A Computational Biology Framework

Creating a platform for biomedical engineers to efficiently do their research

T.P.A. Beishuizen (0791613)
Biomedical Engineering - Computational Biology
Data Engineering - Information Systems
Eindhoven, University of Technology
Email: `t.p.a.beishuizen@student.tue.nl`

November 14, 2017

Contents

1 Introduction

At the Computational Biology department (cBio) of Biomedical Engineering (BME), many requests are made to analyse gathered data. This data usually stems from research in hospitals, but can also be from other BME groups and publicly available. Currently a standard is missing to efficiently analyse those data sets. With the vast number of data sets that are available, such a standard in the form of a framework on data analysis would be valuable. It would speed up projects and give them a higher chance to succeed the goal, due to improved efficiency. Before a framework can be made however a research must be done on all aspects that influence a research.

First an extensive background on important topics for such a framework will be discussed. Four different parts are explained why they are important for the creation of such a framework. These parts are: biomedical data (data used for analysis), data analysis goal (how does a goal influence the choice of analysis), data analysis frameworks (which programs and frameworks can be used for analysis) and at last biomedical knowledge (what engineers of BME already know about data analysis). After the extensive background research, a research question will be formulated with several sub-questions for each of the four parts and a hypothesis as an answer for each of the four questions.

2 Background

Biomedical engineering can be seen as a specific part of engineering with a wide variety of topics. These topics can be theoretical, non-experimental undertakings, but also state-of-the-art applications. Not only research and development can be used, but also implementation and operation. Combining all of these different parts in one definition is hard. [1] For this project, the focus is mainly on research and development and a layout for such a biomedical project can be given.

When a biomedical engineer starts a project, at the start usually only a data set and the research goal are known. To achieve that certain goal from the data set, several different aspects influence the project's course and development. At first obviously the data itself is a big part of such an influencer as the research is restricted to limitations from it. Examples of such restrictions are multidimensionality, set size, data heterogeneity and missing feature values. The other obvious influencer is the main research goal. Since the biomedical engineer wants to achieve a certain goal, the approach outcome must match that goal for the research to be successful. Most goals are focused around either data mining, extracting relations from available data, or modelling, creating a model within data features. A third influencer is the availability of data analysis frameworks. The steps to take from data to goal do not only include an approach, but also a framework to execute it. The choice of a certain framework has a big impact on the project, as each one of them has its own advantages and disadvantages. The two most well known frameworks within BME are MATLAB and Python, however some engineers have used R, Java or C++ and there are still other possibilities. A last big influencer is the biomedical knowledge. What experience the scientist already has with similar projects can greatly influence the choice of approach and framework. Knowledge of the supervisor and publicly known information on the research subject from books and articles also influence the approach, as already known outcomes do not have to be researched again.

Knowledge discovery is a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from large collections of data [18]. One of the knowledge discovery steps is the data mining step concerned with actual extraction of knowledge from data. Design of a framework for a knowledge discovery process is important. Researchers have described a series of steps that constitute the KD process, which range from few steps that usually include data collection and understanding, data mining and implementation, to more sophisticated models like the nine-step model proposed by Fayyad et al. [19], or the six-step DMKD process model by Cios et al. [13] and Cios and Moore [14]. The latter model has been successfully applied to several medical problem domains [24,40]; it added several extensions to the CRISP-DM model [17]. [2]

2.1 Biomedical data

One of the core issues in biomedical data analysis and mining is the so-called ‘curse of dimensionality’ [5–7], particularly the biomedical data are characterized by relatively few instances and presented in a high-dimensional feature space. Irrelevant features not only lead to insufficient classification accuracy, but also add extra difficulties in finding potentially useful knowledge [8,9]. Excluding irrelevant features facilitates data visualization and improves the understanding of the computational models, the feature selection has thus become one of the main sub-fields in biomedical data mining [10–12]. In addition, appropriate feature selection is able to reduce the requirements of measurement and storage and thus minimize the cost in database storage and management [10,13]. In the context of classification, the main goal of feature selection is to search for an optimal feature subset from the initial feature set that lead to improved classification performance and efficiency in generating classification model. [3]

