

Mining Medical Records for Computer Aided Diagnosis

R. Bharat Rao, Romer Rosales, Stefan Niculescu, Sriram Krishnan
Luca Bogoni, Xiang S. Zhou, Balaji Krishnapuram

Siemens Medical Solutions, 51 Valley Stream Parkway, Malvern, PA, USA

1 Introduction

Over the last five years, a new generation of medical data mining tools have dramatically impacted the health care industry by improving the diagnosis of medical diseases and by reducing the time pressure on physicians and nurses. Our demonstration highlights three products for the the health care industry, showcasing the potential of novel data mining technologies to save lives on a large scale. During the demonstration, our products will use real-life (de-identified) patient data, in an effort to convey the practical and theoretical challenges unique to data from the medical domain.

2 Early-stage diagnosis of colorectal cancer

2.1 Background

Colo-rectal cancer (CRC) affected 147,000 patients in the US in 2004, and of them 57,000 died. Unlike many other form of cancers, CRC is removable if it is found at an early stage. In its early stage it manifests itself as colonic polyp. The recommendation is that each individual over age 50 undergo optical colonoscopy so that any polyp may be removed and to repeat the procedure after 10 years if negative and a more frequent review if any polyps are found. The prevalence in the general population is roughly 5% to 8%, with only 10% of these showing any signs of cancer (adenomatous polyps).

Virtual colonography (VC), also known as *CT Colonography* (CTC), was introduced as a means to address problematic cases which could not be accurately diagnosed by earlier methods like *optical colonoscopy* (OC). It was predicted that CTC could be used as a screening tool so that only patients with positive finding from CTC would be sent to OC.

2.2 Results of clinical studies

However, radiologists need substantial training to perform CTC, and it is a long procedure. As a result, in clinical practice, non experts readers often show a substantially lower sensitivity with CTC (75% or less on medium to large sized polyps) [1]. However, in large clinical trials on 145 individuals, our *computer aided detection* (CAD) system can accurately diagnose patients based on CTC images, with a sensitivity of around 90% for medium and large sized polyps [2]. Further, when inexperienced radiologists are assisted by our CAD product in clinical studies, they decrease their false positive rate by 66%, yet improve their sensitivity on medium

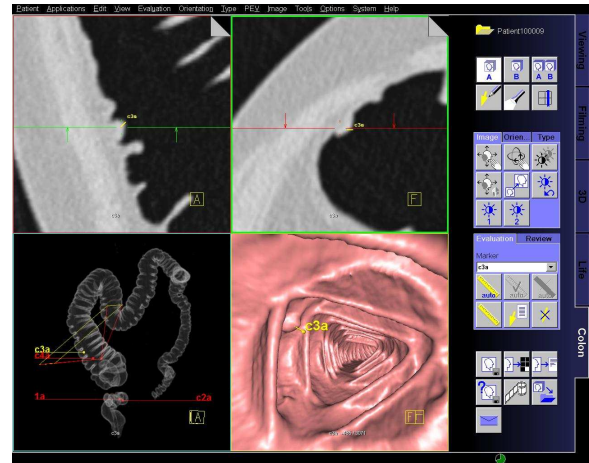


Figure 1: Polyp detection in virtual colonoscopy

and large sized polyps to 97% [1]. Thus CAD assisted inexperienced radiologists were clinically shown to diagnose polyps as accurately as experienced radiologists. Figure 1 shows screen shots from the software that will be demonstrated.

2.3 Novel technical contributions

Although the product development is still ongoing, many new theoretical and technical contributions have already been proposed. Our demonstration will highlight the following novel contributions from the fields of machine learning, data mining and computer vision.

Batch-wise classification of non-iid data: Unlike most algorithms that assume the data to be drawn iid, we exploit inter-sample correlations to improve the accuracy while classifying a set of samples simultaneously [3].

Multiple instance learning: A novel, convex-hull based algorithm finds optimal classifiers when the disease status of regions of images is imputed (guessed statistically) based on proximity to radiologist marks in training images [4].

Curvatures pattern feature descriptions [5]: Computes the principal curvatures on the surface and characterizes patterns of curvature, with the intuition that polyps are ellipsoidal

hence have an inflection belt of curvature around whereas other structures have different patterns.

Divergent Gradient Field Response analysis [6]: fast computation of the gradient field everywhere on the volume and identification of point of divergence of gradient (spheroids). The intuition is that principal directions of planar patches are well-aligned spatially, unlike those of spheroids.

