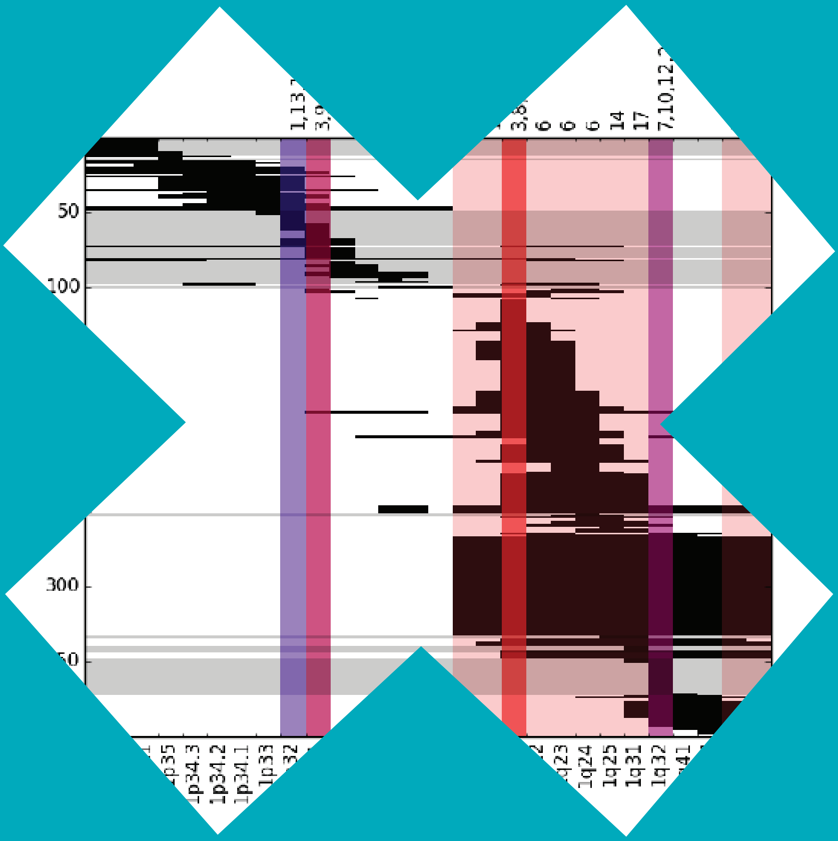
Department of Information and Computer Science

Probabilistic Modelling of Multiresolution Biological Data

Prem Raj Adhikari



Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 157/2014

Probabilistic Modelling of Multiresolution Biological Data

Prem Raj Adhikari

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the school on 21st November 2014 at 12 noon.

Aalto University
School of Science
Department of Information and Computer Science
Parsimonious Modelling Research Group

Supervising professor
Professor Samuel Kaski

Thesis advisor
D.Sc. (Tech.) Jaakko Hollmén

Preliminary examiners
Prof. Dr. Marko Bohanec, Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Olli Yli–Harja, Tampere University of Technology, Tampere, Finland

Opponent
Asst. Prof. Jeroen de Ridder, Delft University of Technology, Delft, The Netherlands

Aalto University publication series
DOCTORAL DISSERTATIONS 157/2014

ISBN 978-952-60-5903-7 (printed)
ISBN 978-952-60-5904-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5904-4

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Prem Raj Adhikari

Name of the doctoral dissertation
Probabilistic Modelling of Multiresolution Biological Data

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 157/2014

Field of research Information and Computer Science

Manuscript submitted 10 June 2014 **Date of the defence** 21 November 2014

Permission to publish granted (date) 14 August 2014 **Language** English

□ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

When the measurements from the ever improving measurement technologies are accumulated over a period of time, the result is a collection of data in different representations. However, most machine learning and data mining algorithms, in their standard form, are designed to operate on data in a single representation only.

This thesis proposes machine learning and data mining algorithms to analyse data in different representations with respect to resolution within a single analysis. The novel algorithms proposed to analyse multiresolution data are in the field of probabilistic modelling and semantic data mining. First, different deterministic data transformation methods are proposed to transform data across different resolutions. After the data transformation, the resulting datasets in same resolution are integrated and modelled using mixture models.

Second, similar mixture components in a mixture model are merged one by one repetitively to generate a chain of mixture models. A new fast approximation of the Kullback Leibler divergence is derived to determine the similarity of the mixture components. The chain of generated mixture models are useful for comparison purposes, for example, in model selection. Third, mixture components in different resolutions are iteratively merged to model multiresolution data generating models in each modelled resolution that incorporate information from data in other resolutions.

Fourth, a single multiresolution mixture model with multiresolution mixture components is proposed whose mixture components independently have the capabilities of a Bayesian network. Finally, a three part methodology consisting of clustering using mixture models, rule learning using semantic subgroup discovery, and pattern visualisation using banded matrices is developed for comprehensive analysis of multiresolution data.

The multiresolution data analysis methods presented in this thesis improve the performance of the methods in comparison with their single resolution counterparts. Furthermore, the developed methods aim to make the results understandable to the domain experts. Therefore, the developed methods are useful additions in the analysis of chromosomal aberration patterns and the cancer research in general.

Keywords Mixture Models, Chromosomal Aberrations, 0-1 Data, Model Selection

ISBN (printed) 978-952-60-5903-7 **ISBN (pdf)** 978-952-60-5904-4

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 168 **urn** http://urn.fi/URN:ISBN:978-952-60-5904-4

Preface

> *A man's life is merely a collection of events, building one upon the other. When all the events are tallied; the triumphs; the failures; the mistakes, their sum makes up the man.*
>
> — Neal Cassady
>
> *The Last Time I Committed Suicide, 1997*

The work presented in this thesis was performed at Parsimonious Modelling (PM) research group at the Helsinki Institute for Information Technology (HIIT), Department of Information and Computer Science (ICS) in Aalto University School of Science (AaltoSCI). I am deeply indebted to my instructor D.Sc.(Tech.) Jaakko Hollmén for his enthusiastic engagement in my research and his invaluable streams of advice, ideas, and support ranging from the minute technical details to the overall research ideas, as well as living a life. I am also grateful to Prof. Samuel Kaski for his supervision of the thesis.

I am thankful to Helsinki Doctoral Programme in Computer Science — Advanced Computing and Intelligent Systems (Hecse) for funding the last three years of my research. I am also thankful to HIIT and Finnish Centre of Excellence for Algorithmic Data Analysis Research (ALGODAN) for duration 2008–2013 which is itself funded by the Academy of Finland. They funded the first year of my graduate studies and together with Hecse, they also funded numerous conference trips and research visits in Finland and abroad. The research visits and conference trips have given me invaluable experiences in understanding research challenges and knowledge of the state of the art in my research area. It also provided me an opportunity to communicate with researchers both formally and informally helping me to network and collaborate.

A big share of thanks also goes to my colleagues in the PM research group (Mikko, Janne) and the whole ICS, HIIT, and ALGODAN including the support staffs for providing splendid working and research environment. I would like to thank everyone involved in teaching and organising the courses I attended for my graduate, and postgraduate degree. The courses have helped me prepare for my research and understand the state of the art in my field of research.

Collaboration with fellow researchers in the field has given me a deeper understanding in teamwork, work ethics, and work environment. Moreover, I have gained ideas from discussion with Prof. Nada Lavrač, Anže Vavpetič, Jan Kralj, Bimal Babu Upadhyaya, and Chen Meng. I thank the pre–examiners Prof. Dr. Marko Bohanec and Prof. Olli Yli–Harja for taking their precious time to review this thesis and providing invaluable comments to improve the thesis. I am grateful to Asst. Prof. Jeroen de Ridder, for the honour of having him as the opponent for my dissertation.

Outside of work, I would also like to thank my mates Gautam, Sandeep, Roshan, Bishal, and Subash. If it weren't for you guys, I might have graduated sooner but without you, I would have never graduated. My sincere gratitude also goes to my parents, family, and relatives for their everlasting encouragement and motivation. Last but not the least, I thank my wife, Manju, for being by my side, not only for the fun times, but also the ups and downs, and for being my support and my inspiration in all aspects of my life.

Espoo, October 22, 2014,

Prem Raj Adhikari

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Prem Raj Adhikari, Jaakko Hollmén. Patterns from Multiresolution 0–1 data. In *Jilles Vreeken, Nikolaj Tatti, and Bart Goethals, Editors, UP '10, ACM SIGKDD Workshop on Useful Patterns*, Washington DC, ACM, New York, NY, USA, Pages 8–16, July 25, 2010, DOI: 10.1007/s10844-013-0282-3, July 2010.

II Prem Raj Adhikari, Jaakko Hollmén. Fast Progressive Training of Mixture Models for Model Selection. *Journal of Intelligent Information Systems*, IN PRESS, Springer, DOI: 10.1007/s10844-013-0282-3, Published Online: December 2013.

III Prem Raj Adhikari, Jaakko Hollmén. Multiresolution Mixture Modeling using Merging of Mixture Components. In *Proceedings of Fourth Asian Conference on Machine Learning (ACML 2012)*, In Steven C.H. Hoi and Wray Buntine Editors, Volume 25 of Journal of Machine Learning Research—Proceedings Track, pages 17–32, November 4–6, 2012, Singapore, URL: http://jmlr.csail.mit.edu/proceedings/papers/v25/adhikari12.html, November 2012.

IV Prem Raj Adhikari, Jaakko Hollmén. Mixture Models from Multiresolution 0–1 Data. In *Proceedings of Sixteenth International Conference on Discovery Science (DS 2013)*, Johannes Fürnkranz, Eyke Hüllermeier, and Tomoyuki Higuchi, Editors, Volume 8140 of Lecture Notes in Computer Science, Springer–Verlag, Berlin Heidelberg, pages 1–16, October 6–9, 2013, Singapore. DOI: 10.1007/978-3-642-40897-7_1 , October 2013.

V Prem Raj Adhikari, Anže Vavpetič, Jan Kralj, Nada Lavrač, Jaakko Hollmén. Explaining mixture models through semantic pattern mining and banded matrix visualization. In *Proceedings of Seventeenth International Conference on Discovery Science (DS 2014)*, Sašo Džeroski, Panče Panov, Dragi Kocev, Ljupčo Todorovski, Editors, Volume 8777 of Lecture Notes in Computer Science, Springer International Publishing Switzerland 2014, pages 1-12, October 8–10, 2014, Bled, Slovenia. DOI: 10.1007/978-3-319-11812-3_1, October 2014.

Author's Contribution

Publication I: "Patterns from Multiresolution 0–1 data"

Generally, mixture models and pattern mining algorithms can handle only single resolution data in their standard form. We propose different deterministic data transformation methods to transform datasets across different resolutions facilitating the integration of datasets. The integrated datasets are in single resolution. We then use pattern mining algorithms such as the maximal frequent itemset and probabilistic modelling methods such as mixture models to identify and compare the patterns and performance of the algorithms in different resolutions of data.

Forming of the original idea and designing of the methodology for the research are performed jointly by the authors. The current author implemented and performed all the experiments and wrote most of the manuscript. The second author suggested the corrections to the manuscript. The current author also presented the contribution at the conference.

Publication II: "Fast Progressive Training of Mixture Models for Model Selection"

Expectation Maximisation (EM) algorithm is a popular algorithm to learn the maximum likelihood parameters of the mixture model. However, EM algorithm requires apriori knowledge of the number of component distributions in the mixture model to learn the maximum likelihood parameters of the mixture model. This is often unknown apriori in most situations. In this publication, we propose an algorithm to efficiently train a series of mixture models each with different number of mixture components suitable for comparisons during model selection.

The authors are jointly responsible for the original idea of the contribution. The current author performed all the experiments and wrote most of the manuscript. The second author suggested corrections on the manuscript. The current author also presented an earlier version of this contribution [2] in a conference.

Publication III: "Multiresolution Mixture Modeling using Merging of Mixture Components"

In this contribution, we propose an algorithm to model multiresolution data by merging the similar components from different mixture models in different resolutions. The mixture models are generated in each data resolution separately but they incorporate the information from the data in other resolutions.

The current author is responsible for forming the original idea, and methodology of the work. The current author also performed the all experiments and wrote most of the manuscript. The second author provided useful suggestions and corrections to the manuscript. The current author also presented the research in the conference.

Publication IV: "Mixture Models from Multiresolution 0–1 Data"

In this contribution, we propose a multiresolution mixture model consisting of multiresolution mixture components. The structure of multiresolution mixture components are determined from the domain ontology which is known apriori. The individual mixture components provide the functionality of Bayesian networks.

The authors are jointly responsible for the original idea and designing the methodology for the research. The current author performed the experiments and wrote most of the manuscript. The second author suggested corrections to the manuscript. The current author also presented the contribution in the conference.

Publication V: "Explaining mixture models through semantic pattern mining and banded matrix visualization"

In this contribution, we propose three part exploratory approach to analyse multiresolution data. The three parts consist of clustering using mixture model, extracting rules from clusters using semantic data mining, and simultaneous visualisation of the clusters from mixture models and the rules from semantic data mining algorithm using banded matrices. The semi–automated methodology proposed in the contribution aims to provide exhaustive analysis of a complex real world multiresolution data.

The authors are jointly responsible for the idea and designing of the research methodology. The current author implemented the clustering part of the three part explanatory process. Writing the paper was a collaborative effort of all the authors.

List of Abbreviations

3Vs	Volume, Velocity, and Variety
aCGH	array Comparative Genomic Hybridization
AIC	Akaike Information Criterion
BAC	Bacterial Artificial Chromosome
BIC	Bayesian Information Criterion
cDNA	complementary Deoxyribonucleic Acid
CGH	Comparative Genomic Hybridization
CNV	Copy Number Variation
CPD	Conditional Probability Distribution
DNA	Deoxyribonucleic Acid
EB	exabytes
EM	Expectation Maximization
E–Step	Expectation Step
FISH	Fluorescence In Situ Hybridization
FP	False Positives
G–banding	Giemsa banding
GMM	Gaussian Mixture Model
GrC	Granular Computing
ICL	Integrated Classification Likelihood
IID	Independent and Identically Distributed
ISCN	International System for Human Cytogenetic Nomenclature
kbp	Kilo base Pairs
KL	Kullback Leibler
LHC	Large Hadron Collider
MAFIA	MAximal Frequent Itemset Algorithm
MAP	Maximum a Posteriori Probability
Mbp	Mega base Pairs

MCMC	Markov Chain Monte Carlo
MDL	Minimum Description Length
MGMM	Multiresolution Gaussian Mixture Model
MLE	Maximum Likelihood Estimate
MM	Malignant pleural Mesothelioma
MPSS	Massive Parallel Signature Sequencing
mRNA	Messenger Ribonucleic Acid
M–Step	Maximization Step
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
SOM	Self–Organizing Maps
TMA	Tissue Microarrays
TS–SOM	Tree Structured Self–Organizing Maps
TP	True Positives
WHO	World Health Organization

Chapter 1

Introduction

> *Data does not equal information; information does not equal knowledge; and, most importantly of all, knowledge does not equal wisdom. We have oceans of data, rivers of information, small puddles of knowledge, and the odd drop of wisdom.*
>
> — Henry Nix
>
> *Keynote address, AURISA, 1990*

Synopsis

This chapter conceptualises the topic of this dissertation with respect to the methodology and application. The chapter also covers the motivations for research, contributions of the thesis to the scientific community, and organisation of the chapters of the thesis.

1.1 Data Explosion

Dictionary definition of data is a piece of information that ranges from the values or measurements of quantitative and qualitative variables to the description of objects or phenomenon [37]. In computing terms, data is any digitally stored information. Throughout history, data was universal, and found everywhere. However, only employees generated data in computing terms by keying in the handwritten information. Nowadays, users generate data on their own, for example, social network statuses or photos, thereby increasing the amount of data produced. Furthermore, new machines such as automatic climatic conditions recorder and technologies such as Large Hadron Collider (LHC) produce colossal amount of data [104]. This astronomical increase in the amount of data is referred to as big data [104, 108]. Modern science revolves around the methods and ways to analyse the data generated in their field to stimulate scientific discoveries.

Production of data these days is such humongous that it surpasses the estimates of Moore's Law [123]. For example, 5 exabytes (EB) (1 EB= 10^{18} bytes = 1 billion gigabytes) of data was generated from the dawn of civilisation until 2003. Today, we create 5EB of data every two days [140]. Three properties: Volume, Velocity, and Variety (often referred to by 3Vs) define the big data. The volume of data and speed at which they arrive and leave the real time systems provide challenges in big data analysis. In addition, variety in the collected data also poses considerable challenges to research in big data.

Over the years, measurement technology has progressed enormously, and produces variety of data in addition to the large volumes of data because each cycle of improvement in measurement technology produces data in a different representation. The variety is the aspect of big data that is closest to the topic of this thesis. Nowadays, individual dataset in the sets of datasets often have higher dimensionality, d, than the number of samples, N, i.e., $d \gg N$. Therefore, challenge in big data analysis is large temporal, and/or spatial data dimensions which results in the curse of dimensionality [17]. Traditional algorithms succumb to the challenges posed by small sample high dimensional datasets. Therefore, it is imperative to develop novel methods to analyse multiple datasets, i.e., sets of datasets in different representations within a single analysis.

Biology is one of the largest producer of big data which necessitates novel computational methods to analyse such wealth of data and to convert data to knowledge and wisdom [74, 111]. There are varieties of biological phenomena often interlinked with one another making variety aspect of big data prevalent in biological data source. This tremendous increase in biological data coupled with the variety is impossible to interpret using visual analysis. Instead, it requires novel computational methods for thorough understanding of the biological phenomenon. The growth of biological data has produced both opportunities and challenges for researchers to develop algorithms and analysis methods in computational domain to extract biological meaning from vast amounts of data.

