Improving clinical record visualization recommendations with Bayesian stream learning
Position Paper

Pedro Pereira Rodrigues[1,2], Cláudia Dias[2,3], and Ricardo Cruz-Correia[2,3]

[1] LIAAD - INESC Porto, L.A., Portugal
[2] Faculty of Medicine of the University of Porto, Portugal
[3] CINTESIS - University of Porto, Portugal
{pprodrigues,camila,rcorreia}@med.up.pt

Abstract. Clinical record integration and visualization is one of the most important abilities of modern health information systems (HIS). Its use on clinical encounters plays a relevant role in the efficacy and efficiency of healthcare. However, integrated HIS of central hospitals may gather millions of clinical reports (e.g. radiology, lab results, etc.). Hence, the clinical record must manage a stream of reports being produced in the entire hospital. Moreover, not all documents from a patient are relevant for a given encounter, and therefore not visualized during that encounter. Thus, the HIS must also manage a stream of events of visualization of reports, which runs in parallel to the stream of documents production. The aim of our project is to provide the physician with a recommendation of clinical reports to consider when they log in the computer. Our approach is to model relevance as the probability that a given document will be accessed in the current time frame. For that, we design a data stream management system to process the two streams, and Bayesian networks to learn those probabilities based on document, patient, department and user information. One of the biggest challenges to the learning problem, so far, is that no negative examples are produced by the stream (i.e. there are no record of documents not being visualized) leading to a one-class classification problem. The aim of this paper is to clearly present the setting and rationale for the approach. Current work is focused on both the stream processing mechanism and the Bayesian probability estimation.

1 Introduction

The identification of clinically relevant information should enable an improvement in user interface design and in data management. However, it is difficult to identify what information is important in daily clinical care, and what is used only occasionally. The main problem addressed by this project is how to estimate the relevance of healthcare information in order to anticipate its usefulness at a specific point of care. In particular, we want to estimate the probability of a piece of information being accessed during a certain time interval, taking into account the type of data, the context where it was generated and is needed and the type of users who access it, and to use this probability to prioritize the information.

1.1 Relevance of clinical documents

In the healthcare domain, and especially in critical and acute care, the age of data is one of the factors often used to assess data relevance, making new information more relevant to the current search. Some authors have categorized *old data* as data at least three days old. However, in a previous study, the authors have examined for how long are clinical documents used by health professionals in a hospital environment, and how this is associated with document content and the context of information request [5]. Those results show that some clinical reports are still used after one year regardless of the context in which they were created, although significant differences exist in reports created in distinct encounter types. The authors conclude that the usage of past patient data (data from previous hospital encounters) varied significantly according to the setting of healthcare and document content, which contradicts the definition of *old data* used in previous studies. Hence the need to define better rules for recommending documents in encounters.

1.2 Setting

The current setting is a central hospital that has several departmental information systems that produce clinical documents (e.g. radiology reports and lab results) that might be relevant for the practice of healthcare. The access to this documents is better achieved by a centralized information system that integrates all the different departmental systems, aggregating the documents that are most relevant for the current encounter [4]. These sources can be modeled as data streams. A data stream is an ordered sequence of instances that can be read only once or a small number of times [10], using limited computing and storage capabilities. Hence, there are two data streams being produced in parallel:

- a stream of documents being produced; each element is a new document (or a new version of an existing one); and
- a stream of visualization events; each element is an event of visualization of a previously produced document.

The first stream is gathered by integrating documents from heterogeneous information systems, so we might end up with syncronization issues. In the current setting, the central hospital information system receives an increasing rate of 200+ documents per hour (as seen in Figure 1), relative to the daily 5300+ patients in the hospital (including inpatients, outpatients and emergency rooms) which need to be processed. This created a pool of 8M+ documents (produced since 2004 from 400K+ patients). However, due to constant document revisions, "only" 2.9M+ are actually available for visualization (active documents), whilst 5M+ are previous versions.

