Proceedings of the 7th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2012)
H. Yang, D. Zeng, O. E. Kundakcioglu, eds.

Application of Bayesian Co-clustering Methods in Healthcare Fraud Detection

Tahir Ekin
Department of Computer Information Systems and Quantitative Methods
Texas State University
San Marcos, TX
t_e18@txstate.edu

Refik Soyer
Department of Decision Sciences
The George Washington University
Washington, DC

Toros Caglar
United Healthcare Group
Irvine, CA

Abstract

We investigate the utilization of Bayesian co-clustering methods to identify fraud in healthcare insurance systems. Bayesian co-clustering (BCC) algorithm is employed in order to detect the visiting patterns among healthcare users and providers. It involves assignment of each provider-user pair to a co-cluster while allowing mixed membership for any provider and user. Upon completion of this analysis, we have the likelihood of belonging to a given co-cluster for each provider and user. The providers who are found to have a high probability of sharing users with an unusual set of providers are flagged to be investigated for potential fraudulent activity.

Keywords: Bayesian co-clustering, Healthcare fraud detection, Unsupervised data mining, Gibbs sampling

Introduction

The size of the expenditures and insufficient tracking have made governmental health care insurance sector an attractive area for fraudulent activities. In the United States, medical fraud and abuse have resulted in significant additional costs, estimated to between 60 and 200 billion dollars annually [1]. In addition to the financial loss, medical fraud decreases the overall effectiveness of the system in providing quality care to deserving patients. Therefore, in recent years efforts to detect fraud such as detailed audits of the healthcare claims have gained importance. However, audits which are generally done by domain experts have not been very useful in detecting not obvious fraudulent activity within a huge set of claims. That is the reason that statistical methods have been applied to investigate healthcare fraud. These methods include neural networks, decision trees, Bayesian networks and genetic algorithms [2]. Our paper investigates the utilization of another statistical method, Bayesian co-clustering, in identifying billing patterns involving healthcare users and providers.

Methodology and Data

Data mining methods can be classified in two main categories in their application to healthcare fraud detection: supervised and unsupervised methods. Supervised methods include but are not limited to Bayesian networks, decision trees, genetic algorithms, neural networks, machine learning and fuzzy methods [2]. These require fraudulent claims that are pre-labelled by domain experts, so most of these methods are tested with a specific data set. Therefore their use may not be extended to other data sets since claims data can be dynamic and heterogeneous [3]. Unsupervised methods do not have this requirement and their objective is to find the outliers and detect patterns in the data.
According to the survey paper [2], application of unsupervised methods in healthcare fraud detection is limited and there is a gap in the literature for adaptive fraud detection in modeling the dynamic patterns of the data. In this paper, we introduce the use of Bayesian co-clustering techniques to address this need. We investigate if we can detect patterns within provider-user pairs that may signal irregular, potential fraudulent activity. Particularly, we focus on the modeling of dyadic data for which the use of data mining approaches have increased in recent years. For example, users' rates for movies in recommendation systems can be utilized to figure out which users and movies can be grouped together and lead to recommendations for users. Co-clustering methods have emerged as one of the most powerful tools to model these type of problems. Traditional co-clustering allows individual rows and columns of a data matrix to belong to only one cluster, which has been remedied by the Bayesian co-clustering (BCC) approach [4]. Our study is one of the first studies that BCC approach is applied in healthcare context.

In our application, data is not disclosed since it contains private health information. Depending on the number of users and providers obtained in the data preparation step, we may choose to limit the dataset by concentrating on the users and providers that are more significantly utilized than others. We have the claim information and transform it into a workable form, so-called the visitation matrix. It is simply a binary matrix composed of a collection of indicators that specify whether the users have claims associated with the providers of interest. For example, in a matrix cell that corresponds to provider P and user U, we have a 1 if U has at least one claim that has P as the rendering provider. Otherwise, the cell will be a 0. After generating the visitation matrix for the users and the providers, we employ the Bayesian co-clustering algorithm to investigate the behavioral patterns in the data set and group providers that render services to similar users. The clustering is performed purely based on visiting behaviours.

