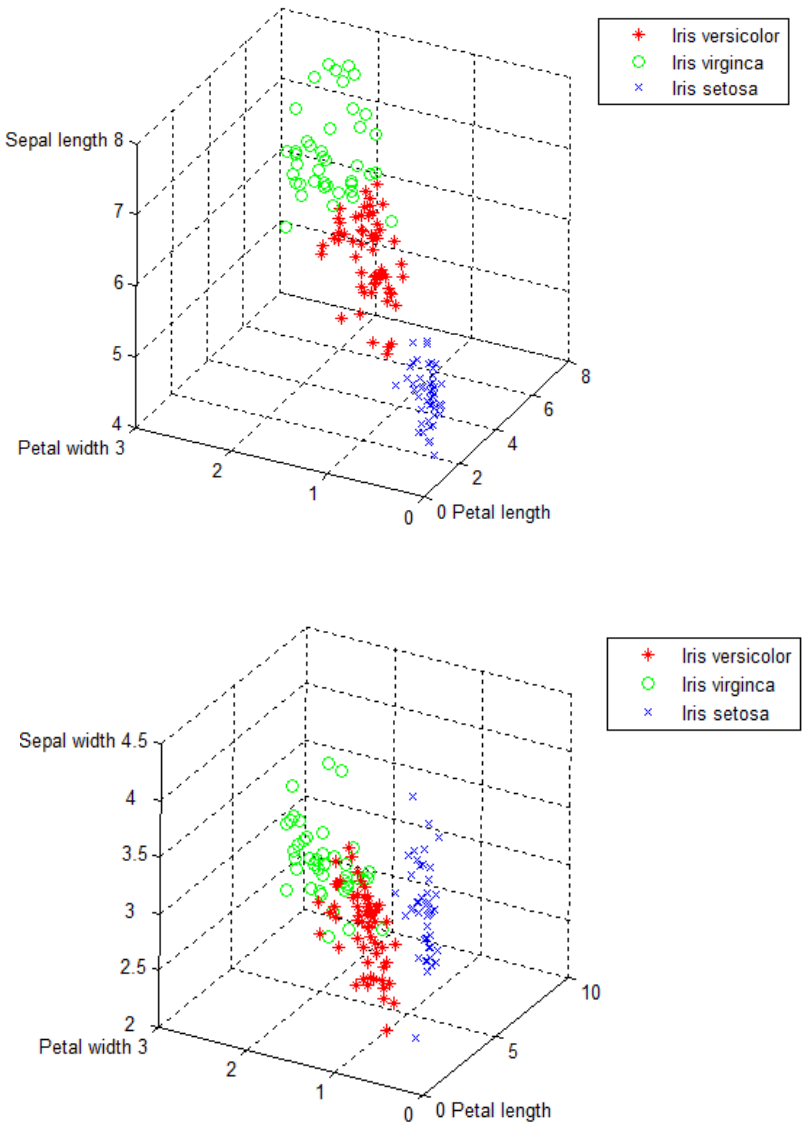
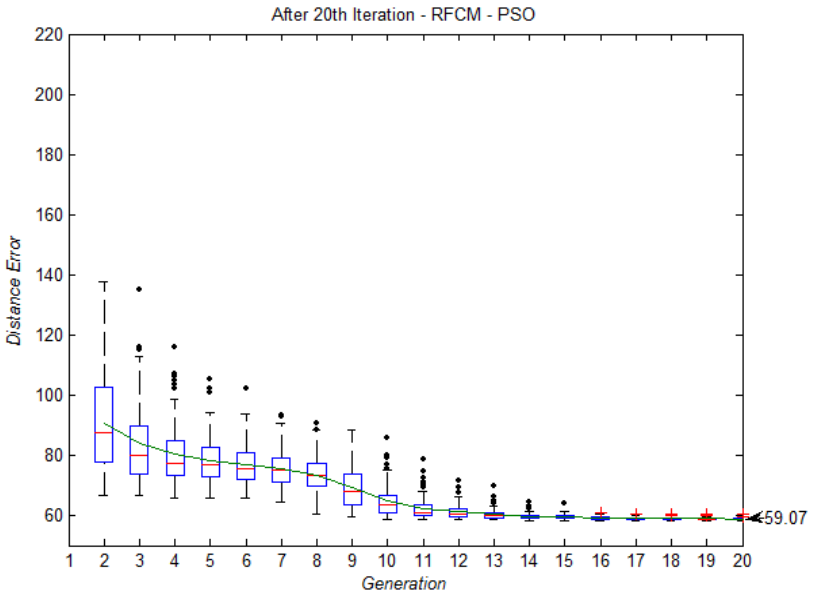
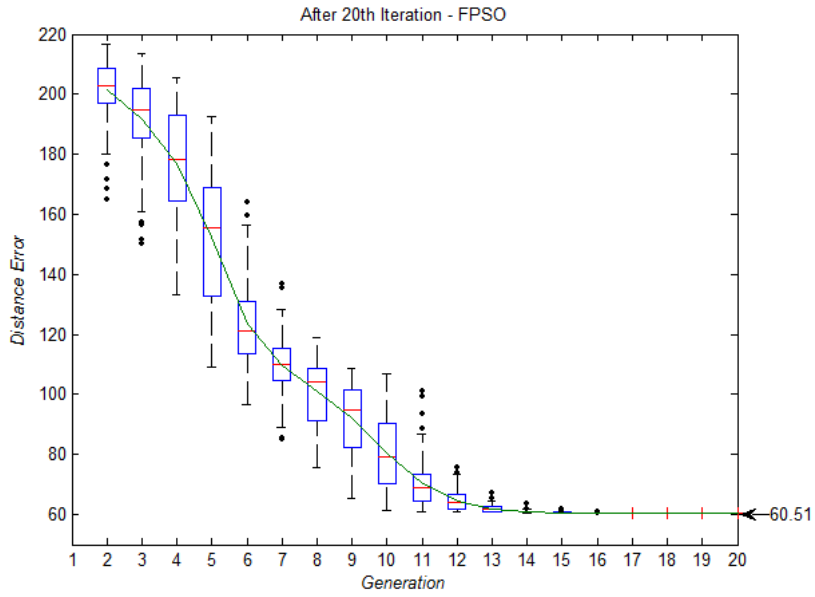


Figure 4-2. Scatter plot result of RFCM – PSO on Iris (Cluster = 3, feature = 4).



(a)



(b)

Figure 4-3. The clustering performance in terms of Distance Error (DE). When convergent condition met, the DE value for RFCM – PSO (a) and FPSO (b) is 59.07 and 60.51 respectively.

Figure 4-3 shows an example of the performance on RFCM – PSO and FPSO in the Iris dataset by minimizing the distance error of all the contained objects, considering cluster = 3. For all algorithms 100 independent runs per generation have been performed. Recorded in every generation steps, the distance error shows the convergence of particles in a single swarm. The distance error (DE) is calculated by the mean distance deviation of every single object to the centroid it belongs to after clustering. Depicted in Figure 4-3, the proposed RFCM – PSO outperforms the prevalent FPSO in terms of the smaller mean DE error in every generation, lower IQR, faster convergence speed and less outliers.

A well performed clustering algorithm does not only support on its property of anti-noise or the resultants of less outlier in clusters, but also on its capability of sampling scale-invariance. Applications of most clustering algorithms provide plausible results only on low dimensionality and small population dataset. The handling with sparse and skewed distributions of the samples in a certain clustering space remains a challenge. When sampling scale in a research population, is relatively small then, the higher the dimension of the attributes, the less accuracy and efficiency of the clustering algorithm will perform. Given the definition of DB Index and Dunn Index, the value of both DB and Dunn should be invariant in spite of a change in the of sampling scale since they have the same overall population. In other words, when different selection of a subset of individuals from within one same research population takes place, as the estimation of the performance of the clustering algorithm, DB and Dunn Index should produce stable results.

The CMC dataset (cf. Table 4-2) is employed to test the capacity of scale-invariance of each algorithm. We utilize five different scales in the sampling population, i.e. of 300, 600, 900, 1200 and the full population of 1473 instances. To assess a valid estimation of median and to derive acceptable standard errors from a complex and high-dimensional population, the bootstrapped sampling approach is being used. For each different scale, 100 bootstrap runs have been independently applied. The results for each run are summarized and calculated in terms of their max- and minimum, average value and standard deviation. From Table 4-3, one can be seen that the smallest standard deviation of DB and Dunn values are observed on the proposed RFCM – PSO. This result draws a conclusion that the proposed method has acceptable and steady clustering results when sampling scale are differentiated, although skewed and sparse distribution of sample instances are encountered.

In Table 4-4, the performance of the different PSO hybridized clustering algorithms on the selected benchmark datasets are compared in terms of DB and Dunn index. For all five benchmark sets, every separate algorithm is applied and the value of DB and Dunn are computed respectively. Since the KPSO algorithm produces non-convergent results in the DB and Dunn value of the Glass and Wine datasets, thus

these have not been included in Table 4-4. The results reported here, however convincingly confirm that the proposed method conducts more promising compared to the recognized methods.

Table 4-3: Scale invariant evaluation results on IRIS dataset

Algorithm	DB_{max}	DB_{ave}	DB_{std}	$Dunn_{min}$	$Dunn_{ave}$	$Dunn_{std}$
KPSO	0.1457	0.1027	0.2835	13.4826	14.3864	0.9336
FPSO	0.0594	0.0958	0.0820	20.4155	20.6475	0.1453
FCM - FPSO	0.0595	0.0592	0.0817	20.3119	20.5035	0.1409
RPSO	0.0590	0.0586	0.0789	19.4677	19.9972	0.3349
RFCM - PSO	**0.0547**	**0.0542**	**0.0757**	**21.3450**	**21.4935**	**0.0867**

Table 4-4: Performance evaluation with different dataset (average)

Dataset	KPSO		FPSO		FCM - FPSO		RPSO		RFCM - PSO	
	DB	Dunn	DB	Dunn	DB	Dunn	DB	Dunn	DB	Dunn
Iris	0.241	3.130	0.248	3.567	0.245	3.568	0.249	3.556	**0.216**	**4.382**
Glass	-	-	0.648	0.125	0.644	0.128	0.458	0.197	**0.441**	**0.238**
CMC	0.0732	15.405	0.0594	20.615	0.0592	20.616	0.058	20.007	**0.0544**	**21.494**
Wine	-	-	0.00129	472.4	0.00129	474.4	0.00140	411.8	**0.00124**	**483.9**
WBCD	0.0127	18.8	0.00476	297.6	0.00476	297.9	0.00479	267.3	**0.00475**	**301.2**

4.5. Conclusion

In this chapter, we have briefly discussed the evolution of clustering techniques based on Particle Swarm Optimization. A literature survey revealed that there is an enormous increase in the popularity of PSO based clustering techniques. In a short review the rough and fuzzy clustering technique is introduced. Thereafter we present a novel and efficient hybrid method, namely the Rough Fuzzy C-means and PSO (RFCM - PSO) clustering approach. The performance of proposed method is compared with the K- means PSO (KPSO), Fuzzy PSO (FPSO), Fuzzy C- means FPSO (FCM - FPSO) and Rough PSO (RPSO) algorithm. The reported results show

that our approach outperforms the rest of the methods in terms of its efficiency, reliability and solution quality based on geometrical DB and Dunn Index.

The contribution of this chapter is in the development of a hybridized methodology, which carefully integrates rough and fuzzy c-means approach and the particle swarm optimization algorithm. In a clustering problem, the membership of fuzzy set enables efficient handling of overlapping partitions, the lower and upper sets of rough theory deal with uncertainty, vagueness and incompleteness in class definition; while PSO has a tenable quality to be more accurate in searching a best solution from candidate sets as well as avoiding being trapped into local optima.

4.6. References

[1] Achankunju, M., Pushpalakshmi, R., Kumar, A. V. A., 2013. Particle swarm optimization based secure QoS clustering for mobile ad hoc network. In Communications and Signal Processing (ICCSP), 2013 International Conference on (pp. 315–320). doi:10.1109/iccsp.2013.6577066

[2] Ahmadyfard, A., Modares, H., 2008. Combining PSO and k-means to enhance data clustering. In 2008 International Symposium on Telecommunications, IST 2008 (pp. 688–691). doi:10.1109/ISTEL.2008. 4651388.

[3] Alam, S., Dobbie, G., Koh, Y. S., Riddle, P., & Ur Rehman, S., 2014. Research on particle swarm optimization based clustering: A systematic review of literature and techniques. Swarm and Evolutionary Computation, 17, 1–13. doi:10.1016/j.swevo.2014.02.001.

[4] Bezdek, J. C., Pal, N. R., 1998. Some new indexes of cluster validity. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 28, 301–315. doi:10.1109/3477.678624.

[5] Chen, C. Y., Ye, F., 2004. Particle swarm optimization algorithm and its application to clustering analysis. In 2004 IEEE International Conference on Networking, Sensing and Control (pp. 789 – 794 Vol.2).

[6] Cui, X., Potok, T. E., Palathingal, P., 2005. Document clustering using particle swarm optimization. In Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE (pp. 185–191). doi:10.1109/SIS.2005.1501621.

[7] Das, S., Abraham, A., Sarkar, S. K., 2006. A Hybrid Rough Set--Particle Swarm Algorithm for Image Pixel Classification. 2006 Sixth International Conference on Hybrid Intelligent Systems (HIS'06). doi:10.1109/HIS. 2006.264909.

[8] Davies, D. L., Bouldin, D. W., 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1, 224–227. doi:10.1109/TPAMI.1979.4766909.

[9] Dunn, Joseph C. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." (1973): 32-57.

