Proceedings of the 6th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2011)
P. Qian, Y. Zhou, C. Rudin, eds.

Predicting Medical Conditions with Bayesian Hierarchical Rule Modeling

Tyler H. McCormick
Department of Statistics
University of Washington
Seattle, WA
tylermc@stat.washington.edu

Cynthia Rudin
MIT Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA

David Madigan
Department of Statistics
Columbia University
New York, NY

Abstract

We propose a statistical modeling technique, called the Hierarchical Association Rule Model (HARM), that predicts a patient's possible future medical conditions given the patient's current and past history of reported conditions. The core of our technique is a Bayesian hierarchical model for selecting predictive association rules (such as "*condition 1 and condition 2* $\rightarrow$ *condition 3*") from a large set of candidate rules. Because this method "borrows strength" using the conditions of many similar patients, it is able to provide predictions specialized to any given patient, even when little information about the patient's history of conditions is available.

Keywords: Hierarchical Bayesian Modeling, Association Rules, Healthcare, Personalized Medicine

Introduction

The emergence of large-scale medical record databases presents exciting opportunities for data-based personalized medicine. Prediction lies at the heart of personalized medicine and in this paper we propose a statistical model for predicting patient-level sequences of medical conditions. We draw on new approaches for predicting the next event within a "current sequence," given a "sequence database" of past event sequences [10, 11]. Specifically we propose the Hierarchical Association Rule Model (HARM) that generates a set of *association rules* such as *dyspepsia and epigastric pain*$\rightarrow$ *heartburn*, indicating that dyspepsia and epigastric pain are commonly followed by heartburn. HARM produces a ranked list of these association rules. Built-in explanations represent a particular advantage of the association rule framework—the rule predicts heartburn *because* the patient has had dyspepsia and epigastric pain.

In our setup, we assume that each patient visits a healthcare provider periodically. At each encounter, the provider records time-stamped medical conditions experienced since the previous encounter. In this context, we address prediction problems such as: (i) given data from a sequence of past encounters, predict the next condition that a patient will report; (ii) given basic demographic information, predict the first condition that a patient will report; (iii) given partial data from an encounter (and possibly prior encounters) predict the next condition.

Though medical databases often contain records from thousands or even millions of patients, most patients experience only a handful of the massive set of potential conditions. This patient-level sparsity presents a challenge for predictive modeling. Our hierarchical modeling approach addresses this challenge by borrowing strength across patients.

This work extends that of Rudin et al. [11] who formalized sequential event prediction and used a non-hierarchical association rule algorithm. Dumouchel and Pregibon [6] also presented a Bayesian analysis of interesting item sets (which could be extended to rules); their method did not address sequential data.

Generally speaking, HARM yields a prediction algorithm for sequential data that can be used for a variety of additional applications. The algorithm can be used as a recommender system (*e.g.*, for vendors such as Netflix or amazon.com). It can be used to predict the next move in a video game in order to design a more interesting game, or it can be used to predict the winners at each round of a tournament (*e.g.*, the winners of games in a football season). These applications possess the same basic structure as the condition prediction problem: a database consisting of sequences of events, where each event is associated to an individual entity (medical patient, customer, football team). As future events unfold in a new sequence, we use the set of past events to predict the next event.

Association rule mining and Bayesian inference

This work presents a new approach to association rule mining by determining the "interestingness" of rules using a particular (hierarchical) Bayesian estimate of the probability of exhibiting condition b, given a set of current conditions, a. We will first discuss association rule mining and its connection to Bayesian shrinkage estimators. Then we will present our hierarchical method for providing personalized condition predictions.

An *association rule* in our context is an implication $a \to b$ where the left side is a subset of conditions that the patient has experienced, and b is a single condition that the patient has not yet experienced since the last encounter. Ultimately, we would like to rank rules in terms of "interestingness" or relevance for a particular patient at a given time.

Two common determining factors of the "interestingness" of a rule are the "confidence" and "support" of the rule [2, 9]. The confidence of the rule $a \to b$ for a patient is defined as the number of times conditions a and b were experienced divided by the number of times conditions a were experienced, or equivalently $\hat{P}(b|a)$. The support of set a is the number of times conditions a were experienced, and is proportional to $\hat{P}(a)$. When a patient has experienced a particular set of conditions only a few times, a new single observation can dramatically alter the confidence $\hat{P}(b|a)$ for many rules. This problem occurs commonly in our clinical trial data, where most patients have reported fewer than 10 total conditions. The vast majority of rule mining algorithms address this issue with a minimum support threshold to exclude rare rules, and the remaining rules are evaluated for interestingness (see reviews by [12, 7]). However, using a minimum support threshold excludes "nuggets," rules with low support but high confidence. This is problematic, for instance, when a condition that occurs rarely is strongly linked with another rare condition; it is essential not to exclude the rules characterizing these conditions. In our data, the distribution of conditions has a long tail, where the vast majority of events happen rarely: out of 1800 possible conditions, 1400 occur less than 10 times. These 1400 conditions are precisely the ones in danger of being excluded by a minimum support threshold.