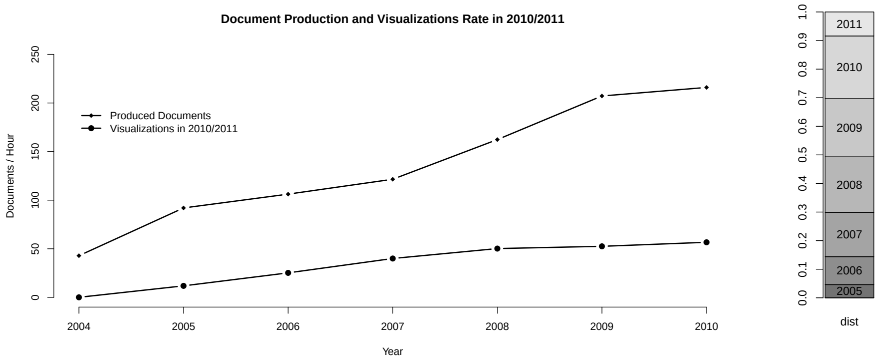

Fig. 1. *Left:* Document production rate from 2004 to 2010 (data for 2011 not shown to improve readability) and visualization rate of those documents (in 2010 and 2011). *Right:* Distribution of visualizations in 2010 and 2011, by date of document production.

The second stream is controlled in the centralized integrating information system, tracking information on visualizations performed by 4850+ users. Older documents tend to be less visualized in encounters but, as seen in Figure 1, nearly half of the visualizations in 2010 and 2011 targeted documents older than January 2009, so old documents cannot be discarded. During an encounter with a patient, the pool of documents that one of the currently active 2375 users might access is 537.9 (average number of active documents per active patient). Clearly, the user cannot be presented with a list of 530+ documents, so a ranking is needed to prioritize the most relevant.

1.3 Learning problem

Overall, this is one clear setting of data streams in medical scenarios, on which machine learning techniques can be applied to improve healthcare. In order to select relevant documents for visualization in a precise encounter, we need to take care of both the age and the information related to those documents. Either way, this setting is an uncertain one. Thus, the best way to model this relevance is to estimate the probability that the document is going to be visualized in the current time period. In machine learning, uncertainty is usally well modeled using Bayesian approaches [16]. Therefore, we seek to develop Bayesian networks to estimate the probability of a certain clinical document being visualized in the near future, and use this probability to rank the list of possible documents related to the current encounter. The fact that no negative examples are present (only visualization events) creates a harder setting for learning, so we are aiming at Bayesian stream learning for one-class classification [13]. The one-class classification problem is different from the conventional binary/multi-class classification problem in the sense that the negative class is either not present or not properly sampled. This is, at least in medical settings and at the best of our knowlege, uncharted research territory.

1.4 Aim and outline

The aim of this work is to extend the health information system responsible for visualization of clinical documents at point of care, providing the physician with a recommendation of clinical reports to consider when in the presence of a patient. Basically, we propose to study how to estimate the relevance of healthcare information in a particular setting, aiming to use this to create adaptive user interfaces. Specifically, our objectives are:

- to collect and prepare log data from hospital information systems usage, to feed our two data streams;
- to study the factors associated with the relevance of clinical documents;
- to define an algorithm to estimate the relevance of a particular document for a precise encounter;
- to implement an adaptive user interface based on a ranked list of recommended documents per encounter.

The paper is organized as follows. Next section presents background knowledge on electronic health records, learning from data streams and Bayesian networks in healthcare. Then, section 3 presents our approach for i) the data stream processing, ii) the estimation strategy, iii) the incremental learning and iv) the recommendation generator. Section 4 ends the exposition with future work.

2 Background

This work is related with three different areas of research: medical informatics, especially devoted to electronic health records; Bayesian learning from data streams; and one-class classification (check [13] for a survey on this problem).

2.1 Electronic Health Records

The practice of medicine has been described as being dominated by how well information is collected, processed, retrieved, and communicated [1]. Patient records, the patient and published evidence are the three information sources needed to practice evidence-based medicine [22].

