Machine Learning for Personalized Medicine: Will This Drug Give Me a Heart Attack?

Jesse Davis JDAVIS@CS.WASHINGTON.EDU
Department of Computer Science and Engineering, University of Washington

Eric Lantz LANTZ@CS.WISC.EDU
David Page PAGE@BIOSTAT.WISC.EDU
Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison

Jan Struyf JAN.STRUYF@CS.KULEUVEN.BE
Department of Computer Science, Katholieke Universiteit Leuven

Peggy Peissig PEISSIG.PEGGY@MARSHFIELDCLINIC.ORG
Humberto Vidaillet VIDAILLET.HUMBERTO@MCRF.MFLDCLIN.EDU
Michael Caldwell CALDWELL.MICHAEL@MCRF.MFLDCLIN.EDU
Marshfield Clinic Research Foundation, Wisconsin

Abstract

This paper presents a significant application domain in health informatics. With the advent of electronic medical records, researchers could have access to clinical records for patients participating in various research studies. Access to this data will allow researchers to pose and investigate the following types of questions. 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? This paper focuses on a case study with a real-world database. The specific task is to learn a statistical model to indicate which patients on Cox-2 inhibitors, such as Vioxx^{TM}and Celebrex^{TM}, are at substantial risk for heart attack.

1. Introduction

Personalized medicine represents a significant application for the health informatics community. Its objective can be defined as follows:

- Given: A patient's clinical history
- Do: Create an individual treatment plan

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. There have also been several dramatic examples of patient variation in response to drugs such as Vioxx^{TM}and Coumadin^{TM}. These cases have highlighted the need for tools that can help a doctor more accurately determine which drug and dosage to prescribe a patient.

From a technical perspective, personalized medicine presents 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. For example, see Figure 1.

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, things like the use of over-the-counter

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

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.

Analyze large amounts of data. As more clinics switch to electronic medical records, the amount of data available for analysis will exceed the capability of current machine learning techniques.

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.

The remainder of this paper presents a case study with a real-world database. The specific task is to learn a statistical model to indicate which patients on selective Cox-2 inhibitors, such as Vioxx™and Celebrex™, are at substantial risk for heart attack.

2. Case Study

Non-steroidal anti-inflammatory drugs, known as NSAIDs, are used to treat pain and inflammation. Many NSAIDs, such as Aleve™and Advil™, work by blocking the Cox-1 and Cox-2 pathways. While these medications can effectively alleviate pain, prolonged use may result in gastrointestinal problems as a consequence of blocking the Cox-1 pathway (Simmons et al., 2004). The selective Cox-2 inhibitor hypothesis is that a drug that only (i.e., selectively) blocks the Cox-2 pathway will have the same benefits of traditional NSAIDs, while eliminating the side effects. To this end, drugs such as Vioxx™, Bextra™and Celebrex™were introduced to the American drug market between 1998 and 2001. While widely prescribed, several of these medications were removed from the market due to concerns that they resulted in an increased risk of myocardial infarctions (heart attacks) (Kearney et al., 2006). The goal of our case study can be defined as follows:

- Given: A patient's clinical history
- Do: Predict whether the patient will have a myocardial infarction (MI)

2.1. Methodological Issues

One of the central issues we needed to address for our case study is what data should we include in our analysis.

Using all available patient data introduces several problems. One, care must be taken to remove the class attribute from the data. For example, removing all instances of the diagnosis code for MI is incorrect. It is important to retain all data for a given patient prior to the first selective Cox-2 inhibitor prescription. Two, data for the positive cases must be cut off after diagnosis of an MI. Data collected after the MI should not be included in the model. Three, other features in the data are highly correlated with MI, such as being administered a lab test to diagnosis an MI. In short, a non-trivial set of information must be removed from the data in order to obtain reasonable results.

Even cutting off data for a patient immediately before an MI raises two important issues. First, the model might not be relevant for deciding whether to prescribe a patient a selective Cox-2 inhibitor. Second, we will have uniformly more data for the non-MI cases, which introduces a subtle confounding factor in the analysis. For example, consider a drug recently introduced into the market. Under this scheme, patients on selective Cox-2 inhibitors who had MIs before the drug came on the market could not have taken the drug. Thus, it could introduce a spurious correlation between taking the drug and not having an MI. In other words, the drug could incorrectly be flagged as being protective against MI.