[10] Eberhart, R., Kennedy, J., 1995. A new optimizer using particle swarm theory. MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science. doi:10.1109/MHS.1995.494215.

[11] Hathaway, R. J., Bezdek, J. C., 1995. Optimization of clustering criteria by reformulation. IEEE Transactions on Fuzzy Systems, 3. doi:10.1109/ 91.388178.

[12] Izakian, Hesam, and Ajith Abraham. "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem." Expert Systems with Applications 38.3 (2011): 1835-1838.

[13] Ji, Z., Sun, Q., Xia, Y., Chen, Q., Xia, D., Feng, D., 2012. Generalized rough fuzzy c-means algorithm for brain MR image segmentation. Computer Methods and Programs in Biomedicine, 108(2), 644–55. doi:10.1016/ j.cmpb.2011.10.010.

[14] Liu, H. C., Yih, J. M., Wu, D. B., Liu, S. W., 2009. Fuzzy C-mean clustering algorithms based on picard iteration and particle swarm optimization. In 2008 International Workshop on Education Technology and Training and 2008 International Workshop on Geoscience and Remote Sensing, ETT and GRS 2008 (Vol. 2, pp. 838–842). doi:10.1109/ ETTandGRS.2008.375.

[15] Maji, P., Pal, S. K., 2007. RFCM : A Hybrid Clustering Algorithm Using Rough and Fuzzy Sets, 80, 475–496.

[16] Mitra, S., Barman, B., 2008. Rough-Fuzzy Clustering : An Application to, 300–307.

[17] Mitra, S., Pal, S. K., Mitra, P., 2002. Data mining in soft computing framework: A survey. IEEE Transactions on Neural Networks, 13, 3–14. doi:10.1109/72.977258.

[18] Omran, M. G. H., 2004. Particle Swarm Optimization Methods for Pattern Recognition and Image Processing by Doctor. PhD Thesis.

[19] Pal, S. K., Ghosh, A., Shankar, B. U., 2000. Segmentation of remotely sensed images with fuzzy thresholding, and quantitative evaluation. International Journal of Remote Sensing. doi:10.1080/01431160050029567.

[20] Pawlak, Z., 1982. Rough Sets. International Journal of Computer and Information Science, 11, 341–356. doi:10.1109/TITB.2009.2017017.

[21] Rana, S., Jasola, S., Kumar, R., 2011. A review on particle swarm optimization algorithms and their applications to data clustering. Artificial Intelligence Review, 35, 211–222. doi:10.1007/s10462-010-9191-9.

[22] Wang, L., Liu, Y., Zhao, X., Xu, Y., 2006. Particle Swarm Optimization for Fuzzy c-Means Clustering. In The Sixth World Congress on Intelligent Control and Automation, 2006. WCICA 2006 (pp. 6055–6058).

[23] Xiao, X., Dow, E. R., Eberhart, R., Miled, Z. B., Oppelt, R. J., 2003. Gene clustering using self-organizing maps and particle swarm optimization. Parallel and Distributed Processing Symposium 2003 Proceedings International, 00, 10. doi:10.1109/IPDPS.2003.1213290.

[24] Yang, F., Sun, T., Zhang, C., 2009. An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization. Expert Systems with Applications, 36, 9847–9852. doi:10.1016/ j.eswa.2009.02.003.

[25] Kaur, Amanpreet, and M. D. Singh. "An overview of pso-based approaches in image segmentation." Int J Eng Technol 2.8 (2012): 1349-1357.

Chapter 5

A Systematic Study on One Dimensional Gel Electrophoresis Image Analysis

This chapter is based on the following publication:

Cai, F., S. Liu, P. ten. Dijke, and F. J. Verbeek. "Image Analysis and Pattern Extraction of Proteins Classes from One-Dimensional Gels Electrophoresis." J. Bioscience, Biochemistry and Bioinformatics11: 1106-1113. vol. 7, no. 4, pp. 201-212, 2017.

Chapter summary

In this chapter, we focus on estimating the practical performance of fuzzy systems on the data analysis within the scope of protein/DNA phenotypic study. In detail, we are going to address the following research questions.

1. Can the 1-dimension gel electrophoresis data be quantitatively and accurately assessed using newly developed fuzzy-logical based algorithm and fuzzy systems?

2. Can we identify the essentials of protein/DNA, and validate the results of gel electrophoresis from published reports?

Following the workflow of data analysis (cf. Figure 1-2), this chapter is divided into two major sections. First, the design of the fuzzy systems and its solutions are demonstrated. Each fuzzy-logic and unsupervised computing processing step is illustrated and their motivations behind this design are explained. In the context of the applied fuzzy systems and, heterogeneous methodologies are integrated into a global picture thereof. Second, the data from electrophoresis experiments are qualitatively and quantitatively evaluated. The variations in the bands/lanes are derived from numerical measurements. These results are then compared and discussed with other experiments as described in literature.

5.1. Introduction

Mixtures of proteins can be separated and visualized by Sodium-dodecyl sulphate (SDS)-polyacrylamide gel electrophoresis (PAGE); this is a classical tool for protein analysis [23]. Combining this analysis with Western blotting and probing, the filter with specific antibodies, or the extraction of protein from gel and mass spectrometric (MS) analysis, make it a very powerful tool to determine relative quantities and identification of proteins. In addition, prior to SDS-PAGE, proteins can be fluorescently labeled and the resulting images can be captured by a flatbed scanner equipped for fluorescence. During protein sample preparation, protease inhibitors should be taken into consideration to prevent degradation of proteins; on the gel, these proteases appear as faster running protein fragments.

A popular separation technique, capable of fast and easy analyzing less complex samples, is the high-resolution 1-dimensional (D) gel electrophoresis. Proteins, as obtained from cell lysates, are usually dissolved in a SDS containing buffer and are boiled before loading onto the polyacrylamide gel. Subsequently, on the authority of the molecular weight of proteins, they are charged by force to migrate through the gel under the influence of an electric field. Using this method employing SDS in sample buffer, there is, for most proteins a good correlation between polypeptide length and charge. The latter is running the samples under so-called denaturing conditions. On the contrast, proteins can also be separated under non-denaturing conditions (proteins are then still in their folded state); but then they are not only separated by molecular weight but also by their shape. On a gel, multiple samples are loaded along with molecular standards. The gel, referred to a matrix instance used to contain and separate target molecules, is stained (for instance, by coomassie brilliant blue or silver) and then visualized by a lightbox; alternatively, the fluorescently labeled proteins can also be visualized by a laser scanner. Afterward, the resulting gel images consists of several vertical lanes equal to the number of wells in which the protein samples were loaded; and a number of horizontal bands corresponding to proteins or fragments thereof, reflecting the amounts and characteristics of individual proteinaceous components.

The banding patterns and the relative differential intensities of the bands can be converted into graphical, numerical and tenable formats through image processing and analysis techniques. In this manner computing with intelligent techniques prevents subjective and tedious image interpretation; this otherwise may lead to reproducibility issues. With respect to the analysis of gel electrophoresis profiles, the image processing requires three main steps [27] (BBS): (1) Background correction; (2) Bands detection, matching and quantification; (3) Similarity clustering analysis.

Several software systems have been developed for the automated analysis of profile images acquired from gel electrophoresis techniques [1-4]. Some of these platforms are semi-automatic and locate 1D mean profiles on peak/minima valley as either bands or noise for bands selection, and lack quantification on the digital description of bands. Other systems identify and classify lanes and/or bands via employing simple texture features that result in an unambiguous matching and grouping [5]. Nevertheless, these approaches are unable to generally face the challenges of extracting hidden (complexity) patterns within the bands/lanes expression in gel electrophoresis images via just visual examination, or simple BBS processing steps. However, these approaches henceforth demonstrated challenges to recognize objectives in terms of lane distortion, band deformity (including doublet effect which means two or more bands are too close together) and background noise. Therefore, a systematic yet precise approach in the image processing is required. Hence, based on proposed fuzzy-logic based methodologies (cf. Chapter 2, Chapter 3 and Chapter 4), we introduce an elaborate fuzzy systems so as to improve the BBS step as follow (cf. Figure 5-1):

- Background correction using fuzzy DCBC method [6].
- Normalization and fuzzy feature extraction for lanes/bands.
- Rough fuzzy c-means and particle swarm optimization (RFCM-PSO) hybridized clustering analysis.
- Functional analysis approaches.

In brief, the work with a comprehensive way presented in this chapter contributes to extract qualitative and quantitative information from 1D gels, consisting of back-ground noise subtraction, topographical normalization of bands and lanes, phenotypical description of bands, revealing hidden patterns recognition by dealing with clustering of overlapping and indiscernible information.

The remainder of this chapter is organized as follows: in Section 5.2 we introduce the methodology including image acquisition and processing; i.e., several new innovative algorithms and analysis procedures. After image enhancement, phenotype measurements are obtained on each individual band of all different lanes. Next, the categorization of phenotypic stages using feature extraction and selection is illustrated. The best combination group of features is applied in a clustering technique to address biological questions of interest. The experimental results are presented in Section 5.3 via a case study example, and Section 5.4 concludes this chapter.

5.2. Methodology

Modern gel electrophoresis techniques allow visualizing protein level structures so that these can be specifically subject to analysis. These techniques revolutionized the field of proteomics and biomarker discovery in detecting the changes in protein expression [7]. However, a significant amount of wet laboratory expertise is still required. Application of these techniques in higher volumes is beyond the capacity of manual processing. Therefore, image processing and machine learning are invoked to help recognizing patterns and to provide an automated analysis solution for gel electrophoresis experiments. In this section, we will introduce the image acquisition protocol followed by approaches for image and data analysis.

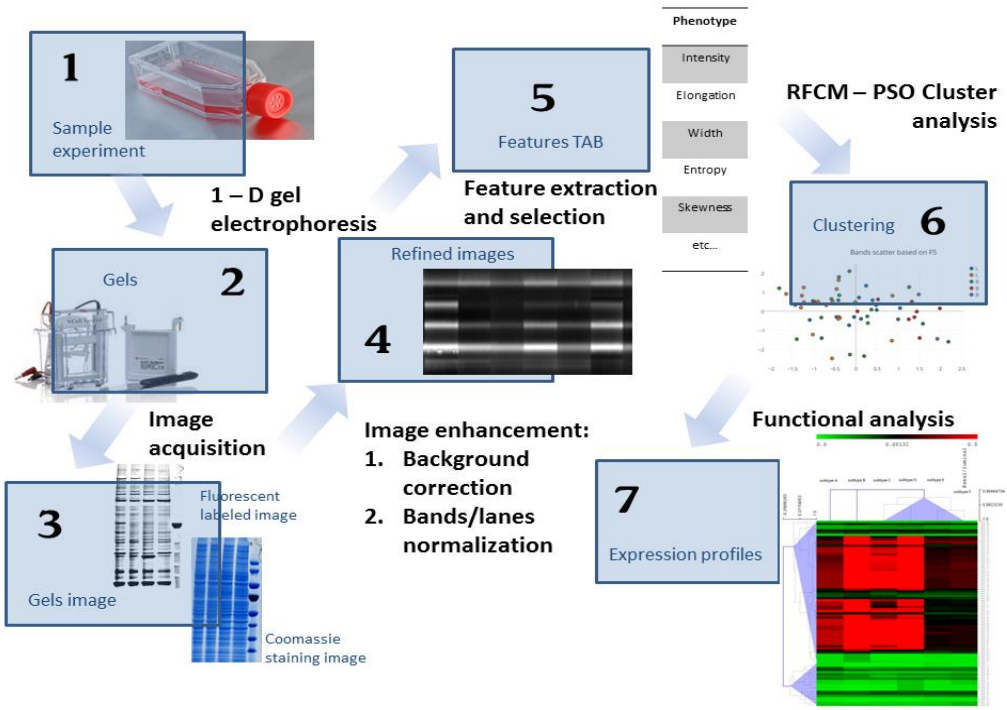


Figure 5-1. Workflow of our 1-D electrophoresis gel image analysis system

A. Image Acquisition

Sample preparation precedes image acquisition (cf. Figure 5-1) and the image of the gel should reflect the differences in composition of the different samples, i.e. each sample represents a particular cell culture. To that end, the samples are labeled with a fluorescently tagged probe. Subsequently, the pre-cast gel (e.g. NuPAGE Bis-Tris) is put into the electrophoresis system (e.g. xCell SureLock Mini-Cell) and in each

well of the gel a sample is loaded. Under an electric current (164-5050 PowerPac) samples migrate over the gel in a linear trajectory and proteins with different molecular weights produce separate bands. Once the current is stopped and the gel is fixed, the result is captured with an imager for the fluorescent signals; i.e. the Typhoon 9410 Gel and Blot Imager as used in this study. Alternatively, after coomassie staining, the gel can be scanned using a lightbox and photo-scanner, i.e. the Microtek ArtixScan F2 scanner. In our experiments described in this chapter, we used images containing fluorescent signals with a spatial resolution of 10 line-pairs/mm, and with a pixel size of 10 μm that produced an image size of 4096x1024 pixels with a 16 bit dynamic range. An example of acquiring images, accompanied with labels and markers is shown in Figure 5-2 and Figure 5-3 (after selection of region of interest) respectively.

