Data mining for better healthcare: A path towards automated data analysis?

Data mining for better healthcare: A path towards automated data analysis?

Tania Cerquitelli*, Elena Baralis*, Lia Morra† and Silvia Chiusano*
*Control and Computer Engineering Dept., Politecnico di Torino, ITALY
Email: name.surname@polito.it
†Email: liamorra@gmail.com

***Abstract*—In today's world, large volumes of medical data are being continuously generated, but their value is severely undermined by our inability to translate them into knowledge and, ultimately, actions. Data mining techniques allow the extraction of previously unknown interesting patterns from large datasets, but their complexity limits their practical diffusion. Data-driven analysis is a multi-step process, in which health care professionals define analysis goals and assess extracted knowledge, while computer scientists tackle the non trivial task of driving the miner system analysis activity. This paper addresses the mining activity from a different perspective. We believe that mining systems should be able to devise which knowledge could be most interesting to users and extract actionable knowledge from large medical datasets, with minimal user intervention. More specifically, mining systems should be capable of (i) devising viable end-goals for a specific dataset (i.e., that yield actionable knowledge to the user and are feasible given the dataset characteristics and size), based on the expertise acquired during the analysis of previous datasets, and (ii) extracting a manageable set of knowledge. Automated data analysis should fuel the next generation of medical mining systems, thus enabling users to automatically mine massive data repositories.**

I. Introduction

The healthcare industry today generates large amounts of complex data from heterogenous sources: electronic patient records, medical reports, hospital resources, medical devices, billing systems, and so forth. In the future, wearable sensors, mobile devices and even social networks will be increasingly used to extract information of potential clinical use. This large amount of data is a key resource that can revolutionize the way healthcare is delivered, promoting more effective research and supporting medical doctors in providing better care for their patients. On the other hand,it can support medical staff and healthcare administrators alike in the need of containing healthcare-related costs.

Data mining emerged during the late 1980s and focused on studying algorithms to find implicit, previously unknown, and potentially useful information from large volumes of data. Data mining activities include studying correlations among data (e.g.,association rules at different levels of abstraction), grouping data with similar properties (e.g., clustering), and extracting information for prediction (e.g., classification, regression). Thus, data mining is an enabling technology to extract meaningful and actionable knowledge from large sets of heterogeneous and complex data. Hidden correlations can emerge: for example, two drugs that separately proved safe during clinical trials, can have dangerous side effects when taken in combination (as was the case for antidepressant Plaxil, and cholesterol-lowering drug Pravachol). While the final decisions rests with health care professionals, medical doctors can benefit from clinical support system that mine medical records and scientific literature for optimizing treatment for a specific patient, based on his/her individual characteristics.

At the present moment, extracting actionable knowledge from data is a multi-step process that requires considerable interaction between the domain expert (who needs to define the end-goal) and a data scientist who tackles the non-trivial task of selecting the optimal techniques to achieve this goal, looking for a good trade-off between quality and execution time. Automated medical data mining systems could improve usability by healthcare professionals, who often do not have extensive experience. Moreover, the end users may not always be aware of the interesting knowledge hidden in their data, and therefore have difficulties in selecting specific end-goals. Nonetheless, a framework of broadly defined analysis, bringing distinctive improvements in various fields related to healthcare management, can be defined, such as (i) providing clinical decision support in determining the best course of treatment for patients with a given disease (precision medicine), (ii) assessing the adherence of medical prescriptions and treatments to relevant clinical guidelines, (iii) predicting and assessing the outcome of medical treatments, (iv) monitoring adverse events and drug effectiveness beyond clinical trials, and (v) planning resource allocation and reduce costs incurred by organizations. Such analysis may be appealing to a wide range of different users, from medical doctors and clinical researchers, to hospital administrators, health insurance companies, and public health agencies. By analyzing the collection of medical records, the data scientist can (i) discover groups of patients with similar clinical history, (ii) identify medical examinations commonly prescribed by physicians to patients with a given disease, (iii) identify which examinations/treatments have the highest patients compliance or better outcome or (iv) discover previously unknown interaction between drugs or medical conditions. For points (i) a clustering-based algorithm can be used [1], while for points (ii)-(iv) a frequent pattern discovering approach can be exploited [2]. These two classes of algorithms are interesting in data mining for their exploratory nature, as they do not require apriori knowledge (such as the target class to be predicted in a classification process), and hence support different and interesting analyses. The domain expert can analyze the obtained results, evaluate the quality of extracted knowledge items and interpret them in view of the specific analysis end-goal.