Our work avoids problems with the minimum support threshold by ranking rules with a shrinkage estimator of $P(b|a)$. These estimators directly incorporate the support of the rule. One example of such an estimator is the "adjusted confidence" [10, 11]:

$$\text{AdjConf}(a \to b, K) := \frac{\text{Number of times conditions } a \text{ and } b \text{ were experienced}}{\text{Number of times conditions } a \text{ were experienced} + K} = \frac{\#(a\&b)}{\#a + K}.$$

The effect of the penalty term K is to pull low-support rules towards the bottom of the list; any rule achieving a high adjusted confidence must overcome this pull through either a high enough support or a high confidence. The adjusted confidence has several nice properties as a measure of "interestingness": first, interestingness is closely related to the conditional probability $P(b|a)$; second, among rules with equal confidence the higher support rules are preferred; third, there is no strict minimum support threshold.

In this work, we extend the adjusted confidence model in an important respect, in that our method shares information across similar patients to better estimate the conditional probabilities. The adjusted confidence is a particular Bayesian estimate of the confidence. Assuming a Beta prior distribution for the confidence, the posterior mean is:

$$\tilde{P}(b|a) := \frac{\alpha + \#(a\&b)}{\alpha + \beta + \#a},$$

where $\#x$ is the support of condition x, and α and β denote the parameters of the (conjugate) Beta prior distribution. Our model allows the parameters of the Binomial to be chosen differently for each patient and also for each rule. This

means that our model can determine, for instance, whether a particular patient is more likely to repeat a condition that has occurred only once, and also whether a particular condition is more likely to repeat than another.

Hierarchical Association Rule Model (HARM)

For a patient i and a given rule, r, say we observe y_{ir} co-occurrences (support for lhs & rhs), where there were a total of n_{ir} encounters that include the lhs (n_{ir} is the support for lhs). We model the number of co-occurrences as Binomial(n_{ir}, p_{ir}) and then model p_{ir} hierarchically to share information across groups of similar individuals.

Define $\mathbf{M}$ as a $I \times D$ matrix of static observable characteristics for a total of I individuals and D observable characteristics, where we assume $D > 1$ (otherwise we revert back to a model with a rule-wise adjustment). Each row of $\mathbf{M}$ corresponds to a patient and each column to a particular characteristic. We define the columns of $\mathbf{M}$ to be indicators of particular patient categories (gender, or age between 30 and 40, for example), though they could be continuous in other applications. Let $\mathbf{M}_i$ denote the ith row of the matrix $\mathbf{M}$. We model the probability for the ith individual and the rth rule p_{ir} as coming from a Beta distribution with parameters π_{ir} and τ_i (previously referred to as α andβ). We then define π_{ir} through the regression model $\pi_{ir} = \exp(\mathbf{M}_i'\boldsymbol{\beta}_r + \gamma_i)$ where $\boldsymbol{\beta}_r$ defines a vector of regression coefficients for rule r and γ_i is an individual-specific random effect. More formally, we propose the following model:

$$\begin{aligned} y_{ir} &\sim \text{Binomial}(n_{ir}, p_{ir}) \\ p_{ir} &\sim \text{Beta}(\pi_{ir}, \tau_i) \\ \pi_{ir} &= \exp(\mathbf{M}_i'\boldsymbol{\beta}_r + \gamma_i). \end{aligned}$$