Consequently, our choice was to cut off data for each patient at the first Cox-2 prescription. This option mitigates a majority of the concerns and problems we encountered due to the longitudinal nature of this data.

2.2. Data

Our data comes from Marshfield Clinic, an organization of hospitals and clinics in northern Wisconsin. This organization has been using electronic medical records since 1985 and has electronic data back to the early 1960's. Furthermore, it has a reasonably stationary population, so clinical histories tend to be very complete. The database contained 77,569 patients who had taken Cox-2 inhibitors, 492 of which later had an MI. From the non-MI group, we subsampled 650 patients for efficiency reasons. We included information from four separate relational tables: lab test results (e.g. cholesterol levels), medications taken (both prescription and non-prescription), disease diagnoses, and observations (e.g. height, weight and blood pressure).

2.3. Empirical Evaluation

The primary objective of the empirical evaluation is to demonstrate that machine learning techniques can predict which patients on Cox-2 inhibitors are substantial risk for MI. We tried many different types of

A.

PatientID	Gender	Birthday
P1	M	3/22/63

B.

PatientID	Date	Physician	Symptoms	Diagnosis
P1	1/1/01	Smith	palpitations	hypoglycemic
P1	2/1/03	Jones	fever, aches	influenza

C.

PatientID	Date	Lab Test	Result
P1	1/1/01	blood glucose	42
P1	1/9/01	blood glucose	45

D.

PatientID	SNP1	SNP2	…	SNP500K
P1	AA	AB		BB
P2	AB	BB		AA

E.

PatientID	Date Prescribed	Date Filled	Physician	Medication	Dose	Duration
P1	5/17/98	5/18/98	Jones	prilosec	10mg	3 months

Figure 1. A simplified clinical database. Table **A** contains information about each patient. Table **B** lists symptoms and disease diagnoses from patient visits. Table **C** contains lab test results. Table **D** has single nucleotide polymorphism (SNP) data for patients, which are included in this study. Table **E** has drug prescription information.

algorithms, which can be divided into (i) propositional learners, (ii) relational learners, and (iii) statistical relational learners.

We looked at a wide variety of feature vector learners. From Weka, we used naïve Bayes and linear SVM (Witten & Frank, 2005) as well as our own implementation of tree-augmented naïve Bayes (Friedman et al., 1997). Additionally, we used the decision trees, boosted decision trees and boosted rules algorithms from the C5.0 (Quinlan, 1987) package. The disadvantage of using these techniques is that they require propositionalizing the data, so we must collapse the data into a single table.

Relational learning allows us to directly operate on the multiple relational tables (Lavrač & Džeroski, 2001). We used the inductive logic programming (ILP) system Aleph (Srinivasan, 2001), which learns rules in first-order logic. ILP is appropriate for learning in multi-relational domains because the learned rules are not restricted to contain fields or attributes for a single table in a database. The ILP learning problem can be formulated as follows:

- Given: background knowledge B, set of positive examples E^+, set of negative examples E^- all expressed in first-order definite clause logic.
- Learn: A hypothesis H, which consists of definite clauses in first-order logic, such that $B \wedge H \models E^+$ and $B \wedge H \not\models E^-$.

In practice, it is often not possible to find either a pure rule or rule set. Thus, ILP systems relax the conditions that $B \wedge H \models E^+$ and $B \wedge H \not\models E^-$. ILP offers another advantage in that domain experts are easily able to interpret the learned rules. However, it does not allow us to represent uncertainty.

Statistical relational learning (SRL) is the subfield of machine learning whose goal is to develop systems that can reason about uncertainty in structured data. We used the SAYU system, which combines Aleph with a Bayesian network structure learner (Davis et al., 2007a). SAYU uses Aleph to propose features to include in the Bayesian network. SAYU uses Aleph to propose features to include in the Bayesian network. Aleph passes each clause (rule) it constructs to SAYU, which converts the clause to a binary feature and adds it to the current training set. Next, SAYU learns a new model incorporating the new feature, and evaluates the model. If the model does not improve, the rule is not accepted, and control returns to Aleph to construct the next clause. In order to decide whether to retain a candidate feature f, SAYU needs to estimate the generalization ability of the model with and without the new feature. SAYU does this by calculating the AUC-ROC on a tuning set. By retaining Aleph,

SAYU also offers the advantage that the constructed features are comprehensible to domain experts.