In this paper we argue towards a new generation of

automated data analytics systems for better healthcare that automatically extract multiple knowledge items from a dataset with minimal user intervention. These systems provide a manageable set of knowledge items which are characterized and ranked in terms of their potential interest to the user, who in turn does not have to deal with the technical details on how such knowledge is actually obtained. Extracted knowledge items will be automatically ranked, based on previous interactions with the domain expert. The user will be able to navigate through the extracted knowledge items, and provide feedback on their quality and interestingness to guide future data analysis sessions.

II. Vision and Challenges

For an automated mining system, the main research goal is to design a engine that, given a set of data, yields actionable knowledge items while hiding the underlying complexity of the data analytics tasks from the end user. This opens a wide range of research issues, such as:

1. What kind of knowledge/analysis could be of interest to the user? Which is the analysis end-goal?
2. Within the large amount of extracted knowledge, which items are most valuable and need to be presented to the user?
3. Which parts of the data need to be mined?
4. Which algorithm drives the mining process?
5. How to set the best configuration of the algorithm?
6. How to navigate and explore the extracted knowledge?

The first two questions represent completely new problems. To the best of our knowledge, none of the available (medical) mining frameworks is capable of automatically extracting interesting knowledge without explicit and detailed request. Also, different efficient techniques have been proposed to process, analyze, and query the huge amount of mined knowledge, but a proactive and adaptive approach to automatically identify the manageable subset of knowledge which is actionable for the domain expert is yet to be pursued. For questions 3-6 extensive research brought many innovative and efficient algorithms customized for a targeted analytics task. Available algorithms can be integrated within the proposed approach.

III. Automated Medical Data Analysis

The components of the envisioned architecture, as well as the interactions between such components, are shown in Figure 1. We will refer to this architecture as ADA-Health (Automated Data Analysis for better Healthcare).

Data characterization and transformation. This is the first step in data mining processes. However, with the increasing range and complexity of available medical data, no single available descriptor is universally appropriate. We focus on the definition of innovative criteria to *model data distributions* by exploiting unconventional statistical indices and underlying data structures (e.g., frequent patterns). Medical datasets are usually generated by a large variety of events, and the features used to model concepts and behavior in heathcare may have large domains, thus resulting in datasets with inherent sparseness and variable distribution. The variability in data distribution increases with data volume, thus further increasing the complexity of data mining. We believe that a dynamic strategy to *cluster data into different and/or disjoint subsets*, each

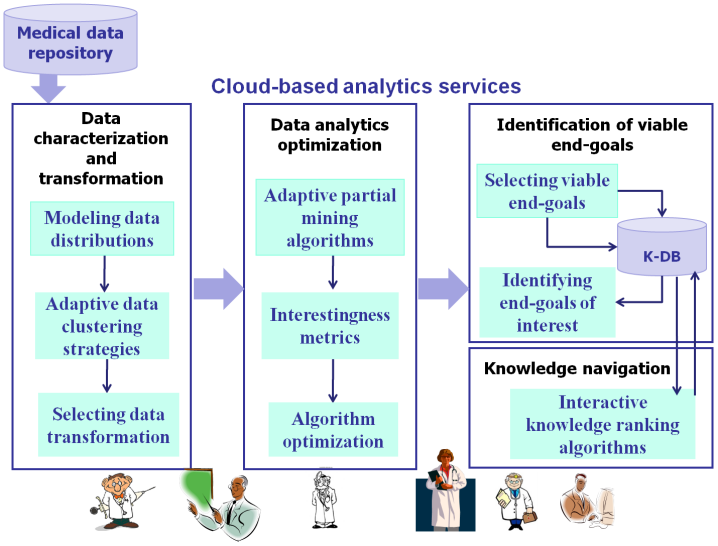

Fig. 1. The ADA-Health system architecture

characterized by a diverse data distribution, is needed to locally apply selected data mining algorithms and better highlight the knowledge hidden in each portion of the dataset. Furthermore, many mining algorithms rely on suitable transformations of input data in order to reduce sparseness, and make the overall analysis problem more efficiently tractable. To this purpose, the ADA-Health architecture includes several techniques to preprocess data and map them into different representation spaces. The main research issue here is to define a totally automatic strategy to *select the optimal data transformation*, which yields higher quality knowledge.

