Intelligent Data Analysis for Knowledge Discovery, Patient Monitoring and Quality Assessment

N. Peek1; S. Swift2
1Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands;
2School of Information Systems, Computing and Mathematics, Brunel University, London, UK

Summary
Objective: To introduce the focus theme of Methods of Information in Medicine on Intelligent Data Analysis for Knowledge Discovery, Patient Monitoring and Quality Assessment.
Methods: Based on two workshops on Intelligent Data Analysis in bioMedicine (IDAMAP) held in Washington, DC, USA (2010) and Bled, Slovenia (2011), six authors were invited to write full papers for the focus theme. Each paper was thoroughly reviewed by anonymous referees and revised one or more times by the authors.
Results: The selected papers cover four ongoing and emerging topics in Intelligent Data Analysis (IDA), being i) systems biology and metabolic pathway modelling; ii) gene expression data modelling; iii) signal processing from in-home monitoring systems; and iv) quality of care assessment. Each of these topics is discussed in detail to introduce the papers to the reader.
Conclusion: The development and application of IDA methods in biomedicine is an active area of research which continues to blend with other subfields of medical informatics. As data become increasingly ubiquitous in the biomedical domain, the demand for fast, smart and flexible data analysis methods is undiminished.

1. Introduction

Twenty years ago, researchers from artificial intelligence (AI) and computer science became increasingly interested in automatic methods for processing, analysing and interpreting large datasets. The reason was that automatic data collection methods were creating vast arrays of data, whose type and size could not be handled by existing methods from statistics and machine learning (ML). Well-known examples are electronically recorded transactions at checkout counters of retailers (“market basket analysis”) [1] and logged user interactions with information systems and websites [2, 3]. This led to the development of a new, interdisciplinary research field that became known under a variety of names such as Data Mining (DM) [4–6], Knowledge Discovery from Databases (KDD) [7, 8] and Intelligent Data Analysis (IDA) [9]. The field has always been closely linked to computer science, statistical modelling, ML, and AI.

Electronic data have also become ubiquitous in the biomedical field, and a corresponding need for analytical methods has...
growth over the years [10]. Examples are the use of routinely collected data in clinical information systems for diagnostic and prognostic purposes [11–15]; interpreting clinical signal and image data, which are rapidly increasing in volume and granularity [16, 17]; secondary use of hospital discharge and insurance claims data for epidemiology, biosurveillance and careflow analysis [18, 19]; integration of clinical data and molecular-biological data from DNA microarrays and other high-throughput techniques [20]; real-time decision support based on ICU monitoring data [21] and interpretation of data from wearable sensors [22, 23].

Since 1996, a series of annual workshops devoted to computational methods for data analysis in biology and medicine has been organised under the name IDAMAP [24] by the IMIA Working Group on Intelligent Data Analysis and Data Mining [25]. The workshops are typically co-located with scientific conferences in general AI, medical AI, or medical informatics and they usually attract 30–50 attendants from diverse fields. Gathering in an informal setting, workshop participants have the opportunity to meet and discuss selected technical topics in an atmosphere that fosters the active exchange of ideas among researchers and practitioners. The most recent workshops were held in Washington, DC, USA (2008 and 2010), Verona, Italy (2009), and Bled, Slovenia (2011). Previously, revised and extended selections of work presented at IDAMAP workshops were published in thematic journal issues [26–30], jointly providing a historical and broad-based compendium of the field.

In this focus theme, we present selected contributions to two recent IDAMAP workshops, held in 2010 and 2011. Six researchers that presented their work at these workshops were invited to write a paper for the focus theme. Each paper was thoroughly reviewed by two or three anonymous referees, and revised one or more times by the authors. The selected papers cover four recurrent and emerging topics in intelligent data analysis, being i) systems biology and metabolic pathway modelling; ii) gene expression data modelling; iii) signal processing from in-home monitoring systems; and iv) quality of care assessment. Each of these topics is discussed below in more detail, including a brief introduction the focus theme papers.

2. Recurrent and Emerging Topics in Intelligent Data Analysis

2.1 Systems Biology and Metabolic Pathway Modelling

Modern biology has often been criticised for being a science that has produced a huge “parts list” for many thousands of biological organisms and not coming up with a procedure for deriving meaning from this vast catalogue [31]. This criticism, whilst being unfair, does have some seeds of truth. Knowing the exact type and location of each part that makes up a car is useful but will not help in diagnosing why it will not start on a cold morning. What are needed are models and techniques to discover how these parts interact and that can be used to answer “what if” scenarios. This is the goal of the rapidly growing new field of Systems Biology [32–34]. This field proposes a holistic systems approach to the understanding of biological systems, using techniques from complexity science, data mining and mathematics. Much of the efforts of Systems Biology have been on the cellular level, looking at the complex interactions between enzymes and other products in signalling, gene regulation and metabolic pathways. Genes can play a role in more than one pathway, creating vast networks of interactions; this makes the process of modelling the pathways extremely complex.