2.4. Feature Construction

The complex, structured nature of patient clinical histories represents several problems for standard machine learning algorithms. First, the data for each patients are spread across multiple relational tables. However, most machine learning techniques assume the data reside in a single table. Second, the rows within a single table can be related. For example, the diagnosis table will contain one entry, or row, for each disease that a patient has been diagnosed with over the course of his life.

Propositionalization is common technique for converting relational data into a suitable format for standard learners (Lavrač & Džeroski, 1992; Pompe & Kononenko, 1995). Such work re-represents each example as a feature vector, and then uses a feature-vector learner to produce a final classifier. However, in our experience propositionalization are very computationally expensive. In fact, in our preliminary experiments, we found that creating a reasonable feature set took on the order of days for each fold. Consequently, we created the following propositional features:

- One binary feature for each lab test, which is true if the the patient ever had that lab test.
- One binary feature for each drug, which is true if the patient ever took the drug.
- One binary feature for each diagnosis, which is true if the patient had this disease diagnosis.
- Three aggregate features (min, max, avg) for each type of observation.

This resulted in 3620 features.

2.5. Results

We performed ten-fold cross validation. For each run of cross-validation, we used information gain to select the top 50 features. We tried selecting the top 100 and top 150 features, but performance seemed to be invariant to this parameter. For SAYU, we used five folds as a training set and four folds as a tuning set. We used tree-augmented naïve Bayes as the structure learner. We scored each candidate feature using area-under the ROC (AUC-ROC) curve. A feature must improve the tune set AUC-ROC by two percent to be incorporated into the model. Additionally, we seeded the initial SAYU model with the 50 highest scoring features on that fold (Davis et al., 2007a).

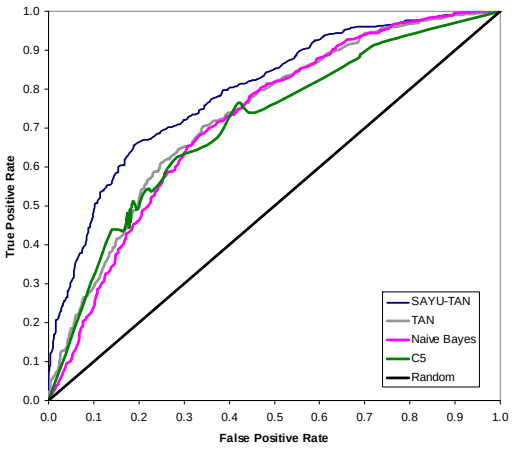

Figure 2. ROC Curves

Table 1. Average area under the ROC curve for the four best methods.

Naïve Bayes	TAN	Boosted Rules	SAYU-TAN
0.7428	0.7470	0.7083	0.8076

Figure 2 shows the ROC curves for the four best methods. All methods do better than chance on this task. Table 1 shows the average AUC-ROC for each method. SAYU clearly dominates the proposition learners for false positive rates less than 0.75. To assess significance, we performed a paired t-test on the per-fold AUC-ROC for each method. No significant difference exists between any of the propositional methods. Aleph, the relational method did extremely poorly and was exceedingly slow, so we did not finish all ten folds. SAYU did significantly better than all the propositional methods. The goal of the empirical evaluation is to present a proof of concept. To this end, these results are extremely promising.

SAYU-TAN mostly likely confers a benefit by integrating the feature induction and model construction into one, coherent process. This confirms recent results (Landwehr et al., 2005; Landwehr et al., 2006; Davis et al., 2005; Davis et al., 2007b) that have empirically demonstrated that an integrated approach results in superior performance compared to the traditional propositionization framework of decoupling these two steps.

2.6. Limitations

However, this limited case study suffers from several drawbacks. Namely, is our model predicting predisposition to MI? To address this question, we plan to

analyze both MI and non-MI patients who did not take Cox-2 inhibitors. If the model built on the Cox-2 patients performs equivalently well on this set of patients, it will indicate that the signal corresponds more to predisposition to MI than to adverse reaction to Cox-2 inhibition. However, this would be a valuable insight in its own right. Yet, selecting patients to include in the no Cox-2 control group is difficult. It will require matching patients between the two groups based on factors such as age, gender and MI risk factors. Furthermore, the MI patients in the control must have experienced the most recent MI during the time period during which the selective Cox-2 inhibitors were prescribed. This is necessary to ensure we have both similar amounts and types of data for the patients in each group.