1.2 Machine Learning and Data Mining

Machine Learning is a core sub–area of artificial intelligence that intersects the discipline of computer science, and statistics. The aim of machine learning is to develop algorithms that learn from the observed data, and use the experience to improve the performance [9, 23, 68, 120]. Machine learning includes a myriad of statistical, probabilistic and optimisation, and induction algorithms that are applicable in different tasks such as classification, regression, clustering, and pattern discovery. Data mining, also known as knowledge discovery, is the process of extracting useful information such as patterns, from unstructured and enormous sets of data by analysing data from different perspectives [67].

Machine learning and data mining complement each other and it is difficult to make a clear distinction between the two. Nonetheless, machine learning algorithms are often used in the data mining process. Machine learning and data mining, although a new discipline, has a large active research community. The community has already developed a cohort of fascinating algorithms and methods to treat the concept classes, and elegant and clever ways to search through databases. Hence, machine learning and data mining methods can address the challenges posed by data intensive disciplines such as biology.

In application areas such as biology, the number of training samples are often limited even in the age of big data. In contrast, the data dimensionality increases considerably. For example, in genetics, number of cancer patients is constant while the new technology can measure the finer units of the phenomenon generating data with large dimensionality. The implication of increasing dimensionality is that, with a limited size of training samples, the performance of the algorithm deteriorates as the number of features increases. This phenomenon is also called Hughes phenomena, or Hughes effect [76] or more generally as a curse of dimensionality [17].

1.2.1 Mixture Model

A mixture model is a probabilistic modelling technique in machine learning and data mining community which models a data distribution under the assumption that all the data points are generated from a mixture of parametric probability distributions [23, 45, 115]. Apart from this assumption of data origination, mixture models are flexible probabilistic models with varying uses such as model based clustering, classification, image analysis, and collaborative filtering in analysis of high dimensional data. Mixture models are suitable for the choice of any probability distributions such as the Gaussian, Bernoulli, Poisson, and Dirichlet. In this thesis, mixture models analyse multiresolution data probabilistic clustering setting. Chapter 3 discusses mixture models in detail.

1.3 Challenges of Multiresolution Data

Measurement of physical phenomenon such as distance, weight, and time started since the time immemorial and has been the cornerstone of our knowledge and learning [92]. Measurement has also become integral part of our everyday life. The inventions and discoveries of the modern world would cease to exist in absence of measurement technology. The measurement technology has been continuously improving over the years. Result of a measurement process is generation of the data. The older generation technologies measure only the coarser unit of the phenomenon generating data in coarse resolution. In contrast, the newer generation technologies measure the finer units of the phenomenon producing the data in fine resolution [44, 54, 110]. Resolution here defines the amount of information in each data sample, i.e., the level of detail.

Multiresolution data is generated when the same phenomenon is measured in different levels of detail [11, 54, 165]. Thus, the multiresolution data describes the same phenomenon in different data representations. Different data representation is a broader challenge in machine learning and data mining community where datasets are represented in different forms such as audio, video, image, table, and text. This thesis concentrates on different data representation only in the context of dimensionality, i.e., datasets are same except for the data dimensionality. Nevertheless, the proposed algorithms and methods can possibly be extended to other different data representations. The measurement of time is one of the simplest illustrations of multiresolution data. We can measure time in fine units such as seconds and minutes producing data at a fine resolution. In contrast, we can also measure time in coarse units such as months, and years producing data in coarse resolution.

Multiresolution data often forms a part of hierarchy as shown in Figure 1.1. For example, the world is a collection of different continents such as Asia, Europe, and Africa. This generates coarser view of data. Similarly, the world is also a collection of different countries such as Singa-

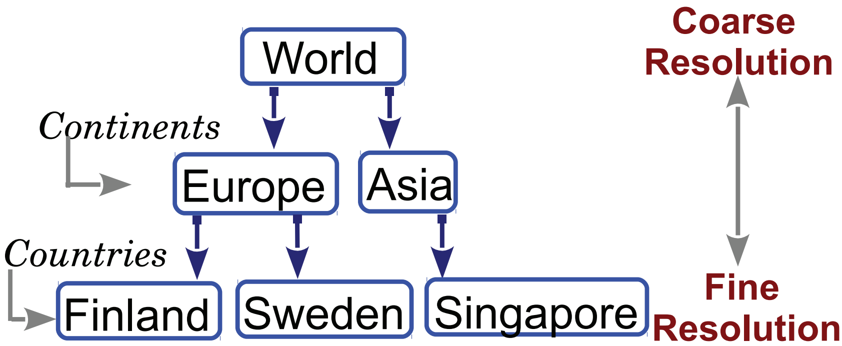


Figure 1.1. Example of part of hierarchy in real world scenario. The figure shows the geographical division as the part of hierarchy which when measured results in multiresolution data. The world is divided into continents and continents into countries.

pore, Finland, and Sweden. These countries can be grouped into different continents. Furthermore, these countries can be divided into municipalities, and the municipalities into streets, and the streets into blocks. This hierarchy forms a multiresolution data which represents a part of hierarchy [139]. This division of the world is just an illustrative example, as the sources of multiresolution data are varied, for example, telecommunications, hydrology, and biology [165]. Chapter 2 discusses multiresolution biological data used in the experiments of the thesis.

1.4 Contributions of the Thesis

This thesis addresses an important challenge encountered in data analysis: what should be done when the data to be analysed are represented differently. The thesis presents different frameworks and methods amalgamating probabilistic modelling and pattern mining domain. The presented methods handle irregular, and heterogeneous division of data in different representations. The major scientific contributions in this thesis are summarised in the following list.

- Different deterministic data transformation methods are proposed to transform the multiresolution datasets from one resolution to another. The transformed datasets in same resolution can be integrated and modelled in same resolution.

- A computationally efficient algorithm is proposed to train a series of mixture models to aid model selection. The trained mixture models in the series differ in number of components but are otherwise similar to each other. This provides an effective means to compare different model selection criteria such as likelihood, AIC, and BIC using different model selection techniques such as cross–validation.

- A mixture modelling solution is proposed to model multiresolution data by merging the mixture components of different mixture models in different resolutions. The proposed mixture modelling solution initially trains a mixture model in each resolution and merges the similar mixture components across different resolutions to incorporate information in multiple resolutions.

- An algorithm that uses domain ontology, known apriori, to determine multiresolution mixture components of the mixture model is proposed to build a single mixture model for multiresolution data. Each individual mixture component is a fully functional Bayesian network.

- A three part methodology is proposed to analyse the multiresolution data blending clustering using mixture models, pattern mining using semantic data mining, and visualisation using banded matrices.

1.5 Organisation of the Thesis

The thesis consists of two parts: an introductory part consisting of six different chapters and publications. In the introductory part, this chapter introduces the research domain, and the Chapter 2 introduces multiresolution data with a focus on cancer genomics, and reviews the previous work in multiresolution analysis and the related areas. Chapter 3 describes mixture models and model selection in mixture models. It also summarises our contribution for efficient training of a series of mixture models (Publication II).

Chapter 4 forms the crux of this thesis and discusses our contributions in multiresolution modelling. First, multiresolution data is modelled using deterministic data transformation methods for data integration (Publication I). Second, multiresolution data is modelled by merging the similar mixture components of different mixture models in different resolutions. The merging of mixture components models the interaction between the models in different data resolutions (Publication III). Third, a multiresolution mixture model having multiresolution mixture components is proposed to analyse the multiresolution data with a single mixture model. Structure of multiresolution components is known from the domain ontology (Publication IV). Finally, a comprehensive solution for the analysis of multiresolution data is provided using three part methodology comprising of clustering, semantic pattern mining, and banded matrices (Publication V). Chapter 6 summarises the findings, presents the conclusions of the research, and also outlines the possible future work related to the topic of the thesis.

CHAPTER 2

MULTIRESOLUTION DATA AND ANALYSIS METHODS

> *Data matures like wine, and the applications like fish.*
>
> — JAMES GOVERNOR
>
> *James Governor's Monkchips, 2007*

Synopsis

This chapter describes the application area and the dataset used in the experiments. The chapter also describes the usefulness of domain ontology in data analysis; and the multiresolution data in the domain of biology. Finally, the chapter also briefly reviews the literature and discusses the related areas of multiresolution modelling.

Human beings are diploid organisms having two homologous copies of each chromosome one each inherited from each parent. Copy Number Variations (CNVs) are structural variations in genome such that a region on the genome will have different copies of DNA [146]. In human beings, normal copy number is two because each child inherits one copy from each parent. Deletion or loss is the condition when the copy number is less than two. Duplication or gain is the condition when the copy number is more than two. Similarly, amplification is the condition when the copy number increases to more than 5. Some of the cancer patients have shown more than hundred copies [158]. There are other different kinds of variations but this thesis concentrates on copy number aberrations.

2.1 Chromosomal Aberrations in Cancer

Cancer is a heterogeneous collection of diseases characterised by abnormal and uncontrolled growth of cells; their ability to migrate to other parts of human body and destroy neighbouring cells and tissues [24]. Cancer rates have been increasing rapidly around the globe. Recent World Health Organisation (WHO) report showed that number of cancer patients escalated to 14.1 million in 2012, and cancer was responsible for 8.2 million deaths in 2012 [147]. The menace of cancer is increasing and WHO estimates that cancer will rise by 57% worldwide in the next 20 years signalling an imminent human disaster. The cost of cancer is also increasing rapidly. In 2010, estimated global cost of cancer reached approximately 963 billion euros [147], which is nine times more than the total budgeted expenditure of Finland.

A wide range of genetic mutations and molecular mechanisms known as chromosomal aberrations are identified as the hallmarks of disorders such as Cancer, Schizophrenia, and infertility [158]. In cancer research, identification and characterisation of chromosomal aberrations are crucial for studying and understanding pathogenesis of cancer. Moreover, study of chromosomal aberrations provides necessary information to select the optimal target for cancer therapy on individual level [91]. Study of chromosomal aberrations also has other clinical applications such as studying multiple congenital abnormalities and identifying the family history of Down syndrome [130].

2.2 Measurement Technology in Biology

Years of evolution and adaptation have made organisms complex biological beings [116]. Ever improving measurement technologies have also provided the facilities to measure the complex phenomena in biology [41]. After the discovery of DNA in 1953 [162], different measurement methods have been proposed to measure the genome. First sequence of lac operator of 24 bp was published twenty years after the discovery of DNA in 1973 [58]. Figure 2.1 summarises the evolution of DNA sequencing technology from the 1973. Initially, different banding methods such as G–banding and Q–banding technologies were developed to produce a visible karyotype by staining the chromosomes [19]. A karyotype here denotes the set of all chromosomes in an organism. Alongside the banding tech-

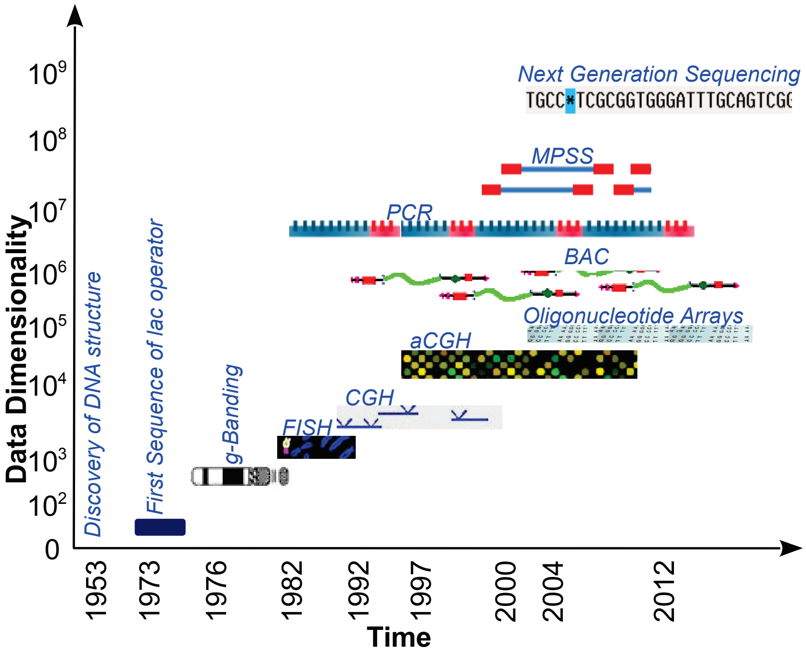


Figure 2.1. Evolution of measurement technology in biology described in terms of their level of detail in measurements and time of usage.

nology, FISH (Fluorescence In Situ Hybridisation) was developed to detect the presence or absence of DNA sequences on chromosomes. Similarly, microarray technologies such as the Comparative Genomic Hybridisation (CGH) [85] and array Comparative Genomic Hybridisation (aCGH) [134] were developed to study the Copy Number Variations (CNV) without requiring culturing of cells. Additionally, Bacterial Artificial Chromosome (BAC) was developed to sequence the genomes of organisms.

Similarly, Oligonucleotide arrays that uses oligos of short lengths (less than 25 bases) were developed to test large number of oligos in presence of smaller number of targets [103]. In addition, promoter arrays were developed to probe thousands of promoter sequences in one array experiment [161]. Besides, Massive Parallel Signature Sequencing (MPSS) was developed to analyse the level of gene expression by identifying and quantifying Messenger Ribonucleic Acid (mRNA) transcripts in the sample [25]. Likewise, Polymerase Chain Reaction (PCR) were developed to amplify one or small number of copies of DNA thereby generating large number of copies of particular DNA sequences useful for biomedical application such as DNA sequencing and diagnostic purposes [12].

Around the beginning of this century new technology known as next generation sequencing (NGS) had resounding impact in DNA sequencing. In [109] and [110], authors summarise the improvement in DNA sequencing which has positive impact on the biomedical research providing high throughput and high resolution techniques to explore, and answer genomewide biological questions. The Carlson curve accurately predicted the doubling time of DNA sequencing technologies measured in terms of cost and performance [27]. Furthermore, the curve illustrates the dramatic decrease in cost which is sometimes hyperexponential and similar dramatic improvements in technology to measure biological phenomenon such as DNA sequencing and synthesis, gene expressions, and protein structures.

These improvement in measurement technology in biology over the period of time produces data in different representation. Consequently, multiresolution data are also present in biology. For example, measurements from an older generation technology (eg. G–banding) can be represented in data with dimensionality in hundreds [19, 143]. In contrast, newer generation technology such as microarray measures the same karyotype generating the data of dimensionality of thousands [85, 134]. In addition, latest technology known as Next Generation Sequencing produces the data with millions of dimensions [109, 138].

The Figure 2.1 shows major changes in sequencing technology. However, within each generation of technology there are several minor improvements. For example, aCGH improves the mapping resolution of 20Mbp (Megabase Pairs) to 100 Kbp (Kilobase Pairs) over its predecessor CGH. Similar methods within a generation of technology also produce data in different resolutions because of improvements within the technology such as microarrays and banding. For example, authors in [155] use microarray data in two resolutions of 44000 and 244000 measurements per microarray measured by Agilent 44B and 244A aCGH platforms to classify different types of leukemias. Similarly, in NGS, different vendors have produced different sequencers for commercial use [102].

Studying the data generated by different technologies above produces wide range of benefits, especially in understanding of the biological phenomenon. Therefore, computational methods have been used to analyse the generated data. The phenomenon of doubling of number of transistors in a chip within 18 to 24 months, often known as Moore's Law [123], has improved the processing power of computers exponentially. Similarly, with the advent of Internet and other communication technologies and protocols; communication systems have also improved dramatically. The data storage capacity is also rapidly rising. These advancements have resulted in improved computing power thus facilitating development of novel algorithms to analyse the generated data.

Pan–cancer Analysis

In addition to the data in different resolutions, efforts have been made to study different cancer types by collecting data from different sources in pan–cancer initiative [127]. The aim of the study is to develop an integrated picture of commonalities, differences, and emergent themes across tumour lineages. The initiative involves multiple datasets and multiple cancers showing possible utility of multiresolution methods in pan–cancer initiative. In the previous research of our research group, we have considered all cancers within a single analysis [125].

2.3 Multiresolution Chromosomal Amplification Data

Similar to the array technology and next generation sequencing, the International System for human Cytogenetic Nomenclature (ISCN) has defined five different resolutions of the chromosome namely: 300, 400, 550, 700, and 850 in G–banding [143]. Each resolution defines the precision in division of karyotype. For example, in coarse resolution, a karyotype is divided into 312 (≈ 300) different regions, i.e., with lower precision. In contrast, in fine resolution, a karyotype is divided into 862 (≈ 850) different regions, i.e., with higher precision compared to resolution 300.

