D.A. Giuse, N.B. Giuse

Division of Biomedical Informatics,
Vanderbilt University Medical Center,
Nashville, USA

Synopsis

Knowledge Processing and Decision Support Systems

This section of the Yearbook contains six articles on Knowledge Processing and Decision Support Systems for various biomedical domains. While the application areas and techniques described in the articles differ considerably, the general thread is the potential for discovery and extraction of clinical information from existing collections of data. Four of the articles have as their primary area of interest the processing of natural-language electronic text, ubiquitous in today's computerized patient records but nonetheless very difficult to harvest properly. The other two articles deal with quantitative data routinely collected during clinical encounters, and with the issues that arise when attempting to classify patients according to some of the data. The idea of classification, in fact, is the central theme that is shared by all articles in this section. Whether the goal is to classify text documents from a large collection according to patient characteristics, or to classify patients requiring coronary artery bypass according to the predicted outcome of the procedure, these articles strive to achieve more accurate and more economical extraction of salient features from unorganized data. The key insight is that manual analysis of large databases, or collections of text records, is both labor-intensive and error-prone, because of well-known limitations in people's ability to absorb and process large numbers of facts.

The six articles exemplify a variety of approaches and have different goals. One article uses a manual simulation method to study the potential for a future computerized application; one evaluates the potential transferability of an existing natural-language processing system to a different institution; one builds a semantic superstructure on top of the UMLS knowledge sources; and three tackle specific classification problems. The software techniques employed range from artificial neural networks to vector-space similarity measures.

Kuilboer et al. [1] investigate the potential suitability of computerized patient records as the basis for a critiquing system. The ultimate goal is the ability to use already collected clinical data to generate helpful advice to clinicians in the course of patient care. While this has been a stated goal of much research in the past, large unresolved issues remain. These issues are not only technical (what kind of computer system would be required to provide helpful clinical advice on a routine, timely basis?) but also, importantly, social. The article examines specifically issues of acceptance (how would care providers react to a potential system's critique?) and impact (would the generated advice be followed, and would it affect patient care?). The authors define the target application as an inquisitive critiquing system, i.e., one that would ask for further information if it determined that available data were not sufficient for competent critiquing. The study described in the article is based on a small number of medical records, and on manual simulation and review rather than on a computer prototype. The study classifies the types of suggestions and critiques that a future system could potentially offer, and determines that comments about drug prescriptions and about the physician's plan of treatment would likely to be the most numerous. The study also determines the potential users' judgment of relevance of each kind of suggestion, and whether the users would tend to agree or disagree with various types of suggestion.

Hripcsak et al. [2] provide an evaluation of the transportability of MedLEE, a text processing system developed at Columbia-Presbyterian Medical Center, to a different institution. Moving a complex software artifact from the developing site to another place, generally involves considerable effort and expense. This problem is especially pronounced for systems that process natural language text (NLP systems), because of the considerable variability in the language used by different sites and different care providers to describe clinical findings. The authors aim at providing a realistic measure of the effort involved in transporting an NLP system, and of how the system performs at the new site. The system evaluated consisted

of two modules, MedLEE proper (a natural language processor that can process textual reports and generate a list of coded findings) and a query engine that, given a list of coded findings, can decide whether particular clinical conditions are present. Thus, the combination of the two modules can scan reports in electronic text form and classify them based on whether they do or do not describe a clinical condition. The article evaluates the porting effort and the effectiveness of the system at processing radiology reports under three different scenarios. In the first scenario, the intact system in use at the originating institution was tested unmodified at the second site. In the second, the NLP module was modified based on a training set from the receiving site, making it better tuned for the clinical language used at that site. In the third scenario, the NLP module was unchanged, but the queries were modified based on a training set of local reports complete with gold-standard classifications by local physicians. To put the system's performance in perspective, the study design also included classification of the reports by various groups of human readers, with comparisons of the sensitivity and specificity achieved by (a) a group of internists; (b) a group of radiologists; (c) a group of lay persons; and (d) the three versions of the system. This intriguing study design allows for comparisons between specialists that are intimately familiar with the language used in the reports (the radiologists), primary users (the internists), people with no medical knowledge but excellent language processing skills (the lay persons), and the computer system. The unmodified system performed acceptably at the new site, but its sensitivity was worse than that of the physicians ($p<0.05$), although it often outperformed the lay persons (not statistically significant). The system with an NLP module adapted to the new site did not perform significantly better. The third version of the system, on the other hand, was not distinguishable from the whole set of physicians, with sensitivity and specificity of 0.86 and 0.98, respectively. The effort required to modify this version was moderate, of the order of a few days (the original system took three person-years to develop). The study also provides an analysis of the disagreement among physicians in interpreting the presence or absence of clinical conditions in the reports, and confirms the well-known difficulty of obtaining good agreement on the interpretation of uncoded reports.

De Bruijn et al. [3] take a decidedly information retrieval (IR) approach to the classification of natural-language texts. As was the case for the first two articles, the main motivation is the potential for automatic retrieval of the reports that match a particular set of clinical conditions. The study uses 7,500 full-text pathology reports, coded with SNOMED as part of the normal billing process, to perform a statistical nearest-neighbor retrieval of the reports that match a given condition more closely. The SNOMED codes are used to evaluate the actual similarity of retrieved reports, and in which rank the desired condition appeared (if at all) within the coded conditions. The evaluation methodology uses a "leave-one-out" technique, in which one report at a time is compared against all the others and the nearest-matching documents are retrieved. For each step, the actual similarity of the retrieved set (as computed from the SNOMED codes) is evaluated. The methodology used is a classic vector-space approach familiar from IR research. The study's most interesting point is the use of the SNOMED codes as an independent evaluation metric that can measure the effectiveness of the nearest-neighbor retrieval scheme.