This problem can also be viewed as a generative mixture modeling problem. It is assumed each row and column to have a mixed membership respectively, from which we generate row and column clusters. Each entry of the data matrix is then generated given the co-cluster. BCC maintains separate Dirichlet priors for rows and columns over the mixed membership and assumes each observation to be generated by an exponential family distribution corresponding to its row and column clusters. Upon completion of the grouping analysis, for each provider and user, we have the likelihood of belonging to any given group. We obtain the so called service probability matrix, which indicates if any groups of users are serviced more often by the members of any particular provider groups. For inference and parameter estimation, a Markov Chain Monte Carlo method, Gibbs algorithm is proposed.

Model

In this section, we introduce the parameters and the model in detail. This co-clustering model assigns each provider-user pair into a co-cluster and investigates if there are unusual pairs to be flagged in these co-clusters.

Notation:

- K: number of clusters for providers
- L: number of clusters for users
- I: number of distinct providers
- J: number of distinct users
- $X_{i,j}$: Data which consists of binary entries; if the i^{th} provider provides service to j^{th} user it is 1, otherwise it is 0
- $Z_{1ijk} = 1$ if provider i belongs to cluster k while serving the j^{th} beneficiary; $Z_{1ijk} = 0$ otherwise
  $Z_{2ijl} = 1$ if beneficiary j belongs to cluster l while being served by the i^{th} provider; $Z_{2ijl} = 0$ otherwise
- Latent variable, $Z_{ijkl} = \{Z_{1ijk}, Z_{2ijl}\}$: denotes membership to the row (provider) and column (user) clusters and constructs the co-cluster (k, l) for the i^{th} provider and j^{th} user.

- $\boldsymbol{\pi_1}$: membership probability vector for providers, with size 1 by K
- $\boldsymbol{\pi_2}$: membership probability vector for beneficiaries, with size 1 by L
- $\boldsymbol{\theta_{kl}}$: service probability parameter, probability of $X_{ij} = 1$ given that $Z_{1ijk} = 1$ and $Z_{2ijl} = 1$
- $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$: fixed hyper-parameters for distributions

Co-clustering model:

The co-clustering algorithm involves assignment of each X_{ij} to a co-cluster Z_{ijkl} defined by the latent pair (Z_{1ik}, Z_{2jl}). Gibbs sampling is proposed to sample from the joint distribution, $p(X, Z_1, Z_2, \pi_1, \pi_2, \theta|\alpha_1, \alpha_2)$. Draws from the full conditional distributions in convergence are used as samples from the joint distribution. To sample from the full conditionals, we use two main models; beta-bernoulli model and multinomial-dirichlet model.

A beta-bernoulli model is used to sample from $p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Z_1}, \boldsymbol{Z_2}, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. $\theta_{k,l}$ is assumed to have a beta prior with fixed hyper-parameters α_{kl} and β_{kl} :

$$\theta_{kl} \sim Beta(\alpha_{kl}, \beta_{kl})$$

$$X_{ij}|Z_{1ijk}, Z_{2ijl}, \theta_{kl} \sim Ber(\theta_{kl})$$

The full conditional distribution of θ_{kl} can be identified within this conjugate framework:

$$\theta_{kl}|X_{ij}, \boldsymbol{Z_1}, \boldsymbol{Z_2} \sim Beta(\alpha_{kl} + \sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}\mathbb{1}(Z_{1ijk} = 1, Z_{2ijl} = 1),$$
$$\beta_{kl} + \sum_{i=1}^{I}\sum_{j=1}^{J}(1 - X_{ij})\mathbb{1}(Z_{1ijk} = 1, Z_{2ijl} = 1))$$

where $\mathbb{1}$ is the indicator function.