Under this model, $E(p_{ir}|y_{ir}, n_{ir}) = \frac{y_{ir}+\pi_{ir}}{n_{ir}+\pi_{ir}+\tau_i}$, which is a more flexible form of adjusted confidence. This expectation also produces non-zero probabilities for a rule even if n_{ir} is zero (patient i has never reported the conditions on the left hand side of r before). This could allow rules to be ranked more highly even if n_{ir} is zero. The fixed effect regression component, $\mathbf{M}_i'\boldsymbol{\beta}_r$, adjusts π_{ir} based on the patient characteristics in the $\mathbf{M}$ matrix. For example, if the entries of $\mathbf{M}$ represented only gender, then the regression model with intercept $\beta_{r,0}$ would be $\beta_{r,0} + \beta_{r,1}\mathbf{1}_{\text{male}}$ where $\mathbf{1}_{\text{male}}$ is one for male respondents and zero for females. Being male, therefore, has a multiplicative effect of $e^{\beta_{r,1}}$ on π_{ir}. In this example, the $\mathbf{M}_i'\boldsymbol{\beta}_r$ value is the same for all males, encouraging similar individuals to have similar values of π_{ir}. For each rule r, we will use a common prior on all coefficients in $\boldsymbol{\beta}_r$; this imposes a hierarchical structure, and has the effect of regularizing coefficients associated with rare characteristics.

The π_{ir}'s allow rare but important "nuggets" to be recommended. Even across multiple patient encounters, many conditions occur very infrequently. In some cases these conditions may still be highly associated with certain other conditions. For instance, compared to some conditions, migraines are relatively rare. Patients who have migraines however typically also experience nausea. A minimum support threshold algorithm might easily exclude the rule "migraines $\rightarrow$ nausea" if a patient hasn't experienced many migraines in the past. This is especially likely for patients who have few encounters. In our model, the π_{ir} term balances the regularization imposed by τ_i to, for certain individuals, increase the ranking of rules with high confidence but low support. The τ_i term reduces the probability associated with rules that have appeared few times in the data (low support), with the same effect as the penalty term (K) in the adjusted confidence. Unlike the cross-validation or heuristic strategies suggested in [10, 11], we estimate τ_i as part of an underlying statistical model. Within a given rule, we assume τ_i for every individual comes from the same distribution. This imposes additional structure across individuals, increasing stability for individuals with few observations.

To rank association rules for the purpose of prediction, however, we need a single estimate for each probability (rather than a full distribution), which we chose as the posterior mean. We carry out our computations using a Gibbs sampling algorithm described in [8]. In [8] we also produce an approximate updating scheme for allowing the algorithm to compute a sequence of predictions in real-time.

Predictive performance in repeated patient encounters

We present results of HARM, with the approximate updating scheme, on co-prescribing data from a large clinical trial. In the trial, each patient visits a healthcare provider periodically. At each encounter, the provider records time-stamped medical conditions (represented by MedDRA terms) experienced since the previous encounter. Thus, each encounter is associated with a sequence of medical conditions. These data are from around 42,000 patient encounters from

about 2,300 patients, all at least 40 years old. The matrix of observable characteristics encodes the basic demographic information: gender, age group (40-49, etc.), and ethnicity. For each patient we have a record of each medication prescribed and the condition/chief complaint (back pain, asthma, etc) that warranted the prescription. For patients with pre-existing conditions, we include these conditions in the patient's list of conditions at each encounter. If a patient reports the same condition more than once during the same thirty day period we only consider the first occurrence of the condition at the first report. If the patient reports the condition once and then again more than thirty days later, we consider this two separate incidents.

As covariates, we used age, gender, race and drug/placebo (an indicator of whether the patient was in the treatment or control group for the clinical trial). We fit age using a series of indicator variables corresponding to four groups (40-49, 50-59, 60-69, 70+). We included all available covariates in our simulation studies. In practice, model selection will likely be essential to select the best subset of covariates for predictive performance. Our experiments consider rules where the left hand side contains either 0 items or 1 item. In other words, we predict using estimates of the marginal probabilities (the support of the item) and probabilities conditional on one previous condition.

We selected a sample of patients by assigning each patient a random draw from a Bernoulli distribution with success probability selected to give a sample of patients on average around 200. For each patient we drew uniformly an integer t_i between 0 and the number of encounters for that patient. We ordered the encounters chronologically and used encounters 1 through t_i as our training set and the remaining encounters as the test set. Through this approach, the training set encompasses the complete set of encounters for some patients ("fully observed"), includes no encounters for others ("new patients"), and a partial encounter history of the majority of the test patients ("partially-observed patients"). We believe this to be a reasonable approximation of the context where this type of method would be applied, with some patients having already been observed several times and other new patients entering the system for the first time. We evaluated HARM's predictive performance using a combination of common and rare conditions. For each run of the simulation, we used the 25 most popular conditions, then randomly selected an additional 25 conditions for a total of 50.