Joubert et al. [4] describe project ARIANE, a French initiative to facilitate the retrieval of information from clinical repositories by creating a conceptual superstructure that can hide from users differences in the underlying terminology or representation. The authors use UMLS as the building block, and develop a collection of semantic structures that encompass the UMLS knowledge sources (specifically, the Metathesaurus and the Semantic Network). The specific goal of this activity is to identify and distinguish the various viewpoints ("contexts") from which the same UMLS concept can be seen, thus making explicit the differences that are necessary to support meaningful querying of a large database for different purposes. The authors describe a set of structures, based on John Sowa's Conceptual Graph notation, that provide an intermediate semantic layer between Metathesaurus concepts and the top-level hierarchy of concepts in the UMLS Semantic Network. This new set of structures is especially tailored to support user queries. Typically, this could be done through the creation of a query graph - based on the original query - that can then be mapped against the enriched UMLS structures and restricted to only the context of the original query. This approach would make the result more specific to the user's information needs in the particular context. The work described in the article should be considered at the theoretical stage, as the authors do not provide any user evaluation or indication of actual system use.

Lippmann et al. [5] investigate the use of artificial neural networks as a mechanism to predict risk-adjusted mortality for patients undergoing coronary artery bypass. The study design is a retrospective analysis of 33 yes/no predictors, using a data set of more than 80,000 CABG patients developed by The Society of Thoracic

Surgeons. The objective is to evaluate the discrimination and calibration of several configurations of a neural network versus a logistic regression analysis and a Bayesian model. Overall, the various classification methods performed quite similarly, with sensitivity and specificity that were not significantly different and that agreed with previously reported results on the accuracy of outcome predictions for CABG patients. Both the neural network models and the Bayesian model tended to behave worse for high-risk patients. The best calibration results were actually obtained by a hybrid "committee classifier", which blended the predictions of the neural network model with those of the logistic regression model. This result does not appear to be generalizable to other classification problems, however. The key conclusion of the study is that, at least for this particular problem, neural networks did not perform significantly differently from two other well-known classification techniques.

Katsuragawa et al. [6] reach a similar conclusion in a different comparison based on a different clinical problem - the classification of digital lung radiographs to distinguish between normal and abnormal lung images. The two techniques evaluated include a rule-based method and a neural network method. The input variables for the evaluation were various indices computed by geometric feature analysis of the radiographs. The indices were fed into a rule-based method previously developed by the authors, and into a three-layer neural network with back-propagation. Additionally, the authors experimented with various logical combinations of the two methods (such as "a chest radiograph is classified as abnormal if either the rule-based method or the neural network method classify it as abnormal". The best-performing combined method performed better than either method alone, although not all differences were statistically significant. The authors speculate that this is because the rule-based method can quickly eliminate the easy cases, and allow the neural network method (which requires substantial training to achieve good performance) to be trained directly on the most difficult cases.

For the most part, the articles in this section evaluate techniques and methods designed to facilitate classification, rather than actual decision-support systems for clinical use. The system described by Hripcsak et al. [2] is the only one in actual use; the study simulates transporting it to a different site, although the system was not actually used by clinicians at the new site. The overall impression from this collection of articles is that the field is exploring different techniques and attempting to evaluate their effectiveness for real-world classification problems. Unfortunately, it does not appear that any one technique will significantly outperform the others; in several of the studies the differences between human classifiers and the various automated techniques were minor. While it is comforting that the automated methods can often perform as well as humans in limited domains, it is not clear that this by itself will make them commonplace. It would seem, then, that we have reached a temporary plateau that is intrinsic to either the nature of the data or the task itself. It may be necessary to either improve dramatically the quality of the available data or to discover entirely new analysis methods, to overcome the current difficulties and realize the original vision of extracting clinical gold nuggets from our vast data repositories.

References

1. Kuilboer MM, Van der Lei J, de Jongste JC, Overbeek SE, Ponsioen B, Van Bemmel JH. Simulating an integrated critiquing system. J Am Med Inform Assoc 1998;5:194-202.
2. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Meth Inform Med 1998; 37:1-7.
3. De Bruijn LM, Hasman A, Arends JW. Automatic SNOMED classification — a corpus-based method. Comput Methods Progr Biomed 1997;54:115-22.
4. Joubert M, Fieschi M, Robert J-J, Volot F, Fieschi D. UMLS-based conceptual queries to biomedical information databases: An overview of the project ARIANE. J Am Med Inform Assoc 1998;52-61.
5. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. Ann Thorac Surg 1997;63:1635-43.
6. Katsuragawa S, Doi K, MacMahon H, Monnier-Cholley L, Ishida T, Kobayashi T. Classification of normal and abnormal lungs with interstitial diseases by rule-based method and artificial neural networks. J Digit Imag 1997;10:108-14.

Address of the authors:
Dario A. Giuse, Nunzia B. Giuse
Division of Biomedical Informatics,
Vanderbilt University Medical Center,
Nashville, TN 37232-8340,
USA