Currently in most hospitals there are great quantities of stored digital data regarding patients, in administrative, clinical, lab or imaging systems. An important challenge is to guarantee the optimal conditions for health professionals to access clinical data while hospital information systems are still being developed. Although great advances have been made over the years, on-demand access to clinical information is still inadequate in many settings, contributing to duplication of effort, excess costs, adverse events, and reduced efficiency. Although it is widely accepted that full access to integrated electronic health records (EHRs) and instant access to up-to-date medical knowledge significantly reduces faulty decision making resulting from lack of information [6], there is still very little evidence that life-long EHRs improve patient care [2].

Shapiro et al. found that, although emergency department doctors believe their patients would benefit from longitudinal records, they only try to obtain such data in 10% of the cases [21]. Furthermore Hripcsak et al. described access rates to WebCIS in the emergency department [11], which indicated that data generated before the current emergency visit are accessed often, but by no means in a majority of times (5% to 20% of the encounters), even when the user was notified of the availability of such data.

Cruz-Correia et al. have done several pilot studies to analyse for how long clinical documents are useful for health professionals in a hospital environment, bearing in mind document content and the context of the information request. The results show that some clinical reports are still used one year after creation, regardless of the context in which they were created, although significant differences existed in reports created during distinct encounter types. The median-life of reports by the type of encounter during which they were created is 1.7 days for emergency, 3.9 days for inpatient and 27.7 days for outpatient encounters. They conclude that the usage of patients past information (data from previous hospital encounters), varied significantly according to the setting of healthcare and content [5].

Also, the amount of digital data produced in the medical imaging departments has increased rapidly in recent years, due mainly to two factors: 1) greater use of additional diagnostic procedures, resulting in a greater number of tests produced and, 2) an increase in the quality of the examinations, which is translated into a greater number of images acquired. Despite great development in digital storage technologies, the cost and effort required to maintain this information online throughout its life cycle can be considerable.

The management of information in these systems is usually implemented using Hierarchical Storage Management (HSM) solutions. This type of solution enables the implementation of various layers which use different technologies with different speeds of access, corresponding to different associated costs. However, the solutions which are currently implemented in Picture Archiving and Communication System (PACS) use simple rules for information management, based on variables such as the time elapsed since the last access or the date of creation of information, not taking into account the likely relevance of information in the clinical environment.

Classifying the relevance of information based only on the time elapsed since the date of acquisition is clearly inefficient. It is expected that the need to consult an examination at a given time will be dependent on several factors beyond the date of the examination, such as type of examination and the patient's pathology. Thus, a system that uses more factors to identify the relevance of information at a given time would be more efficient in managing the information that is stored in fast memory and slow memory.

2.2 Machine learning from data streams

What distinguishes current data from earlier one are automatic data feeds. We do not just have people who are entering information into a computer. Instead, we have computers entering data into each other [17]. Thus, there are applications in which the data is modeled best not as persistent tables but rather as transient data streams.

A data stream is an ordered sequence of instances that can be read only once or a small number of times using limited computing and storage capabilities. The data elements in the stream arrive online, being potentially unbounded in size. Once an element from a data stream has been processed it is discarded or archived. It cannot be retrieved easily unless it is explicitly stored in memory, which is small relative to the size of the data streams. These sources of data are characterized by being open-ended, flowing at high-speed, and generated by non stationary distributions [8,9].

In online streaming scenarios, predictions are usually followed by the real label value in a short future (e.g., prediction of next value of a time series). Nevertheless, there are also scenarios where the real label value is only available after a long term, such as predicting one week ahead electrical power consumption [18]. Learning techniques which operate through fixed training sets and generate static models are obsolete in these contexts. Faster answers are usually required, keeping an anytime data model and enabling better decisions, possibly forgetting older information.

The sequences of data points are not independent, and are not generated by stationary distributions. We need dynamic models that evolve over time and are able to adapt to changes in the distribution generating examples [8]. If the process is not strictly stationary (as most of real-world applications), the target concept may gradually change over time. Hence data stream mining is an incremental task that requires incremental learning algorithms that take drift into account [7].

Hulten et al. [12] presented desirable properties for data stream learning systems. Overall, they should process examples at the rate they arrive, use a single scan of data and fixed memory, maintain a decision model at any time and be able to adapt the model to the most recent data. Successful data stream learning systems were already proposed for both prediction [18] and clustering [3,19]. All of them share the aim to produce reliable predictions or clustering structures.