3 Automated measurements of left ventricular function in the human heart

Each year 330,000 patients suffer sudden cardiac death in the US. Many of these patients can be saved if they have an implantable defibrillator (ICD) installed in them: Studies have shown that patients with a history of myocardial infarction (*i.e.*, heart attack) and left ventricular (LV) dysfunction—*i.e.*, $Ejection\ Fraction \leq 30\%$ —would benefit from such ICDs. Our second demonstration (Fig 2) will focus on a tool for automating the measurement of the ejection fraction, in order to save a significant number of lives each year through the appropriate installation of ICDs.

Reliably delineating the left ventricle for robust quantification requires echocardiographers and sonographers with years of clinical experience. Trabeculations of the myocardium, fast-moving valves, chorda and papillary muscles all make delineation of the left ventricle a challenge. Technical issues, such as the fact that a 2-D plane is acquired on a twisting 3-D object, make this problem even more difficult. Traditional image processing systems simply look for a border between black and white structures in an image. Experienced echocardiographers and sonographers, on the other hand, rely on their experience in viewing a fast-moving object and understanding individual variations in the heart. Thus, they can apply learned knowledge and experience to confidently locate the border of the left ventricle [8].

Based on this insight, Siemens has developed a fundamentally new approach to address quantification in echocardiography. This approach is based on a technology called learned pattern recognition (LPR) which makes it possible to automatically identify the endocardium of the left ventricle and track the border through the heart cycle [7].

4 Data Mining for Clinical Data Analysis

4.1 Background

Many clinical tasks of significance require the search through and analysis of medical patient records. The fundamental analysis task can be reduced to the identification and verification of medical events, conditions, diseases, or in general any (medical) concept. The number of real applications that would benefit from this type of analysis is considerable in number and importance. Examples include: diagnosis, treatment planning, recruitment for clinical trials, medical guideline compliance, quality measure evaluation, and clinical research.

The above fundamental analysis tasks present major challenges. The clinical concepts of interest may vary broadly. Examples include: incidence of a disease; presence of a medical condition; existence of an exam or laboratory test;

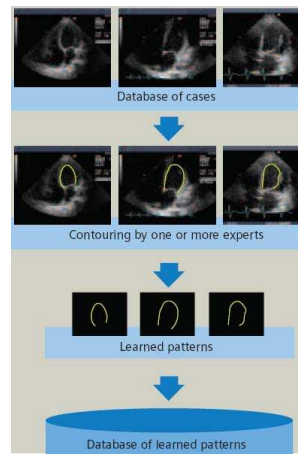


Figure 2: Automated EF measurement

presence of a symptom, habit (e.g., smoking), or risk factor. Sometimes these clinical concepts can be easily obtained from the medical records, which is the case, for example, when they are stored in structured database tables. However, in most cases, this information is hidden in unstructured data, such as lab reports, doctor notes, images, or any of the hundreds of medical documents used by health care centers.

Even when information is stored in structured form, such as in the case of diagnostic codes (ICD-9 codes), it may be unreliable from the clinical point of view. Various studies have shown that the clinical accuracy of ICD codes is only 60%-80%. The principal reason for this is that billing data reflects financial rather than clinical priorities. Financial data alone is insufficient for almost any kind of patient-level clinical decision support. Most of the information about the clinical context is stored as unstructured free text, dictated by physicians at different time points. Efforts to extract key clinical information based on natural language processing alone have met with limited success [9] - and for even slightly complex decisions like guideline eligibility, reliability is very poor.

Institutions often hire trained nurses to manually extract information from a random sample of patients, since a complete analysis is unattainable. Current methods of analyzing medical data fall into 3 classes: mine existing structured (financial) data (unfortunately ICD-9 diagnosis codes are not useful for clinical analysis), retrospective manual data extraction by nurses (costly, time-consuming and only feasible for a small patient cohort) and making clinicians enter structured clinical patient data at the point of care (physician time, ignores legacy data, must decide what to collect in advance, far away from general use).

4.2 Technical contribution

In this part of the demo, we will present various aspects of the REMIND (for Reliable Extraction and Meaningful Inference from Nonstructured Data) project. This project consists of a probabilistic framework for medical data analysis for clinical decision-support from existing medical patient records.

Our approach to inference with this multi-source, both

structured and unstructured, medical data is to model the data as arising from a generative process, and combine prior medical knowledge about this process with observations for a specific patient using Bayesian techniques. The medical prior knowledge is encoded in both a Bayesian Network that relates variables of interest as well as in the form of probabilistic rules, which are used to extract pieces of evidence about these variables from the different available data sources. REMIND's algorithm is a three step process. In an *Extraction* step, probabilistic observations about the variables in the model are gathered from the data sources. Next, in a *Combination* step, each observation is assigned to its corresponding variable and a posterior of the observation vector associated with the variable is computed locally. Finally, in the *Inference* step, the local inferences are propagated across the Bayesian Network that describes the relationships among variables of interest and the posterior probability for the variable vector is computed. The variable vector is then assigned to the maximum aposteriori realization.

