Discovering Latent Structure in Clinical Databases

Jesse Davis
Dept. of Computer Science
Katholieke Universiteit Leuven
3000 Leuven, Belgium
jesse.davis@cs.kuleuven.be

David Page
Dept. of Biostatistics and Medical Informatics
University of Wisconsin-Madison
Madison, WI
page@biostat.wisc.edu

Vitor Santos Costa
Dept. of Computer Science
Universidade do Porto
Porto, Portugal
vsc@dcc.fc.up.pt

Peggy Peissig and Michael Caldwell
Marshfield Clinic Research Foundation
Marshfield, WI
{caldwell.michael, peissig.peggy}@mcrf.mfldclin.edu

There has been a fundamental shift in health care practice with the advent and wide spread use of electronic medical records. Now that the relevant data reside on disk as opposed to paper charts, it is possible to apply machine learning and data mining techniques to this data. Access to this data will allow researchers to pose and investigate a large number of questions. Examples of these questions include: Can one develop a model to predict the efficacy of a potential drug for a given individual? Can clinical data provide insight into which individuals will have adverse reactions to a drug?

From a technical perspective, analyzing this type of data poses many challenges for machine learning and data mining techniques. These obstacles include:

Multiple relations. Each type of data (e.g. drug prescription information, lab test results, etc.) is stored in a different table of a database. Traditionally, machine learning algorithms assume that data are stored in a single table.

Represent uncertainty. The data are inherently noisy. For example, lab test results may vary due to lab conditions and personnel.

Missing/incomplete data. Patients switch doctors and clinics over time, so a patient's entire clinical history is unlikely to reside in one database. Furthermore, crucial information, such as the use of over-the-counter drugs, may not appear in the clinical history.

Schema not designed to empower learning. The clinical databases are designed to optimize ease of data access and billing rather than learning and modeling.

Methodological issues for longitudinal data. Working with data that contains time dependencies introduces several problems. The central problem we had to address in our work was what data was appropriate to include in our analysis.

One challenge, which is the focus of this work, is the presence of latent structure within the data. More specifically, the data contain information about specific medicines taken and disease diagnosis for an individual patient. Yet, what is missing is the connection among groups of drugs (e.g., Aleve and Tylenol are pain killers) or diseases (e.g., hypertension and high-cholesterol pertain to the heart). Consequently, it can be difficult to detect interesting and meaningful patterns present in the data. For example, there may be a large number of people that take a cholesterol medicine and also have certain outcome (e.g., disease diagnosis, adverse drug reaction, etc.). However, the data only tells us which specific medicine a patient has been prescribed and the number of people that take each individual medicine and have the outcome may be small and not meet an interestingness threshold (e.g., support threshold in association rule mining). What is missing is the ability to automatically detect that these medicines are related and group them together. An additional challenge is that it is unclear what is the best way to group together different diseases or drugs. For example, Zocor^{TM}, Mevacor^{TM} and Niacin^{TM} all treat high cholestrol, but operate differently. Here, it may be reasonable to group together all three drugs since they treat the same disease. Alternatively, grouping just

Zocor$^{\mathrm{TM}}$ and Mevacor$^{\mathrm{TM}}$ would be possible as they both belong to the class of drugs known as statins. Furthermore, it may be desirable to group together drugs based on potential side-effects and not the disease they treat. An ideal approach to this problem should not commit to one grouping and should be able to dynamically decide which grouping is best. Furthermore, each entity (e.g., drugs, diseases, etc.) should be able to appear in multiple groupings at once.

We base our approach off the VISTA system [1], which combines first-order rule learning with Bayesian network structure learning. At a high-level, VISTA works by proposing a first-order rule, which is then converted into a binary variable in the Bayesian network. VISTA then learns a new model (i.e., the structure of the Bayesian network) incorporating the new feature, and evaluates the model. If the model does not improve, the rule is rejected, and VISTA reverts back to the old model that does not contain the rule. In order to decide whether to retain a candidate feature f, VISTA needs to estimate the generalization ability of the model with and without the new feature. VISTA does this by calculating the area under the precision-recall curve on a tuning set. VISTA offers several advantages for analyzing clinical data. First, first-order rules can incorporate information from different tables in the database within a single rule. Second, by incorporating each rule into a Bayesian network, we can capture the inherent uncertainty in the data. Third, the learned rules are comprehensible to domain experts.

The key innovation occurs in the first-order rule construction, where we automatically attempt to group together drugs or diseases to learn a more general rule. For example, if a rule only applies to patients who have taken "Zocor", it could be generalized to cover patients who have taken either "Zocor" or "Mevacor." Our approach works as follows. If a rule performs a test on a drug (disease), it attempts to general the test to look at a set of drugs (diseases). The generalization is done by greedily adding drugs (diseases) to the group until no drug (disease) improves the score of the model. Furthermore, it take a macro step by adding a set of drugs (diseases) that have previously been grouped together. This approach offers two key advantages. One, the groupings are driven by what results in more accurate learned model. Two, each drug or disease can appear in multiple different groupings. One complication is the large number of diseases and drugs, which makes it prohibitively expensive to try adding each possible disease or drug to a group. Therefore, we only consider adding the most promising candidates to a group. We have explored several different techniques to identify promising candidates including looking at correlations among drugs or diseases and looking at drugs or diseases in near-miss examples.

As a case study, we focus on the task of predicting which patients that take a selective Cox-2 inhibitor (e.g., Vioxx$^{\mathrm{TM}}$) will have a myocardial infarction (MI) (i.e., a heart attack) while on the drug [2]. To create a set of negative examples, we took patients that were prescribed a selective Cox-2 inhibitor and did not have an MI. Furthermore, we matched the negative examples to have the same age and gender distribution as the positive examples to control for those risk factors. Our data comes from Marshfield Clinic, which is a large multispecialty outpatient clinic located throughout central and northern Wisconsin. This organization has been using electronic medical records since 1985 and has electronic data back to the early 1960's. We found that the proposed approach produced a more accurate model than the baseline approach. Furthermore, we found interesting clustering, such as one that grouped together diagnosis codes for hypertension, diabetes and high cholesterol.

References

[1] Jesse Davis, Irene Ong, Jan Struyf, Elizabeth Burnside, David Page, and Vítor Santos Costa. Change of representation for statistical relational learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2719–2726. AAAI Press, 2007.

[2] Patricia M Kearney, Colin Baigent, Jon Godwin, Heather Halls, Jonathan R Emberson, and Carlo Patrono. Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the risk of atherothrombosis? meta-analysis of randomised trials. *BMJ*, 332:1302–1308, 2006.

Presenter: Jesse Davis
Topics: Data mining
Preference: Oral