Data analytics optimization. When dealing with Big Data collections, easily generated in medical applications, the computational cost of the data mining process (and in some cases the feasibility of the process itself) can potentially become a critical bottleneck in data analysis. To this purpose, a set of *online cloud-based services* for automatic configuration of data analytics will exploit the computational advantages of massively parallel cloud computing. With increasing input data volumes, the amount of extracted knowledge also potentially increases. Thus actionable knowledge may still be hidden in a growing volume of extracted knowledge. To avoid the expensive and resource-consuming procedure of mining the entire dataset when not necessary, *adaptive partial mining strategies* need to be designed. However, the use of partial mining techniques has also other advantages. More specifically, the amount of extracted knowledge increases with the input data cardinality, and interesting knowledge is less likely to emerge. Almost every large dataset with a degree of randomness is likely to contain some spurious correlations. Partial mining strategies will leverage the innovative techniques developed for the characterization of input data, by focusing the analysis on the most significant subsets of input data. More specifically, interesting properties could apply to specific subsets, but not to the general collection. Selecting and independently mining specific subsets can thus bring hidden knowledge to the surface. In analyzing an N-dimensional dataset, partial mining can reduce the dataset along any dimension (*vertical mining*) or by considering different subsets of the input data (*horizontal*

mining). Both approaches are included in the ADA-HEALTH system. The *algorithm optimization* component will include innovative techniques to efficiently optimize a suitable algorithm for a given data mining task. Large parameter spaces need to be explored at different abstraction levels (i.e., end-goal analysis, algorithm and algorithm parameters). It is hard to envision a system capable of evaluating and comparing hundreds of different data mining technique configurations, without being able to effectively and automatically compare and rank their output. To this end, a set of interestingness metrics are needed to assess the quality of knowledge discovered by different algorithm runs. An interesting research direction is the exploitation of different (and possibly a combination of) unconventional statistical and machine learning criteria to evaluate extracted knowledge and drive the optimization process.

Identification of viable end-goals. This is the core and one of the of most innovative contributions of the ADA-HEALTH architecture. We believe that an interesting research direction is based on three key components to drive this complex prediction task, while keeping the overall architecture as general as possible. The key components are (i) a knowledge database storing past user feedback on previously processed datasets (including dataset characterization and extracted knowledge items), (ii) an algorithm to identify viable end-goals, and (iii) an algorithm to select end-goals of interest. More specifically, information about past user interactions and domain-specific expertise (considering differences in physician opinions based on their diverse background and specialization), data statistics as well as the different knowledge items discovered by different data analytics algorithms are stored in a *Knowledge Base (K-DB)* and used to drive the self-learning analysis tasks. The K-DB will be continuously enriched with new health care professionals feedbacks. *Selecting viable end-goals* for a given medical dataset is challenging because it is strongly affected by the characteristics of the dataset and also by differences in physician opinions, due to their diverse background and specialization. An interesting research direction is the discovery of a set of formal rules able to predict the feasible analysis end-goals on a given dataset, which could be based on the information stored in the K-DB. Many different approaches (e.g., data mining approaches, logic programming, statistics models) can be exploited for this complex and challenging task. After the selection of viable end-goals, the ADA-HEALTH system *identifies which end-goals are more interesting for a specific user*. This can be addressed again as a classification problem, thus, the model is trained by previous user interactions with the ADA-HEALTH system. The larger the number of previous user interactions, the more accurate the classification model will be.

Knowledge navigation. ADA-HEALTH also includes an *interactive knowledge ranking algorithm* (e.g., statistics methods, data mining algorithms) which will help to select, among a set of knowledge items, which ones are most interesting for a user. Based on user feedbacks, the algorithm dynamically adjusts the way and order how knowledge items are organized and presented to the user. A *user interface* allows interactive presentation and navigation of the extracted knowledge items.