Recent advances in the area of metabolic pathway modelling have concentrated in data integration: looking at pathway data, clinical data and gene expression data; this is discussed and applied in the paper by Pavlidis et al. in this focus theme [35]. The authors use both pathway data and gene expression data to look at the behaviour of genes that play a role in multiple pathways. Pathway modelling in itself has problems, such as missing or incompatible data, small volumes of data and all of the issues arising with conducting expensive genetic level experiments. Interest in this area has “exploded” due to the availability of web based repositories such as Kegg [36], Reactome [37] and the NCBI based Entrez [38]. The pathway models themselves for these pathways are as complex as they are numerous, for example the well tried techniques of Petri Nets have been used with some success, but more recently the focus has been on the use of differential equations. Here measures of activity of components within a cell are described in terms of interacting decay rates which give rise to systems of differential equations [39] which are structurally complex and have numerous parameters (kinetics). Differential equation based models have had wide ranging success in modelling physical systems in general, but have tended to have been used in areas where data is more abundant in sample size, such as electrical or power systems. This is where Intelligent Data Analysis has started to play a major part. Model selection, model fitting and model verification have been widely studied in the IDA community and the experiences are now being put to use in Pathway Modelling to address the afore mentioned issues. In the paper by Patterson et al. in this issue [40], IDA techniques have been used to complement image processing techniques to understand a component of a complex trafficking pathway. Cell images are used to construct trajectories of a component called vesicles and Hidden Markov models are used to model them dynamically. These models are then clustered on their parameters to see if different drug treatments behave in different ways.

Pathway modelling and Systems Biology are new and exciting area which are rapidly growing, with applications in the development of new drugs and treatments. This path is going to be a long and arduous one, but never the less one worth taking.

2.2 Gene Expression Data Modelling

As mentioned above, gene expression data have been used to assist in the modelling and understanding of pathways. This technology has now been around for over a decade and was revolutionary in the way that it
could provide measures of gene activity for thousands of genes at a time [41]. The gene expression levels of whole pathways and even small organisms could be measured by placing all of the genes onto a single glass slide, stimulating them to produce RNA and then using image processing techniques to measure each gene’s activity. The technology started off being extremely expensive, but recent advances have seen an enormous decrease in the cost. Repositories such as the EBI [42] and NCBI [38] provided global access to publicly funded dataset resulting in a plethora of papers covering this topic, see [43] for a review of a large number of typical papers. IDA and ML techniques have been applied in abundance in the analysis of these datasets which were typically extremely high in dimensionality, small in sample size, noisy and high in missing observations. Advanced image processing techniques and noise reduction methods have to be applied to the slides even before a dataset could be looked at.

It quickly became apparent that classification techniques had a lot to offer in the way of modelling the data and linking it to clinical data. There are numerous papers reporting the link between a selection of genes and conditions such as cancer. Additionally, these techniques could be used to infer the function of unknown genes, using the notion of “guilt by association”, see e.g. [43, 44]; here you assume that if a gene of unknown function is clustered with a number of genes that all have the same function, e.g. DNA repair, then there is evidence to support the hypothesis that the unknown gene has the same function. One problem is that there may be hundreds of genes related to a condition, but a very small number of samples. Many of the classification models simply cannot be used or would have a low level of accuracy or confidence. This is where gene subset selection comes in. Subsets of genes can be derived from the dataset that produce a better classification model than considering all of the data at once. However these techniques, especially those based on wrapping the search around a classifier, can take an extremely long time to run. Other approaches to selecting sets of “interesting genes” have been proposed, a new and interesting one is found in this focus theme [45]; here the authors use Principle Components Analysis to select subsets of genes for predictive modelling using multiple gene expression datasets.

All of the problems associated with gene expression data will probably have to be revisited with the advent of RNA-Seq [46] techniques (aka next generation sequencing and deep sequencing; widely regarded as being the successor for gene expression data) and we can expect a whole new set of problems to rise; interesting times are ahead!

2.3 Quality of Care Assessment

The remaining two topics in this focus theme relate to clinical healthcare rather than biology. Since the publication of two landmarks reports by the Institute of Medicine [47, 48], quality improvement and patient safety have dominated the healthcare research agenda. The reports led to wide awareness that evidence from scientific studies often does not reach medical practice, that medical care often fails to reach quality standards laid down in clinical practice guidelines and that preventable medical errors are responsible for substantial numbers of adverse events in routine care. As a result, a wide variety of quality improvement strategies is nowadays being applied in clinical practice. Examples are educational interventions, audit and feedback, financial and regulatory incentives, and computerised decision support at the point of care [49, 50]. A common element in many quality improvement strategies is assessment of the quality of care in clinical practice, by monitoring measurable aspects of quality [51]. Usually, a distinction is made between structural elements (e.g., number of available beds), processes (e.g., adherence to guidelines and protocols) and outcomes of care [52]. Frequently-used outcome measures are patient mortality, occurrence of adverse events, and recurrence of disease. In this focus theme, Stiglic and Kokol [53] present an instance-based subgroup discovery method that is applied to administrative hospital discharge data in order to identify diagnostic groups with increased risks of adverse outcomes (e.g., a postoperative infection).