Figure 2.2 shows five different resolutions in chromosome 21 according to the ISCN standard. The figure depicts the division of regions and chromosome nomenclature with an example in chromosome 21. Chromosome 21 is chosen for visualisation because it is the smallest chromosome. Chromosome 21 is divided into 8, 8, 10, 12, and 14 regions in resolution 300, 400, 550, 700, and 850. The nomenclature of the regions and their division in different resolutions are irregular and hierarchical [143]. Some regions are undivided whereas other regions are divided into different number of regions. For example, the regions 21p12 and 21p13 are undivided in all the resolutions where as the region 21q22 is divided into 3 and 5 different regions in resolution 550 and 850. This division of karyotype

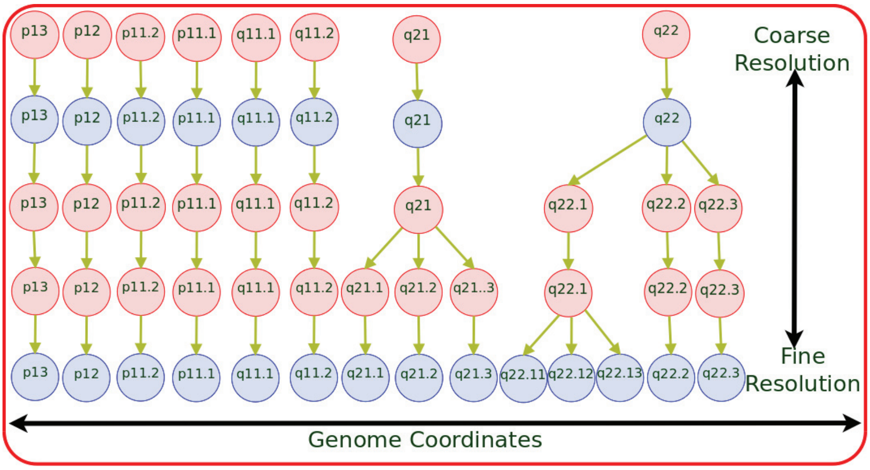

p13 p12 p11.2 p11.1 q11.1 q11.2 q21 q22
Coarse Resolution
p13 p12 p11.2 p11.1 q11.1 q11.2 q21 q22
p13 p12 p11.2 p11.1 q11.1 q11.2 q21 q22.1 q22.2 q22.3
p13 p12 p11.2 p11.1 q11.1 q11.2 q21.1 q21.2 q21..3 q22.1 q22.2 q22.3
p13 p12 p11.2 p11.1 q11.1 q11.2 q21.1 q21.2 q21.3 q22.11 q22.12 q22.13 q22.2 q22.3
Fine Resolution
Genome Coordinates

Figure 2.2. A typical relationship between multiple resolutions of genome. Figure shows chromosome 21 in five different resolutions of genome as defined by ISCN standard. The division is irregular, and hierarchical but consistent because of the ISCN standard. Chromosome 21 is chosen for the clarity of the presentation because it is the smallest chromosome. Y–axis denotes different resolutions of genome while x–axis denotes spatial coordinates (different regions) of the genome. Figure is adapted from Publication IV.

in different levels of detail allows measurement technologies to generate data in multiple resolutions. Each chromosomal region in coarse resolution is related with a chromosomal region in fine resolution with a one to many relationship. Given the measurements of same subject in two different resolutions, the aberrations should be consistent with each other, i.e., the aberrations should be the same except for some measurement errors.

For the experiments, two different datasets were available in coarse resolution and fine resolution. Researchers at the University of Helsinki compiled a dataset of chromosomal amplification in coarse resolution reading through the literature published between 1992–2002 [94]. All 838 journal articles were read through manually. The data describes the chromosomal amplification patterns of 4590 cancer patients in coarse resolution, i.e., resolution 400 where a karyotype is divided into 393 different regions. Similarly, data in fine resolution extracted from [13, 14] describes chromosomal amplification in fine resolution, i.e., resolution 850 where a karyotype is divided into 862 different regions. Since the cancer patients were not the same, there is no direct correspondence between data samples in two different resolutions. Therefore, most of the analysis methods discussed in this thesis consider unsupervised methods which learn the hidden structure in the data without the help of the class labels [23, 120]. If the measurements were available from the same cancer patients in two

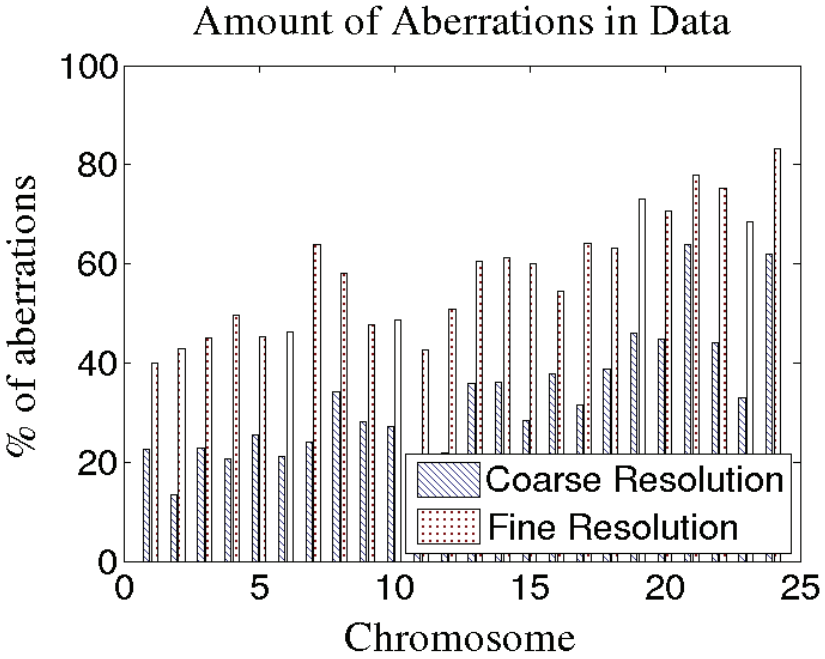


Figure 2.3. Amount of aberrations in each chromosome in two datasets in different data resolutions. The bar diagram shows that chromosomes in fine resolution are comparatively more aberrated than the coarse resolution.

different resolutions, we can expect consistent matching in aberrations except for measurement errors.

In the coarse resolution data, a total of 26527 (out of 1,803,870) matrix elements are aberrated which accounts for approximately 1.5% of the total matrix elements in the dataset. In all our experiments, we process the data chromosomewise to reduce the data dimensionality and with the expectation of finding chromosome specific patterns to describe different cancers. When the data is divided into each chromosome, there are some samples which do not show aberration in any of the chromosomal regions. Such data samples are deleted as we are interested in modelling chromosomal aberration patterns and their relation to cancer, not their absence. Therefore, number of samples and data dimensionality in each chromosome is different. We therefore calculate percentage of aberrations in each chromosome in each resolution for comparison purposes. Figure 2.3 depicts the amount of aberrations in both coarse resolution and fine resolution data. Data in fine resolution shows more aberration than the data in the coarse resolution. The percentage of aberrations are approximately 50% overall, while the minimum and maximum are approximately 15% and 80% respectively.

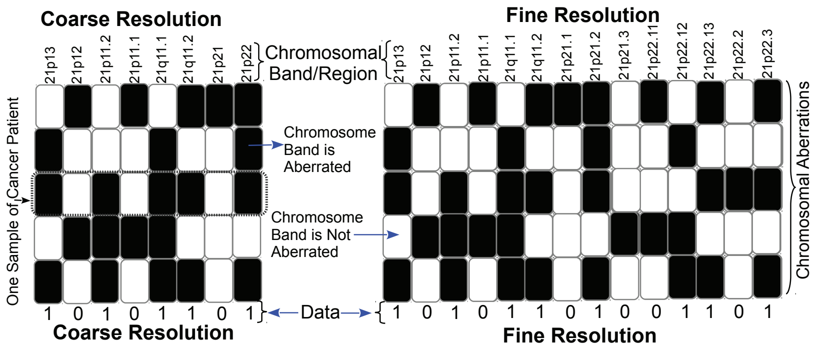


Figure 2.4. Visualisations of data describing chromosome 21 in two different resolutions: 300, and 850. Each sample, i.e., each row denotes one cancer patient and each column determines a chromosomal region. The black colour denotes presence of amplification and white col or denotes the absence of amplification. The two different panels in the figure depict the same phenomenon measured at different resolutions. Some chromosomal regions (variables or features in machine learning terms) such as 21p21 in left panel have been divided into different regions such as 21p21.1, 21p21.2, and 21p21.3 in the right panel. Figure is adapted from Publication II.

Figure 2.4 depicts five samples of data from chromosome 21 in both the coarse and the fine resolution. In the Figure 2.4, rows denote the cancer patients and the spatial coordinates on the X–axis denote the chromosomal region. In addition, white colour denotes value of zero (0), i.e., the absence of amplification, and black colour denotes the value of one (1), i.e., amplification in that specific region of genome for that specific cancer patient. The left panel of the Figure 2.4 shows that one region 21p21 in coarse resolution is divided into 3 regions in the fine resolution: 21p21.1, 21p21.2, and 21p21.3 as shown in the right panel of the figure. In contrast, some of the regions such as 21p13 and 21p12 are same in both coarse and fine resolution. Some regions are undivided while other regions are divided into varying number of regions. Nevertheless, the division is consistent because of the ISCN standard. Detailed description of the amplification dataset in coarse resolution can be found in [125].

2.4 Ontology of Multiresolution Data

The concept of ontology transcends back to the dates of noble philosophers Aristotle, Parmenides, and Jacob Lorhard, who used the term ontology in the philosophical context to describe the state of being, and reality [34]. Recently, the term ontology has found its prominence in computer and information science community. In computer science community, ontology is the mechanism for explicit description of the conceptualisation of the knowledge represented in the knowledge base [63, 151].

Ontology is a popular methodology to describe the semantics of the data in machine learning and data mining community [132]. Recent studies have shown that relevant additional knowledge enhances the knowledge discovery process of empirical data [132]. Expansion of semantic web and increasing availability of domain knowledge as ontologies has resulted in growth of semantic data. Semantic data mining algorithms address the challenge of mining abundance of knowledge encoded in domain ontologies constrained by the heuristics computed from the empirical data [157].

Multiresolution data conceptualises one of the essential ontological dichotomies of universals and particulars in metaphysics [57, 139]. The data in the coarse resolution can be conceptualised as universals whereas data in fine resolution can be conceptualised as particulars. Therefore, we can use ontological information in modelling multiresolution data as in Publication IV and Publication V.

Biological systems are complex consisting of many interwoven subsystems that effect the functionalities of each other [89]. As a result, chromosomal amplifications can effect, and be effected by other biological phenomenon. Furthermore, cancer is a multifactorial disease and the heterogeneity of cancer also suggests that biological phenomenon besides chromosomal aberration can catalyse the development of cancer. For this reason, additional background knowledge in biology was used to enhance the comprehensive analysis of chromosomal amplification datasets and to help understand the phenomenon of cancer. The additional knowledge used in the analysis of multiresolution data are the taxonomy of hierarchy of chromosomal regions, the cancer genes, virus integration sites, fragile sites, and amplification hotspots in Publication V. Only taxonomy of hierarchy of regions is used as background ontology in Publication IV.

The mutations in genes resulting to a larger extent by “acquired mutation” and to a lesser extent by “germline mutation”, known as cancer genes, are one of the most prominent causes of cancer [49]. Authors in [49] have listed the cancer genes and compared them to the complete gene set revealed by the human genome sequence. Similarly, fragile sites are non-randomly distributed loci on human chromosome that show a constriction or a gap and increased frequency of chromosome breakage under the conditions of partial replication stress [42, 141]. The fragile sites are often found rearranged in cancers [60]. Virus integration sites are the locations in chromosome where the viral Deoxyribonucleic acid (DNA) inserts into host cell DNA [88]. Viruses are responsible for approximately 12% of cancers [88, 172]. Amplification hotspots are frequently amplified chromosomal locations in cancer patients identified using computational modelling in [125]. The semantic data mining methods use these additional knowledge to enhance the knowledge discovery process in Publication V in semantic subgroup discovery framework.

2.5 Pattern Mining

Pattern mining is a popular branch of data mining that aims to extract interesting, relevant, and meaningful patterns from the data [66, 67]. Frequent itemset mining is one of the first and most popular pattern mining algorithm. Itemsets are a set of items or columns in a 0–1 dataset having high concentration of 1s and are used as patterns in a 0–1 dataset [152]. Let $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n$ be the attributes (items) of a dataset, $\mathcal{D}$, and σ be the given support. A frequent itemset is a set $\mathcal{F}$ of items of $\mathcal{D}$ such that at least a fraction of σ of the rows of $\mathcal{D}$ have 1 in all columns of $\mathcal{F}$ [4, 106].

Anti–monotone property of frequent itemset suggests that if an itemset is frequent, then all its subsets are also frequent [51]. Hence frequent itemsets generate a larger number of patterns making it difficult to report and interpret the results. Maximal frequent itemset ameliorates challenges posed by larger number of patterns in frequent itemsets. An itemset is maximal frequent if none of its immediate supersets is frequent [26]. We use maximal frequent itemset in Publication I to compare and report patterns across different resolutions.

Similarly, association rule is a popular data mining methodology to determine the interesting relations between variables based on different measures of interestingness [5, 72, 93, 133]. Most initial studies in association rule mining focused on finding interesting patterns from the large databases in an unsupervised setting. Nevertheless, association rules have been used in classification [82, 101]. Continuing with the research on association rules and classification, domain of subgroup discovery has emerged as a popular data mining methodology for labelled data. Subgroup discovery aims at finding interesting rules from the data that best describe the target variable [53, 71, 129]. Additionally, contrast set mining aims to learn the variables that differentiates one group of target variables from the rest, i.e., the most discriminating sets of variables [16, 129].

Semantic data mining method is a branch of data mining that uses taxonomies and ontologies of background data to improve the performance of algorithms [98, 156, 157]. Semantic data mining has recently gained research interest in pattern mining community because of the availability of large amount of data in the form ontologies encoded in semantic web [98]. Especially, the additional knowledge are abundantly available in biology as discussed in Section 2.4. In Publication V, we use semantic data mining algorithm to explain the clustering results obtained by probabilistic clustering using background knowledge discussed in Section 2.4.

2.6 Related Work in Multiresolution Data Analysis

Multiresolution analysis and modelling research community is growing steadily because of the pragmatic approach in dealing with datasets in different representation within a single analysis and also because of the increasing availability of multiresolution data in different application areas [11, 69, 80]. For instance, authors in [136] have improved the efficiency of boosting algorithms in regression and classification, using the model–driven and data–driven multiresolution strategy. Similarly, multiresolution trees have been used for object recognition in homogeneous data based on recursive neural networks [21]. In addition, multiresolution visualisations have been used to visualise large volumes of complex data using semantic analysis to infer increasing levels of meaning from the data [79]. Similarly, tree structured self–organising maps (TS–SOM) have been proposed in the literature as a multiresolution representation of several self–organising maps (SOMs) [95].

Multiresolution Probabilistic Models

Multiresolution modelling has also received research interest in probabilistic modelling domain. Most of the focus in this thesis has been the use of probabilistic models, namely mixture models, to analyse multiresolution data. Traditional machine learning and data mining methods, such as mixture models, are unable to analyse multiresolution data in their standard form because of the difference in representations of data in different resolutions. The only approach to model multiresolution data is to model each resolution separately and at best compare the results. Nevertheless, multiresolution models have found their usage in the literature, especially, in the image processing domain. For example, multiresolution kd–trees have been used to improve the performance of mixture models and reduce the cost associated with the Expectation Maximisation (EM) algorithm [122]. Similarly, multiresolution kd–trees have also been used to build robust models against the outliers using the EM algorithm [128]. Similarly, multiresolution binary trees have been used to store probability values efficiently both in terms of time and memory [18].

Authors in [124] have improved the performance of Gaussian Mixture Model (GMM) using wavelet subbands with an additional feature of incorporating variable number of components in the GMM. The GMM in [124] can use any multiresolution based decomposition for background suppression. Authors in [166] show that Multiresolution Gaussian Mixture Model (MGMM) adapts to smooth motions. The authors then apply the MGMM to estimate the visual motion. Similarly, authors in [117] propose efficient algorithms to learn a mixture of trees model in a maximum likelihood and Bayesian network framework for discrete multidimensional domains.

Related Areas

Multiresolution analysis and modelling shares commonality with various research areas and applications. The following sections briefly review the work on multiresolution modelling in the relevant research areas.

Multiscale Analysis and Scale space Theory

Multiresolution modelling is often synonymously used in literature with the scale space theory [99] and also multiscale analysis [163]. In image processing domain, pyramid structures generated after successive smoothing, and subsampling produces a multiscale representation [99]. Similarly, in scale space theory a scale parameter, t, handles images at different scales. Scale space representation, an improvement over multiscale representation, has an ability to compute representation at a desired scale. Authors in [8] address an important challenge in cancer research by identifying densely connected components of known and putatively novel cancer genes in protein protein interaction networks using a multiscale diffusion kernel. The results in [8] demonstrate the importance of multiscale analysis as the putative cancer genes and network are significant at different diffusion scales. Similarly, authors in [38] detect statistically significant co–mutations in multiple independent insertional mutagenesis screens. The significance is estimated on multiple scales and results are visualised in scale space thus providing valuable supplementary information on the putative cooperation. Multiscale analysis and scale space theory also provide functionalities to address the challenges of image representation at different resolutions. Similarly, a family of methods known as super–resolution has been used to increase the resolution of images and videos [119]. Generally, both multiscale and scale space methods work in model domain. However, multiresolution methods developed in this thesis are the result of multiresolution challenges arising in the data domain.

Wavelets

Wavelets are appropriate methods to describe the mathematical phenomenon such as functions and signals at different levels of resolution [105]. Wavelet analysis have been popular tool in multiresolution analysis [81]. However, the classical wavelets based techniques are useful in regular, consistent, and homogeneous setting. Hence, wavelets cannot directly handle the irregularities in the chromosomal amplification data.

Learning from Multiple Sources

Similar to multiresolution modelling, learning from multiple sources aims to ameliorate the problem of curse of dimensionality, or Hughes effect by exploiting any related additional datasets such as earlier measurement experiments [36]. Unlike multiresolution modelling, the additional datasets may only be weakly related to the analysed dataset. The paradigm of learning from multiple sources is extended to the paradigms of multiview [150], multiway [78], and multitask learning [29].

