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 21, 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 data analysis.

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 tools (which ones are usable) and at last biomedical knowledge (what engineers of BME and third parties 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, missing feature values and population handling. 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 tools. The steps to take from data to goal do not only include an approach, but also a tool to execute it. The choice of a certain tool has a big impact on the project, as each one of them has its own advantages and disadvantages. The two most well known tools 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.

Previous research projects on data mining have called for a model how to retrieve patterns from data collections. Frameworks to effectively do that have been proposed, usually with a number of steps.[2] These suggested frameworks do not specifically fit the cBio group though for being to broad [3] or being to specific. [4] A customized framework is needed.

2.1 Biomedical Data

A big aspect of choosing how to set up the data analysis is the data itself. The amount of data in the biomedical world is growing at an enormous rate, faster than biomedical engineers can analyse. Due to this rapid growth being uncontrollable, several additional challenges arose, aside

being more than the biomedical world can handle. These challenges are mainly focused around data volume, dimensionality, complexity, heterogeneity and quality.[5]

Collecting data because it is possible can make data sets bigger than needed. Both in number of instances and features, data sets can be harder to understand or analyse when more is available.[5] This volume problem usually is tackled by taking sub-populations of the complete set. These subsets can either be focused around a part of the population (gender, age, race) or taken at random to still represent all of it. Due to the efficiency of analysis techniques and the rise in computational speed of servers[6], volume on its own becomes less of an issue. Volume does however become an issue when combining with heterogeneity and quality. [7, 8]

Not all data sets have a high number of instances that cause a big data volume. Sometimes there are relatively few instances, while the number of features is proportionally high. Usually many of those features are not relevant enough for the research, however are still used for testing. Trying to remove features that are not important, will greatly help finding relations between the others and create more knowledge about the research topic. Lowering the number of features also makes the data volume go down, so analysis should be easier. Mainly an optimal features set should be selected to obtain the best results. [9]

Biomedical data can also be very complex. Useful results may be present, however it can be very hard to obtain it. Examples of complex data are images, several biomedical signals and temporal data. Details of the useful results that are present in images can for example be very hard to detect, the temporal data can vary quite much over time and the biomedical signals can be hard to combine with static biomarkers. This aspect can benefit from exchanging knowledge with other research areas that specialize in mining of those complex data sets. [7, 10]

The biggest challenge encompasses aligning different data sets. No standard for data sets is available and therefore data sets differ greatly from each other. Data is weakly structured or even unstructured [8] and variables are processed differently due to other protocols or the collectors' preference of representation.[11] Also the variety of data is hard to combine when sources are fundamentally different. When parts of the data are images, another part is a table from the laboratory and a third part is textual remarks of the doctor, standardizing merging those three is much harder than merging three lab sets. Those merges are also very prone to errors, as imprecisions can be vastly different between those data sets. No tool can work directly with these raw data sets and preprocessing must almost definitely occur beforehand.[2, 7]

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. [12]

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]

It has been generally recognized that patient record management systems is highly desired in clinical settings (Heathfield and Louw, 1999; Jackson, 2000; Abidi, 2001). The major reasons include physicians' significant information needs (Dawes and Sampson, 2003) and clinical information overload. Hersh (1996) classified textual health information into two main categories: patient-specific clinical information and knowledge-based information, which includes research reported in academic journals, books, technical reports, and other sources. Both types of information are growing at an overwhelming pace. [5]

In particular, exploring the associations among all the different pieces of information in these data sets is a fundamental problem to develop reliable medical tools based on data-driven approaches and machine learning. [13]

Although existing models demonstrate great promises (e.g. [7–11]), predictive tools based on machine learning techniques have not been widely applied in medicine [12]. In fact, there remain many challenges in making full use of the biomedical data, owing to their high-dimensionality, heterogeneity, temporal dependency, sparsity and [13]

2.2 Data Analysis Goals

2.2.1 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. [8]

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]. [11]

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 [14]

2.2.2 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]. [12]

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. [15]

Once these data have been collected, we can create a data warehouse and retrieve special case data or apply data mining to find hidden facts and rules [10]

Data mining aims to analyze a set of given data or information in order to identify novel and potentially useful patterns [5]

Especially in high-dimensional biomedical data sets, it is of great importance to find underlying patterns in the data. In this work we present an intuitive and easy-to-compute method that finds groups of similar objects by utilizing a network structure of the data. The key idea is that every object first ranks its neighbors and in a second step, the overlap of nearest neighbors between all objects is computed and stored in a sparse adjacency matrix. [16]

2.2.3 Machine Learning

A machine learning technique gaining increasing recognition and popularity in recent years is the support vector machines (SVMs). SVM is based on statistical learning theory that tries to find a hyperplane to best separate two or multiple classes (Vapnik, 1998). This statistical learning model has been applied in different applications and the results have been encouraging. For example, it has been shown that SVM achieved the best performance among several learning methods in document classification (Joachims, 1998; Yang and Liu, 1999). SVM is also suitable for various biomedical classification problems, such as disease state classification based on genetic variables or medical diagnosis based on patient indicators. [5]