2.3 Machine learning in healthcare

The application of data mining and machine learning techniques to medical knowledge discovery tasks is now a growing research area. These techniques vary widely and are based on data-driven conceptualizations, model-based definitions or on a combination of data-based knowledge with human-expert knowledge [16].

The definition of clinical decision support systems is now a major topic since it may help the diagnosis, the prognosis of rate of mortality, the prognosis of quality of life, or even treatment selection. However, the complicated nature of real-world biomedical data has made it necessary to look beyond traditional biostatistics [14] without loosing the necessary formality. For example, naive Bayesian approaches are closely related to logistic regression [20]. Hence, those systems could be implemented applying methods of machine learning [16], since new computational techniques are better at detecting patterns hidden in biomedical data, and can better represent and manipulate uncertainties [20].

Traditional statistical methods require that the model structure is given and only probabilistic information is learned from biomedical evidence, in the form of data, whereas machine-learning approaches enable that both the structure of the models and the probabilistic information are evidence-based[14,16]. Bayesian approaches have an extreme importance in these problems as they provide a quantitative perspective [15]. Moreover, Bayesian statistical methods allow taking into account prior knowledge when analyzing data, turning the data analysis into a process of updating that prior knowledge with biomedical and health-care evidence [14]. However, only after the 90's we may find evidence of a large interest on these methods, namely on Bayesian networks, which offer a general and versatile approach to capturing and reasoning with uncertainty in medicine and health care [15].

Given their improved management of uncertainties, Bayesian networks have been successfully applied in healthcare domains [15]. Bayesian networks can be seen as an alternative to logistic regression where statistical dependence and independence are not hidden in approximating weights, rather explicitly represented by links in a network of variables [14]. They describe the distribution of probabilities of one set of variables, making possible a two-fold analysis: a qualitative model and a quantitative model, presenting two types of information for each variable.

On a general basis, a Bayesian network represents a joint distribution of one set of variables, specifying the assumption of independence between them, with the inter-dependence between variables being represented by a directed acyclic graph. Each variable is represented by a node in the graph, and is dependent of the set of variables represented by its ascendant nodes; a node X is a ascendant of another node Y if exists a direct arc from X to Y [16]. To give more representational power to the relations represented by the arcs of the graph, it is necessary to associate values to it. The matrix of conditional probability is given for each variable, describing the distribution of probabilities of each variable given its ascendant variables.

After the qualitative and quantitative models are constructed, the next step, and one of the most important, is how to calculate the new probabilities when new evidence is introduced in the network. This process is called inference and works as follows. Each variable has a finite number of categories greater than or equal to two. A node is observed when there is knowledge about the state of that variable. The observed variables have a huge importance because with conditional probabilities they define the prior probabilities of the non observed variables. With the joint probabilities we can calculate the marginal probabilities of each unobserved variable, adding for all categories the probabilities that the variable is in the desired state [15].

3 Optimizing the visualization of the clinical record

This section exposes our approach to the problem, presenting the data stream management system, the relevance estimation strategy, the Bayesian learning model, and the documents recommender system.

3.1 Data stream management system

Currently, a lot of patient information is accessible to healthcare professionals at the point of care. In some cases, the amount of information is becoming too large to be readily handled by humans or to be efficiently managed by traditional storage algorithms. Most Hospital information systems record the actions performed by users – the *log file.* These logs are kept for audit purposes but can give insights into the information needs of healthcare professionals in a particular situation. The study of these logs should allow us not only to describe how the systems were used, but may also be useful to predict future use of the system and of the data items it contains. To this latter task, there are two data streams being produced in parallel: the stream of documents being created, and the stream of visualization events.