Data Fusion

The domain of data fusion shares a common ground with the domain of multiresolution modelling. Data fusion integrates multiple data and knowledge depicting the same real world phenomenon in a single, logical, precise, and useful knowledge base [61]. Data fusion techniques are often used to combine data from multiple sensors in such a way that the inference from the combined data is better than that from individual sensors. Data integration approaches have also been widely used in bioinformatics domain. For example, authors in [62] have proposed integrated database and software system that enables retrieval and visualisation of biological relationships across heterogeneous data sources. Similarly, authors in [87] combined data from complementary Deoxyribonucleic Acid (cDNA) arrays and tissue microarrays (TMA) to study the molecular changes in malignant pleural mesothelioma (MM). The study shows that novel proteins associated with cell adhesion are expressed either directly or as a regulatory mechanism in MM. The process of data fusion takes place at the different stages of analysis but it is a common practice to merge the data at the earliest stage of analysis in a single resolution. Data fusion techniques have also been used in multiresolution analysis, especially in remote sensing [28].

Granular Computing

Granular computing (GrC) has roots in multiresolution modelling [10]. GrC is a multidisciplinary field of study comprising of theories, methodologies, and tools to analyse data using the granules in data [170]. Granular computing aims to divide data into different intrinsic resolutions to solve a problem which resembles with multiresolution modelling framework.

CHAPTER 3

MIXTURE MODELS AND MODEL SELECTION

> *The purpose of models is not to fit the data but to sharpen the questions.*
>
> — SAMUEL KARLIN
>
> 11^{th} *R A Fisher Memorial Lecture (1983)*

Synopsis

This chapter introduces mathematical foundation and formulation of mixture models. The chapter also discusses the model selection in mixture models. This chapter also discusses one of the associated publications where we propose a computationally efficient algorithm to train a series of mixture models to aid model selection procedure.

Classical probability distributions such as Gaussian, Bernoulli, and Poison provide methods for probabilistic modelling of data [160]. However, in the real world scenario, a single probability distribution cannot emulate the complexity in the data. Nevertheless, a combination of sufficiently large number of probability distributions can possibly emulate complexity in the data. Such combination of multiple classical probability distributions forms a mixture model. Formally, mixture models are semiparametric latent variable models that model a complex data distribution by weighted sum of different probability distributions [23, 45, 115].

The probability distributions within a mixture model, known as component distributions, describe the observations present in the data. The formulation of mixture model involves determining the number of components in the mixture model, their associated distribution, and identification of the component generating the specific data sample [115]. Mixture models are often used in hard clustering analysis as in this thesis. In hard clustering, only one component is responsible to generate a specific data sample. Mixture models also provide the option of learning soft clustering. In soft clustering, a data sample belongs to more than one cluster with a certain degree of association [23]. A standard formulation of the mixture model assumes that the samples are independent and identically distributed (IID). Under the assumption that data originates from a known number of components, J, the probability density of a mixture model can be expressed as the weighted sum of its component distributions as:

$$p(x) = \sum_{j=1}^{\mathrm{J}} \pi_j P_j(x \mid \boldsymbol{\theta}_j), \tag{3.1}$$

where j indexes the component distributions. In the Equation (3.1), the mixing proportion (mixing or mixture coefficient) is denoted by π_j for the j^{th} component in the mixture model. It determines the weight of the component distribution in the overall mixture model. Mixing proportions satisfy the property of convex combination such that $\int p(x)dx = 1$, $\pi_j > 0$, and $\sum_{j=1}^{J} \pi_j = 1$ [45]. Similarly, the parameters $\boldsymbol{\theta}_j$ in Equation (3.1) denotes the parameters of the j^{th} component distribution of the mixture model. Application area dictates the choice of distributions, which in literature is dominated by the distributions from exponential family such as Gaussian, and Dirichlet [115]. In this thesis, Bernoulli distribution is the preferred distribution because the datasets are 0–1 datasets describing chromosomal amplifications.

3.1 Mixture Models of 0–1 Data

Finite mixture model of multivariate Bernoulli distributions for a dataset, $\boldsymbol{X}$, of dimensionality, d, are parametrized by $\boldsymbol{\Theta} = \{J,\ \{\pi_j, \Theta_j\}_{j=1}^{J}\}$. The dataset, $\boldsymbol{X}$, consists of samples $\mathrm{x}_1, \ldots, \mathrm{x}_N$ in such a way that $\boldsymbol{X} = \{\mathrm{x}_1, \ldots, \mathrm{x}_N\}$. Replacing the general probability distribution function with the distribution of choice, i.e., Bernoulli distribution, a mixture model of multivariate Bernoulli distribution can be mathematically expressed as [45, 167]:

$$p(\mathbf{x} \mid \mathbf{\Theta}) = \sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_i} (1-\theta_{ji})^{1-x_i}, \tag{3.2}$$

where j indexes the components, and i indexes the data dimensionality. x_i denotes the data point such that $x_i \in \{0, 1\}$. The parameter of a random variable θ_{ji} denotes the probability of the variable taking the value 1 in i^{th} dimension of the j^{th} component. We can collect all the random variables in a component in a vector, Θ_j such that $\Theta_j = [\theta_{j1}, \theta_{j2}, \theta_{j3}, \ldots, \theta_{jd}]$. Similarly, we can collect all the parameters of the mixture model including mixture components in a matrix, $\mathbf{\Theta}$ such that $\mathbf{\Theta} = \{J, \{\pi_j, \Theta_j\}_{j=1}^{J}\}$. The parameter values that maximise the log–likelihood function of the parameters can be defined using maximum likelihood principal [23] as:

$$\mathcal{L}(\mathbf{\Theta} \mid \boldsymbol{X}) = \sum_{n=1}^{\mathrm{N}} \log \left[\sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_{ni}} (1-\theta_{ji})^{1-x_{ni}} \right]. \tag{3.3}$$

The EM algorithm can be used to learn the maximum likelihood parameters of mixture model of Bernoulli distributions by maximising the likelihood in the Equation (3.3) [167].

3.2 Expectation Maximisation Algorithm

Expectation Maximisation (EM) algorithm is an iterative algorithm to determine the maximum likelihood (MLE) or maximum a posteriori (MAP) estimates of the parameters of latent variable models [39, 114]. The EM algorithm is a popular algorithm for learning model parameters in probabilistic latent variable models by maximising the marginal likelihood. The iterations of EM algorithm alternate between Expectation step (E–Step) and Maximisation Step (M–Step).

E–step estimates the posterior probability of each component for every data point. Whereas, M–step updates the model parameters for next iteration. Iterations between E and M step produce a succession of monotonically increasing sequence of log–likelihood values for the parameters $\tau = 0, 1, 2, 3, \ldots$ regardless of the starting point $\{\pi^{(0)}, \Theta^{(0)}\}$ [114].

3.3 Model Selection in Mixture Models

Model selection is the process of selecting a model of optimal complexity for the given set of (finite,training) data [32, 68]. In the statistics literature, model selection is the process of selecting a specific model from a plethora of choices [84]. For example, in classification, model selection may refer to choosing a classification algorithm from different classification algorithms such as Naive Bayes, Decision Trees, and Support Vector Machines. The focus in this thesis is modelling of a heterogeneous chromosomal amplification dataset. Mixture models are the model of choice because of their ability to model heterogeneity and their clustering capabilities. The choice of mixture models is also motivated by their ability to learn the structure of the data better than most other methods because each component distributions capture dominant patterns in the data. Furthermore, mixture models are scientifically proven as learning of mixture models involve well studied statistical inference techniques.

In this thesis, model selection refers to the model structure selection or complexity selection which determines the flexibility of the model to fit or explain the data. In other words, model selection in this context refers to choosing an appropriate level of model complexity in the selected class of model, i.e., mixture model. The complexity parameter in mixture model is the number of component distributions in the mixture model. Model selection, therefore, is the selection of number of components in the mixture model [47].

EM algorithm requires apriori knowledge of the number of components in the mixture model to learn the maximum likelihood parameters from the data [114]. However, the number of component distributions are often unknown apriori. Furthermore, one of the major objectives of machine learning and data mining challanges in the real world can often be restricted to determining the number of components in the mixture model. Hence, model selection is essential to learn a mixture model using the EM algorithm.

A mixture model with large number of mixture components produces larger value for the log–likelihood in Equation (3.3). However, a mixture model with large number of mixture components also overfits the data, and generalises poorly on the future unseen data. Additionally, mixture models with large number of components increase complexity in training of mixture models with respect to both time and memory. In contrast, a mixture model with smaller number of mixture components underfits the data, and is unable to adequately represent the underlying data structure. Therefore, model selection aims to optimise this tradeoff between too simple and complex models.

Related work in Model Selection in Mixture Models

A plethora of criteria and methods have been proposed in the literature to determine the optimal number of mixture components in a mixture model [115]. For example, authors in [30], [46], and [131] provide comprehensive review of deterministic, stochastic and resampling criteria for model selection. Deterministic criteria consists of Akaike Information Criterion (AIC) [6], Bayesian Information Criterion (BIC) [142], Minimum Description Length (MDL) [137], and integrated classification likelihood (ICL) [22]. Similarly, stochastic methods includes Markov Chain Monte Carlo (MCMC) [20], and resampling methods includes bootstrapped likelihood ratio test [112]. Similarly, authors in [168] propose a robust approach against model misspecification leading to a better fitting mixture density based on minimum Hellinger distances. In addition, the authors in [31] and [75] use penalised likelihood method for model selection in mixture model.

Data likelihood is often used as the measure of the quality of mixture models [144]. A well trained mixture model with appropriate number of mixture components estimates the underlying data distribution better and produces high likelihood values for the unseen data. In addition, cross–validation have been popular model validation technique in the literature [56, 121, 149]. Hence, in this thesis we use cross–validated log–likelihood as a criteria for model selection.

3.4 Fast Progressive Training of Mixture Models

The EM algorithm is sensitive to initialisation and susceptible to local optima [114, 169]. One solution to avoid local optima is to run the EM algorithm from different random initialisations and select the model with highest likelihood as the global optimum. Similarly, another solution is to take the average of different runs as general performance of the model [153]. Furthermore, the EM algorithm is computationally expensive because of its slow monotonic convergence property [114]. Therefore, multiple restart strategy is popular method in literature where the EM algorithm is run only for a small number of steps, i.e., not until convergence, generating large number of models. Among those models, the model with maximum likelihood can be selected to continue training until convergence [33].

Similarly, different sophisticated algorithms have been proposed to alleviate the problem of local optima in EM algorithm, for example, using splitting and merging of mixture components [86, 154]. In Publication II, we use merging of mixture components as in [154] to train a series of mixture models. The aim is to aid the model selection algorithm to select a model of appropriate complexity, not to avoid local optima. We train multiple models with highest number of component distributions and select the best models among them to start the chain of mixture models by merging the similar mixture components. The training strategy to generate the chain of mixture models resembles backward elimination methodology in feature selection literature [64]. We initially start with large number of mixture components and progressively merge the similar components until the number of components is 1. We use an approximation of Kullback Leibler (KL) divergence as a measure of similarity between the two components in the mixture model.

3.4.1 Kullback Leibler Divergence and Approximation

Kullback Leibler (KL) divergence is a nonsymmetric measure of difference between two probability distributions [96]. The KL divergence between two given probability distributions P and Q on a finite set $\boldsymbol{X}$ is symmetrized by adding the KL divergence from P to Q and Q to P [83].

$$\begin{aligned}\mathcal{D}_{KL}(P||Q)+\mathcal{D}_{KL}(Q||P) &= \sum_i P(i) log \frac{P(i)}{Q(i)} + \sum_i Q(i) log \frac{Q(i)}{P(i)} \\ &= \sum_i \left[\{P(i)-Q(i)\} log \frac{P(i)}{Q(i)} \right], \qquad (3.4)\end{aligned}$$

where i indexes all possible combinations of data elements. Extending the KL divergence in Equation (3.4), the KL divergence between two components of a mixture model for data of dimension, d, indexed by k for two component distributions θ and β have been derived in [2] as:

$$\begin{aligned}KL_{\theta\beta} &= \sum_{i=1}^{2^d} \left[\left\{ \prod_{k=1}^{d} \left(\theta_k^{x_{ik}} (1-\theta_k)^{(1-x_{ik})} \right) - \prod_{k=1}^{d} \left(\beta_k^{x_{ik}} (1-\beta_k)^{(1-x_{ik})} \right) \right\} \right. \\ &\left. \cdot \sum_{k=1}^{d} log \ \frac{\theta_k^{x_{ik}} (1-\theta_k)^{(1-x_{ik})}}{\beta_k^{x_{ik}} (1-\beta_k)^{(1-x_{ik})}} \right], \qquad (3.5)\end{aligned}$$

where x_{ik} denotes an element in kth dimensionality of ith sample in the data matrix. The Equation (3.5) is the sum of a large number of elements. If the dimensionality of the data is 5 then we iterate 32 times and when the dimensionality is 20, we iterate more than a million times (1,048,576). Moreover, the number of comparisons in a mixture model having J components for data of dimensionality d is $2^d J^2$ which is computationally expensive. Therefore, in Publication II, we derive a computationally efficient approximation of the KL divergence as:

$$KL_{\theta\beta} = \sum_{i\in x^*} \left\{ \prod_{k=1}^{d} \left(\theta_k^{x^*_{ik}}(1-\theta_k)^{(1-x^*_{ik})}\right) - \prod_{k=1}^{d} \left(\beta_k^{x^*_{ik}}(1-\beta_k)^{(1-x_{ik})}\right) \right\}, \quad (3.6)$$

where $X^* = \{x^* : x^* \in \underline{\overline{\boldsymbol{X}}}\}$ is a set of all the unique data samples present in the dataset denoted by $\underline{\overline{\boldsymbol{X}}}$. Here, the summation is approximated only with the samples present in the data. Similarly, we remove the fraction containing the log term from Equation (3.5). In Publication II, we are primarily interested in determining the two closest component distributions in a mixture model. We are not necessarily interested in the exact minimum values of KL divergence between two component distributions in a mixture model. These approximations can inaccurately identify two components as most similar to each other while they differ considerably in the full and accurate KL divergence.

The inaccuracies are in the form of selection of two dissimilar components in mixture models to merge. However, in Publication II, we show that our approximation is good estimate of the full KL divergence in terms of matching the two most similar components distributions. Our approximations, as reported in Publication II, is considerably more accurate (twenty five times) than random matching of the components. Moreover, our approximation are 10,000 times faster than full KL divergence for the data dimensionality twenty. Nevertheless, we compensate for any mismatches by retraining the mixture models after merging the mixture components. The aim of the methodology described in Publication II is not to propose any new model selection criteria but to propose an efficient methodology to train a series of mixture models. The models in the series are similar to each other except for the number of mixture components.

3.4.2 Series of Mixture Models

In the algorithm proposed in Publication II, first, we train a large number of mixture models with large number of mixture components (20 in our experiments). Second, we then calculate the approximated KL divergence among all the pairs of mixture components. The two components with minimum approximated KL divergence are then merged as in [154].

The process of merging of mixture components is iterative and continues until the number of components is 1. Mathematically, the merging of the mixing proportions of two candidate component distributions $\pi_{klmin,1}$, and $\pi_{klmin,2}$ to generate a single component distribution π_{merged} can be expressed as:

$$\pi_{merged} = \pi_{klmin,1} + \pi_{klmin,2}. \tag{3.7}$$

Merging the mixture components using Equation (3.7) preserves the properties of mixing proportions such that they have to sum to 1. Similarly, we can merge the parameters of two candidate mixture components $\Theta_{klmin,1}$ and $\Theta_{klmin,2}$ weighted with their mixing components to generate parameters for merged component Θ_{merged} as:

$$\Theta_{merged} = \frac{\pi_{klmin,1} \times \Theta_{klmin,1} + \pi_{klmin,2} \times \Theta_{klmin,2}}{\pi_{klmin,1} + \pi_{klmin,2}}. \tag{3.8}$$

The parameters of merged component distributions in Equation (3.8) also satisfy the properties of probability of a random variable, θ such that $0 \leq \theta \leq 1$. The mixture model obtained after merging the mixture components is retrained before next iteration of merge operation. This progressive training and merging results in a series of mixture models as shown in the Figure 3.1.

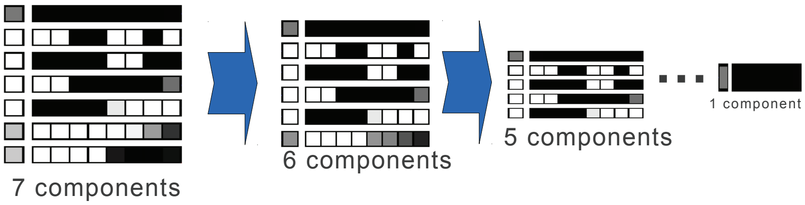


Figure 3.1. Series of mixture models resulting from the progressive merging of the mixture components and retraining of the mixture model. Reprinted with permission from Publication II.

The Figure 3.1 shows snapshot of our algorithm in Publication II where two components in a mixture model with 7 components are merged to generate a mixture model with 6 components. Similarly, mixture models with one less components than the previous model are generated by merging two most similar component distributions until the number of components is 1. The principal focus in Publication II is generating series of mixture models for model selection and not on avoiding local optima or proposing a new model selection criteria.

This series of mixture models can be used with any model selection criteria such as cross–validation, AIC, BIC, and MDL to choose a model of suitable complexity. In our earlier research [2], we have used ten–fold cross–validation to select model of appropriate complexity. We calculate likelihood of each mixture model in the series on both training and validation sets. We then select the model that generalises the best on the validation set taking parsimony into account, i.e., if two models produces comparable results, we select the simpler model [171]. In addition to the gain in computational efficiency, the simple models are also easier to interpret, and understandable to the domain experts [73].