IV. PRELIMINARY DEVELOPMENT AND RESULTS

Here we presented a preliminary implementation of the ADA-HEALTH system to show both the feasibility and potential of the envisioned system. Preliminary results were obtained on a real, anonymized dataset of diabetic patients. It contains the examination log data of 6,380 patients (age range 4-95 years) with overt diabetes, covering the time period of one year, for a total of 95,788 records. Each record contains at least a unique patient identifier, and the type and date of every exam. 159 different types of examinations are present, including regular checkups as well as more specific diagnostic tests for complications with varying degrees of severity (e.g. cardiovascular complications, blindness). This dataset, albeit small, is characterized by an inherently sparse distribution, which is typical of big data collections.

A. Preliminary development

A preliminary implementation of four basic components of ADA-HEALTH has been developed, tailored to two exploratory data mining algorithms: a clustering algorithm [3] (discussed here) and a pattern-based discovery approach [2].

The current implementation of *selecting data transformation* includes a single pre-processing block capable of tailoring a given dataset to a Vector Space Model (VSM) representation, which is particularly suited to handle sparse datasets. We developed a preliminary implementation of an *adaptative partial mining strategy based on a horizontal partial mining technique*. At each step, a larger portion of data is analyzed. In the case of clustering, horizontal partial mining is implemented by running K-means on different subsets, as well as on the complete collection; the quality of each result was evaluated by means of the overall similarity index [4]. The overall similarity, selected as an *interestingness metrics*, measures the cluster cohesiveness by computing the internal pairwise similarity of patients within each cluster, and then taking the weighted sum over the whole cluster set.

We also developed a *basic algorithm optimization* component to assess the quality of the knowledge extracted through a clustering algorithm (i.e., a center-based algorithm such as K-Means), which is based on a combined approach exploiting both a traditional quality index (e.g., Sum of Squared Error - SSE [4]) and a classification technique. Given a dataset and a clustering algorithm, our technique performs several runs of the mining activity with varying parameters (e.g. different numbers of clusters), thus obtaining several different cluster sets. The SSE index measures the cluster cohesion for center-based clustering techniques as the total sum of squared errors over all the objects in the collection, where for each object the error is computed as the squared distance from the closest centroid. The smaller the SSE, the better the quality of discovered clusters. However, as the number of classes increases, the SSE decreases, because smaller and more cohesive clusters are identified [5]. In contrast, in many medical applications the actual number of interesting clusters is usually small. Thus, a trade-off is needed between the number of clusters and their significance: too few clusters will lead to poor data grouping (i.e. large SSE), whereas too many clusters will result in small, poor quality and possibly not significant clusters. A classifier was then built to assess the robustness [6]

of clustering results by means of different quality metrics (such as accuracy, precision, recall), using the same input features of the clustering algorithm, and the class label assigned by the clustering algorithm itself as target. The higher the classification metrics, the better the robustness and overall quality of discovered clusters. In our first implementation, we used decision trees as classification model. Clustering parameters are choosen in order to optimize the values of all considered metrics.

Finally, we designed and implemented a *preliminary version of the K-DB* on a cluster of MongoDBs by classifying, with the help of a domain expert, knowledge items discovered by applying both a clustering-based algorithm and a pattern-based discovery approach. The complete data model consists of six collections, which store (1) the original dataset, (2) the transformed dataset after preprocessing and data transformation, (3) statistical descriptors to model the data distribution, (4-5) interesting and selected knowldege items discovered through different data mining algorithms, and (6) user interaction feedbacks. With the support of a physician, each knowldege item will be enriched with a degree of interestingness {high, medium, low}. This apriori knowledge, together with user interaction feedbacks can easily support (i) the selection of actionable and interesting knowledge and (ii) the prediction of a degree of interestingness based on previous interactions by means of a classification algorithm.