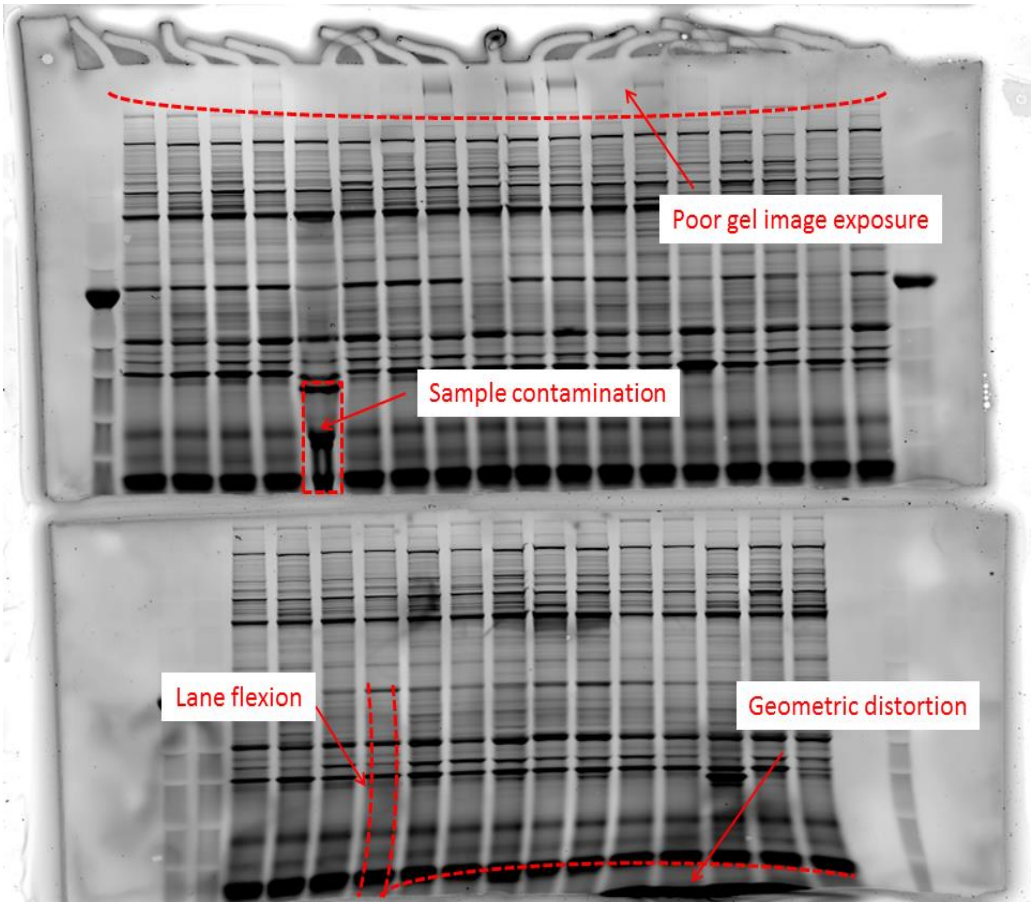
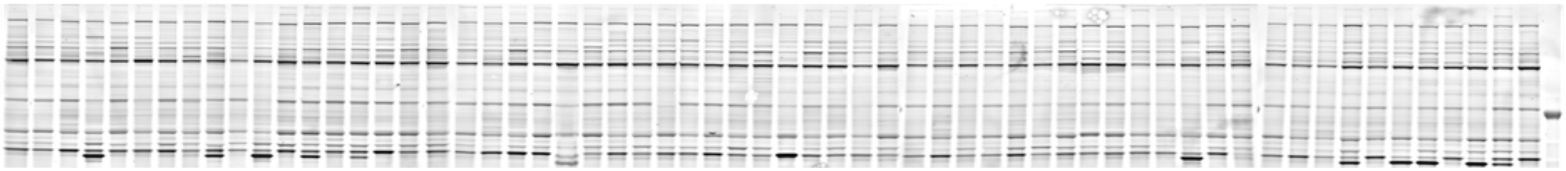


Figure 5-2. A sample of acquiring raw gel electrophoresis images that reveals common challenges for imaging processing, including geometric distortion, cell line flexion, low contrast and noisy illumination.



#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62
LINE	SUM190PT	CAMA-1	MDA-MB-231	HCC1806	SUM229PE	BT20	HCC70	HCC1143	HCC1937	MDA-MB-468	HCC1395	SUM149PT	HCC1569	HCC1954	HCC202	EVSA-T	MDA-MB-453	SK-BR-3	CAMA-1	MDA-MB-231	SUM225CWN	SK-BR-5	OCUB-F	UACC893	HCC1419	MPE600	ZR75-30	MDA-MB-330	HCC2218	BT474	MDA-MB-361	UACC812	SUM185PE	MDA-MB-134VI	MDA-MB-175VII	SUM52PE	CAMA-1	MDA-MB-231	BT483	CAMA-1	MDA-MB-415	MCF-7	SUM44PE	HCC1500	ZR75-1	T47D	MDA-MB-435s	SK-BR-7	SUM159PT	SUM1315MO2	CAMA-1	MDA-MB-231	Hs578T	MDA-MB-157	MDA-MB-231	MDA-MB-436	HCC38	SUM102PT	HCC1187	HCC1599	OCUB-M	
TYPE	LUMINAL	LUMINAL	BASALB	BASAL	BASAL	BASALA	BASALA	BASALA	BASALA	BASALA	BASALB	BASALB	BASALA	BASALA	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	BASALB	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	BASALB	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	LUMINAL	BASAL	BASAL	BASAL	BASAL	LUMINAL	BASALB	BASALB	BASALB	BASALB	BASALB	BASALB	BASALB	BASALA	BASALA	
subtype Riaz et al.	luminal-erbb2+	luminal	normal-like/claudin-low	basal-like	basal-like	basal-like	basal-like	basal-like	basal-like	basal-like	basal-like	basal-like	E Rneg-erbb2+	E Rneg-erbb2+	luminal-erbb2+	luminal-erbb2+	luminal-erbb2+	luminal-erbb2+	luminal	normal-like/claudin-low	luminal-erbb2+	luminal	luminal	luminal-erbb2+	luminal-erbb2+	luminal	luminal	luminal-erbb2+	luminal	luminal	luminal	luminal	luminal	luminal	luminal	luminal	normal-like/claudin-low	luminal	luminal	luminal	luminal	luminal	luminal	E Rneg-erbb2+	luminal	luminal	normal-like/claudin-low	luminal	normal-like/claudin-low	normal-like/claudin-low	normal-like/claudin-low	normal-like/claudin-low	normal-like/claudin-low	normal-like/claudin-low	normal-like/claudin-low	normal-like/claudin-low	normal-like/claudin-low	normal-like/claudin-low			MARKER
ER	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+	+	+	+	+	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	
PgR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
ERBB2	++	-	-	-	-	-	-	-	-	-	-	-	++	++	++	++	++	++	-	-	++	++	++	++	++	++	++	++	++	++	++	++	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
EGFR	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-	-	+	-	-	-	-		-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	+	+	+	+	-	+	+	+	+	+							
CK5	-	-	-	+	+	+	+	+	+	+	+	+	-	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-							

Figure 5-3. Image data acquired from 1 Dimensional Gel Electrophoresis and its relevant marker. The number in each column represents the imaging cell line, and w.r.t its essentials (including name of cell line, cell type, reactor, etc.) in row.

B. Image Enhancement

A problem that arises in gel electrophoresis imaging is the introduction of information that is not part of the original signal. This part of the information should be considered as noise and outlier. Images acquired from the optical detection system may inevitably suffer from the various sources of systematic and experimental variation, through which the "true" information is masked. Hence, a cropping of regions of interest, background correction and data normalization are required.

Region cropping. Before loading images into preprocessing track, bands of interest are firstly resolved in a region that this part of image can be isolated for further analysis by manually cropping and adhering (cf. Figure 5-3 from Figure 5-2, this processing is typically operated by biologists). Cropping process should also be performed to remove regions of gel that show sample contamination and extreme distortion of cell line which could interfere with bands detection.

Background correction. The adjustment or removal of the background signal should be performed to accurately quantify the fragments present in the gel image; i.e. the true signal. Approaches that have been suggested for such background correction, include global minimum subtraction from time domain, signal-pass filtering in wavelet domain or frequency domain and processing using mathematical morphology. A more comprehensive discussion on an amount of approaches can be found in [8] [9].

In our system, an estimate from the background illumination is produced by a morphological subtraction of the gel image. To that end, the fuzzy and rough concepts are employed [6] to an improved dam-based rolling ball method. In this manner the mutual information shared by foreground and background is balanced for an ideal image $f(x, y)$, the background correction procedure can be described as Equation 5-2 and can also be recalled from Chapter 2.

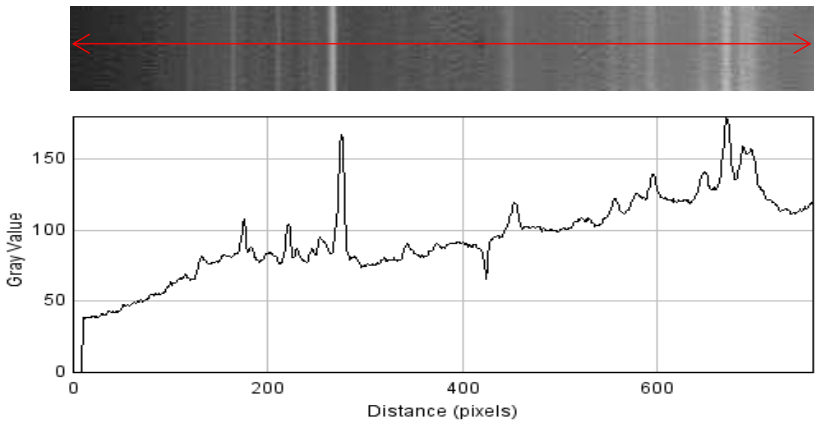
Raw image estimation:

$$\tilde{f}(x,y) = f(x,y) + Background(x,y) \qquad \textbf{Equation 5-1}$$

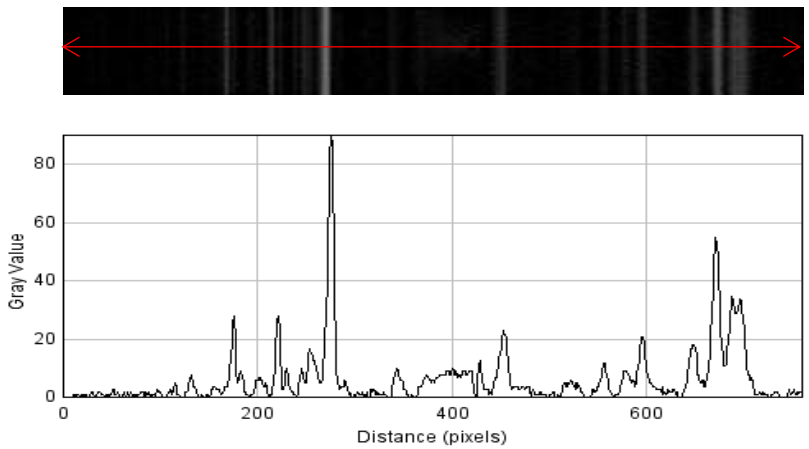
Background subtraction:

$$f(x,y) \approx \tilde{f}(x,y) - Background(x,y) + I(\tilde{f}(x,y), Background(x,y)) \qquad \textbf{Equation 5-2}$$

where I represents the mutual information as defined in [10]. Figure 5-4 shows the comparison of selected raw gel image and enhanced gel image.



(a) Raw gel image

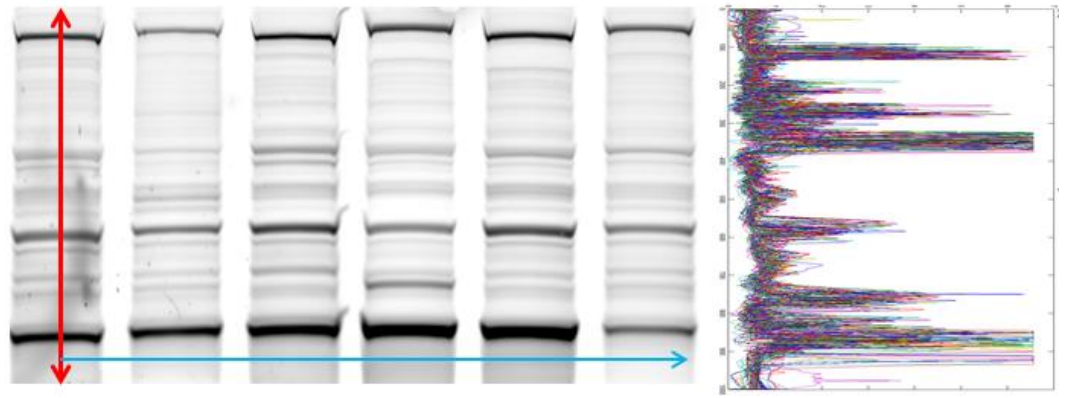


(b) Enhanced image after background correction

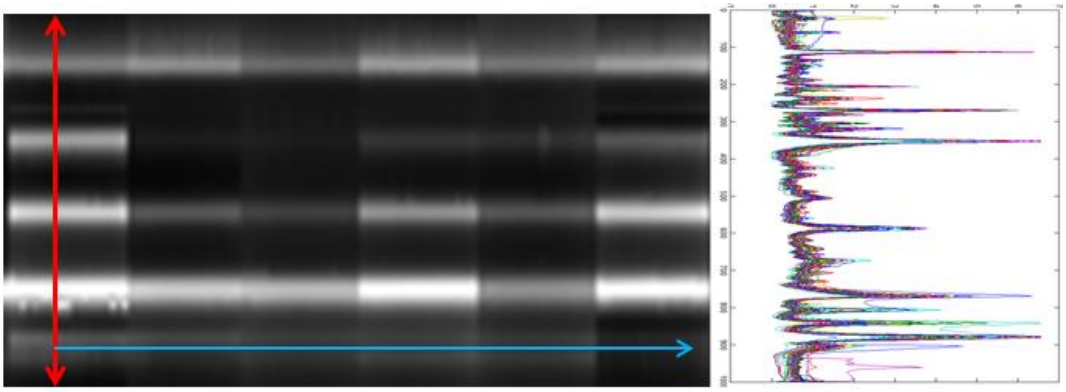
Figure 5-4. 1D profiles (the intensity in y-axis w.r.t the location in x-axis) scanned along with red line.

Data normalization. To make quantitative comparisons between profiles of lanes, and/or position of bands, it is necessary to normalize the distortion. There are two major steps in correcting the distortion of gel electrophoresis images. One is the straightening of vertical flexion of cell line, namely intra-lane alignment of bands. This procedure helps to recognize and relocate cell line. Various methodologies, i.e. interpolation based [28] and grid based [29], are reported to efficiently deform image shape. In this case study, the flexion in cell lines takes place out of our interested region, and thus can be neglected or slightly aligned from horizontal normalization.

Another normalization procedure in our study is the horizontal bands assignment, where several approaches have been described [20] [21]. We employ the Sparse Dynamic Time Warping (SDTW) method [11] to yield optimal conjunct alignments, as it is very efficient and can maintain the ability of searching for a more optimal solution. By this mean, all the separate bands are relocated into a parallel line where the corresponding bands have the same positions as in the different cell lines.



(a)



(b)

Figure 5-5. Bands normalization procedure. The second column is the profile plots on red arrow that along blue direction, and profiles are normalized to have a similar distribution after alignment (drifts are aligned). (a) Original profiles in a gel, e.g. fluorescently labeled gel. (b) Mean of the warped image after band alignment (background between cell line-band is eliminated).

Segmentation of the bands. The aligned profiles of each of the lanes are put up for analysis by separating all of its bands. It performs as an extraction of region of interest mask that image information is ready to be converted to numerical matrix. For each band, a (binary) mask is computed by first finding a (local) peak value from the centerline of the lane, and then established the neighboring (local) valleys for every peak (cf. Figure 5-6). The mask is subsequently obtained from these threshold values for each of the bands in each of the lanes.

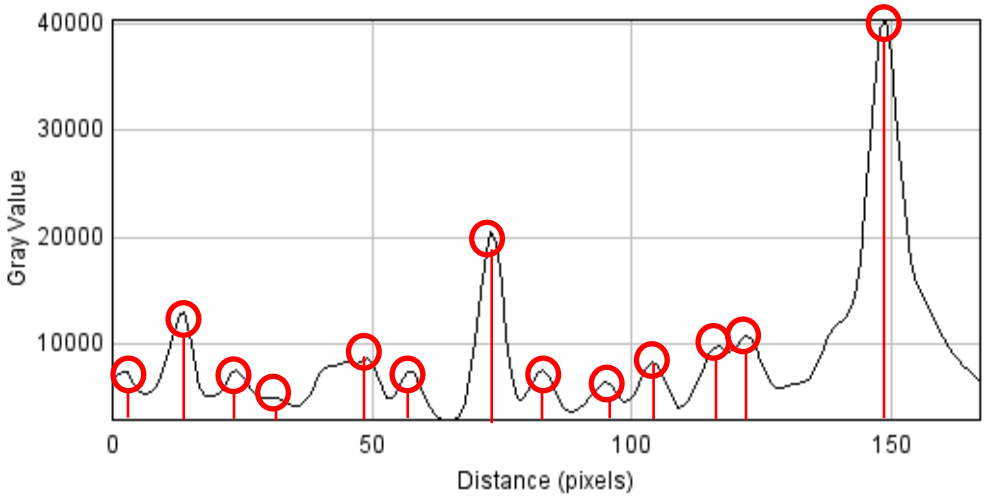


Figure 5-6. Band detection within a cell line. After background subtraction (difference cf. Figure 5-4) and data normalization, the red stems denote peak value on a lane profile, which indicate a possibly location of bands.

C. Feature Extraction and Selection

Feature extraction aims to reduce the data dimensionality and complexity, and therefore its application provides an efficient way to allow more feasible statistical analysis. Multifactorial classes of algorithms can be applied on 2-D gel images [12], for instance, boundary-based techniques, region-based methods, and hybrid methods that combine boundary and region criteria. However, none of the general or optimal procedures for extraction and quantification on 1-D gel images is reported. Targeting on segmented bands of each cell lines, it is crucial to find certain features to represent the characteristics of these bands. This is the key of data analysis of 1D gel electrophoresis images.

Phenotypic features are considered as the composites of observable characteristics or traits for an organism [13], and therefore these are employed in our work. In the

attempt to find prominent phenotypic features to characterize the proteins or fragments, two aspects should be considered: (1) features should be representative and relevant; and (2) features should be robust with respect to the small variations in bands intensities.

Table 5-1. Basic measurements for a phenotype

Feature name	Description
size	The surface area of object
Intensity	Amount of intensity belong to object
Perimeter	The perimeter of object
Circularity	Area-to-perimeter ratio
Extension	Derived from 2nd –order invariants of object [14][15]
Dispersion	Derived from 2nd –order invariants of object [14][15]
Elongation	Derived from 2nd –order invariants of object [14][15]
Orientation	Derived from 2nd –order invariants of object [14][15]

Table 5-2. Texture measurement of a phenotype (x represents the intensity value of one pixel, while $H(x)$ is the histogram of the intensities)

Feature name	Expression	Description
Avg	$f_1 = \mu$	Average intensity in a region of object.
Std	$f_2 = \sqrt{\sum_x (x - f_1)^2 H(x)}$	Standard deviation of intensity in a region of object.
Smoothness	$f_3 = 1 - \frac{1}{(1 + f_1^2)}$	Relative smoothness of intensity in a region.
Skewness	$f_4 = \sum_x (x - f_1)^3 H(x)$	Deviation from symmetry of mean intensity
Uniformity	$f_5 = \sum_x H^2(x)$	Sum of squared elements in histogram
Entropy	$f_6 = -log_2 H(x) \sum_x H(x)$	Statistical measure of uncertainty

Direct and indirect quantifications include determination of the selected phenotypic measurements (cf. Table 5-1 and Table 5-2), in which each result is calculated from the pixels that define the shape of lanes/bands. This procedure quantifies the information pattern of 1-D gel electrophoresis images into distinct measurements, which requires further selection of features. The manner in which prominent features

are chosen to represent the dynamics of fragments-migrating process becomes an important step for identification of the phenotype.

In order to guarantee that all selected features are independent and equal of variance, the Mahalanobis distance [16] is chosen as the probabilistic distance criterion. Subsequently, we employ the methodology, the Fuzzy Criteria in Multi Objective Feature Selection Algorithm [17] (cf. chapter 3), which selects a subset of features from feature-pool that can best predict and describe the data. The resulting solutions from this method lead to a set of candidate feature combinations. This will facilitate the bio-scientist in selecting the proper features to predict a biological phenomenon and provide a guideline for new experimental design. In this case study, the selected features based on the best performance in the approach from multi-objective optimization, are band width, band intensity standard deviation, and lane skewness. The original 1-D gel electrophoresis images usually have different widths of bands (cf. Figure 5-1) at different positions reflecting the molecular weights of fragments. The "band intensity standard deviation" is a global index for a detected band; while the "lane skewness" is a vertical descriptor to understand the deviation from symmetry (cf. Figure 5-5 (a) and (b)) as fragments migrate downstream.

D. Information Clustering Analysis

The measurement information is summarized into a matrix of statistics that represents the patterns information. To date, various pattern finding procedures have been settled. However, for research implementation the information clustering is particularly important. In order to address biological questions accordingly, two issues come up: how to partition sets of samples that contain various features into groups among a large number of bands or lanes from electrophoresis gels; and how to figure out different patterns amidst samples with indistinguishable information and features.

To tackle these issues, we present [18] (cf. Chapter 4) an innovative and efficient approach that is capable of clustering information from overlapping and otherwise indiscernible partitions. This method, a.k.a. Rough Fuzzy C-means and Particle Swarm Optimization hybridization (RFCM-PSO), combines the RFCM clustering algorithm [19] with an optimization technique. In RFCM, the rough approximation sets are employed to constrain the *fuzzifier* membership index. Subsequently, the iterative procedure of partitioning is then minimizing the RFCM objective function. Whereas the optimization of the clustering results occurs, the PSO procedure searches for the global optimum by updating the candidate centroid positions of partitioning solutions. The pseudo-code of RFCM-PSO is shown:

Algorithm *Rough fuzzy c-means and PSO hybrid method*:

Input: fuzzifier, weighting factor, cluster number, and controlling parameters in RFCM
Given: integral population and generation in PSO
Initializing: stochastic centroids, membership matrix, position and velocity at first generation
for each generation **do**
1. training RFCM parameter:
Compute similarity distance for each object to its belonging cluster centroids.
if Rough approximation condition **then**
Reset membership matrix from FCM.
end if
Update centroids.
Update membership matrix.
2. Optimization procedure:
Computing the personal best and global best positions.
Update position and velocity for each particle.
Convergence check: break main loop
end for

By taking the advantage of both the RFCM clustering method and the intrinsic characteristics of PSO, this combined-model can now deal with overlapping partitions, uncertainty and vagueness of information. At the same time, the optimization procedure in PSO demonstrates the ability of searching optimal solutions. In this case study, the information clustering takes place at two aspects: 1) clustering of bands and 2) clustering of cell-lines. In the first aspect, a band of all cell lines is selected. For instance, band number 65 in all 60 cell lines is chosen, then 1x60 samples are obtained and will contain several selected features (e.g. band width, skewness, local entropy). Thereafter via a clustering of bands, we can notify the intrinsic property of a certain band that affect expressions (captured by 1D gel electrophoresis, and represented as different intensity on image data) in a variety of cell lines. Another implementation of the clustering technique is on the cell line. With little labelled cell type (cf. Figure 5-3, some types of cell lines are considered as null or unverified), we utilize results from the decision tree (cf. Figure 5-7), and cluster the cell line based upon their selected feature performance. The result of clustering example is shown in Figure 5-8.