One important property of EM algorithm is that EM algorithm is deterministic for a given initialisation and a given dataset [114]. In other words, if we run EM algorithm on the same data with same initialisation it always converges to the same final model. When the mixture components are merged, the initialisation for the EM algorithm is same. This avoids multiple restarts required in [33] and [153]. Furthermore, EM algorithm converges faster when it is initialised from a merged model than when initialised at random because the merged model resembles the final trained model.

In Publication II, we have shown that EM algorithm converges approximately ten to fifty times faster when initialised from merged model. Similarly, the models produced in the series of models are similar to each other except for the number of components. This allows comparison among similar models for model selection but with different number of components. This avoids the situation when mixture model with some components have been stuck in local minima while models with other components reach global optima. Such situations create a bias in comparison among models with different components in similar vein as 'unfortunate split' in cross–validation.

CHAPTER 4

METHODS FOR MULTIRESOLUTION MODELLING

> *With too little data, you won't be able to make any conclusions that you trust. With loads of data you will find relationships that aren't real... Big data isn't about bits, it's about talent.*
>
> — DOUGLAS MERRILL
>
> *Former CIO and VP of Engineering at Google*

Synopsis

The abundance of multiresolution data and increasing benefits of analysing multiple datasets within a single analysis have given major impetus to the research in multiresolution data analysis. In application areas where division of data across different resolutions is smooth, wavelets [81], multiscale methods [11, 163], and scale space theory [100] have been popularly used to analyse multiresolution data. This chapter discusses the core of the thesis and includes most of the scientific contributions of this thesis. This chapter also summarises four of the five publications contained in this thesis.

4.1 Data Transformation

Standard algorithms, such as mixture models, are unable to model multiresolution data in their standard form. Therefore, in Publication I, we propose data transformation methods to analyse multiresolution data by transforming the data to different resolutions and integrating the data in the same resolution. We can then apply the algorithm on the combined data in a single resolution. The methodology of data transformation integrates data in different resolutions and therefore, resembles fusion techniques [28].

Data transformation methods, also called sampling methods, proposed in Publication I are non–stochastic. Sampling resolution in genomics explains the level of precision for measuring the results of a particular experiment: either global (coarse resolution) or detailed (fine resolution). As discussed in Section 2.3 and also shown in the Figure 2.2, the relationship between different resolutions of chromosome, i.e., correspondence of each of the regions in genome in different resolutions are known apriori [143]. We propose two different data transformation methods to transform data across fine and coarse resolution using the knowledge of the correspondence of chromosomal regions in different resolutions.

1. Upsampling transforms the data from coarse resolution to fine resolution increasing the dimensionality of the data. We make multiple copies of a chromosomal region in coarser resolution to upsample the data in coarse resolution to fine resolution.

2. Downsampling transforms the data resolution from fine resolution to coarse resolution decreasing the dimensionality of the data. We downsample using three different methods: OR–function, Weighted, and Majority Decision. We consider the chromosomal amplification pattern of neighbouring chromosomal regions if the number of aberrated chromosomal regions and the number of unaberrated chromosomal regions are equal.

   (a) In OR–function downsampling, a chromosomal region in the coarse resolution is aberrated if any of the chromosomal regions in the fine resolution is aberrated.

   (b) Division of the regions of a chromosome are highly irregular and the length of a region often differs from the next [143]. In weighted downsampling, a chromosomal region in coarse resolution is aberrated if the total length of the aberrated chromosomal regions is greater than total length of unaberrated chromosomal regions in fine resolution.

(c) In majority decision downsampling, a chromosomal region in the coarse resolution is aberrated if majority of the chromosomal regions in the fine resolution are aberrated.

Experiments on Data Transformation

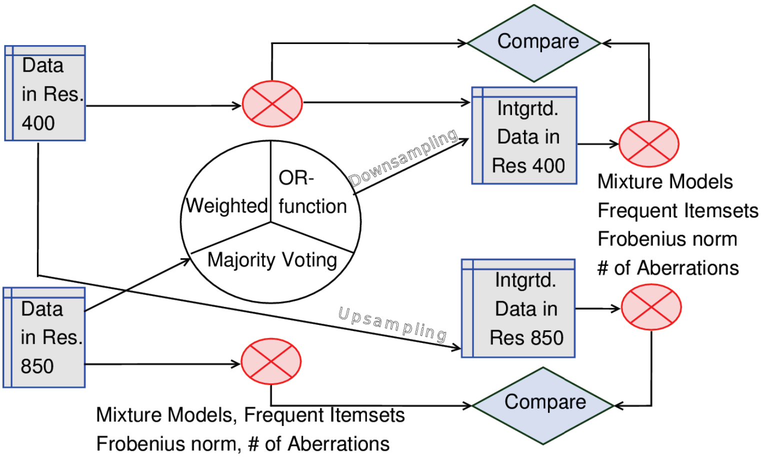


Figure 4.1. Schematic representation of experimental procedure of data transformation methods for multiresolution modelling. First, the data in two different resolutions are transformed to other resolutions. After transformation datasets in the same resolution are integrated. Finally, the algorithm is applied on the integrated dataset. For comparative purposes the algorithm is also applied on the original data before data transformation.

Figure 4.1 depicts the overall experimental procedure where one of the three different downsampling methods transforms the data in fine resolution to coarse resolution. Similarly, a deterministic upsampling method transforms the data in the coarse resolution to the fine resolution. Before data transformation, algorithms such as mixture models, and pattern mining are applied on the data in original resolution. We then integrate the data obtained in same resolution after data transformation. The algorithm is again applied on the integrated data. Finally, we compare the results of the analysis before transformation and after integration in terms of the patterns obtained and model fitting.

Experiments with mixture modelling in different resolutions reported in Publication I show that validation likelihood of the mixture models is higher in the coarser resolution compared to the finer resolution. However, the model selection results are similar across different resolutions as similar number of components are selected in both the coarse and the fine resolution. Although similar number of components are selected, mixture models in coarse resolution produces better likelihood values than the data in fine resolution. In addition, time complexity is higher in the models in the finer resolution. This degradation of performance in fine resolution data can be attributed to the "curse of dimensionality" phenomenon [17], or Hughes effect [76]. Models in coarse resolution are also suitable for understanding and interpreting the results [73].

The results in Publication I also show that the mixture models produce better results on the combined data with the similar number of components than the standalone data in single resolution. This proves the property of mixture model which states that number of components in the mixture model scales with the complexity of the data not with the sample size of the data [65]. The increased sample size arising from the integration of data from other resolution helps nullify the curse of dimensionality and constrains the complexity of mixture models, and avoid overfitting.

MAFIA (MAximal Frequent Itemset Algorithm) [26] was used to mine maximal frequent itemsets in data in the original resolution and the sampled resolution to determine if the data transformation methods preserves the patterns in the data. The results in Publication I show that data transformation across resolutions preserves the maximal frequent itemsets with negligible differences. The negligible differences are expected because data in coarse resolution cannot subsume all the information of the data in fine resolution.

In our earlier research [1], we have also tested the statistical significance of the frequent itemsets (not the maximal frequent itemset) to show that data transformation across different resolution preserves the statistically significant patterns present in the data. In addition, results in [1] also show that statistically significant patterns are also preserved by the generative property of mixture models in all the resolutions. We also compare three different downsampling methods using metrics such as the Frobenius norm [148]. Experimental results in Publication I show that the resulting data produced by three downsampling methods are similar to each other; the variation, if any are negligible.

4.2 Merging of Mixture Components

In Publication III, we use merging of the mixture components of different mixture models in different resolutions to model the multiresolution data. Mixture models can also be used in clustering where component distributions are used as clusters in the data. Proposed multiresolution modelling algorithm resembles clustering aggregation algorithm in [59]. The similarity with clustering aggregation is that we use multiple clustering results, i.e., mixture models to improve the mixture modelling. However, clustering aggregation clusters single dataset generating results as a single clustering solution. In contrast, the proposed multiresolution modelling algorithm analyses different datasets in different resolutions generating clustering solutions in different resolutions.

In the proposed multiresolution modelling algorithm, we first apply mixture models on the data in each resolution separately. Secondly, we iteratively merge the similar mixture components in different mixture models in different resolutions. This is unlike Publication II where we merge the components from the same mixture model. We extend the derivation of fast approximation of Kullback Leibler divergence as a criterion in Publication II to determine the similarity between the mixture components to two mixture models as:

$$\begin{aligned} KL &= \sum_{i \in X^*} \pi_\alpha \prod_{m=1}^{d} \left(\alpha_m^{X^*_{im}} (1-\alpha_m)^{(1-X^*_{im})} \right) \\ &\quad - \sum_{i' \in Y^*} \pi_\beta \prod_{n=1}^{d'} \left(\beta_n^{Y^*_{i'n}} (1-\beta_n)^{(1-Y^*_{i'n})} \right) . \end{aligned} \tag{4.1}$$

The approximation of KL divergence in Equation (4.1) resembles Equation (3.6) but for the two component distributions α and β which are components of two different mixture models in different resolutions. Similarly, X^* and Y^* are the unique samples of data in two different resolutions.

We calculate the pairwise KL divergence between all the components in two mixture models. We then select the similar components using minimum weight bipartite matching algorithm [164] as shown in Figure 4.2. The similar components are merged preserving the properties of component distributions and probability of random values in the mixture model. We iterate the matching shown in Figure 4.2 and merging of mixture components until the KL divergence is small enough. Finally, we retrain the

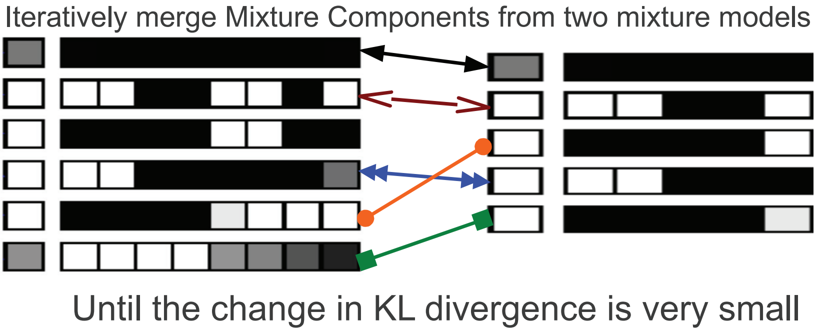

Iteratively merge Mixture Components from two mixture models
Until the change in KL divergence is very small

Figure 4.2. Simplified picture of multiresolution modelling using merging of mixture components. We iteratively merge the similar components from different models until the change in KL divergence is very small. The different arrow shapes show the pairwise similarity of mixture components.

mixture models in each resolution. Although mixture models are generated separately in each resolution, they incorporate information about the data in other resolutions.

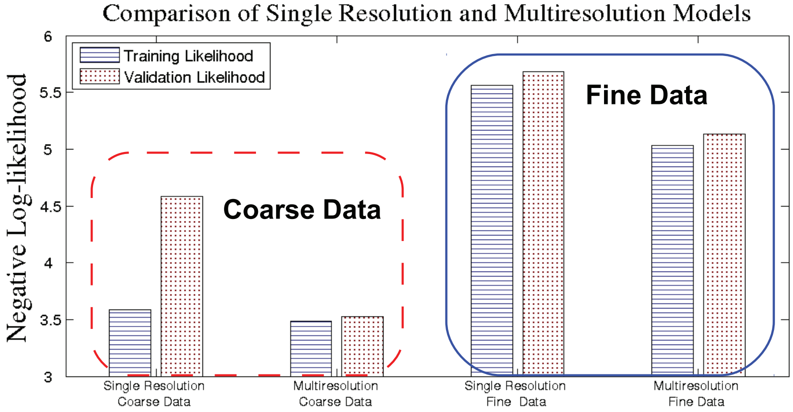


Figure 4.3. Likelihood of multiresolution mixture models trained by merging of mixture components and individually trained mixture models in single resolution. Since the units in Y–axis is the negative log–likelihood, the shorter the bar the better the result. The improvement gained by multiresolution mixture model in the fine resolution is greater than that gained in the coarse resolution. The figure is adapted from Publication III.

The algorithm generates plausible results when the algorithm is experimented on multiresolution chromosomal amplification datasets discussed in Section 2.3. The bar diagram in the Figure 4.3 depicts the improvements gained by multiresolution models over single resolution models. The figure shows training and validation likelihood of the multiresolution and independent single resolution mixture models trained in a 10–fold cross–validation setting. Since the units in Y–axis is negative log–likelihood, the shorter the bar better the result. The Figure 4.3 shows the two different conditions of the likelihood: first, the performance of single resolution, and the multiresolution model on the coarse data, which is enclosed in the dashed rectangle in the left side of the Figure 4.3. Second, Figure 4.3 also shows performance of the single resolution and the multiresolution models on the fine data which is enclosed in solid rectangle in the right side of the figure.

Scrutinising the results inside both the dashed and the solid rectangles in the Figure 4.3, the performance of the multiresolution model is markedly better in the coarse resolution and slightly better in the fine resolution. The improvement in the performance of the multiresolution model in the coarse resolution is greater than that in the fine resolution. This is because the number of data samples is comparatively smaller in the coarse resolution to add more information to the model in the fine resolution. The results also show that the models trained in the multiresolution setting generalises better on the future unseen data. As discussed in Section 4.2, the performance of the models are better in coarse resolution because of the curse of dimensionality. We also performed the two–tailed t–test to ascertain the statistical significance of the result on the data likelihood [160]. The results also show that both the validation and the training likelihoods are statistically significant when the significance level, α, is 0.1.

4.3 Multiresolution Mixture Components from Domain Ontology

Multiresolution data often forms hierarchical structure as discussed in Chapter 2. The domain ontology used in this thesis is known apriori from the application area. Consequently, we can exploit this structural information from the application area to determine the relationships between data resolutions. Therefore, we can determine the structure of the Bayesian network as shown in the Figure 4.4 with some realistic assumptions for computational efficiency. For this reason, we do not learn the structure of Bayesian networks in our contribution. The assumptions are that the data features in the coarse resolution form the root and branches near the root of the Bayesian network. Similarly, the data features in the finer resolutions form the branches towards the leaves and the leaves

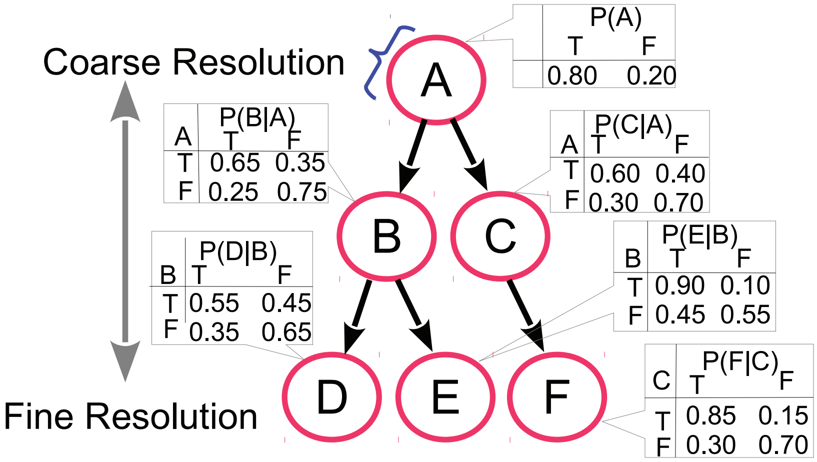


Figure 4.4. A structure of the Bayesian network from the apriori domain knowledge shown in Figure 1.1. The figure shows both Bayesian Network with nodes and edges; and the associated conditional probability tables. The figure is adapted from Publication IV.

of the Bayesian network. Additionally, we can assume that the directed arrows originate from the features in the coarse resolution.

Figure 4.4 shows a Bayesian network generated from the hierarchical structure of data depicted in the Figure 1.1. In the real world, although the hierarchical structure as shown in the Figure 1.1 are known, data in all the resolutions in the structure may not be available. Nevertheless, Bayesian networks have been known for their prowess in missing value imputation [40]. Therefore, in Publication IV, Bayesian networks in the component distributions are used to impute missing data resolutions. Experimental results in Publication IV show that Bayesian networks satisfactorily imputes missing data resolutions.

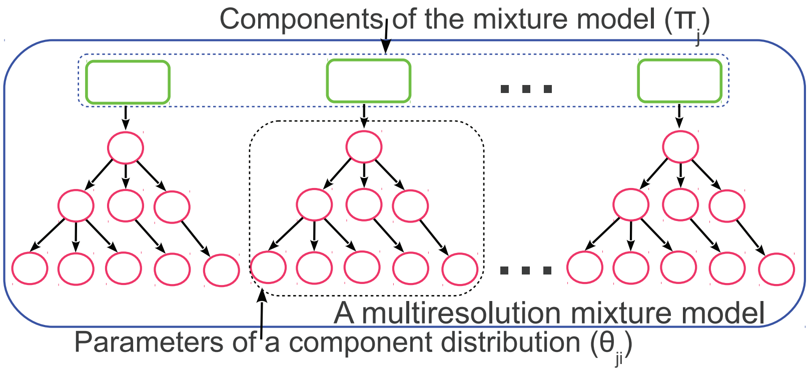

Components of the mixture model (π_j)
A multiresolution mixture model
Parameters of a component distribution (θ_{ji})

Figure 4.5. Structure of multiresolution mixture model whose components are Bayesian networks. The figure is adapted from Publication IV.