4.3 Demonstration with Results of clinical studies

Even though this project is still in progress, a number of pilot studies have been deployed showing excellent results. Our experimental results include successful deployments of our methods on populations reaching up to 6 million patients. These fall into two application areas, even though technically, they are based on the same fundamental task explained above:

- Automated Analysis for Quality of Care: Our system is at least as accurate as nurses; moreover, it can be applied to the entire relevant patient population, leading not only to more accurate quality estimates, but also to actionable information that can improve patient care.
- Automated Patient Identification for Clinical Trials: Our results show that we can automatically identify patients for trials with high accuracy and results show a many fold improvement in patient enrollment compared to the traditional recruitment process.

We plan to demonstrate the different technical aspects that help us achieve these results. This will be done with the help of a data-browsing tool (Fig. 3). This tool has been designed to help understand and go through the different steps taken by our system to arrive to the inferences made about each concept. This tool shows an actual application to health care, related to real hospital quality measures (HQM). The application involves identifying, by looking at the patient record, whether pre-defined guidelines were followed for the care of each patient hospitalized for heart failure.

5 Conclusion

We believe Computer Aided Diagnosis is becoming an increasingly important area for intelligent computer systems. In this paper we have summarized several projects our group has been focusing on. The primary goal of this project demonstration is to draw attention to the applications and challenges faced by the field and highlight the strong potential that machine learning and data mining approaches have in helping to solve these problems. This demonstration intends to achieve

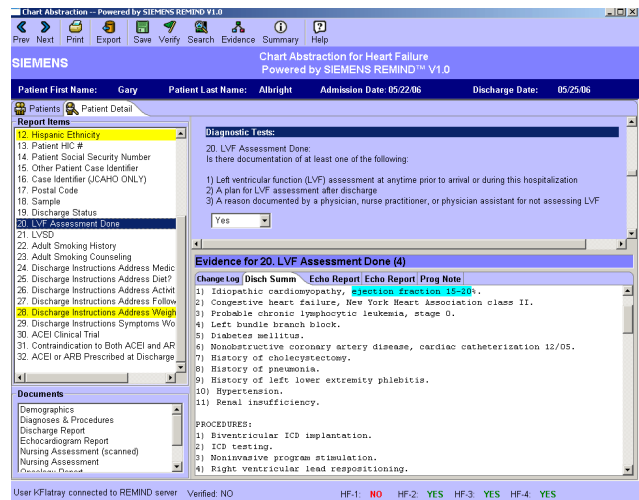


Figure 3: Data browsing tool for hospital quality measurement

these goals by combining visualization of real medical data with technical description of the algorithms employed; thus, we expect it to be of interest to the broad KDD audience.

References

- [1] T. Mang, P. Peloschek, C. Plank, A. Maier, C.J. Herold, L. Bogoni, and W. Schima. Effect of computer aided detection as a second reader in multidetector ct colonography: A multiobserver study. *Proceedings of the European Conference on Radiology*, 2006.
- [2] A. Graser, L. Bogoni, M.F. Reiser, and C.R. Becker. Performance of a prototype computer-aided detection (cad) system in mdct colonography. *Proceedings of the European Conference on Radiology*, 2006.
- [3] V. Vural, G. Fung, B. Krishnapuram, J. Dy, and R. B. Rao. Batch-wise classification with applications in cad. *Proceedings of the European Conference on Machine Learning*, 2006.
- [4] G. Fung, M. Dundar, and et al. Multiple instance learning for computer aided diagnosis. *Submitted to NIPS*, 2006.
- [5] A. Jerebko, S. Lakare, P. Cathier, S. Periaswamy, and L. Bogoni. Symmetric curvature patterns for colonic polyp detection. *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, 2006.
- [6] L. Bogoni, P. Cathier, M. Dundar, A. Jerebko, S. Lakare, J. Liang, S. Periaswamy, M.K. Baker, and M. Macari. Computer-aided detection for CT colonography: a tool to address a growing need. *The British Journal of Radiology*, 78, 2005.
- [7] X. S. Zhou, D. Comaniciu, and A. Gupta. An information fusion framework for robust shape tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2005.
- [8] B. Georgescu, X. S. Zhou, D. Comaniciu, and A. Gupta. Database-guided segmentation of anatomical structures with complex appearance. *IEEE Conf. Computer Vision and Pattern Recognition*, 2005.
- [9] C. L. Leibson, J. M. Naessens, R. D. Brown, and J. P. Whisnant. Accuracy of hospital discharge abstracts for identifying stroke. *Stroke*, 25:2348 – 2355, 1994.