A multinomial-dirichlet model is used to sample from $p(\boldsymbol{Z_1}, \boldsymbol{Z_2}, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}|\boldsymbol{X}, \boldsymbol{\theta}, \alpha_1, \alpha_2)$. The full conditional distributions of $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_2}$ can be identified within the described conjugate framework. Membership probabilities are assumed to have dirichlet priors:

$$\boldsymbol{\pi_1} \sim Dir(\alpha_{1k}; k = 1, .., K)$$

$$\boldsymbol{\pi_2} \sim Dir(\alpha_{2l}; l = 1, .., L)$$

Latent variables $\mathbf{Z_1}$, $\mathbf{Z_2}$ can be sampled from

$$\pi(Z_{1ijk} = 1, Z_{2ijl} = 1|\boldsymbol{\theta}, \pi_1, \pi_2, (\boldsymbol{Z})^{-i,j}) \propto (\theta_{kl})^{X_{ij}}(1 - \theta_{kl})^{1-X_{ij}}\pi_{1k}\pi_{2l}$$

where $(\boldsymbol{Z})^{-i,j}$ denotes all Z other than $\boldsymbol{Z_{1ij}}$ and $\boldsymbol{Z_{2ij}}$.

The full conditional distributions of the membership probabilities are:

$$\boldsymbol{\pi_1}|\mathbf{Z_1} \sim \mathbf{Dir}(\alpha_{\mathbf{1k}} + \{\sum_{\mathbf{i=1}}^{\mathbf{I}}\sum_{\mathbf{i=1}}^{\mathbf{J}} \mathbb{1}(\mathbf{Z_{1ijk}} = \mathbf{1})\})$$

$$\boldsymbol{\pi_2}|\mathbf{Z_2} \sim \mathbf{Dir}(\alpha_{\mathbf{2k}} + \{\sum_{\mathbf{i=1}}^{\mathbf{I}}\sum_{\mathbf{i=1}}^{\mathbf{J}} \mathbb{1}(\mathbf{Z_{2ijl}} = \mathbf{1})\})$$

In convergence, this Gibbs sampling based Markov Chain Monte Carlo algorithm provides the samples drawn from the joint distribution of these parameters and variables. Therefore, we can make inference with probabilistic statements indicating the potential of fraud for a particular claim.

Conclusion

This grouping method allows to determine the potential patterns among providers and users. As we have the likelihood of belonging to a given co-cluster for each provider and user, we can flag the providers who are found to be have a high probability of sharing users with an unusual set of providers. This model can be used as a tool for detection of fraud as it forms meaningful profiles for peer groups. For instance, a pediatrician may be flagged for billing similar users with cardiologists as the user profiles for these physicians are expected to be very different. Our model is an attempt for adaptive fraud detection which helps to deal with the potential heterogeneity of the data.

There are possible extensions to improve the power of this model. Covariates can be used within a logit function to model the membership probabilities. A dynamic cluster membership over time can also be considered by using a Markov modulated co-clustering model. Better handling of sparse matrices and more efficient parameter estimation can make this model more efficient. The main limitation of the utilization of Bayesian co-clustering methods in healthcare fraud detection seems to be increasing computational time in the case of many providers and users. Overall, the use of Bayesian co-clustering methods is promising in the area of healthcare fraud detection.

References

[1] Center for Medicaid and Medicare Services, 2010, "2010 Financial Report".

[2] Li, J., Huang, K-Y., Jin, J. and Shi, J., 2008, "A survey on statistical methods for health care fraud detection," Health Care Management Science, 11: 275-287.

[3] Musal, R., 2010, "Two models to investigate Medicare fraud within unsupervised databases," Expert Systems with Applications, 37 (12): 8628-8633.

[4] Shan, H. and Banerjee, A., 1998, "Bayesian co-clustering," Proceedings of IEEE ICDM, 530-539.