The algorithm was used to iteratively predict the conditions revealed at each encounter. For each selected patient, starting with their first test encounter, and prior to that encounters' first condition being revealed, the algorithm made a prediction of c possible conditions, where $c = 3$. For patients with no previous encounters, both n_{ir} and y_{ir} are zero. For these patients, we use the posterior means of the parameters estimated from the training set to make the first prediction. The algorithm earned one point if it recommended the current condition before it was revealed, and no points otherwise. Then, y_{ir} and n_{ir} were updated to include the revealed condition. This process was repeated for the patient's remaining conditions in the first encounter, and repeated for each condition within each subsequent encounter. We then moved to the next patient and repeated the procedure.

The total score of the algorithm for a given patient was computed as the total number of points earned for that patient divided by the total number of conditions experienced by the patient. The total score is the average proportion of correct predictions per patient. We repeated this entire process (beginning with selecting patients) 500 times. We compared the performance of HARM (using the same scoring system) against an algorithm that ranks rules by adjusted confidence, for several values of K. We also compared with the "max confidence minimum support threshold" algorithm for different values of the support threshold θ, where rules with support below θ are excluded and the remaining rules are ranked by confidence. Neither of these algorithms utilize information across patients.

Figure 1 shows the results, as boxplots of the distribution of scores for the entire collection of partially-observed, fully observed, and new patients. Paired t-tests comparing the mean proportion of correct predictions from HARM to each of the alternatives had p-values for a significant difference in our favor less than 10^{-15}. In other words, HARM has statistically superior performance over all K and θ; *i.e.*, better performance than either of the two algorithms even if their parameters K and θ had been tuned to the best possible value. For all four values of K for the adjusted confidence, performance was slightly better than for the plain confidence ($K = 0$). The "max confidence minimum support threshold" algorithm (which is a standard approach to association rule mining problems) performed poorly for minimum support thresholds of 2 and 3. This poor performance is likely due to the sparse information we have for each patient. Setting a minimum support threshold as low as even two eliminates many potential candidate rules.

The main advantage of our model is that it shares information across patients in the training set. This means that in early stages where the observed y_{ir} and n_{ir} are small, it may still be possible to obtain reasonably accurate probability estimates, since when patients are new, our recommendations depend heavily on the behavior of previously observed similar patients. This advantage is shown explicitly through Figures 1(b) and 1(c), which pertain to partially-observed

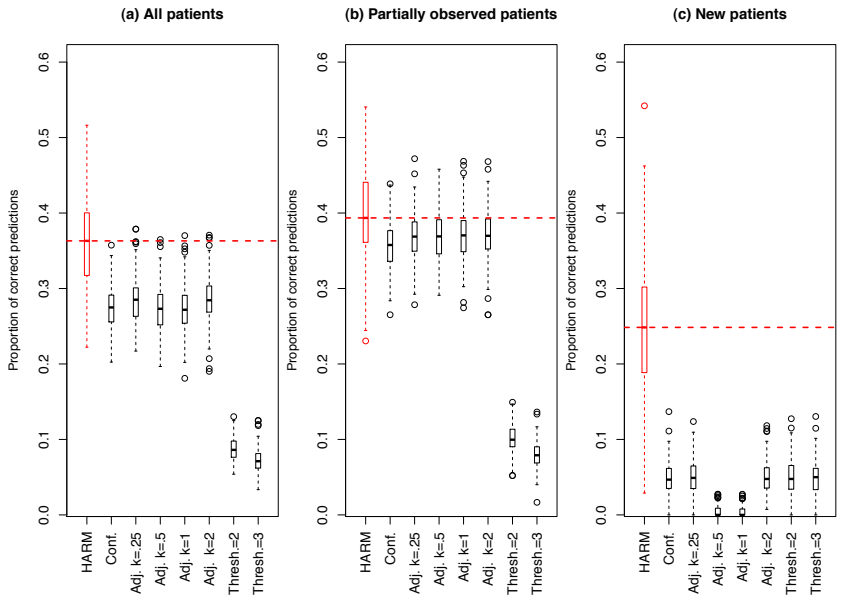

Figure 1: Predictive performance for (a) all patients, (b) partially-observed patients, (c) new patients. Each boxplot represents the distribution of scores over 500 runs. For (a), each run's score (an individual point on a boxplot) is based on a sample of approximately 200 patients. For (b) and (c), each point is based on a subset of these ~ 200 patients.

and new patients, respectively. The advantage of HARM over the other methods is more pronounced for new patients: in cases where there are no data for each patient, there is a large advantage to sharing information.