Figure 4.5 depicts the structure of the proposed mixture model in Publication IV for data in multiple resolutions. The three solid rectangles on the top represent different mixture coefficients, π. Similarly, the three network of nodes denote the three component distributions. Each node defines a parameter of the component distributions, θ. The structure of the component distribution is determined from the domain knowledge. Since, the structure of Bayesian network is known, the parameters of these Bayesian networks can be learned in the maximum likelihood framework [9]. If some of the data are missing, we need some assumptions to learn the parameters of the Bayesian networks. One of such similar assumption is founded from the Potts model [15, 135] where we estimate the CPD of the child (C) given parent (P) as: $P(C \mid P) = 0.9$.

After learning the Bayesian networks, and imputing the missing values, the next step is to learn the mixture models. First of the challenges confronting the learning of mixture model is the model selection, i.e., determining the optimal number of component distributions [47]. Similarly, learning the parameters of the component distributions involves learning the parameters of those networks. In general framework for the EM algorithm, we can assign only a single probability value to a node in the mixture model [39]. However, each variable in Bayesian network consists of minimum of two probability values denoting the CPD of the nodes. Hence, we learn the mixture model in the two step procedure. First, we learn the the parameters of individual Bayesian networks in the framework of Bayesian networks [9, 70]. Second, we transform the networks to vectors to learn the parameters of mixture model using the EM algorithm as in [153].

In addition to the multiresolution chromosomal amplification datasets discussed in Section 2.3, we have in Publication IV experimented with a simulated dataset that allows observation of complete data without missing resolutions. The bar diagram in the Figure 4.6 displays the performance of the multiresolution mixture model trained in a 10–fold cross–validation setting and also three different single resolution mixture models trained individually in each resolution. Since units used in the Y–axis is negative log–likelihood, the shorter the bar, better the result. The Figure 4.6 shows two different conditions of likelihood: training and validation. However, the results do not depict change in training and validation likelihood during model selection instead they show the difference in training and validation likelihoods after the selection of components.

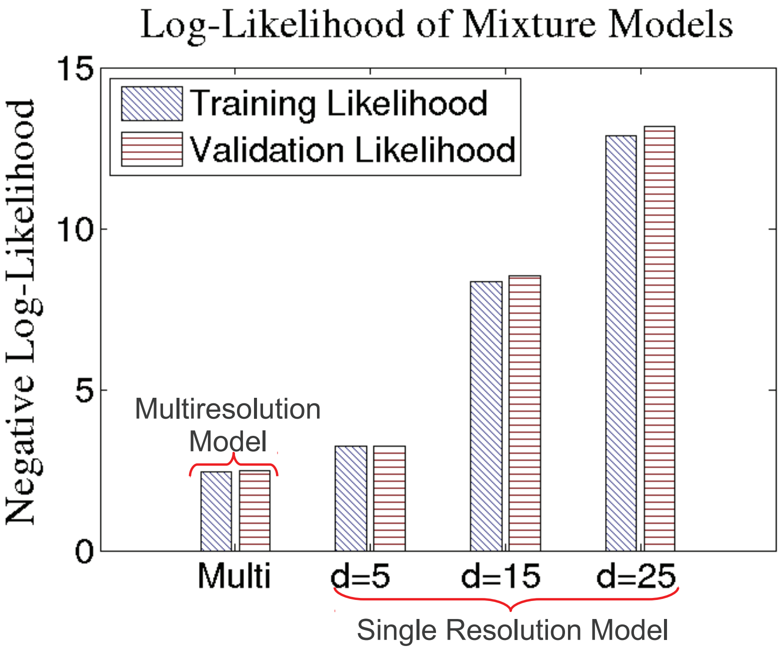


Figure 4.6. Likelihood of single resolution and multiresolution mixture model on simulated dataset. Since the units in Y–axis is the negative log–likelihood, shorter the bar better the result. The performance of multiresolution mixture models surpasses that of all the single resolution models. Reprinted with permission from Publication IV.

The Figure 4.6 shows that the performance of the multiresolution mixture model is markedly better than the three single resolution models. Log–likelihood is comparatively poor in dimensionality of 15, and 25 because of the larger data dimensionality demonstrating curse of dimensionality. The likelihood of the proposed multiresolution model is better than the data with the smallest dimensionality of five in single resolution. The results show that proposed multiresolution mixture model produces plausible results in addition to providing single analysis solution for the data in multiple resolutions.

4.4 Multiresolution Semantic Subgroup Discovery

As discussed in Section 2.4, semantic data mining methods have been gaining popularity in the data mining domain. Similarly, banded matrices have also found usage in data mining domain [43, 55]. In Publication V, we comprehensively analyse multiresolution data using a three stage methodology depicted in Figure 4.7. In the contribution, we explain the clustering generated by mixture models using semantic data mining methods, and visualise the clusters and the semantic rules using the banded matrices.

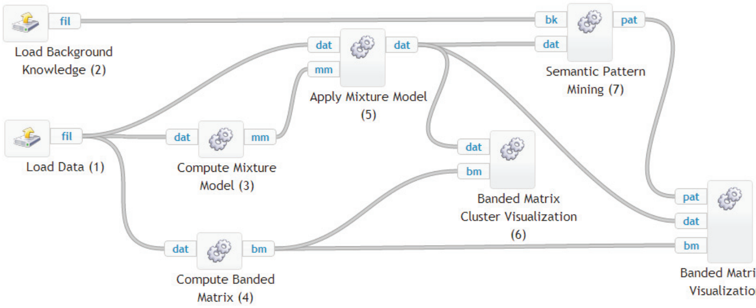

fil
Load Background Knowledge (2)
fil
Load Data (1)
dat
mm
Apply Mixture Model (5)
dat
dat
mm
Compute Mixture Model (3)
dat
bm
Compute Banded Matrix (4)
bk
dat
pat
Semantic Pattern Mining (7)
dat
bm
Banded Matrix Cluster Visualization (6)
pat
dat
bm
Banded Matrix Rule Visualization (8)

Figure 4.7. The workflow for comprehensive analysis of multiresolution data using a combination of probabilistic model based clustering, semantic data mining, and banded matrices. Reprinted with permission from Publication V.

Figure 4.7 depicts the working of the three part methodology. The figure shows that input to the methodology is the empirical data and additional background knowledge. The additional background knowledge is used by the semantic data mining algorithm to supplement the analysis of the empirical data. As cancer is a heterogeneous and multifactorial disease [90], we use additional background knowledge with an aim to better understand and interpret the results. The additional knowledge provided to the semantic data mining algorithm comprises of fragile sites [42, 141], cancer genes [49], amplification hotspots [125], and virus integration sites [88, 172]. Finally, the taxonomies of hierarchical regions of chromosomes discussed in Section 2.3 are also used as additional background knowledge so that semantic data mining methods are able to analyse multiresolution data.

Mixture models provide an ability to cluster the data considering the components in the mixture model as a cluster [113, 118]. We train the mixture model in a ten–fold cross–validation setting taking parsimony into account [153]. The results produced by mixture models are complex to explain to the application area specialist. Efforts have, however, been made in the past to make the results understandable to the domain experts [73]. In Publication V, we explain the clusters with the rules generated by semantic data mining algorithms and visualisation produced by banded matrices. The cluster labels generated using clustering from mixture model are used as class labels in semantic pattern mining algorithm along with the additional background knowledge. We use general purpose

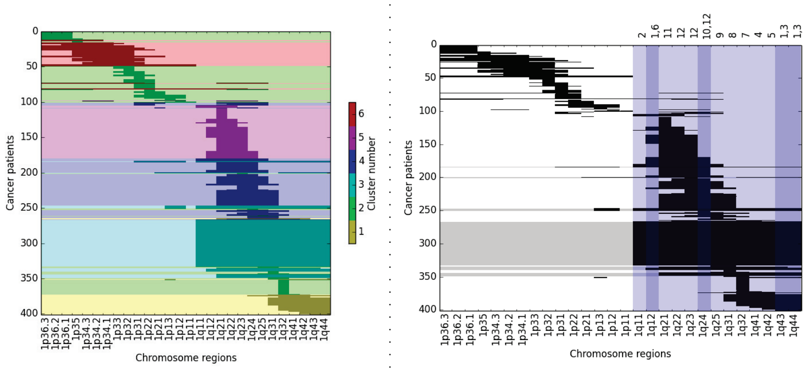


Figure 4.8. The comprehensive analysis of multiresolution data using a combination of probabilistic model based clustering, semantic pattern mining, and banded matrices. Figure on the left panel depicts the clusters overlayed on the banded structure of data. Similarly, figure on the right panel depicts the both clusters and semantic rules overlayed on the banded structure of the data. For the clarity of presentation, the figure on the right depicts only cluster 3 and the rules explaining only cluster 3. Figures are adapted from Publication V.

semantic subgroup discovery system, Hedwig, to find a hypothesis (a predictive model or a set of descriptive patterns) in domain ontology terms, given the training data and the domain knowledge in the form of ontologies [157]. Hedwig, for instance, is developed by the collaborators in Jožef Stefan Institute in Slovenia who are the co–authors in Publication V.

We use the constrained banded matrices [55] to visualise the data. In chromosomal amplification data, the matrices are constrained because the columns denote the specific and unchangeable chromosome regions. We, therefore, shuffle only the rows, i.e., only the samples and not the columns. We then overlay the cluster information along the rows of the banded matrix as shown in the left panel of the Figure 4.8 showing clear distinction among different clusters. In addition, we overlay the rules generated using the semantic subgroup discovery method as shown in the right panel of the Figure 4.8. The visualised rules are tabulated in Table 4.1. The numbers above the rules on top right corner denote the position of rules in the table. A darker hue means that specific region in chromosome appears in more than one rule denoted by more than one position of the rules in the table. Overlaying all the clusters and the rules for each of the clusters will clutter multitude of information on a single figure compromising the understandability of the visualisation. Therefore, we first visualise all the clusters in the data overlaying it on a banded matrix as shown in the left panel of the Figure 4.8. Second, we visualise only a single cluster

#	Rules for cluster 3	TP	FP	Precision
1	`Cluster3(X)` ← `1q43-44(X)` ∧ `1q12(X)`	81	0	1.00
2	`Cluster3(X)` ← `1q11(X)`	78	9	0.90
3	`Cluster3(X)` ← `1q43-44(X)`	88	26	0.77
4	`Cluster3(X)` ← `1q41(X)`	88	28	0.76
5	`Cluster3(X)` ← `1q12(X)`	81	43	0.65
6	`Cluster3(X)` ← `1q32(X)`	88	52	0.63
7	`Cluster3(X)` ← `1q31(X)`	87	54	0.62
8	`Cluster3(X)` ← `1q25(X)`	88	64	0.58
9	`Cluster3(X)` ← `1q24(X)`	88	97	0.48
10	`Cluster3(X)` ← `1q21(X)`	88	134	0.40
11	`Cluster3(X)` ← `1q22-24(X)`	88	149	0.37
12	`Cluster3(X)` ← `HotspotSite(X)`	88	222	0.28
13	`Cluster3(X)` ← `CancerSite(X)`	88	245	0.26
14	`Cluster3(X)` ← `FragileSite(X)`	88	259	0.25

Table 4.1. Rules induced for 3 using semantic data mining algorithm Hedwig.

and the rules describing that cluster as shown in the right panel of the Figure 4.8.

The left panel of the Figure 4.8 distinctly shows different clusters proving the credibility of the clustering results. Similarly, the rules visualised in the right panel of the Figure 4.8 identify the amplifications in chromosomal regions that are responsible for certain cluster (cluster 3) and consequently, specific groups of cancer as reported in [126]. In addition, the rules generated by semantic data mining algorithm provide additional insights into the clustering solutions. For example, from the left panel of the Figure 4.8 cluster 3 is denoted by the pronounced amplification in regions 1q11-q44. The rule: Rule 1: `Cluster3(X)` ← `1q43-44(X)` ∧ `1q12(X)` characterises 81 out of 88 data samples that are in cluster 3 showing that amplifications in regions 1q43–44 and 1q12 characterises cluster 3 and related cancers with good coverage and precision. Results show that whole region of 1q11–44 need not be aberrated to discriminate that specific cluster of cancers. This provides insights into the data and improvements in the understandability of the amplification to the domain experts.

CHAPTER 5

DISCUSSION

> *Learning is not attained by chance, it must be sought for with ardor and attended to with diligence.*
>
> — ABIGAIL ADAMS
>
> *Letter to John Quincy Adams (1780)*

Synopsis

The work in the thesis focused on the analysis of multiresolution 0–1 data. The application area of choice was chromosomal aberrations patterns in cancer genomics defined in multiple resolutions. The proposed algorithms, mixture models and semantic data mining, for analysis of multiresolution data are experimented on the chromosomal aberrations data with plausible results. Furthermore, an efficient method to train a chain of mixture models was proposed to aid model selection in mixture models. Multiresolution modelling methods, and model selection in mixture models are discussed in this chapter along with their applicability, limitations, and possible future directions of work. The future directions of work discussed in this chapter concerns specific methods discussed and developed in the thesis. The future work section in Chapter 6, i.e. Section 6.2, discusses the overall future work in the multiresolution modelling domain.

5.1 Model Selection in Mixture Models

Model selection is an age–old problem in statistics and machine learning [7, 97]. In Publication II, we do not propose a new model selection criteria but a computationally efficient method to train a series mixture models differing only in the number of components. The proposed method provides additional facilities of computational efficiency, and similarity of the mixture models in the chain except for the number of components. Therefore, the method is suitable for comparison in model selection. The experiments performed on the three datasets provide evidence of its efficiency and suitability in model selection. Furthermore, the proposed mixture model for Bernoulli distributions can be seamlessly extended to other distributions such as the Gaussian distribution.

The proposed method is sensitive to local optima while learning mixture model via EM algorithm [114, 169]. We try to address the challenges of local optima by training multiple mixture models once for largest number of mixture components before merging the similar components. The best mixture model among the trained mixture models is selected to calculate the KL divergence among mixture components. The most similar components, i.e., the the pair of components with the minimum KL divergence are progressively merged to generate a chain of mixture models. However, this does not guarantee that the EM algorithm reaches global optimum. Avoidance of local minima is still an open research problem in optimisation and also the EM algorithm. Nevertheless, effectiveness, efficiency, and seamless scalability of the proposed method makes the proposed method, the method of choice for training mixture models for model selection.

5.2 Multiresolution Analysis and Modelling of 0–1 Data

Algorithms and methods to study and analyze multiresolution data forms the crux of the thesis. The proposed algorithms complement each other and specific algorithm fulfills the requirements of a specific application. Nevertheless, ample possibilities and challenges for future improvements identified in the proposed algorithms and methods are discussed in the subsequent paragraphs.

5.2.1 Data Transformation for Multiresolution Analysis

The data transformation methods deterministically transform the data across different resolutions in such a way that data in different resolutions can be integrated in a single resolution. The integrated data in single resolution can then be analyzed using a method of choice because the data is of the same dimensionality. In Publication I, we experiment with mixture models and pattern mining algorithms generating credible results for multiresolution chromosomal aberrations data.

The data transformation methods are suitable for analysis requiring high processing speed and robustness. One of such application area is stream data mining [3, 50, 52, 159] where the requirements are efficient processing and robustness in analysis against minor changes occurring in the data. Data transformation methods are efficient because their computation is simple and are robust against small changes and outliers; for the data transformation methods are deterministic given the structure of the multiresolution phenomena. Furthermore, data transformation methods are suitable for applications requiring single resolution models for multiresolution data. In hindsight, the data transformation methods lack probabilistic interpretation. Adding stochasticity in those methods is a possible future work, for example, with foundations on Potts model [15, 135].

5.2.2 Merging of Mixture Components

In Publication III, we model multiresolution data by generating mixture models in each resolution separately in such a way that the models in each resolution incorporate the information from other data resolutions. The experiments with chromosomal aberrations data show multiresolution mixture models incorporating the interactions between data resolutions produce better results compared to the individually trained single resolution models. The method is suitable for application areas that require models in each level of processing resolution such as image processing, and computer vision [145]. Furthermore, experiments in Publication III have shown that merging of mixture components also helps in avoiding local optima when experimented on the two single resolution models.

Merging of mixture components from different mixture models aids in modelling interaction among the mixture models in different resolutions. An approximation of symmetric KL divergence is used to compare the similarity of the components in the mixture model. The similar components are then merged. However, the convergence analysis of KL divergence is not studied in detail in Publication III. Furthermore, upsampling and downsampling of the parameters of the mixture model adds another complexity to the methodology. Additionally, improvements of the multiresolution mixture model and avoidance of local optima in single resolution mixture model have been verified only by the empirical experiments. However, solid mathematical foundations and the proofs for the improvement are missing. One direction of future work could focus on mathematical proofs for the empirical evidence in merging components for multiresolution modelling.

5.2.3 Multiresolution Mixture Components

In Publication IV, a single multiresolution mixture model with multiresolution mixture components are proposed and experimented using multiresolution chromosomal aberrations dataset. Only a single multiresolution model is generated in Publication IV, which is unlike Publication III where a model is generated for each data resolution. The individual mixture components provide the functionality of Bayesian networks. The proposed model is suitable for the situations requiring generative modelling prowess of probabilistic models. In Publication IV generative property of the Bayesian network helps imputing the missing resolutions of the data. Furthermore, the proposed multiresolution mixture model could be applicable in any domain where the network structure in the multiresolution data is consistent across the dimensionality, for example, in the image processing domain.