The first stream is gathered by integrating documents from heterogeneous information systems, which should be modeled according to the *insert-delete* or *turnstile* model [17], allowing that observations might be updated or deleted by future events (some clicnical documents are subject of validation and revision, deactivating the previous versions of that document). The second one is controlled in the centralized integrating information system, but it only tracks information on visualizations, so it should be modeled according to the *insert-only* or *time series* model [17], since the event of visualization is not subject of deletion. However, given the predictive task of our system, we could consider the *accumulative* or *cash-register* model [17], where each observation is an increment to a given sum, i.e. the counters of the number of visualizations of each document in the current time period.

3.2 Estimating relevance of clinical documents

By applying regression methods or other modeling techniques it is possible to identify which factors are associated with the usage or relevance of patient data items. These factors and associations can then be used to estimate data relevance in a specific future time interval. We expect to find considerable differences in the relevance or median-life of information depending on several factors, such as the origin of the data (e.g. lab department, emergency department), the age of patient, and the current patient diagnosis. We also expect there to be an exponential decay in the use and thus relevance of each patient's data over time [5].

As previously discussed, we rely on the Bayesian networks ability to model uncertainty to estimate the probability of a given clinical document to be visualized. After defining the set of patient variables P and the set of other factors F that are associated with that relevance, we shall create a Bayesian network with $||P|| + ||F|| + 2$ nodes, where the remaining nodes are the age of the document and the class (visualized or not). Given the discrete characteristics of variables in the Bayesian network, age of document needs to be categorized into contiguous intervals. According to previous work [5], this categorization should be defined with exponential intervals, to model the relevance decay with time. Moreover, they can be defined as percentiles of visualization events in previous data.

3.3 Incremental learning the Bayesian network

When a document is visualized (i.e. a new observation is created in the visualization stream) a learning example can be created, computing the age of the document with real creation and visualization dates. This learning example can thus be fed to the Bayesian network. However, there are no records of non-visualization of documents, so negative learning examples are never created. Sophisticated approaches shall be tested to solve this issue:

1. for each positive example representing a visualization event, create as many negative examples for the time periods elapsed since the last visualization (or creation); or
2. define a purely one-class classification learner.

From our setting, we can easily extract information on previous visualization of a given clinical document (the log file includes this data) so we will follow the first approach.

3.4 Generating the recommendations

The goal is to achieve the following setup. A doctor-patient encounter (e.g. emergency, outpatient consultation, inpatient consultation) requires visualization of clinical documents. These could be documents from the same patient (patient's medical history) or documents from patients with similar characteristics or diagnosis. Given the huge amount of available clinical documents, the information system should list only the most relevant for that encounter.

Several approaches to rank the recommendations could be followed, and certainly they will be studied. At first, we shall consider the probability of a clinical document being visualized as the single rank variable. However, if we need to include documents from different patients, we might end-up biasing the results by considering erroneous visualizations. Naively, the probabilities of visualization of all possible documents need to be computed everytime the system needs to list the recommendations, which will turn the process infeasible given the streaming setup. There are, at least, three possible ways to solve this issue:

1. at every visualization event, probability of visualization of only that document is updated;
2. at every visualization event, probability of visualization of all documents is updated;
3. at every ranked list request, relevance of that patient's or similar patient's documents is updated;

This is the least solved part of the system, and future work is expected. Nevertheless, we feel that it is important to keep this target in mind because it may condition the possible paths that previous modules are going to traverse.

4 Future steps and expected impact of the system

Future work is concentrated on: a) defining a data stream management system, b) defining the factors that influence the relevance of clinical documents, c) build learning models to estimate the relevance of a single document for a given encounter, d) generate recommendations based on a ranking, and e) develop and test the prototype with real data. To our knowledge, the use of machine learning techniques to support graphical user interfaces and management storage systems in healthcare information systems is novel and could be an important contribution to science and likely to be incorporated into commercial products in the future. The results of this research are also likely to have an important impact on the quality of healthcare by further increasing the usability and intelligence of existing information systems.

Acknowledgments

This work is supported by Portuguese Foundation for Science and Technology (FCT) under projects OPTIM (PTDC/EIA-EIA/099920/2008) and KDUDS (PTDC/EIA-EIA/98355/2008).