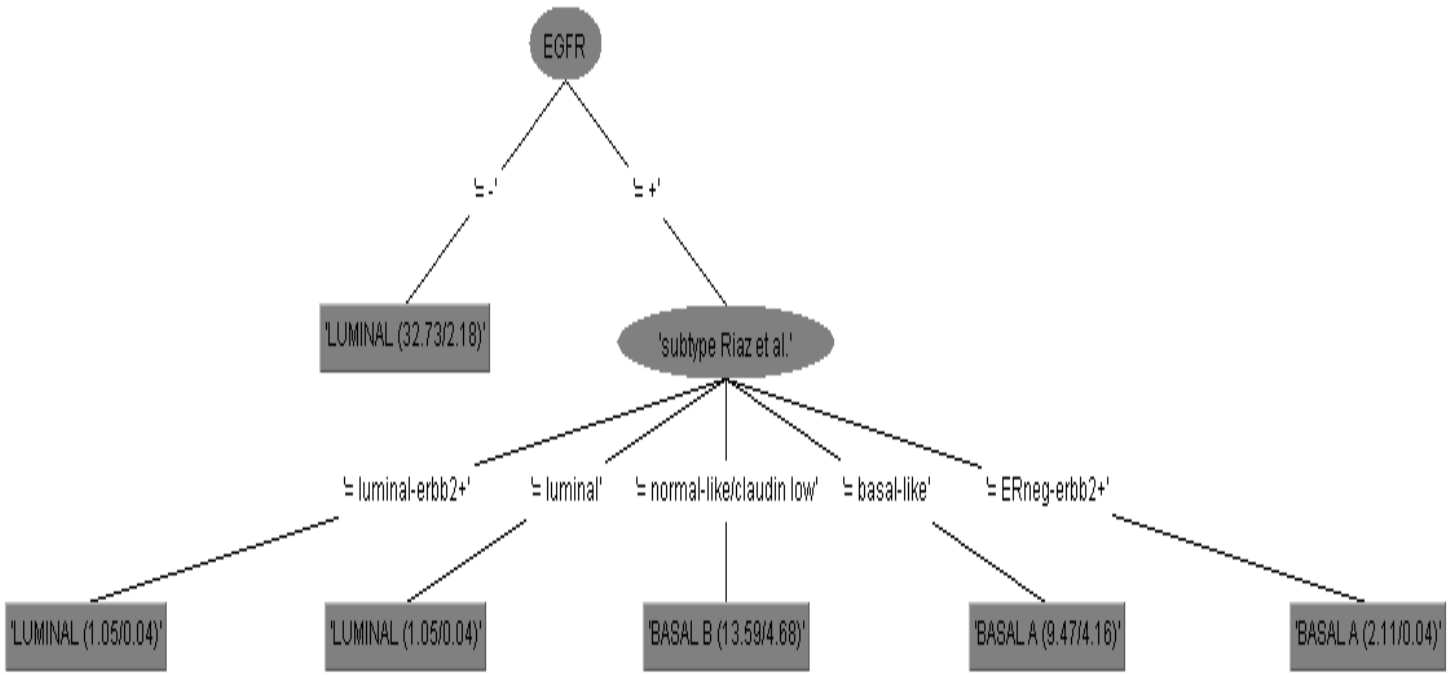
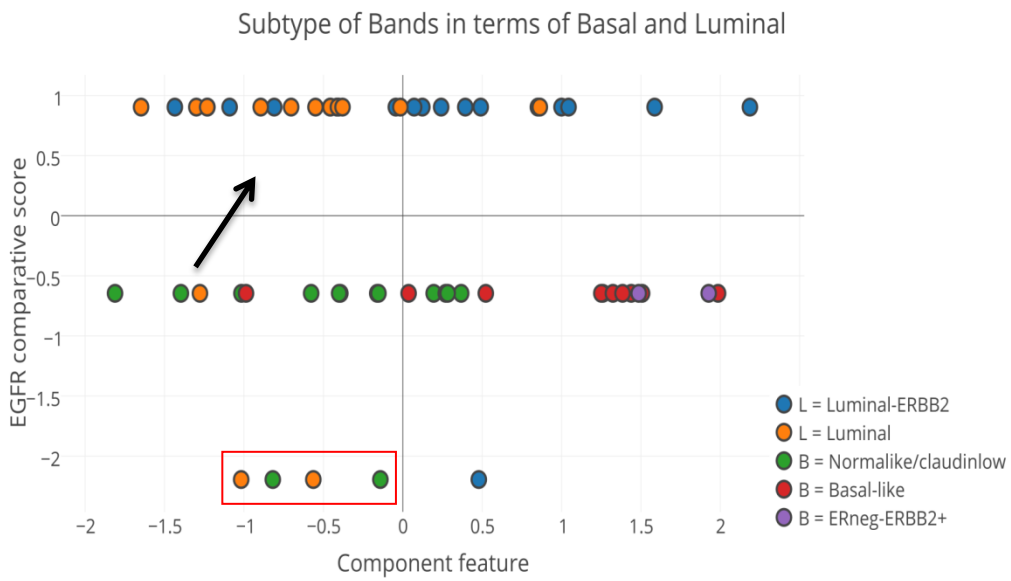
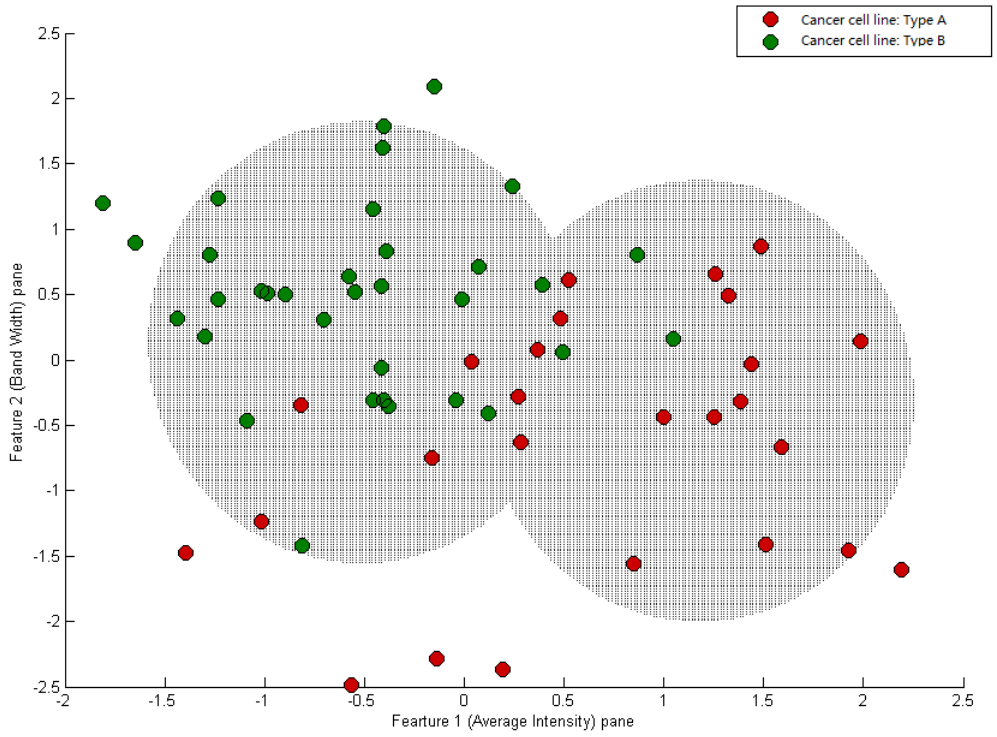


Figure 5-7. Decision tree view of chances that consequent clustering outcome may be significantly affected by cell line types w.r.t their subtype and reactor.


Subtype of Bands in terms of Basal and Luminal
EGFR comparative score
Component feature
L = Luminal-ERBB2
L = Luminal
B = Normalike/claudinlow
B = Basal-like
B = ERneg-ERBB2+

(a)


Cancer cell line: Type A
Cancer cell line: Type B
Feature 2 (Band Width) pane
Fearture 1 (Average Intensity) pane

(b)

Figure 5-8. Clustering result of genetically distinct cancer cell lines. (a) All 60 cell lines should be clustered into two types, either one band should belong to "Basal" or "Luminal". In Figure 3, cluster L contains all the subtypes that belong to "Luminal", and, the rest of subtypes are clustered to B cloud, namely the "Basal" cloud. However, the Line T47D is occurred in the B cloud when it has quite similar feature properties as its clustered neighbors. Moreover, lines laid in red region are indiscernible, though they can also have a similar feature property. When checking the line information table (cf. figure 5-2), the abovementioned lines are number 57 to 60 respectively, and with a void score for their corresponding EGFR value. Here, 27 of basal bands are compared with 33 different luminal bands based on the difference in their intrinsic feature. (b) We combine the selected features together, and perform a clustering analysis against the information table in terms of cell line type Basal and Luminal (type A in red dot means basal, type B in green dot means Luminal). Every green and red dot represents an integration of features from detected bands which belong to a list of cell lines (LCLs). Features are normalized, and cell lines are categorized unsupervised. For this particular clustering result, the F-score is 0.92 using the feature pool as reported in section 5-2.D.

5.3. Measurements and Results

In In this section, we apply our approaches to illustrate the fuzzy system research strategy on a dataset of cancer cell-lines; the cell-lines have different genotypes and from gene expression profiling, it becomes clear that they should be distinguished in different subgroups.

To estimate and investigate how quantitatively and accurately the proposed fuzzy system can be employed on interpreting biological questions, dedicated solutions are conducted and analyzed on this dataset. The experiment consists of 60 cell-lines that are loaded on the PAGE gels in 60 separate lanes. After running the gel, they will result in a sequence of images that contains 60 lanes with 85 detected bands. For each band the features are measured (cf. Section 5-2.C). The preprocessing and normalization of the gel lanes make it possible to directly compare the bands for position and intensity.

We aim at finding a pattern in a series of gel images to characterize groups of cell-lines. To this end, we formulate a null hypothesis (H_0) and test if the statistical inference of the underlying distribution can be considered significant. In other words, attributes (features) of bands, within the scope of different cancer cell-lines, will be carefully measured by applying T-test schema. It is different from what typically used in student T-test, where samples contain only one or two attributes. In this case study, we propose a more comprehensive Hoteling T-square test [22] employing multi-variates (cf. features from Section 5-2.C) to obtain an accurate measurement

of the significance. Ultimately, the resulting p-value is used to weight the strength of the evidence, demonstrating the significance of protein/DNA expressions in different cell type groups. We have found several bands that perform significantly in either cell-type, subtype, or the reaction of regulator (cf. Table 5-3 and Table 5-4). Each band of all cell lines are compared separately, and their statistical performance are shown in ranges of either $p \leq 0.05$ or $p \leq 0.01$ (where bands have p value ≤ 0.01 are involved in those bands with p value ≤ 0.05). The latter value of p indicates a more significance in expression (effect) of the group (population). Table 5-3 gives an example of one series gel images, where the significances of detected bands are explored.

In Table 5-4, accordingly, a significance test of Basal and Luminal cell line is conducted using top 5 best descriptors, e.g. intensity standard deviation, intensity, entropy, skewness, and width. For either Luminal or Basal cell lines, the detailed and numerical attributes of tested bands are averaged (Total) and compared with Luminal group average (Luminal) and Basal group average (Basal). When the total expression of a band (all features are counted) is close to the expression of Luminal group (same band), we consider this band is Luminal-significant; and vice versa for the band of Basal-significant. Alternatively, the value of the features in Table 5-4 with N/A means a specific band has no expression in such groups. For example, the number 77 band (cf. Table 5-4), as it is recognized as UCHL1 band [26] that has higher significance in Basal group but lower significance in Luminal group. A bar plot of UCHL1expression in two separate experiment is shown in Figure 5-9. Gels containing both 20 cell lines and 60 cell lines (partially overlap for validation) are conducted on the experiments, where both T-test results depict a high significance, $p \leq 0.05$. The visualization of employed feature, particularly in intensity, denotes that band UCHL1 in some cell lines has higher significant expression in basal group, and its average intensity in luminal group is lower compared with the basal one. The combination of the selected features provides a guide-line for clustering and statistical analysis of bands with respect to corresponding cell lines. These results have the same comparisons and validations in either mass spectrometry and/or reports [23] [26].

Table 5-3. Significant bands description in different type of cell line (85 detected bands in total)

	ER	PgR	ERBB2	EGFR	CK5	Basal/Luminal
Bands description (number)	14	46	Null	Null	29	49
P≤0.01	69	54				
Bands description (number)	24	24	7	17	8	71
P≤0.05	46	25	23	46	57	77
	66	26	27		64	78
	71	36	47		74	80
		42	81		75	
		51				
		56				
		82				
		80				

Table 5-4. Significance performance of bands with p ≤0.05 (basal/ luminal cell type)

Band Nr.	Type(ave)	Int Std	Intensity	entropy	skewness	width
# 49	Basal	10.461	315.509	-295.555	2.729	1.259
	Total	10.803	180.136	-159.980	1.564	0.750
	Luminal	N/A	N/A	N/A	N/A	N/A
# 71	Basal	N/A	N/A	N/A	N/A	N/A
	Total	22.274	238.182	-599.722	2.303	2.408
	Luminal	40.498	433.058	-109.040	4.187	4.378
# 77	Basal	198.614	3776.476	-538.034	30.059	17.518
	Total	178.773	4776.549	-846.869	32.663	18.95
	Luminal	161.353	5467.407	-109.478	34.079	19.909
# 78	Basal	115.256	3255.756	-291.145	9.448	9.740
	Total	117.027	4232.853	-403.740	8.965	11.008
	Luminal	117.144	4897.807	-489.061	8.560	11.803
# 80	Basal	27.359	2358.341	-175.230	13.616	5.314
	Total	13.108	1736.411	-86.669	6.697	2.800
	Luminal	1.449	1227.559	-14.210	1.037	0.742