The mixture components used as a Bayesian network model the dependency among the nodes in the network. In addition each node requires at least two probability values describing the probability of the node given the probability of its parent node [9]. In contrast, the EM algorithm assumes IID distributions for the samples [114]. Additionally, the EM algorithm provides only a single probability value for a node, i.e., probability of a random variable (θ) taking the value 1; two if you consider 1 - the given probability ($1 - \theta$). Hence, we transform the nodes to a vector representation to learn the mixture models via the EM algorithm. For this reason, future work in multiresolution mixture modelling could be to develop EM algorithm to directly learn the parameters of mixture models when the components are not vectors but a network. Furthermore, transformation network representation in multiresolution mixture model to vectors and then learning the mixture models using the EM algorithm requires structural similarity of networks used as the different mixture components. Therefore, the future work could focus on relaxing this requirement.

5.2.4 Multiresolution Analysis by Semantic Data Mining

In Publication V, we propose a three part methodology for comprehensive analysis of the multiresolution data. We use clustering results from the mixture models as the labels for the semantic data mining algorithm. The additional background knowledge consists of taxonomy of hierarchy of regions, fragile sites, virus integration sites, amplification hotspots, and cancer genes. We use banded matrices to visualize the clusters from mixture models and the rules from semantic data mining algorithm. The proposed method is suitable for both labeled and unlabeled data as cluster indices can be used as class labels in semantic data mining. Furthermore, banded matrix provides the visualization aspect to the analysis for detailed study of the data. Thus, the method is also suitable for rigorous analysis of multiresolution data.

Every system in the world is connected with one another and each system effects the other system. Consequently, understanding one system can help understand another system better. In this scenario, knowledge or understanding of one system can be used as a background information to understand another system. These methods are applicable in bioinformatics as interacting systems produce different datasets. Similarly, the proposed methodology could be applicable in natural language processing [107] because the additional background knowledge in natural language processing are available in form of ontologies such as the semantic web.

The three part methodology proposed in Publication V takes as an input only data in a single resolution. Multiresolution analysis is achieved by using the taxonomy of multiresolution hierarchy as an additional background knowledge to the methodology. In the future, the semantic pattern mining algorithms can be developed to include data in multiple resolutions simultaneously in addition to the taxonomy of hierarchy of regions.

CHAPTER 6

SUMMARY AND CONCLUSIONS

> *A story has no beginning or end: arbitrarily one chooses that moment of experience from which to look back or from which to look ahead.*
>
> — GRAHAM GREENE
> *The End of the Affair (1951)*

Synopsis

This chapter summarizes the contributions of the thesis and draws conclusion from the research. The chapter also discusses the overall future research perspectives in multiresolution modelling domain.

6.1 Summary

In traditional machine learning and data mining scenario data analysed is from a single source represented in a single resolution. In current age of big data, the challenge is to analyse massive set of datasets, i.e., the challenge is to analyse multiple datasets within a single analysis. The multiple datasets can be available in different representations. Analysis of data in multiple representations needs methods and algorithms suitable for different situations and application areas. Analysis of data in multiple representations within a single analysis framework also caters the needs of data hungry algorithms.

The work in this thesis has concentrated in developing algorithms and methods to address the challenges in modelling data in multiple representations. In this thesis, multiple representations aspect is provided by the data represented in multiple resolutions. The algorithms especially covers mixture models and semantic data mining methods. Different methods and algorithms have been developed to analyse multiresolution data suitable for different situations and application areas.

The data transformation methods proposed in the thesis transforms data across different resolutions to integrate datasets in different resolutions providing an opportunity to analyse data in a single resolution. Additionally, a computationally efficient algorithm to train a series of mixture models to aid model selection algorithms is developed in the thesis. Similarly, an algorithm based on merging of mixture components to model multiresolution data produces models in each resolution incorporating information from other data resolutions. In addition, a multiresolution mixture model uses the domain knowledge to design multiresolution mixture components which are individually functional as Bayesian networks. Furthermore, a semantic data mining algorithm developed in this thesis uses knowledge of hierarchy of multiresolution data and other background knowledge to extract rules from the data. The algorithms and methods provide plausible improvements in multiresolution data analysis compared to the individual analysis in the single resolution data.

6.2 Future Work

The multiresolution analysis methodology developed in this thesis are at its initial stage. The thesis forms the foundations for multiresolution modelling and the algorithms and methods proposed in the thesis need further research on the scope and general applicability. The methods are tested only on datasets such as the chromosomal aberrations datasets, publicly available datasets, and simulated datasets. However, the methods have not been developed as a tool with rigorous testing for general applicability. The improvements necessary for each of the developed methods and algorithms are discussed in Chapter 5. This section discusses the future improvements in overall multiresolution analysis domain. It includes developing the EM algorithm to learn the multiresolution components of the mixture models. The EM algorithm used in this thesis learns the maximum likelihood parameters when networks were arranged as vectors.

Throughout the thesis, mixture models are used in hard clustering setting, i.e., one sample is only associated with one component distribution generating the maximum posterior probability. Mixture models can also be used in a soft clustering setting where posterior probability can be used to assign a sample to more than one component distribution. Soft clustering setting is beneficial in the chromosomewise analysis of chromosomal aberrations data because some cancer samples with the same known cancer labels can be grouped in two different clusters. Soft clustering of chromosomal aberrations data can also be justified because of the heterogeneous nature of cancer.

In chromosomewise analysis, two exactly similar cancer samples can be labelled as two different cancers because other chromosomes that are likely to discriminate cancers will be ignored in the current analysis. Furthermore, we have 73 different types of cancer labels for data in coarse resolution. Therefore, we can use multiclass classification to analyse the data. In a broader context, multiresolution multiclass classification can be a way forward in analysis of multiresolution data.

We need to consider multiresolution data because of the large number of cancer types and smaller number of samples making multiclass classification a challenging task. Furthermore, labels are unavailable for data in fine resolution. In such situations, learning from ambiguous labels [77] or partial labels [35] using clustering labels or the cancer types can help in the analysis of chromosomal aberrations data. Finally, analysis of multiresolution modelling also requires visualisation of the data as well as the results. Therefore, visualisation is also another direction for future work. In Publication V, we use banded matrix to visualise rules and cluster only in single resolution. Initial ideas to visualise multiresolution can borrow from a popular visualisation method in information visualisation known as the Fish eye view [48]. Similar to multiresolution data, Fish eye view also visualises data, providing users a detailed and also a global view.

Bibliography

> *If I have seen further, it is by standing on the shoulders of giants.*
>
> — ISAAC NEWTON
> *In a letter to his rival Robert Hooke (1676)*

[1] P. R. Adhikari and J. Hollmén. Preservation of Statistically Significant patterns in Multiresolution 0-1 Data. In T. Dijkstra, E. Tsivtsivadze, E. Marchiori, and T. Heskes, editors, *Proceedings of the 5th IAPR International Conference on Pattern Recgnition in Bioinformatics*, volume 6282 of *Lecture Notes in Computer Science*, pages 86–97, Nijmegen, The Netherlands, September 2010. Springer Berlin / Heidelberg.

[2] P. R. Adhikari and J. Hollmén. Fast Progressive Training of Mixture Models for Model Selection. In J.-G. Ganascia, P. Lenca, and J.-M. Petit, editors, *Proceedings of Fifteenth International Conference on Discovery Science (DS 2012)*, volume 7569 of *Lecture Notes in Artificial Intelligence*, pages 194–208. Springer-Verlag, October 2012.

[3] C. C. Aggarwal. *Data Streams: Models and Algorithms*, volume 31 of *Advances in Database Systems*. Springer, illustrated edition, 2007.

[4] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, New York, NY, USA, 1993. ACM.

[5] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, San Francisco, CA, USA, 12-15 September 1994. Morgan Kaufmann Publishers Inc.

[6] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, 1973.

[7] T. Ando. *Bayesian Model Selection and Statistical Modeling*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 2010.

[8] S. Babaei, M. Hulsman, M. J.T. Reinders, and J. de Ridder. Detecting recurrent gene mutation in interaction network context using multi–scale graph diffusion. *BMC Bioinformatics*, 14:29, January 2013.

[9] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[10] A. Bargiela and W. Pedrycz. *Granular Computing: An Introduction*. The Springer International Series in Engineering and Computer Science. Springer US, 2003.

[11] T. J. Barth, T. Chan, and R. Haimes. *Multiscale and Multiresolution Methods: Theory and Applications*. Lecture Notes in Computational Science and Engineering. Springer, 2002.

[12] J. M. S. Bartlett and D. Stirling. A Short History of the Polymerase Chain Reaction. In J. M.S. Bartlett and D. Stirling, editors, *PCR Protocols*, volume 226 of *Methods in Molecular Biology*, pages 3–6. Humana Press, 2003.

[13] M. Baudis. Online database and bioinformatics toolbox to support data mining in cancer cytogenetics. *Biotechniques*, 40(3):269–272, March 2006.

[14] M. Baudis. Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta-analysis of chromosomal CGH data. *BMC Cancer*, 7(1):226, December 2007.

[15] R. J. Baxter. *Exactly Solved Models in Statistical Mechanics*. Dover Publications, January 2008.

[16] S. D. Bay and M. J. Pazzani. Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

[17] R. E. Bellman. *Adaptive control processes - A guided tour*. Princeton University Press, Princeton, New Jersey, U.S.A., 1961.

[18] D. Bellot and P. Bessiére. Approximate Discrete Probability Distribution Representation using a Multi–Resolution Binary Tree. In *Proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence, 2003*, ICTAI '03, pages 498–503, 2003.

[19] P. A. Benn and M. A. Perle. Chromosome staining and banding techniques. In D. E. Rooney and B. H. Czepulkowski, editors, *Human Genetics: A Practical Approach*, pages 57–84. IRL Press, 1992.

[20] H. Bensmail, G. Celeux, A. E. Raftery, and C. P. Robert. Inference in model-based cluster analysis. *Statistics and Computing*, 7(1):1–10, 1997.

[21] M. Bianchini, M. Maggini, and L. Sarti. Object Recognition Using Multiresolution Trees. In D-Y Yeung, J. T. Kwok, A. Fred, F. Roli, and D. Ridder, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, volume 4109 of *Lecture Notes in Computer Science*, pages 331–339. Springer Berlin Heidelberg, 2006.

[22] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, July 2000.

[23] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Secaucus, NJ, USA, 2006.

[24] J. F. Bishop. *Cancer facts : a concise oncology text*. Harwood Academic Publishers, Amsterdam, The Netherlands, 1999.

[25] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, 18(2000):630–634, June 2000.

[26] D. Burdick, M. Calimlim, and J. Gehrke. MAFIA: a maximal frequent itemset algorithm for transactional databases. In D. Georgakopoulos and A. Buchmann, editors, *Proceedings of 17th International Conference on Data Engineering, 2001*, pages 443–452, 2001.

[27] R. Carlson. The pace and proliferation of biological technologies. *Biosecurity and bioterrorism : biodefense strategy, practice, and science*, 1(3):203–214, 2003.

[28] D. B. Carter. *Analysis of multiresolution data fusion techniques*. PhD thesis, Virginia Polytechnic Institute and State University, 1998.

[29] R. Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, July 1997.

[30] G. Celeux. Mixture Models for Classification. In R. Decker and H-J. Lenz, editors, *Advances in Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 3–14. Springer Berlin Heidelberg, 2007.

[31] J. Chen and A. Khalili. Order Selection in Finite Mixture Models With a Nonsmooth Penalty. *Journal of the American Statistical Association*, 103(484):1674–1683, 2008.

[32] V. S. Cherkassky and F. Mulier. *Learning from Data: Concepts, Theory, and Methods*. John Wiley & Sons, Inc., New York, NY, USA, first edition, 1998.

[33] D. M. Chickering and D. Heckerman. Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. *Machine Learning*, 29(2-3):181–212, November 1997.

[34] S. M. Cohen. Aristotle's Metaphysics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford Online, spring 2014 edition, 2014.

[35] T. Cour, B. Sapp, and B. Taskar. Learning from Partial Labels. *Journal of Machine Learning Research*, 12:1501–1536, July 2011.

[36] K. Crammer, M. Kearns, and J. Wortman. Learning from Multiple Sources. *Journal of Machine Learning Research*, 9:1757–1774, August 2008.

[37] "datum data". *Merriam-Webster's dictionary of English usage*. Merriam-Webster, Springfield, Massachusetts, 2002.

[38] J. de Ridder, J. Kool, A. Uren, J. Bot, L. Wessels, and M. J. T. Reinders. Co–occurrence analysis of insertional mutagenesis data reveals cooperating oncogenes. *Bioinformatics*, 23(13):i133–41, July 2007.

[39] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

[40] M. Di Zio, M. Scanu, L. Coppola, O. Luzi, and A. Ponti. Bayesian networks for imputation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(2):309–322, May 2004.

[41] U.S. DOE. *New Frontiers in Characterizing Biological Systems: Report from the May 2009 Workshop*. DOE/SC-0121. U.S. Department of Energy, Office of Science, 2009.

[42] S. G. Durkin and T. W. Glover. Chromosome Fragile Sites. *Annual Review of Genetics*, 41(1):169–192, 2007.

[43] L. Eldén. *Matrix Methods in Data Mining and Pattern Recognition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007.

[44] Electronics, Electrical Engineering Laboratory (National Institute of Standards, and Technology). *Measurements for competitiveness in electronics [microform] / prepared by the Electronics and Electrical Engineering Laboratory*. The Laboratory ; National Technical Information Service [distributor Gaithersburg, MD : Springfield, VA, first edition, 1993.

[45] B. S. Everitt and D. J. Hand. *Finite mixture distributions*. Chapman and Hall, London; New York, 1981.

[46] M. A. T. Figueiredo and A. K. Jain. Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, March 2002.

[47] C. Fraley and A. E. Raftery. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 41(8):578–588, 1998.

[48] G. W. Furnas. Generalized Fisheye Views. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '86, pages 16–23, New York, NY, USA, 1986. ACM.

[49] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nature reviews. Cancer*, 4(3):177–183, Mar 2004.

[50] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy. Mining Data Streams: A Review. *SIGMOD Record*, 34(2):18–26, June 2005.

[51] A. Gallo, P. Miettinen, and H. Mannila. *Finding Subgroups having Several Descriptions: Algorithms for Redescription Mining*, chapter 29, pages 334–345. SIAM, 2008.

[52] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A Survey on Concept Drift Adaptation. *ACM Computing Survey*, 46(4):44:1–44:37, March 2014.

[53] D. Gamberger and N. Lavrač. Expert-Guided Subgroup Discovery: Methodology and Application. *Journal of Artificial Intelligence Research*, 17(1):501–527, 2002.

[54] M. Garland. Multiresolution Modeling: Survey & Future Opportunities. In *Eurographics '99 – State of the Art Reports*, pages 111–131, 1999.

[55] G. C. Garriga, E. Junttila, and H. Mannila. Banded structure in binary matrices. *Knowledge and Information Systems*, 28(1):197–226, 2011.

[56] S. Geisser. A Predictive Approach to the Random Effect Model. *Biometrika*, 61(1):101–107, 1974.

[57] G. Giancarlo, G. Wagner, and M. Marten Sinderen van. A formal theory of conceptual modeling universals. In *1st Intl. Workshop on Philosophy and Informatics, WSPI 2004*, volume 112 of *CEUR workshop proceedings*. DFKI, 2004.

[58] W. Gilbert and A. Maxam. The Nucleotide Sequence of the lac Operator. *Proceedings of the National Academy of Sciences*, 70(12, Part I):3581–3584, December 1973.

[59] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, 1(1):341–352, March 2007.

[60] T. W. Glover, M. F. Arlt, A. M. Casper, and S. G. Durkin. Mechanisms of common fragile site instability. *Human Molecular Genetics*, 14(Supplement 2):R197–R205, 2005.

[61] I. R. Goodman, R. P. Mahler, and H. T. Nguyen. *Mathematics of Data Fusion*. Kluwer Academic Publishers, Norwell, MA, USA, 1997.

[62] P. V. Gopalacharyulu, E. Lindfors, C. Bounsaythip, T. Kivioja, L. Yetukuri, J. Hollmén, and M. Orešič. Data integration and visualization system for enabling conceptual biology. *Bioinformatics*, 21(Suppl.1):i177–i185, 2005.

[63] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199 – 220, 1993.

[64] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.

[65] R. Hammoud and R. Mohr. Biometrics: Promising frontiers for emerging identification market. Technical Report 3905, Unité de recherche INRIA Rhǒne-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France), March 2000.

[66] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, January 2007.

[67] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. Adaptive Computation and Machine Learning Series. MIT Press, 2001.

[68] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, second edition, February 2009.

[69] T. X. He. *Wavelet Analysis and Multiresolution Methods*. Lecture Notes in Pure and Applied Mathematics. Taylor & Francis, 2000.

[70] D. Heckerman. A Tutorial on Learning With Bayesian Networks. In M. I. Jordan, editor, *Learning in graphical models*, pages 301–354. MIT Press, USA, 1999.

[71] F. Herrera, C. J. Carmona, P. González, and M. Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, 29(3):495–525, 2011.

[72] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for Association Rule Mining — a General Survey and Comparison. *SIGKDD Explorations Newsletter*, 2(1):58–64, June 2000.

[73] J. Hollmén and J. Tikka. Compact and understandable descriptions of mixture of Bernoulli distributions. In M.R. Berthold, J. Shawe-Taylor, and N. Lavrač, editors, *Proceedings of the 7th International Symposium on Intelligent Data Analysis (IDA 2007)*, volume 4723 of *Lecture Notes in Computer Science*, pages 1–12, Ljubljana, Slovenia, September 2007. Springer-Verlag.

[74] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. S. Pierre, S. Twigger, O. White, and S. Y. Rhee. Big data: The future of biocuration. *Nature*, 455(7209):47–50, 2008.