B. Preliminary results

We tested the preliminary ADA-HEALTH implementation on a real, anonymized dataset containing examination histories for diabetic patients. The goal was to find groups of patients with similar examination history. The data transformation block through the VSM model generates a unique vector for each patient, representing his/her examination history (i.e. number of times he/she underwent each examination). The ADA-HEALTH partial mining technique identifies meaningful data subsets. In the first iterations, only the most frequent examinations are included, then the subset is expanded by incrementally adding more specific tests perfomed on lower numbers of patients. Three incremental runs have been analysed by considering up to 20%, 40% and 100% of the total number of examination types (corresponding to 70%, 85% and 100% of the original row data), hence reducing the cardinality of the feature space while retaining the total number of patients. In each run, the examination types were chosen in decreasing order of frequency within the original raw data. Based on the overall similarity measures, the results are promising, as performances on only 85% of row data are comparable to those obtained on the entire dataset, regardless of the number of clusters. This could be due many reasons; for instance, some examination types are probably correlated (e.g. they could be prescribed in conjunction or are needed to monitor/diagnose the same condition), while other could be rarely prescribed and thus have little impact on clustering (but could affect other types of analyses such as outlier detection). For a fixed number of clusters, the overall similarity decreases as the number of exams is reduced. ADA-HEALTH selects the optimal subset size based on the percentage difference between the overall similarity value calculated on the subset, and that calculated on the complete dataset: in this example, 85% of raw data yields a percentage difference less than 5%.

K	SSE	Accuracy	AVG Precision	AVG Recall
6	3098.32	87.79	90.82	77.3
7	2805	87.93	86.93	78.52
8	2550	90.41	92.51	79.72
9	2482.36	88.75	71.03	57.62
10	2205	87.49	70.53	51.06
12	2101.6	85.45	64.29	43.80
15	1917.2	75.18	75.98	55.93
20	1534	82.11	52.59	33.43

TABLE I. OPTIMIZATION METRICS

The ADA-HEALTH optimizer module in the case of K-Means needs to optimize only the K parameter (i.e. the number of clusters). Based on our previous results, only a subset of the original dataset was used (85% of the original raw data). For each K value, the resulting cluster set is evaluated by computing the SSE, and a decision tree (classification model) is built to evaluate its quality; 10-fold cross validation was used to evaluate the classification model. Table I shows (i) the SSE value and (ii) accuracy, average precision and average recall, all computed on the different clusters sets obtained by varying K. Based on the SSE index, good values for K are in the range from 8 to 20. The best value for the classification metrics were obtained for K equal to 7 and 8, where the accuracy values are 88% and 90% respectively. However, the average recall and precision values are higher for $K = 8$ (80% and 92% respectively). ADA-HEALTH automatically selects $K = 8$ that corresponds to the best overall classification results.

V. CONCLUSIONS

This paper presents a challenging vision for a self-configuring medical analytics system capable of predicting viable end-goals for a specific dataset, based on the expertise acquired during the analysis of previous datasets, and extracting a manageable set of knowledge. As each and every of the ADA-HEALTH building blocks presents a number of open research questions, a possible basic approach to tackle some of the emerging issues has been proposed on a real life example to prove the feasibility of our vision. We thus expect that such systems could dramatically boost the applicability and user-friendliness of data mining in the medical scenario.

REFERENCES

[1] D. Antonelli, E. Baralis, G. Bruno, T. Cerquitelli, S. Chiusano, and N. Mahoto, "Analysis of diabetic patients through their examination history," *Expert Syst. Appl.*, vol. 40, no. 11, pp. 4672–4678, Sep. 2013.

[2] D. Antonelli, E. Baralis, G. Bruno, L. Cagliero, T. Cerquitelli, S. Chiusano, P. Garza, and N. A. Mahoto, "MeTA: Characterization of Medical Treatments at Different Abstraction Levels," *ACM TIST*, vol. 6, no. 4, p. 57, 2015.

[3] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.

[4] Pang-Ning T. and Steinbach M. and Kumar V., *Introduction to Data Mining*. Addison-Wesley, 2006.

[5] T. Cerquitelli and E. D. Corso, "Characterizing thermal energy consumption through exploratory data mining algorithms," in *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, Bordeaux, France, March 15-18, 2016*, 2016.

[6] G. Bruno, T. Cerquitelli, S. Chiusano, and X. Xiao, "A clustering-based approach to analyse examinations for diabetic patients," in *ICHI '14*, 2014.