Several other issues exist which are much harder to quantify. In particular, it is highly likely that the training data contains false negatives. For example, some patients may have a future MI related to Cox-2 use.

3. Lessons Learned

SRL is the best approach. We found that using techniques from SRL lead to improvements over both standard propositional learning techniques as well as relational learning techniques. On this task, the ability to address the first two challenges listed in the introduction and simultaneously handle multiple relations and represent uncertainty appears to be crucial for good performance.

Comprehensibility is crucial. The ability to learn a comprehensible model provided the basis for interacting with medical collaborators. Our collaborators love to see learned rules and a large portion of meetings were devoted to reviewing rules discovered by C5 and SAYU. In particular, they found it very useful to look at the number of positive and negative examples covered by each rule. In fact, we were often asked to look at altering the rules by adding or deleting preconditions in order to determine how this would effect coverage. It is impossible to have this type of interactions with a techniques such as SVMs. In fact, SVMs were sufficiently incomprehensible that we quickly abandoned investigating them as a potential model. Producing comprehensible models lead to our ability to incorporate new background knowledge. Additionally, it helped us address the methodological issues about what data to include in our analysis.

We are not there yet. Biomedical applications hold promise, but we need better accuracy for practical use. Incorporating genetic data, acquired through high-throughput screening, may significantly improve performance, and is important direction for future work.

Acknowledgments

Jesse Davis and Jan Strufy's work on the data analysis took place while both were at the University of Wisconsin-Madison. JD and EL are supported by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program at UW-Madison (NLM 5T15LM007359).

References

Davis, J., Burnside, E., Dutra, I., Page, D., & Costa, V. S. (2005). An integrated approach to learning Bayesian networks of rules. *Proceedings of the 16th European Conference on Machine Learning* (pp. 84–95). Porto, Portugal.

Davis, J., Burnside, E., Dutra, I. C., Page, D., Ramakrishnan, R., Costa, V. S., & Shavlik, J. (2007a). Learning a new view of a database: With an application to mammography. In L. Getoor and B. Taskar (Eds.), *An Introduction to Statistical Relational Learning*. MIT Press.

Davis, J., Costa, V. S., Ray, S., & Page, D. (2007b). An integrated approached to feature construction and model building for drug activity prediction. *Proceedings of the 24th International Conference on Machine Learning*. Corvallis, Oregon.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian networks classifiers. *Machine Learning*, *29*, 131–163.

Kearney, P. M., Baigent, C., Godwin, J., Halls, H., Emberson, J. R., & Patrono, C. (2006). 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.

Landwehr, N., Kersting, K., & Raedt, L. D. (2005). nFOIL: Integrating Naive Bayes and FOIL. *Proceeding of the 20th National Conference on Artificial Intelligence* (pp. 795–800). Pittsburgh, Pennsylvania.

Landwehr, N., Passerini, A., Raedt, L. D., & Frasconi, P. (2006). kFOIL: Learning simple relational kernels. *Proceedings of the 21st National Conference on Artificial Intelligence*. Boston, Massachusetts.

Lavrač, N., & Džeroski, S. (1992). Inductive learning of relations from noisy examples. In S. Muggleton (Ed.), *Inductive Logic Programming*, 495–516. Academic Press.

Lavrač, N., & Džeroski, S. (Eds.). (2001). *Relational Data Mining*. Springer.

Pompe, U., & Kononenko, I. (1995). Naïve Bayesian classifier within ILP-R. *Proceeding of the 4th International Workshop on Inductive Logic Programming* (pp. 417–436). Toyko, Japan.

Quinlan, J. (1987). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Simmons, D., Botting, R., & HLA, T. (2004). Cyclooxygenase isozymes: The biology of prostaglandin synthesis and inhibition. *Pharmacological Reviews*, *56*, 387–437.

Srinivasan, A. (2001). *The Aleph Manual*.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.