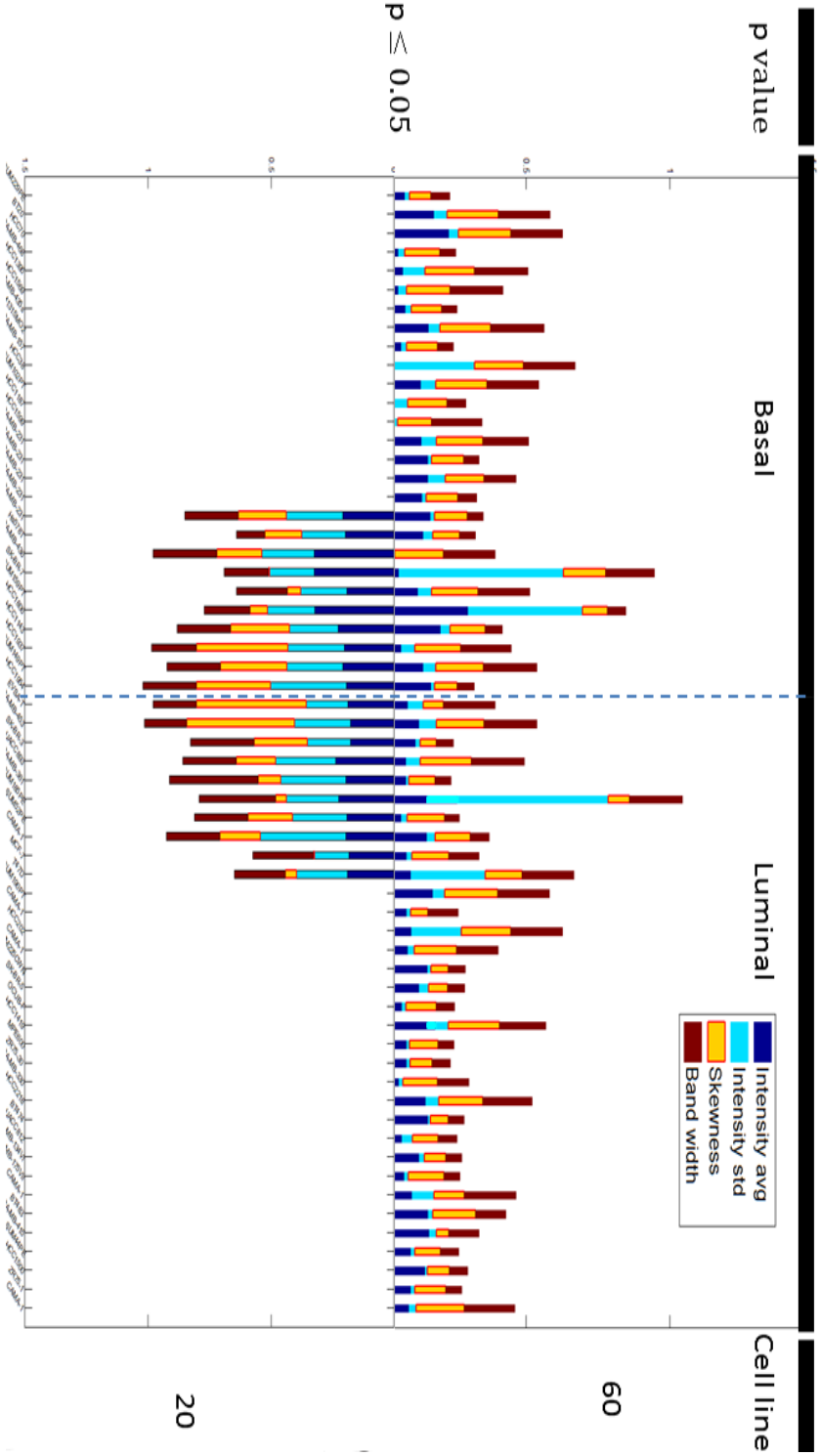
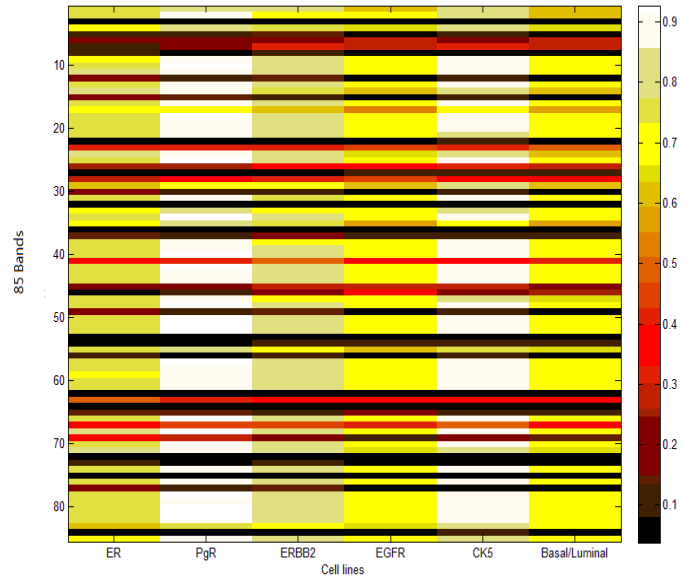
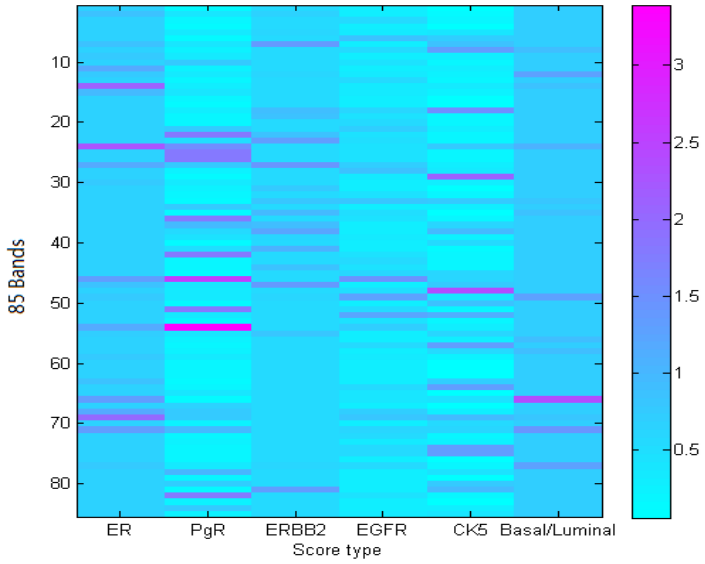


Figure 5-9. Features visualization of band UCHL1 in two gels (20 cell lines and 60 cell lines).



(a)

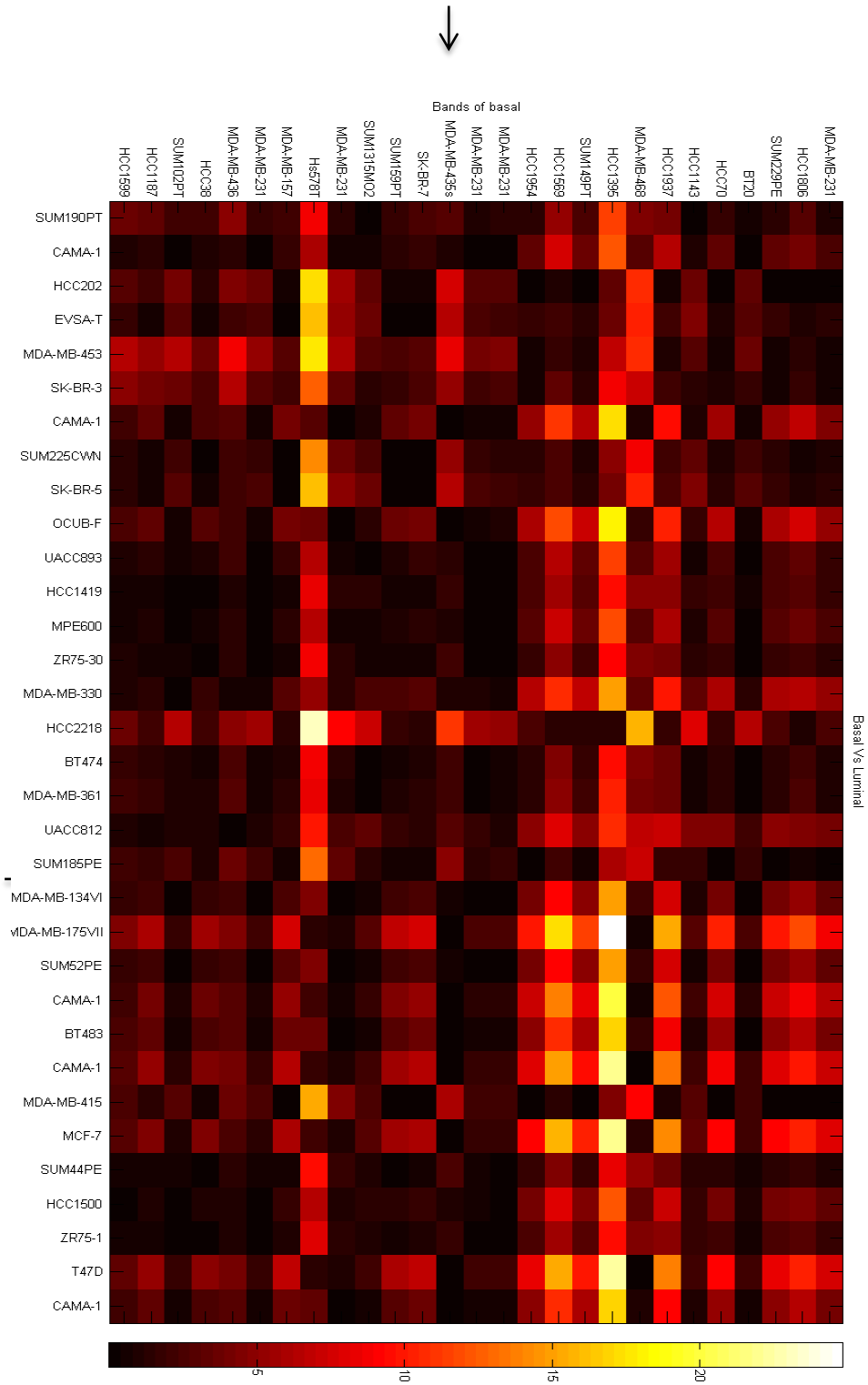


(b)

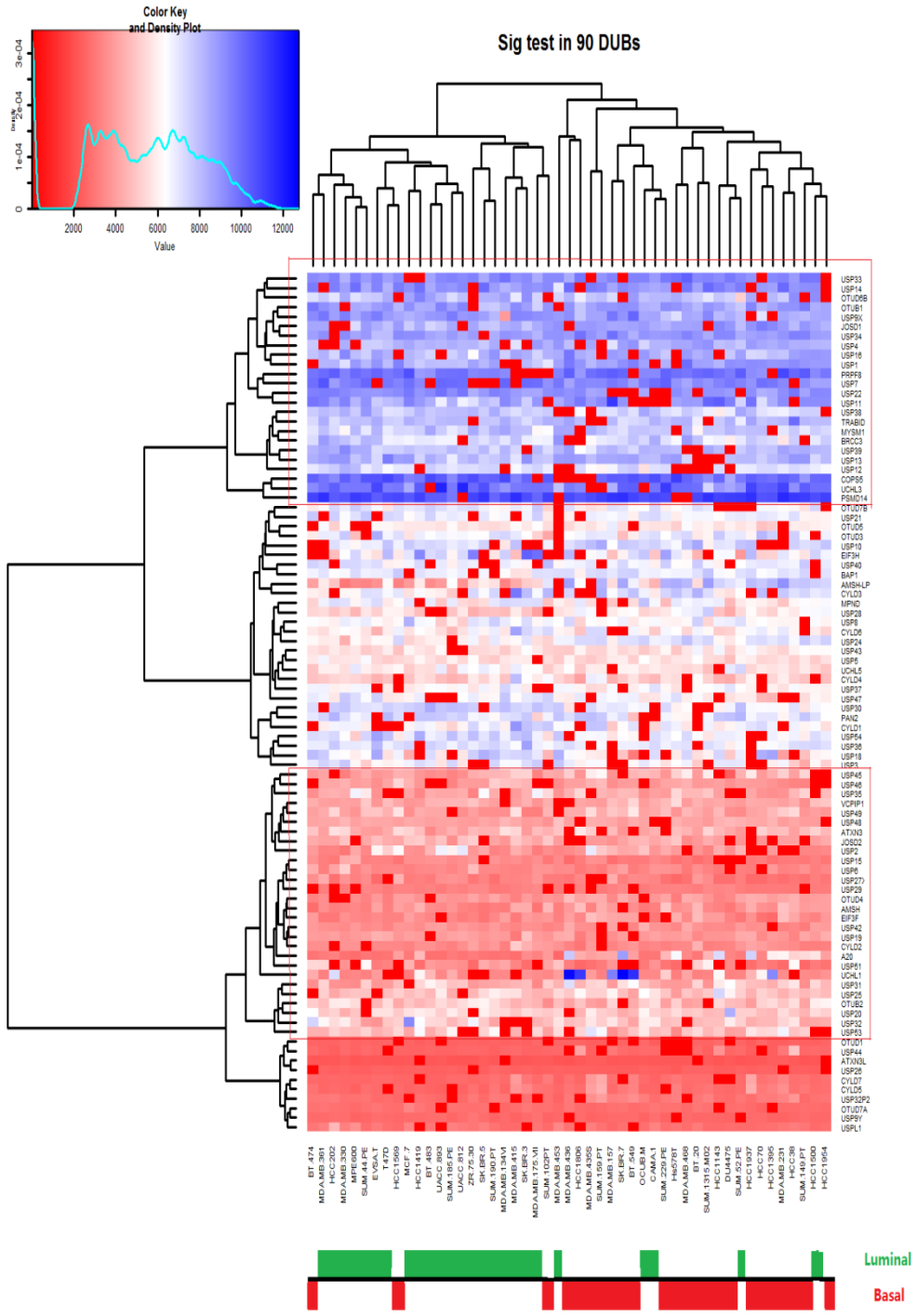
Figure 5-10. Heat map of statistical tests on six cell-line type respectively. The y-axis represents the number of detected bands and six different types in x-axis are investigated for: (a) F-accuracy test in terms of classification using the feature information of bands; (b) Resulting p-value (log-transformed) from Hoteling T-square significant test. Brighter values indicate a higher accuracy (significance) in the classes.

Figure 5-10 (a) and (b) show the results of F-test and T-test of all 60 lanes with their detected 85 bands in one visualization. The similarity of the band with respect to its corresponding cell lines can be categorized under an unsupervised condition utilizing the proposed fuzzy system pipeline and the accuracies of bands clustering are verified by F-test. Additionally, for a particular cell line subtype, the entire sets of cell-lines are recognized as two groups, e.g. ER positive and ER negative, PgR positive and PgR negative, ERBB2 positive and ERBB2 negative, CK5 positive and CK5 negative, as well as cancer type basal and type luminal.

The result of the F-test, as depicted in Figure 5-10 (a), reveals that the ability for subtype-clustering of each separate band varies with subtypes in the population of the cell-lines (a priori knowledge) as well as with the gel analysis results. The intrinsic properties of the bands are directly related to the cell subtype that they originate from. Some bands exhibit extremely low F-values in all the six subtypes which originate from proteins indicating an absent or are not activated in the cell under the conditions of the experiment. In Figure 5-10 (b), it is shown that bands can be reported to have significant differences (p-value < 0.01) from a classification into the six subtypes. This indicates that the detected bands of proteins or DNA, can represent a uniqueness expression profiles.