The accuracy of a learning system needs to be evaluated before it can become useful. Limited availability of data often makes estimating accuracy a difficult task (Kohavi, 1995). Choosing a good evaluation methodology is very important for machine learning systems development. [5]

Machine learning is a general-purpose method of artificial intelligence that can learn relationships from the data without the need to define them a priori [13]

2.3 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]. [11]

Resources for Studying Statistical Analysis of Biomedical Data and R [17]

One aspect in comparing learning machines with each other deserves specific attention. If various machines are trained on training data, their performance can only be compared in a fair manner by applying all machines to the same test data. In this case, the above described procedures lead to valid estimates. [10]

2.4 Biomedical Knowledge

The traditional approach in biomedical science has been knowledge-driven and aimed at generating hypotheses from domain knowledge in a top-down fashion. Within biomedical data mining, one of the most interesting aspects is the exploitation of domain knowledge and the integration of different data sources in the data analysis process. As a matter of fact, data analysis is strongly empowered by the knowledge available in electronic format, which can be either already formalized, say through ontology and annotation repositories, or still informal but novel, as, for example, the one reported in Pubmed abstracts and papers[10]

Perhaps the most distinctive feature that accompanies medical data analysis is knowledge; data analysis in medicine strives to discover new and useful knowledge, while using available knowledge to guide the process and incorporate it into discovered models. In these terms, we perform data analysis to extract new findings that either refine or supplement existing knowledge on the problem domain, a population of patients, or a specific patient under consideration. [18]

In data analysis, knowledge is represented either implicitly or explicitly. By explicit knowledge we refer to knowledge already established, formalized and coded within some knowledge base. Such knowledge is used in some way in the process of data analysis. Researchers in machine learning often refer to such knowledge as "background knowledge," and use it in learning either in data preprocessing, feature selection or modeling. While this approach seems very promising in medical applications, the number of methods, tools and applications of such an approach are few. Most applications of data mining in medicine, for instance, have focused on building models directly from data and do not consider any explicitly represented knowledge in the process. These applications leave the interpretation of the models and the placement of any new information found in this way within the context of the available knowledge to analysts and domain experts. [18]

•The process of feature selection, or determining the set of variables that are thought best, a priori, to contribute to an analysis depends on an intimate familiarity with the problem domain. Numerous methods are available for reducing the feature space. Some are cognitive, such as expert panel verification, while others are computational. Results of either type of approach may be combined with existing knowledge to select most informative and interesting features that cover the problem from various angles. •The selection of preprocessing methods that may be required for certain data mining tasks depends on expertise in using not only the preprocessing methods but the data mining tools that may require the use of such methods. Discretization of continuous variables is an example of such a method. While the knowledge required for this task would appear to be primarily analytic, problem domain knowledge is needed as well, since some methods (such as discretization) can cause the loss of important scientific or clinical information. •The selection of an appropriate modeling methodology would also appear to be primarily in the analytic domain. However, the contribution of clinical or scientific knowledge to the modeling process cannot be underestimated. Examples of this are knowing when and how to model feature interactions. The selection and creation of these models requires collaboration between domain and analytic experts, due to the domain-specific realities of complex interactions and analysis-specific realities (and constraints) of computational complexity. •Finally, the selection of an appropriate methodology for presenting the results of an analysis requires expertise in such methodologies, but also in the clinical and scientific domain experts to whom the presentation is targeted. The

primary goal of these methods is to reveal discovered knowledge in the optimal way to facilitate its communication to all members of the research team.[18]

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] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *et al.*, "Knowledge discovery and data mining: Towards a unifying framework.," in *KDD*, vol. 96, pp. 82–88, 1996.

[4] K. J. Cios, A. Teresinska, S. Konieczna, J. Potocka, and S. Sharma, "A knowledge discovery approach to diagnosing myocardial perfusion," *IEEE Engineering in Medicine and Biology Magazine*, vol. 19, no. 4, pp. 17–25, 2000.

[5] H. Chen, S. S. Fuller, C. Friedman, and W. Hersh, *Medical informatics: knowledge management and data mining in biomedicine*, vol. 8. Springer Science & Business Media, 2006.

[6] D. Blythe, "Rise of the graphics processor," *Proceedings of the IEEE*, vol. 96, no. 5, pp. 761–778, 2008.

[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] 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.

[9] 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.

[10] R. Bellazzi, M. Diomidous, I. N. Sarkar, K. Takabayashi, A. Ziegler, A. T. McCray, *et al.*, "Data analysis and data mining: current issues in biomedical informatics," *Methods of information in medicine*, vol. 50, no. 6, p. 536, 2011.

[11] 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.

[12] 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.

[13] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, p. bbx044, 2017.

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

[15] 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.

[16] J. E. Vogt, "Unsupervised structure detection in biomedical data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, pp. 753–760, July 2015.

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

[18] B. Zupan, J. H. Holmes, and R. Bellazzi, "Knowledge-based data analysis and interpretation," *Artificial Intelligence in Medicine*, vol. 37, no. 3, pp. 163–165, 2006.