Raw medical data are voluminous and heterogeneous. Medical data may be collected from various images, interviews with the patient, laboratory data, and the physician’s observations and interpretations. All these components may bear upon the diagnosis, prognosis, and treatment of the patient, and cannot be ignored. This heterogeneity requires high capacity data storage devices and new tools to analyze such data. he physician’s interpretation of images, signals, or any other clinical data, is written in unstructured free-text English, that is very difficult to standardize and thus difficult to mine. Nearly all diagnoses and treatments in medicine are imprecise, and are subject to rates of error. Another unique feature of medical data mining is that the underlying data structures of medicine are poorly characterized mathematically, as compared to many areas of the physical sciences. Because of the sheer volume and heterogeneity of medical databases, it is unlikely that any current data mining tool can succeed with raw data. In any large database, we encounter a problem of missing values. The medical data set may contain redundant, insignificant, or inconsistent data objects and/or attributes. [2]

First, in many cases, the quality of data in the biomedical and healthcare fields is inferior to that found in other fields because of many reasons: (1) Medical data inevitably contains many missing values [Ichise, R., and Numao Learning, M., First-order rules to handle medical data. NII Journal]. This occurs because even patients with the same disease(s) do not always undergo identical examinations and lab tests (due to different ages, symptoms, family histories, and/or risks of complications) which results in different, and sometimes more or less, data sets being generated. In addition, medical data often contains time-series attributes (meaning that dates of examinations and lab tests are very important from a clinical perspective) so researchers must handle these data sets with special consideration of the time element. (2) Because hospital information systems or hospital databases are primarily designed for financial/billing purposes and not for medical/clinical purposes [24, 25], it can be especially challenging to obtain high quality data for clinical data mining. (3) In the United States, for example, many hospitals do not use full (i.e., no-paper) EMR systems. Thus, much of medical data (especially lab test results) are paper-based which, in turn, results in medical data that are often incomplete in terms of electronic availability [26]. In addition, much of the historical patient data are paper-based or in scanned-digital format so those data cannot be used for data mining without significant data preparation. [4]

There is the pressing need to combine vast amounts of diverse data, including structured, semi-structured and weakly structured data and unstructured information [11]. [5]

The problem is that these data sets are characterized by heterogeneous and diverse features. Individual data collectors prefer their own different schema or protocols for data recording, and the diverse nature of the applications used results in various data representations. Data heterogeneity and diverse dimensionality issues then become major challenges if we are trying to enable data aggregation by combining data from all sources [8], [9]. [6]

Interactive visual methods have been utilized within a wide spectrum of domains. In biomedicine, visualization is specifically required to support data analysts in tackling with problems inherent in this domain [19–21]. These can be summarized in three specific and general challenges: Challenge 1: Due to the trend towards a data-centric medicine, data analysts have to deal with increasingly growing volumes and a diversity of highly complex, multi-dimensional and often weakly-structured and noisy data sets and increasing amounts of unstructured information. biomedical data analysts have to deal with results from various sources in different structural dimensions [7]

2.2 Data Analysis Goals

2.3 Preprocessing

Pastrello et al. (2014) [48] emphasize that first and foremost it is important to integrate the large volumes of heterogeneous and distributed data sets and that interactive data visualization is essential to obtain meaningful hypotheses from the diversity of various data (see Figure 1). They see network analysis (see e.g. [49]) as a key technique to integrate, visualize and extrapolate relevant information from diverse data sets and emphasize the huge challenge in integrating different types of data and then focus on systematically exploring network properties to gain insight into network functions. They also accentuate the role of the interactome in connecting data derived from different experiments, and they emphasize the importance of network analysis for the recognition of interaction context-specific features. [5]

The ability to share data effectively and efficiently is the starting point for successful analysis, and thus several attempts have been made to standardize formats for such data exchange: PSI-MI [35], BioPAX [42], KGML, SBML [40], GML, CML, and CellML [30]. [6]