Bands of basal
HCC1599
HCC1187
SUM102PT
HCC38
MDA-MB-436
MDA-MB-231
MDA-MB-157
Hs578T
MDA-MB-231
SUM1315MO2
SUM159PT
SK-BR-7
MDA-MB-435s
MDA-MB-231
MDA-MB-231
HCC1954
HCC1569
SUM149PT
HCC1395
MDA-MB-468
HCC1937
HCC1143
HCC70
BT20
SUM229PE
HCC1806
MDA-MB-231
SUM190PT
CAMA-1
HCC202
EVSA-T
MDA-MB-453
SK-BR-3
CAMA-1
SUM225CWN
SK-BR-5
OCUB-F
UACC893
HCC1419
MPE600
ZR75-30
MDA-MB-330
HCC2218
BT474
MDA-MB-361
UACC812
SUM185PE
MDA-MB-134VI
MDA-MB-175VII
SUM52PE
CAMA-1
BT483
CAMA-1
MDA-MB-415
MCF-7
SUM44PE
HCC1500
ZR75-1
T47D
CAMA-1
5
10
15
20
Basal Vs Luminal

(a)


Color Key
and Density Plot
Count
Value
Sig test in 90 DUBs
Luminal
Basal

(b)

Figure 5-11. Heat map of the 1D gel electrophoresis phenotypic analysis, where protein/DNA occurrence as analyzed from bands on the gel. (a) Heat map of normalized features of intensity and band width for group basal vs. group luminal. Brighter colors represent a higher response in terms of their phenotypic expression. The arrows indicate a specific type of cancer cell-line in group A/B with strong responses; (b) Hierarchical clustering analysis of essentials of detected bands (vertical) with respect to cancer types (60 cell-lines involved in either basal or luminal group) (horizontal). The colors indicate a degree of correlation between bands and subtypes.

Experiments with gel electrophoresis support the understanding of the relationship between sample groups. A clustering analysis, hereby, targets to find the hidden patterns in data. Hierarchical clustering revealed distinct positive (blue) and negative (red) expression of proteins (cf. Figure 5-11 (b)) in terms of the normalized features quantified from gel electrophoresis images. It helps in validating predicted preference of cell-line type. In Figure 5-11 (a), cell-line groups of different subtypes are compared to examine the differences in the patterns. Two maximum variations of bands in conflict the subtype are pointed with arrows.

5.4. Conclusions

This chapter investigates and illustrates the ability of proposed fuzzy-logic based methodologies and their integrated fuzzy system pipeline for data analysis from 1-D PAGE gel electrophoresis. The adequate processing algorithms and heterogeneous information are thereof composed into a global picture. It is demonstrated via a practical implementation on a series of bio-imaging experiments that this system is reliable and is capable of qualitatively and quantitatively assessing information. Quintessential are the fuzziness background correction, feature extraction and selection of region of mask based upon fuzzy criteria. These elaborated approaches contribute to phenotypic quantification and henceforth unsupervised classification. The pattern extraction and recognition aim to support phenotype analysis. In this chapter, the experiment shows that employing the proposed fuzzy system it can be accomplished by investigating and understanding the identity of proteins characteristics which are distinct/shared between different subgroups of cancer cell-lines. In addition to this case study, in protein characterization, DNA and RNA fragments can also be separated by 1D electrophoresis. In the same manner, the proposed method can also be applied for a systematic analysis of DNA and RNA patterns.

5.5. References

[1] Machad, Alexei, et al. "An iterative algorithm for segmenting lanes in gel electrophoresis images." Computer Graphics and Image Processing, 1997. Proceedings, X Brazilian Symposium on. IEEE, 1997.

[2] Lin, Yun-Liang Yang, et al. "Automatic method to compare the lanes in gel electrophoresis images." Information Technology in Biomedicine, IEEE Transactions on 11.2 (2007): 179-189.

[3] Adiga, PS Umesh, et al. "Automatic analysis of agarose gel images." Bioinformatics 17.11 (2001): 1084-1089.

[4] Kaabouch, Barry Milavetz, et al. "A Novel Automated Analysis System for DNA Gel Electrophoresis Images." In Artificial Intelligence and Pattern Recognition, pp. 36-41. 2007.

[5] Ye, Xiangyun, et al. "A recent development in image analysis of electrophoresis gels." Vision Interface'99, Trois-Rivières 19.21 (1999): 432-438.

[6] Fuyu Cai, and Fons J. Verbeek. "Dam-based rolling ball with fuzzy-rough constraints, a new background subtraction algorithm for image analysis in microscopy." Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on. IEEE, 2015.

[7] Kimberly F. Sellers and Jeffrey C. Miecznikowski. Statistical Analysis of Gel Electrophoresis Data, Gel Electrophoresis - Principles and Basics, Dr. Sameh Magdeldin (Ed.), InTech, (2012): DOI: 10.5772/36959.

[8] Kaczmarek, Krzysztof, et al. "Preprocessing of two‐dimensional gel electrophoresis images." Proteomics 4.8 (2004): 2377-2389.

[9] Vovk, Botjan Likar, et al. "A review of methods for correction of intensity inhomogeneity in MRI." IEEE transactions on medical imaging 26.3 (2007): 405-421.

[10] Papoulis, S. Unnikrishna Pillai, et al. Probability, random variables, and stochastic processes. Tata McGraw-Hill Education, 2002.

[11] Al-Naymat, Javid Taheri, et al. "Sparse dtw: A novel approach to speed up dynamic time warping." Proceedings of the Eighth Australasian Data Mining Conference-Volume 101. Australian Computer Society, Inc., 2009.

[12] Nedenskov Jensen, Bo M. Jørgensen, et al. "Multivariate Data Analysis of Two-Dimensional Gel Electrophoresis Protein Patterns from Few Samples†." Journal of Proteome Research 7.3 (2008): 1288-1296.

[13] Cowin, John Wysolmerski, et al. "Molecular mechanisms guiding embryonic mammary gland development." Cold Spring Harbor perspectives in biology 2.6 (2010): a003251.

[14] Hu, Ming-Kuei. "Visual pattern recognition by moment invariants." information Theory, IRE Transactions on 8.2 (1962): 179-187.

[15] K. Yan, Fons J. Verbeek, et al. (2009), Cell Tracking and Data Analisys of in vitro Tumor Cells from Time--Lapse Image Sequences. In: Proceedings VISAPP 2009. 281-287.

[16] Cai, Fuyu, et al. "Fuzzy Criteria in Multi-objective Feature Selection for Unsupervised Learning." Procedia Computer Science 102 (2016): 51-58.

[17] Jain, Jianchang Mao, et al. "Statistical pattern recognition: A review." Pattern Analysis and Machine Intelligence, IEEE Transactions on 22.1 (2000): 4-37.

[18] F. Cai, F.J. Verbeek. “Rough Fuzzy C-means and Particle Swarm Optimization Hybridized Method for information Clustering Problem”. Journal of Communications. ISSN: 1796-2021 (in press)

[19] Ji, Zexuan, et al. "Generalized rough fuzzy c-means algorithm for brain MR image segmentation." Computer methods and programs in biomedicine 108.2 (2012): 644-655.

[20] Lemire, Daniel. "Faster retrieval with a two-pass dynamic-time-warping lower bound." Pattern recognition 42.9 (2009): 2169-2180.

[21] Salvador, Philip Chan, et al. "Toward accurate dynamic time warping in linear time and space." Intelligent Data Analysis 11.5 (2007): 561-580.

[22] Marozzi, M. Multivariate multi-distance tests for high-dimensional low sample size case-control studies. Statist. Med., (2015), 34: 1511–1526. doi: 10.1002/sim.6418.

[23] Simpson, Erica Golemis, et al. Basic methods in protein purification and analysis. Cold Spring Harbor Laboratory Press, 2009.

[25] Kaur, Amanpreet, and M. D. Singh. "An overview of pso-based approaches in image segmentation." Int J Eng Technol 2.8 (2012): 1349-1357.

[26] Marcotte, Richard, et al. "Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance." Cell 164.1 (2016): 293-309.

[27] Intarapanich, Apichart et al. “Automatic DNA Diagnosis for 1D Gel Electrophoresis Images Using Bio-Image Processing Technique.” BMC Genomics 16.Suppl 12 (2015): S15. PMC. Web. 28 July 2017.

[28] Schaefer, Scott, Travis McPhail, and Joe Warren. "Image deformation using moving least squares." ACM transactions on graphics (TOG). Vol. 25. No. 3. ACM, 2006.

[29] Igarashi, Takeo, Tomer Moscovich, and John F. Hughes. "As-rigid-as-possible shape manipulation." ACM transactions on Graphics (TOG). Vol. 24. No. 3. ACM, 2005.

Chapter 6

Conclusions and Outlook

The main contributions of this thesis, underpinning bioinformatics studies for which the application of image analysis is important, and are summarized as:

1. Uncover common fuzzy principals that applied across many unsupervised computing systems, and highlight the features that adaptively address biological questions.
2. Dedicated pipelined methodologies are proposed for formalizing, measuring and modeling the uncertainty and diversity in bioinformatics studies.
3. By means of extensions and derivatives of soft computing, this work enlarges the depth of biological investigation, and engages width of dimensions as well.

6.1. Conclusions

In this thesis, we focus on exploring solutions for bio-imaging data analysis using fuzzy-computing paradigms. To gain further insight into the solutions, experiments are conducted on several publicly available datasets representing multiple disciplines and practical problems. The performance of solutions is then examined comprehensively and carefully using sorts of convincing index and evaluation strategies. The case study in this thesis further demonstrates the possibility of producing objective understandings of the identity of characteristics in proteins/DNA that are distinct/shared among different groups in phenotypic measurements. The conclusions of the chapters in this thesis are presented in the next sections.

Chapter 2 Biological Image Background Correction

In this chapter, we challenge a common interference in bio-imaging data, known as the inhomogeneity of illumination in the background. This effect leads to information contamination and loss of accuracy in both qualitative and quantitative work. A robust approach, the Dam-Constrained Background Correction (DCBC) is proposed to reduce the deviations before image datasets are subject to further analysis track. The fuzzy membership function in this approach, together with the rough set constraint, innovates constructing a morphological “dam” in the image, and thereby prevents over-segmentation of the background. The experimental results demonstrate the potential of the application on bright-field and fluorescence images, since further analysis will no longer be hampered by undesired vignetting.

Chapter 3 Feature Selection Strategy in Region of Interest Mask

In this chapter, we present a novel feature selection strategy, namely the Fuzzy Criteria in Multi Objective Feature Selection Algorithm (FC-MOFSA) that selects potential predictors in the feature pool. In FC-MOFSA, a fuzzy-defined entropy measurement is used in filter-approach, meanwhile the fuzzy criteria referred to as Correlation Membership Measurement (CMM), is employed as a cost-function in wrapping-approach in the feature selection process. The proposed CMM criterion highlights the essentials in sparse and skewed datasets clustering procedure by ambiguously estimating the difference between all pairs of points in different clusters. Aiming at converting digital signals to numerical feature descriptors, this strategy further optimizes the tasks of handling data in large volumes, and data clarification without jeopardizing the quality of information.

Chapter 4 Unsupervised Information Classification and Analysis

In this chapter, we present a dedicated classification algorithm, the Rough Fuzzy C-Means and Particle Swarm Optimization (RFCM-PSO). In the information clustering process, the concept employed in the RFCM handles with uncertain, vague and sparse data processing. PSO is able to optimize searching procedure and avoids the results from being trapped into local optimum. To demonstrate the potential of this method, a number of state-of-the-art algorithms are compared. The extensive numerical analysis and reported results indicate the good performance in revealing hidden patterns and exploring relations or attributes between clusters.

Chapter 5 A Systematic Study on One Dimensional Gel Electrophoresis Image Analysis

In this chapter, we present the feasibility and capability of proposed fuzzy-system and unsupervised computing based methodologies pipeline in protein phenotypic quantification and subsequent classification. The mixtures of protein in a cell lysate can be separated, visualized and analyzed by classical one dimensional (1D) gel electrophoresis. Subsequently, the resulting gel images can consist of several vertical lanes (number of wells in which the protein samples were loaded), and a number of horizontal bands (corresponding to proteins or fragments thereof). The data reflects the amounts and characteristics of individual proteinaceous components. This case study is one of the first attempts to reveal the intrinsic properties of bands from which cell subtype they are originated.

The work in this chapter, compared to the conventional phenotypic analysis system, improves gel data analysis into a four-stage strategy. Dedicated algorithms ensure

that the measurements are summarized into a matrix of primary representing the information of patterns. Therefore, the employment of fuzzy systems plays an important role in unambiguously uncovering patterns hidden in the data, yet yielding precise evaluation. In terms of unsupervised computing, it is also objective that the correlation of data clusters is qualitatively assessed.

6.2. **Outlook**

This thesis has explored and discussed solutions for addressing biological questions by building fuzzy logic-based and rule-based systems. The results are promising, but should be refined and expanded in further studies. In the next paragraphs we elucidate what aspects need additional attention.

Data acquisition

Modern bio-imaging systems are often synchronized with other computer or human controlled equipment and software. This on-the-fly system allows auto- or semi-automated adaption for experimental preparation, e.g. focusing lens, position of specimen, signal acquisition and transformation. However, the quality of data obtained via this adaption in controlling system is always subjected from heuristic mechanisms of objects and intrinsic algorithms. This mechanism can introduce either frequency noise (cf. Chapter 5) or out-of-focus background illumination (cf. Chapter 2). Additionally, based upon biological experimental design, the resolution of acquiring data is often sacrificed in exchange of a better image quality, or acquisition speed. These issues are hardware related and could be significantly improved by the development of imaging techniques.

Methodology design

The fuzzy systems, involved in soft computing paradigms, has been successfully employed to interpret biological questions. There are several reasons accounting for the trend of its increasing use in bio-sciences, and a few aspects could also be further improved. The most significant reason for accelerating a bioinformatics study is that fuzzy theory permits approximation instead of high-precision which sometimes arrives at an expensive computational budget. Optimization algorithms could play a very important role in helping fuzzy systems search the “best available” solutions from sparse and skewed data. As an ingredient from fuzzy rule-based model, adequate optimization strategies might further reduce computational overhead. Second, but not less important, is that fuzzy theory reduces complexity and thereby simplifies system modeling when empirical knowledge and behavior are not known. The concept of fuzzy systems offers a simple but effective way to arrive at a

definitive conclusion on the basis of ambiguous, imprecise, noisy or missing input information. Moreover, this mechanism can benefit most from new advances in deep learning. For instance, if a few ground truth (labels) are known, then a sophisticated weakly supervised fuzzy systems might have a significant improvement on accuracy and efficiency compared to the unsupervised ones. Another possibility to improve the proposed system is that the methodologies of measurements should meet biological description. This means a question-driven experiment design is strongly recommended.