One of the earliest fields where quality-of-care assessment was performed is critical care, typically focusing on mortality outcomes. Many countries have started initiatives to compare average mortality outcomes between intensive care units (ICUs) from different hospitals [54, 55]. However, patients who are older, are more severely ill, or have more co-morbidities generally respond worse to medical treatment and therefore have poorer chances of survival than other patients. Therefore one needs to adjust for patient case-mix before comparing mortality outcomes. This is done by constructing a model that predicts the probability of each patient’s survival, based on their demographic data, prior disease history, reason for ICU admission and physiological status at the time of ICU admission. The construction, validation, and use of the predictive models for case-mix adjustment is an important field of research. Although logistic regression analysis is the most popular modelling approach for the models [56, 57], a number of alternatives have been explored based on methods from ML and IDA [58–61]. As time progresses, predictive models for case-mix adjustment may become outdated when ICU treatment strategies are subject to policy changes. This topic is largely unaddressed in the literature to date. In this focus theme, Minne et al. [62] compare the “shelf life” of a classification tree model and a logistic regression model for ICU case-mix adjustment using methods from statistical process control.

2.4 Signal Processing from In-home Monitoring Systems

While institutional medicine developed in the 20th century, in previous centuries almost all ailments were cared for at home. A reverse development is foreseen for the decades to come, driven by societal needs to reduce healthcare spending, individual preferences to live longer independently at home, and by technological advances fostering the availability digital broadband communication in most households [63]. It was already shown in the Kaiser Permanente Tele-Home Health Research Project [64] that remote video technology in the home setting, for real-time interaction between nurses and
patients, can yield a considerable reduction of healthcare costs by decreasing the number of hospitalisations. More recently, initiatives have started to exploit the use of sensor networks in the domestic environment to continuously monitor physiological parameters and daily activities. A sensor network typically consists of several types of wired or wireless sensors mounted in different places throughout the resident’s apartment, including motion sensors, bed sensors, and temperature sensors. Alwan et al. [65] showed that care efficiency can be increased by giving professional caregivers remote access to the monitoring data from domestic sensor networks, by lowering the amount of billable interventions and hospital days.

A more ambitious approach was taken in the Liverpool Telecare project [66] which used a domestic sensor network to monitor activity levels and behavioural patterns, combined with intelligent data analysis methods to automatically detect situations of cause for concern. In such situations, characterised by departure from the individual’s normal behaviour (e.g., abnormal long lack of activity), a direct connection into Liverpool’s existing call centre infrastructure was used to initiate action by caregivers. The system used adaptive algorithms to create personalised thresholds, based on an individual’s behavioural patterns, for abnormal activity detection. As activity levels vary considerably between individuals, the approach aimed to create thresholds that were optimal for the individual in question, the time of day and the room occupied.

Generally speaking, automatic classification of domestic sensor data requires signal processing methods such as filtering of the acquired signals to remove measurement artefacts, feature extraction in the time and frequency domains [67], and subsequent classification of pattern based on the extracted features. As an increasing number of studies investigate in-home monitoring systems, a corresponding rise in methodological studies is seen [68]. Thge full article in this focus theme, by Popescu and Mahnot [69], investigates the automatic, early detection of mental illnesses such as dementia from behavioural patterns captured by domestic sensor networks. They address a particular problem in this field, arising from the fact that any classification algorithm will need annotated signals to derive the meaning of specific signal patterns. Unambiguous annotation of individual sensor data from in-home monitoring networks is however extremely laborious and not always possible. For this reason, Popescu and Mahnot explore the use of multiple instance learning, a classification approach which does not require all individual sensor data to be annotated but can work with annotated "bags" (sets) of data.

3. Conclusion

The development and application of IDA methods in biomedicine is an active area of research which continues to blend with other subfields of medical informatics. As data become increasingly ubiquitous in the biomedical domain, the demand for fast, smart and flexible data analysis methods is undiminished. In this focus theme, we present six methodological studies that exemplify the challenges and solutions of the field.

Acknowledgement

The editors would like to thank their co-organizers of the IDAMAP workshops in Washington (Kirk Philips) and Bled (Allan Tucker, John Holmecs, Riccardo Bellazzi). In addition, they are grateful to all reviewers of the papers in this focus theme.

References

322 Editorial