References

1. Barnett, O.: Computers in medicine. JAMA: the journal of the American Medical Association 263(19), 2631 (1990)
2. Clamp, S., Keen, J.: Electronic health records: Is the evidence base any use? Medical Informatics and the Internet in Medicine 32(1), 5–10 (2007)
3. Cormode, G., Muthukrishnan, S., Zhuang, W.: Conquering the divide: Continuous clustering of distributed data streams. In: Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007). pp. 1036–1045 (2007)
4. Cruz-Correia, R., Vieira-Marques, P., Ferreira, A., Almeida, F., Wyatt, J.C., Costa-Pereira, A.: Reviewing the integration of patient data: how systems are evolving in practice to meet patient needs. BMC Medical Informatics and Decision Making 7(1), 14 (2007)
5. Cruz-Correia, R., Wyatt, J., Dinis-Ribeiro, M., Costa-Pereira, A.: Determinants of frequency and longevity of hospital encounters' data use. BMC Medical Informatics and Decision Making 10, 15 (2010)

6. Dick, R., Steen, E.: The Computer-based Patient Record: An Essential Technology for HealthCare. National Academy Press (1997)
7. Gama, J., Medas, P., Castillo, G., Rodrigues, P.P.: Learning with drift detection. In: Bazzan, A.L.C., Labidi, S. (eds.) Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA 2004). Lecture Notes in Artificial Intelligence, vol. 3171, pp. 286–295. Springer Verlag, São Luiz, Maranhão, Brazil (October 2004)
8. Gama, J., Rodrigues, P.P.: Data stream processing. In: Gama, J., Gaber, M.M. (eds.) Learning from Data Streams - Processing Techniques in Sensor Networks, chap. 3, pp. 25–39. Springer Verlag (2007)
9. Gama, J., Sebastião, R., Rodrigues, P.P.: Issues in evaluation of stream learning algorithms. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009). pp. 329–337. ACM Press, Paris, France (2009)
10. Guha, S., Meyerson, A., Mishra, N., Motwani, R., O'Callaghan, L.: Clustering data streams: Theory and practice. IEEE Transactions on Knowledge and Data Engineering 15(3), 515–528 (2003)
11. Hripcsak, G., Sengupta, S., Wilcox, A., Green, R.: Emergency department access to a longitudinal medical record. Journal of the American Medical Informatics Association 14(2), 235–238 (2007)
12. Hulten, G., Spencer, L., Domingos, P.: Mining time-changing data streams. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 97–106. ACM Press (2001)
13. Khan, S.S., Madden, M.G.: A survey of recent trends in one class classification. In: Artificial Intelligence and Cognitive Science - 20th Irish Conference. Lecture Notes in Computer Science, vol. 6206, pp. 188–197. Springer (2010)
14. Lucas, P.: Bayesian analysis, pattern analysis, and data mining in health care. Current Opinion in Critical Care 10, 399–403 (2004)
15. Lucas, P., van der Gaag, L., Hanna, A.: Bayesian networks in biomedicine and health-care. Artificial Intelligence In Medicine 30, 201–214 (2004)
16. Mitchell, T.M.: Machine Learning. McGraw-Hill, international edn. (1997)
17. Muthukrishnan, S.: Data Streams: Algorithms and Applications. Now Publishers Inc, New York, NY (2005)
18. Rodrigues, P.P., Gama, J.: A system for analysis and prediction of electricity load streams. Intelligent Data Analysis 13(3), 477–496 (June 2009)
19. Rodrigues, P.P., Gama, J., Pedroso, J.P.: Hierarchical clustering of time-series data streams. IEEE Transactions on Knowledge and Data Engineering 20(5), 615–627 (May 2008)
20. Schurink, C., Lucas, P., Hoepelman, I., Bonten, M.: Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units. Lancet Infectious Diseases 5, 305–312 (2005)
21. Shapiro, J.S., Gathers, S., Kannry, J., Kushniruk, A.W., Kuperman, G.: Survey of emergency physicians to determine requirements for a regional health information exchange network. In: AMIA Spring Congress (2006)
22. Wyatt, J.C., Wright, P.: Design should help use of patients' data. Lancet(British edition) 352(9137), 1375–1378 (1998)