Related Works

Four relevant works on Bayesian hierarchical modeling and recommender systems are those of DuMouchel and Pregibon [6], Breese, Heckerman, and Kadie [3], Condliff et al. [4] and Agarwal, Zhang, and Mazumder [1]. DuMouchel and Pregibon deal with the identification of interesting itemsets (rather than identification of rules). Specifically, they model the ratio of observed itemset frequencies to baseline frequencies computed under a particular model for independence. Breese, Heckerman, and Kadie use Bayesian methods to cluster users in a recommender systems application, and also suggests a Bayesian network. Condliff et al. present a hierarchical Bayesian approach to collaborative filtering that "borrows strength" across users. Neither Condliff et al. nor Breese, Heckerman, and Kadie aim to model repeat purchases (recurring conditions). Agarwal, Zhang, and Mazumder also build a personalized recommender system that models item-item similarities. Their model uses logistic regression for estimating p_{ir} rather than using π_{ir} and τ_i. This has the advantage of being a simpler model, but loses the interpretability our model has through using association rules. It also loses the potential advantage of estimating only conditional probabilities involving few variables.

As far as we know, the line of work by Davis, Chawla, Christakis, and Barabasi [5] is the first to use an approach from recommender systems to predict medical conditions, though in a completely different way than ours; it is based on vector similarity, in the same way as Breese, Heckerman, and Kadie. (Also see references in [5] for background on collaborative filtering.)

Conclusion and Future Work

We have presented a hierarchical model for ranking association rules for sequential event prediction. The sequential nature of the data is captured through rules that are sensitive to time order, that is, $a \rightarrow b$ indicates conditions a are

followed by conditions b. HARM uses information from observably similar individuals to augment the (often sparse) data on a particular individual; this is how HARM is able to estimate probabilities $P(b|a)$ before conditions a have ever been reported. In the absence of data, hierarchical modeling provides structure. As more data become available, the influence of the modeling choices fade as greater weight is placed on the data. The sequential prediction approach is especially well suited to medical condition prediction, where experiencing two conditions in succession may have different clinical implications than experiencing either condition in isolation.

Model selection is important for using our method in practice. There are two types of model selection required for HARM: the choice of covariates encoded by the matrix $\mathbf{M}$, and the collection of available rules. For the choice of covariates in $\mathbf{M}$, standard feature selection methods can be used, for instance, a forward stagewise procedure where one covariate at a time is added as performance improves, or a backward stagewise method where features are iteratively removed. To perform model selection on the choice of rules, it is possible to construct analogous "rule selection" methods as one might use for a set of covariates. A forward stagewise procedure could be constructed, where the set of rules is gradually expanded as prediction performance increases.

Acknowledgements: This work was partially supported by a Google PhD fellowship in statistics. We would like to acknowledge support for this project from the National Science Foundation under grant IIS-1053407.

References

[1] Agarwal, D., Zhang, L. and Mazumder, R. Modeling item-item similarities for personalized recommendations on Yahoo! front page. *Annals of Applied Statistics*, 2011. Forthcoming.

[2] Agrawal, R., Imieliński, T. and Swami, A. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

[3] Breese, J., Heckerman, D., and Kadie, C. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty and Artificial Intelligence*, pages 43–52, 1998.

[4] Condliff, M., Lewis, D., Madigan, D., and Posse, C. Bayesian mixed-effects models for recommender systems. In *Proceedings of the ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation*, 1999.

[5] Davis, D., Chawla, N., Christakis, N., and Barabasi, A. Time to CARE: a collaborative engine for practical disease prediction. *Data Mining and Knowledge Discovery*, 20(3):388–415, May 2010.

[6] DuMouchel, W. and Pregibon, D. Empirical Bayes screening for multi-item associations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 67–76, 2001.

[7] Geng, L. and Hamilton, H. Choosing the right lens: Finding what is interesting in data mining. In *Quality Measures in Data Mining*, pages 3–24. Springer, 2007.

[8] McCormick, T., Rudin, C. and Madigan, D. Bayesian hierarchical rule modeling for predicting medical conditions. *SSRN eLibrary, number 1736062*, 2011.

[9] Piatetsky-Shapiro, G. Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. J. Frawley, editors, pages 229–248. AAAI Press, 1991.

[10] Rudin, C., Letham, B., Kogan, E. and Madigan, D. A learning theory framework for association rules and sequential events. *SSRN eLibrary, number 1868446*, 2011.

[11] Rudin, C., Letham, B., Salleb-Aouissi, A., Kogan, E. and Madigan, D. Sequential event prediction with association rules. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011.

[12] Tan, P., Kumar, V. and Srivastava, J. Selecting the right interestingness measure for association patterns. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.