To sum up, fuzzy logic and unsupervised computing based systems, constitute a very promising analysis direction in the field of bioinformatics. The experiments in this thesis carefully and comprehensively investigate how this proposed system could be applied to biological questions. We have accomplished a good start in this direction and further progress from research efforts can potentially be anticipated.

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 acknowledges 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

Samenvatting (Dutch Summary)

In dit proefschift wordt de toepassing van het fuzzy systeem onderzocht in het bijzonder in het domain van bioinformatica. In de afgelopen tientallen jaren is de belangstelling voor bioinformatica bij zowel biologen als computerwetenschappers fors toegenomen. Dit geldt voor biologen omdat zij in hun onderzoek grote hoeveelheden data produceren die op een robuste en accurate manier moeten worden geanalyseerd. Dit geldt voor computerwetenschappers omdat zij ingaan op de uitdaging toegevoegde waarde voor deze grote data volumes te creëren door het ontwikkelen en ontwerpen van methodologieën waarmee zulke accurate en robuste analyses kunnen worden gerealiseed.

Data uit het onderzoek in de levenswetenschappen zijn in oorssronkelijke ruwe vorm, notoir vaag, dubbelzinnig, onnauwkeurig, en soms missen er data of zijn er delen niet bekend. Om met het begrip dat data niet perfect zijn om te kunnen gaan, is het nodig dat er veschillende nieuwe benaderingen worden bestudeerd en geïmplementeerd. Het Fuzzy Systeem – gebaseerd op zogenaamde “vage” logica - is een relatief nieuwe heuristieke benadering die de geschikt is voor het simplificeren van een anders complexe beslissing, door het toestaan van meerdere hypotheses in de data analyse. Met andere woorden, de logica van het fuzzy systeem erkent het feit dat significantie de belangrijkste factor in het modelleren van data is; dit terwijl in andere benaderingen aan precisie meer waarde wordt toegekend.

De fuzzy systeem benadering wordt geacht potentie te hebben voor data analyse, vandaar dat het onderzoek beschreven in dit proefschift de intentie heeft de nadruk te leggen op het ontwerpen van efficiënte en betrouwbare heuristieke oplossingen voor analyse om daarmee verborgen informatie in biologische experimenten te ontdekken. Ons onderzoek heeft als doel een speciale analysemethodiek te bouwen gebaseerd op fuzzy systemen; in de onderzoekshoofdstukken van dit proefschrift worden daartoe drie verschillende perspectieven uitgewerkt. Uiteindelijk integreren we deze drie verschillende benaderingen van fuzzy systemen in een analysemethodiek die wordt geïllustreerd aan de hand van een voorbeeld studie.

(1) We hebben fuzzy systemen gebruikt voor de correctie van achtergrondbelichting. Een microscoopbeeld bevat ruis en andere effecten die niet bijdragen aan het daadwerkelijke signaal dat wordt gemeten. Behalve willekeurige ruis die wordt veroorzaakt door de electronica van het apparaat, wordt het beeld beïnvloed door ongelijke belichting. Om de invloed van deze ongelijke belichting te reduceren moet er een achtergrondcorrectie worden uitgevoerd.

In Hoofdstuk 2 van dit proefschrift wordt een achtergrondcorrectie methode voorgesteld die gebaseerd is op mathematische morfologie en deze wordt

samengevoegd met randvoorwaarden uit de fuzzy logica, waarmee het achtergrondverloop wordt geëlimineerd. Vergeleken met de algemeen gebruikte EMI methode en het, in commercieële software veel gebruikte, "rolling ball" algoritme, demonstreert de door ons ontwikkelde DCBC methode robuuste prestaties met beelden die karakteristiek zijn voor biomedische microscopie.

(2) We hebben fuzzy systemen bestudeerd in relatie tot kenmerkselectie (eng. feature selection) en verwijderen van overtollige data. De procedure van kenmerkselectie speelt een belangrijke rol in het omzetten van fenotypische waarnemingen naar een matrixvorm voor statistische berekingen. Hierbij dienen twee aspecten onder ogen gezien te worden: (i) selectie van de kleinste hoeveelheid kenmerken waarmee de dataset het best kan worden beschreven en waaruit voorspellingen kunnen worden afgeleid; (ii) reductie van de dimensionaliteit van de dataset door het verwijderen van overbodige informatie.

In Hoofdstuk 3 van dit proefschrift wordt een filter-gestuurde aanpak voor kenmerkselectie aangedragen, waarin criteria uit de fuzzy logica worden gebruikt als één van de kostenfuncties. Daarbij wordt een zogenaamd multi-objectief evolutionair algoritme gebruikt voor het produceren van optimale oplossingen langs de Pareto grens. Deze aanpak verschilt van andere strategieën, daar de door ons werk geïntroduceerde methode voorziet in een set van kandidaat kenmerk combinaties. Met deze combinaties kan de besluitvormer zijn voordeel doen door de voor het specifieke geval meest waardevolle combinatie te kiezen.

(3) We hebben het gebruik van fuzzy systemen verkend voor het clusteren van informatie. Onder de patroonextractie methoden, i.e. samenvatting, associatie en predictie, is het clusteren van informatie van groot belang; dat geldt voor onderzoek alsook voor de dagelijkse praktijk. Dit is met name het geval voor een dataset die weinig of geen labels heeft.

In Hoofdstuk 4 van dit proeftschrift hebben we een nieuwe clustering methode verwezenlijkt waarbij de zogenaamde fuzzy rough c-means benadering en het deeltjes zwerm optimalizatie (PSO) algoritme worden gecombineerd. Deze combinatie integreert tot een zinnig globaal resultaat. Het concept van fuzzy logica kan omgaan met onzekerheid, vaagheid en partities die in de dataset overlappen, terwijl het PSO algoritme helpt met het zoeken naar vrijwel optimale oplossingen.

(4) We hebben fuzzy systemen onderzocht als onderdeel van een speciale analysemethodiek, dit is een samenvoeging van de benaderingen die in de vorige hoofdstukken behandeld zijn. De toepasbaarheid wordt gedemonstreerd aan de hand van een casus.

In Hoofdstuk 5 van dit proefschrift beschrijven we een case studie die gericht is op een overkoepelende data analyse van een 1-dimensionale gel-electroforese experiment. De voorgestelde analysemethodiek gebaseerd op fuzzy systemen heeft laten zien in staat te zijn tot nauwkeurige extractie van kenmerken uit beelden van eiwitgels die karakteristiek voor kankercellen. De beelden van de eiwitgels zijn verre van ideaal, vandaar dat een aantal correcties moeten worden toegepast zodat er juiste metingen uit deze beelden kunnen worden verkregen. Deze metingen kunnen worden gebruikt om de significante verschillen tussen kankercel-lijnen, en de daarbij horende groepen en subgroepen te onderscheiden en te karakteriseren. Dit draagt bij aan het verdere begrip met betrekking tot de ontwikkeling van kanker. Daarenboven kan de analysemethodiek worden gebruikt voor het verkijgen van begrip in eiwit/DNA patronen en hun karakteristieken zoals die in gelbeelden worden gezien.

Concluderend, het onderzoek beschreven in dit proefschrift heeft als doel het onderzoeken van fuzzy systemen in combinatie met patroonherkenning in het onderzoeksveld van de bioinformatica. Uit het bestuderen van bestaande methodologieën is een ensemble van fuzzy logica en unsupervised algortimen ontworpen en geïntgreerd in een analysemethodiek waarmee vragen met een oriëntatie in de biologie en in patroonherkenning kunnen worden begrepen en aangepakt.

论文扼要 (Chinese Summary)

本文主要研究了模糊系统在生物信息学领域内的应用。在过去的几十年里，越来越多的科研工作者将目光投入到生物信息学领域。因为在生物学领域，生物实验数据需要被精准而又不失鲁莽地分析；而在计算机领域，对不同类型的数据进行有价值的分析和开发以及设计出高效的处理算法是具有挑战的研究难点。

通常生命科学等领域研究得到的数据不精确或存在缺失，使其具有模糊性和未知性。针对生物数据的特性，模糊系统被应用到相关的研究中。模糊系统是一种较新的启发式算法，相较于其他更加关注计算结果精准度的算法，其更加注重计算结果的重要性。模糊系统是通过在分析数据时提出更多的假设条件，从而有效的简化分析数据中的复杂且非唯一性问题。

本文着重于设计高效可靠的启发式分析算法去挖掘并理解隐藏在生物实验数据中的重要信息。本文设计了一个基于模糊系统的数据分析流程，并在文中不同章节处详细阐述了模糊系统在不同场景的应用及其前景。最后，通过案例分析来展示本文所提出的数据处理的流程。本文亮点如下：

（1）本文使用模糊算法处理图像背景的噪声。通常，显微成像的数字图像会受到例如电子设备引起的工频噪声，或者实验过程中光照不均带来的叠加伪影的影响。所以，为了解决以上问题，一种新的基于图像形态学和模糊及粗糙系统的图像背景去除算法在本文第二章中被提出。与大多数使用的 EMI 方法以及商业软件中频繁使用的滚球算法相比，本文提出的 DCBC 算法被验证对于传统生物显微图像更具有鲁棒性。

（2）本文对模糊系统在数据中的特征提取和冗余信息剔除做了学习和研究。这个过程对于数据分析中的表象型数据到经典的统计数据模型的转换是至关重要的：一，该过程可以选择最少的特征去最优地描述整个数据集；二，通过去除冗余的信息来降低数据集的维度，从而一定程度上减少后期分析的计算量。在本文第三章中，一个基于模糊标准函数的“过滤”并“包装” 模式的特征提取算法被提出。与此同时，文中还采用了多目标优化函数来帮助原算法在“帕洛托”前沿寻找最优解析解。本文所提出的算法与一般特征提取算法不同之处在于，文中的算法可以给出一组最优解的解集让决策者根据自己的经验、需求选择相对应的可以描述原数据的特征集合。

（3）本文对模糊系统在信息聚类分析中的应用进行了研究。因为在所有模式提取流程中，即信息概述、信息联合、信息预测和信息聚类等，其中信息聚类是至关重要的一个环节；而且由于生活中的数据大多数是不带有或是带有少量标签信号，所以该聚类分析的方法论在科学研究及日常生产管理中都有着十分

广泛的应用。在本文第四章，通过结合模糊粗糙聚类算法和粒子群算法，提出了一种全新的且对于获得全局最优解非常敏锐的聚类算法。在这个新的算法中，模糊粗糙的概念可以有效地处理原数据中不确定的、模糊的以及重叠覆盖的成分；与此同时，粒子群理论又可以帮助聚类算法找到所有不同聚类结果中的最优解。

（4）本文基于前几章的研究内容与所设计的算法，整合并提出了一个新的以模糊系统为蓝本的数据处理流程。该处理流程包括图像的预处理（第二章），图像特征提取和选择分析（第三章）以及数据的聚类分析（第四章），并且该分析流程的可行性研究将通过一个特有的案例进行分析。在本文第五章，通过对一维电泳胶体成像数据的分析和研究来进一步了解文中提出的模糊分析系统对癌症细胞（蛋白和 RNA 等）数据处理流程的帮助。对这些电泳胶体成像结果的处理和分析有助于提取，分析和鉴别不同的细胞系（类）在图像中的表达特征。而通过研究图像中细胞蛋白/RNA 的位移和表达变化可以帮助理解并预测案例中关于癌症细胞的发展。

综上所述，本篇论文旨在研究与模式识别相结合的模糊系统在生物信息学领域中的应用。通过对已有的文献的考究，结合模糊逻辑与无监督学习的概念搭建了一个高效且鲁莽的数据处理流程平台，并在文末的案例分析中得以用来理解研究以生物学或模式识别为导向的问题。

Curriculum Vitae

Fuyu Cai, was born on November 16th, 1988, in Zhanjiang, Canton, the People's Republic of China. In 2007, he started his bachelor study in Biomedical and Information Engineering at the Northeastern University (China), focusing on 2D blood vessel (vein) image analysis. He received the bachelor degree under the supervision of Prof. Dean. Yan Kang and co-supervision of Prof. Bart M. ter Haar Romeny and Dr. Han van Triest.

In 2011, he received a full-grant National Scholarship to start his master study in Biomedical and Information Engineering at the Northeastern University (China). For his mater thesis, in collaboration with local hospital, he focused on cardiovascular blood fluid dynamic simulation based on ANSYS Workbench, and successfully verified Fractional Flow Reserve (FFR) score in a non-invasive environment. He received his master degree under the supervision of Prof. Dean. Yan Kang and co-supervision of Prof. Shouliang Qi.

In 2013, he received a full-grant Chinese Scholar Council (CSC) scholarship to start his Ph.D. research in Imaging and Bioinformatics group, Leiden Institute of Advanced Computer Science, Leiden University (the Netherlands), under the supervision of Prof. Fons. J. Verbeek. In collaboration with Aging and Signal Transduction group, Cell Molecular Biology, Leiden University Medical Center, his Ph.D. research focused on the designs of fuzzy system and unsupervised learning algorithms for bioinformatics data analysis.


Day Month Year
12 / 12 / 2017
20 / 08 / 2013
θ
Emmm ...
This is the
'Fuzzy' Math!
Fuyu, what is the duration?
...
This is?
Progress Report