Kernel methods [6–8] incorporate important distinctions from traditional statistical pattern recognition approaches, which typically involve an analysis of object clustering within a measurement space. Rather, kernel-based approaches implicitly construct an embedding space via similarity measurements between objects, within which the classification (e.g. via an SVM) or regression takes place. The dimensionality of the space is dictated by number of objects and the choice of kernel rather than the underlying feature dimensionality. Kernel methods thus provide an ideal basis for combining heterogeneous medical information for the purposes of regression and classification, where data can range from hand-written medical notes to MR scans to genomic micro array data ; the way in which missing 'intermodal' data is combined in within the kernel-based framework depends on the authors' neutral point substitution method. A neutral point is defined as a unique (not necessarily specified) instantiation of an object that contributes exactly zero information to the classification [8]

2.3.1 Data Mining

There are several definitions of the term data mining [2] (Larose's book introduces several definitions). One of the most widely-used definitions states that "data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [3]. As this definition implies, the goal of data mining is to gain novel and deep insights and unprecedented understanding of large datasets (often accumulated for operational purposes) which can then be used to support decision making. Data mining can also enable the generation of scientific hypotheses from large experimental data sets and from biomedical literature [4, 5]. Data mining has matured into one way of addressing the growing availability of digital data and the gap between that data availability and the use of knowledge derived from them [6, 7]. [4]

In response to the huge demand of data mining for target discovery in the 'omics' era, this review explicates various data mining approaches and their applications to target discovery with emphasis on text and microarray data analysis. [9]

2.4 Data Analysis Frameworks

One of the technologies that can help in carrying out the DMKD process is XML (eXtensible Markup Language) [6]. All formatted text documents consist of text and markup. Markup is the set of commands, or tags, placed within the text, that control spacing, pagination, linkages to other documents, font style, size, color, and foreign alphabets. On the Internet, the most popular markup language is the Hypertext Markup Language (HTML). In HTML, each start-tag begins with ¡and ends with¿; each end-tag begins with ¡/and ends with¿. Thus, for example, the sequence ¡B¿ ... TEXT ... ¡/B¿ causes the computer monitor or printer to display ... TEXT ... in boldface. [2]

The ability to share data effectively and efficiently is the starting point for successful analysis, and thus several attempts have been made to standardize formats for such data exchange: PSI-MI [35], BioPAX [42], KGML, SBML [40], GML, CML, and CellML [30]. [6]

Resources for Studying Statistical Analysis of Biomedical Data and R [10]

2.5 Biomedical Knowledge

3 Research Question

3.1 Hypotheses

References

[1] J. D. Bronzino and D. R. Peterson, *Biomedical engineering fundamentals.* CRC press, 2014.

[2] K. J. Cios and G. W. Moore, “Uniqueness of medical data mining,” *Artificial Intelligence in Medicine*, vol. 26, no. 1, pp. 1 – 24, 2002. Medical Data Mining and Knowledge Discovery.

[3] Y. Peng, Z. Wu, and J. Jiang, “A novel feature selection approach for biomedical data classification,” *Journal of Biomedical Informatics*, vol. 43, no. 1, pp. 15 – 23, 2010.

[4] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, “Data mining in healthcare and biomedicine: A survey of the literature,” *Journal of Medical Systems*, vol. 36, pp. 2431–2448, Aug 2012.

[5] A. Holzinger and I. Jurisica, *Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions*, pp. 1–18. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[6] D. Otasek, C. Pastrello, A. Holzinger, and I. Jurisica, *Visual Data Mining: Effective Exploration of the Biological Universe*, pp. 19–33. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[7] C. Turkay, F. Jeanquartier, A. Holzinger, and H. Hauser, *On Computationally-Enhanced Visual Analysis of Heterogeneous Data and Its Application in Biomedical Informatics*, pp. 117–140. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[8] D. Windridge and M. Bober, *A Kernel-Based Framework for Medical Big-Data Analytics*, pp. 197–208. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[9] Y. Yang, S. J. Adelstein, and A. I. Kassis, “Target discovery from data mining approaches,” *Drug Discovery Today*, vol. 17, no. Supplement, pp. S16 – S23, 2012. Strategic Approach to Target Identification and Validation: A Supplement to Drug Discovery Today.

[10] M. Kobayashi, *Resources for Studying Statistical Analysis of Biomedical Data and R*, pp. 183–195. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.