[75] T. Huang, H. Peng, and K. Zhang. Model Selection for Gaussian Mixture Models. *arXiv preprint arXiv:1301.3558*, 2013.

[76] G. Hughes. On the Mean Accuracy of Statistical Pattern Recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, January 1968.

[77] E. Hüllermeier and J. Beringer. Learning from Ambiguously Labeled Examples. *Intelligent Data Analysis*, 10(5):419–439, September 2006.

[78] I. Huopaniemi, T. Suvitaival, J. Nikkilä, M. Orešič, and S. Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 26:i391–i398, 2010.

[79] A. Hussain and A. Visilsoanathan. Multiresolution Semantic Visualization of Network Traffic. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 364–367, Washington, DC, USA, 2011. IEEE Computer Society.

[80] A. Iske. *Multiresolution Methods in Scattered Data Modelling*. Springer Berlin Heidelberg, first edition, 2004.

[81] B. Jawerth and W. Sweldens. An Overview of Wavelet Based Multiresolution Analyses. *SIAM Review*, 36(3):377–412, September 1994.

[82] V. Jovanoski and N. Lavrač. Classification Rule Learning with APRIORI–C. In P. Brazdil and A. Jorge, editors, *Progress in Artificial Intelligence*, volume 2258 of *Lecture Notes in Computer Science*, pages 44–51. Springer Berlin Heidelberg, 2001.

[83] B. H. Juang and L. R. Rabiner. A probabilistic distance measure for Hidden Markov Models. *AT&T Technical Journal*, 64(2):391–408, February 1985.

[84] J. B. Kadane and N. A. Lazar. Methods and Criteria for Model Selection. *Journal of the American Statistical Association*, 99(465):279–290, March 2004.

[85] A. Kallioniemi, O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *SCIENCE*, 258(5083):818–821, October 30 1992.

[86] G. Karčiauskas. Learning of Latent Class Models by Splitting and Merging Components. In P. Lucas, J. A. Gómez, and A. Salmerón, editors, *Advances in Probabilistic Graphical Models*, volume 214 of *Studies in Fuzziness and Soft Computing*, pages 235–251. Springer Berlin Heidelberg, 2007.

[87] E. Kettunen, A. G. Nicholson, B. Nagy, J. K. Seppänen, T. Ollikainen, G. Ladas, V. Kinnula, M. Dusmet, S. Nordling, J. Hollmén, D. Kamel, P. Goldstraw, and S. Knuutila. L1CAM, INP10, P-cadherin, tPA and ITGB4 over-expression in malignant pleural mesotheliomas revealed by combined use of cDNA and tissue microarray. *Carcinogenesis*, 26(1):17–25, September 2005.

[88] J. D. Khoury, N. M. Tannir, M. D. Williams, Y. Chen, H. Yao, J. Zhang, E. J. Thompson, F. Meric-Bernstam, L. J. Medeiros, J. N. Weinstein, et al. The Landscape of DNA Virus Associations Across Human Malignant Cancers Using RNA-seq: An Analysis of 3,775 Cases. *Journal of Virology*, 87(16):8916–8926, 2013.

[89] P. M. Kim. *Understanding Subsystems in Biology through Dimensionality Reduction, Graph Partitioning and Analytical Modeling*. PhD thesis, Massachusetts Institute of Technology, February 2003.

[90] R. A. King, J. I. Rotter, and A. G. Motulsky, editors. *The Genetic Basis of Common Diseases.* Oxford Monographs on Medical Genetics. Oxford University Press, second edition, 2002.

[91] I. R. Kirsch. *The causes and consequences of chromosomal aberrations*. CRC Press, 1993.

[92] H. A. Klein. *The Science of Measurement: A Historical Survey*. Dover Books on Mathematics. Dover Publications, 2012.

[93] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding Interesting Rules from Large Sets of Discovered Association Rules. In N. R. Adam, B. K. Bhargava, and Y. Yesha, editors, *Proceedings of the Third International Conference on Information and Knowledge Management*, CIKM '94, pages 401–407, New York, NY, USA, 1994. ACM.

[94] S. Knuutila, K. Autio, and Y. Aalto. Online Access to CGH Data of DNA Sequence Copy Number Changes. *American Journal of Pathology*, 157(2):689–689, August 2000.

[95] P. Koikkalainen. Progress with the Tree-Structured Self-Organizing Map. In A. G. Cohn, editor, *In Proceedings on 11th European Conference on Artificial Intelligence (ECAI)*, pages 211–215. European Committee for Artificial Intelligence (ECCAI), John Wiley & Sons, Ltd., 1994.

[96] S. Kullback and R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

[97] P. Lahiri, editor. *Model Selection*. Institute of Mathematical Statistics Lecture Notes Monograph. Institute of Mathematical Statistics, illustrated edition, 2001.

[98] N. Lavrač, A. Vavpetič, L. Soldatova, I. Trajkovski, and P. Kralj Novak. Using Ontologies in Semantic Data Mining with SEGS and g-SEGS. In

T. Elomaa, J. Hollmén, and H. Mannila, editors, *Proceedings of the International Conference on Discovery Science (DS '11)*, volume 6926 of *Lecture Notes in Computer Science*, pages 165–178. Springer Berlin Heidelberg, 2011.

[99] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224–270, 1994.

[100] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[101] B. Liu, W. Hsu, and Y. Ma. Integrating Classification and Association Rule Mining. In *Proceedings of the 4th international conference on Knowledge Discovery and Data mining (KDD'98)*, pages 80–86. AAAI Press, August 1998.

[102] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012:1–11, 2012.

[103] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680, 1996.

[104] C. Lynch. Big data: How do your data grow? *Nature*, 455(7209):28–29, September 2008.

[105] S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[106] H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. In U. M. Fayyad and R. Uthurusamy, editors, *AAAI Workshop on Knowledge Discovery in Databases (KDD-94)*, pages 181–192, Seattle, Washington, 1994. AAAI Press.

[107] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, illustrated edition, 1999.

[108] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, June 2011.

[109] E. R. Mardis. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, June 2008.

[110] E. R. Mardis. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, February 2011.

[111] V. Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–260, June 2013.

[112] G. J. McLachlan. On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):318–324, 1987.

[113] G. J. McLachlan and K. E. Basford. Mixture models. Inference and applications to clustering. *Statistics: Textbooks and Monographs, New York*, 1, 1988.

[114] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley series in probability and statistics. Wiley, second edition, 2008.

[115] G. J. McLachlan and D. Peel. *Finite mixture models*, volume 299 of *Probability and Statistics – Applied Probability and Statistics Section*. Wiley, 2000.

[116] D. W. Mcshea. Complexity and evolution: What everybody knows. *Biology and Philosophy*, 6(3):303–324, July 1991.

[117] M. Meila and M. I. Jordan. Learning with Mixtures of Trees. *Journal of Machine Learning Research*, 1:1–48, Oct 2000.

[118] V. Melnykov and R. Maitra. Finite Mixture Models and Model–Based Clustering. *Statistics Surveys*, 4(1):80–116, 2010.

[119] P. Milanfar. *Super-resolution imaging*. CRC Press, 2010.

[120] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[121] F. Monsteller and J. W. Tukey. Data analysis, including statistics. In G. Lindzey and E. Aronson, editors, *Handbook of Social Psychology*, volume 2, chapter 10, pages 80–203. Addison-Wesley, Reading, MA, 1968.

[122] A. Moore. Very Fast EM-based Mixture Model Clustering Using Multiresolution KD–trees. In M. Kearns and D. Cohn, editors, *Advances in Neural Information Processing Systems*, pages 543–549. Morgan Kaufman, April 1999.

[123] G. E. Moore. Cramming More Components onto Integrated Circuits. *Electronics*, 38(8):114–117, April 1965.

[124] D. Mukherjee, Q. M. J. Wu, and T. M. Nguyen. Multiresolution Based Gaussian Mixture Model for Background Suppression. *IEEE Transactions on Image Processing*, 22(12):5022–5035, 2013.

[125] S. Myllykangas, J. Himberg, T. Böhling, B. Nagy, J. Hollmén, and S. Knuutila. DNA copy number amplification profiling of human neoplasms. *Oncogene*, 25(55):7324–7332, November 2006.

[126] S. Myllykangas, J. Tikka, T. Böhling, S. Knuutila, and J. Hollmén. Classification of human cancers based on DNA copy number amplification modeling. *BMC Medical Genomics*, 1(15), May 2008.

[127] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, September 2013.

[128] S-K. Ng and G. J. McLachlan. Robust Estimation in Gaussian Mixtures Using Multiresolution Kd-trees. In C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen, editors, *Proceedings of the 7th International Conference on Digital Image Computing: Techniques and Applications*, pages 145–154. CSIRO Publishing, 2003.

[129] P. Novak, N. Lavrač, and G. I. Webb. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *Journal of Machine Learning Research*, 10:377–403, February 2009.

[130] G. Obe and Vijayalaxmi. *Chromosomal alterations: methods, results, and importance in human health*. Springer, 2007.

[131] A. Oliveira-Brochado and F. V. Martins. Assessing the Number of Components in Mixture Models: a Review. FEP Working Papers 194, Universidade do Porto, Faculdade de Economia do Porto, November 2005.

[132] P. Panov. *A Modular Ontology of Data Mining*. Doctoral dissertation, Jožef Stefan International Postgraduate School, July 2012.

[133] G. Piatetsky-Shapiro. Discovery, Analysis, and Presentation of Strong Rules. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.

[134] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20:207–211, 1998.

[135] R. B. Potts. Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(01):106–109, 1952.

[136] C. K. Reddy and J-H. Park. Multi-resolution Boosting for Classification and Regression Problems. In T. Theeramunkong, B. Kijsirikul, N. Cercone, and T-B. Ho, editors, *Advances in Knowledge Discovery and Data Mining*, volume 5476 of *Lecture Notes in Computer Science*, pages 196–207. Springer Berlin Heidelberg, 2009.

[137] J. Rissanen. Modeling By Shortest Data Description. *Automatica*, 14(5):465 – 471, 1978.

[138] S. W. Roh, G. C. J. Abell, K-H. Kim, Y-D. Nam, and J-W Bae. Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends in Biotechnology*, 28(6):291–299, 2010.

[139] B. Russell. On the Relations of Universals and Particulars. *Proceedings of the Aristotelian Society*, 12:1–24, 1911.

[140] S. Sagiroglu and D. Sinanc. Big data: A review. In *International Conference on Collaboration Technologies and Systems (CTS), 2013*, pages 42–47, 2013.

[141] M. Schwartz, E. Zlotorynski, and B. Kerem. The molecular basis of common and rare fragile sites. *Cancer Letters*, 232(1):13–26, 2006.

[142] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, March 1978.

[143] L. G. Shaffer and N. Tommerup. *ISCN 2005: An International System for Human Cytogenetic Nomenclature(2005) Recommendations of the International Standing Committee on Human Cytogenetic Nomenclature*. Karger, 2005.

[144] P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72, 2000.

[145] M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis, and machine vision*. Thomson, Toronto, 3 edition, 2008.

[146] P. Stankiewicz and J. R. Lupski. Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61(1):437–455, 2010.

[147] B. W. Stewart and C. P. Wild, editors. *World Cancer Report 2014*. International Agency for Research on Cancer (IARC) Nonserial, 2008.

[148] G. W. Stewart. *Matrix Algorithms: Volume 1, Basic Decompositions*. Society for Industrial Mathematics, illustrated edition, 1998.

[149] M. Stone. Cross-validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society*, 36(2):111–147, 1974.

[150] S. Sun. A survey of multi–view machine learning. *Neural Computing and Applications*, 23(7–8):2031–2038, 2013.

[151] B. Swartout, R. Patil, K. Knight, and T. Russ. Towards distributed use of large–scale ontologies. In *Proceedings of AAAI Symposium on Ontological Engineering*, pages 138–148, 1996.

[152] N. Tatti. *Advances in Mining Binary Data: Itemsets as Summaries*. PhD thesis, Helsinki University of Technology, Faculty of Information and Natural Sciences, 2008.

[153] J. Tikka, J. Hollmén, and S. Myllykangas. Mixture Modeling of DNA copy number amplification patterns in cancer. In F. Sandoval, A. Prieto, J. Cabestany, and M. Graña, editors, *Proceedings of the 9th International Work-Conference on Artificial Neural Networks (IWANN 2007)*, volume 4507 of *Lecture Notes in Computer Science*, pages 972–979, San Sebastián, Spain, 2007. Springer-Verlag.

[154] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. SMEM Algorithm for Mixture Models. *Neural Computation*, 12(9):2109–2128, 2000.

[155] A. Usvasalo, E. Elonen, U. M. Saarinen-Pihkala, R. Räty, A. Harila-Saari, P. Koistinen, E-R. Savolainen, S. Knuutila, and J. Hollmén. Prognostic classification of patients with acute lymphoblastic leukemia by using copy number profiles identified from array-based comparative genomic hybridization data. *Leukemia Research*, 34(11):1476–1482, November 2010.

[156] A. Vavpetič and N. Lavrač. Semantic Subgroup Discovery Systems and Workflows in theSDM-Toolkit. *The Computer Journal*, 56(3):304–320, 2013.

[157] A. Vavpetič, V. Podpečan, and N. Lavrač. Semantic subgroup explanations. *Journal of Intelligent Information Systems*, 42(2):233–254, 2013.

[158] B. Vogelstein and K. W. Kinzler. *The genetic basis of human cancer*. McGraw-Hill, New York, 2002.

[159] I. Žliobaitė and J. Hollmén. Optimizing regression models for data streams with missing values. *Machine Learning*, page In Press, 2014.

[160] R. Walpole, R. Myers, S. Myers, and K. Ye. *Probability & Statistics for Engineers & Scientists*. Prentice Hall, NJ, USA, nineth edition, 2012.

[161] Y. Wang, J. Hayakawa, F. Long, Q. Yu, A. H. Cho, G. Rondeau, J. Welsh, S. Mittal, I. De Belle, E. Adamson, M. McClelland, and D. Mercola. "promoter array" studies identify cohorts of genes directly regulated by methylation, copy number change, or transcription factor binding in human cancer cells. *Annals of the New York Academy of Sciences*, 1058(1):162–185, November 2005.

[162] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, April 1953.

[163] E. Weinan. *Principles of Multiscale Modeling*. Cambridge University Press, 2011.

[164] D. B. West. *Introduction to graph theory*. Prentice Hall, second (illustrated) edition, 1996.

[165] A. S. Willsky. Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, August 2002.

[166] R. Wilson. MGMM: multiresolution Gaussian mixture models for computer vision. In *Proceedings of 15th International Conference on Pattern Recognition*, volume 1, pages 212–215. IEEE, September 2000.

[167] J. H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5(3):329–350, July 1970.

[168] M-J. Woo and T. N. Sriram. Robust Estimation of Mixture Complexity. *Journal of the American Statistical Association*, 101(476):1475–1486, December 2006.

[169] C. F. J. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, March 1983.

[170] J. T. Yao and Y. Y. Yao. A Granular Computing Approach to Machine Learning. In L. Wang, S. K. Halgamuge, and X. Yao, editors, *Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery: Computational Intelligence for the E-Age, 2 Volumes, November 18-22, 2002, Orchid Country Club, Singapore*, FSDK'02, pages 732–736, 2002.

[171] A. Zellner, H. A. Keuzenkamp, and M. McAleer, editors. *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*. Cambridge University Press, March 2009.

[172] H. zur Hausen. The search for infectious causes of human cancers: Where and why. *Virology*, 392(1):1 – 10, September 2009.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

Aalto-DD80/2014	Toivola, Janne Advances in Wireless Damage Detection for Structural Health Monitoring. 2014.
Aalto-DD105/2014	Parkkinen, Juuso Probabilistic components of molecular interactions and drug responses. 2014.
Aalto-DD108/2014	Faisal, Ali Retrieval of Gene Expression Measurements with Probabilistic Models. 2014.
Aalto-DD110/2014	Virtanen, Seppo Bayesian latent variable models for learning dependencies between multiple data sources. 2014.
Aalto-DD120/2014	Bergström-Lehtovirta, Joanna The Effects of Mobility on Mobile Input. 2014.
Aalto-DD127/2014	Zhang, He Advances in Nonnegative Matrix Decomposition with Application to Cluster Analysis. 2014.
Aalto-DD138/2014	Sovilj, Dušan Learning Methods for Variable Selection and Time Series Prediction. 2014.
Aalto-DD144/2014	Eirola, Emil Machine learning methods for incomplete data and variable selection. 2014.
Aalto-DD149/2014	Äijö, Tarmo Computational Methods for Analysis of Dynamic Transcriptome and Its Regulation Through Chromatin Remodeling and Intracellular Signaling. 2014.
Aalto-DD156/2014	Sjöberg, Mats From pixels to semantics: visual concept detection and its applications. 2014.

Aalto-DD 157/2014

Machine learning and data mining algorithms have recently been very popular as they provide an opportunity to analyze, understand, and extract information and knowledge from the overwhelming amount of data being produced everyday. Similarly, over the years, measurement devices and technologies have also been improving steadily providing us a platform to measure finer details of any phenomenon thus generating data in different representations. However, current state-of-the art algorithms are capable of analyzing data in a single representation from a single data source only thus ignoring any available supplementary information. This thesis presents novel methods to analyze data in different representations within a single analysis. Additionally, methods to speed-up the algorithms proposed in the thesis are also developed. Similarly, the proposed algorithms consider supplementary background information in the analysis of data in different representations.

9 789526 059037

ISBN 978-952-60-5903-7 (printed)
ISBN 978-952-60-5904-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS