Literature Review

Huanye Liu

Identifying fraud and abuse in healthcare claims, doctor's over-prescription of drugs in particular, previously relied heavily on specialized knowledge and forensic skills of human expertise, and thus was normally considered to be a human intensive job. However, the ever-growing number of healthcare claims now calls for the development of a suitable computer-aided methodology for fraud and abuse detection in the healthcare system, which requires extracting domain knowledge from data using computer-aided analysis to replace human expertise from professionals. The implementation of a computer-aided methodology is supposed to be based on various factors related to medical diagnoses, considering all kinds of procedures and treatment protocols and the subtleties of the prescribed medications. All included, these factors would form a high dimensional set of predictors, so we find an appropriate computational model, the Bayesian network, to store and process the claims data and to capture the complex relationship among all these factors for further analysis in an efficient way.

The extensive use of the probabilistic inference methods such as the probabilistic expert systems [1], causal models [2], and Bayesian networks [3] in many other areas of healthcare also motivates us to adopt the Bayesian network to analyze the healthcare claims data. These areas include patient care management [4,5], gene regulation networks [6], diagnosis systems [7], disease and infection [8]and bioinformatics and computational biology [9].

The basic classification of healthcare fraudulent behaviors is described in [10], which also briefly explains why we should use statistical methods for detecting fraud and abuse in various scenarios such as the phantom or duplicate billing, identity theft, prescription forgery, fictitious or deceased beneficiaries, bill padding, and prescription forgery [11].But the method proposed in this paper for fraud and abuse detection in prescription claims data is to identify doctors who are associated with abnormal and excessive prescriptions for specific drug. We do not attempt to develop statistical models for the underlying mechanism of any of these abnormal and excessive prescriptions here; we instead identify the deviations from normative or baseline behavior, considering all factors which characterize the interaction between the patient and the prescriber, and then aggregates over the set of claims for each doctor.

Fraud detection in the healthcare system can be carried out in either the off-line or online modes of analysis [12]. The off-line mode can be used by audit investigators to retrospectively review claims data for identification of fraud or abuse for further investigation. The online mode instead puts emphasis on early detection of a potential fraudulence behavior in order to take actions as early as possible to prevent further abuse or fraud, and thus can help to eliminate unnecessary waste in the healthcare system in time.

The approach we take here has the same nature as other previous works [12–14] on statistical methods for detection of fraud and abuse in various domains such as telecommunications, financial trading, network intrusion, health care and credit card transactions. However, our approach is more of an unsupervised statistical method which does not require labeling fraudulent claims explicitly. In contrast, the supervised or semi-supervised statistical methods may directly model the fraud outcomes in terms of all other related predictors. Due to the ever-growing number of healthcare claims and rapidly-changing nature of fraud and abuse incidences, the labeling of fraudulent claims becomes more and more difficult, which justifies the approach we propose here.

Two previous works [12, 15]also do not require any explicit labeling of fraudulent claims in the analysis, and here we can compare each of them with the method proposed in this paper. For example, [12] used both the off-line and online application modes for prescription fraud detection. For each drug, the previous claims data were used to obtain pairwise occurrence frequencies in the individual prescriptions combined with other factors including age, gender, medical diagnosis, or other co-prescribed medications. For each of these co-occurrence dimensions , the likelihood of fraud was associated with the particular medication in a given prescription claim. This likelihood was assigned a very high value when compared with appropriate reference frequencies while the other factors in the prescription claim of interest small values of the relevant occurrence frequencies. Based on domain expertise, some appropriate thresholds were set for each likelihood score, and claims with likelihood scores above these thresholds were classified as fraudulent claim.

Our approach, compared with the approach in [12], is to identify doctors who exhibit anomalous prescription behavior in the offline mode based on multiple claims instances which can be used to model the normal or expected prescription behavior, and this baseline model for normal behavior should capture the more complex and high-dimensional interactions between healthcare providers and patients recorded in claims data. For example, the medical history of a patient including patient medication profiles which are collected from the anonymized claims data, can play the role of training data for the normalizations of prescription behavior and predicting the expectation of prescription behavior, and this kind of training data should also include prescriber attributes such as their practice specialties, their profiles and pattern of diagnoses and procedures. The reason why we need a relative complex behavioral models for the analysis of sparse, high-dimensional data has been mentioned in the first paragraph with several examples. Finally, the big difference between [12] and our work is that for the claims data set used here, linking individual prescriptions to a corresponding single-valued patient diagnosis is impossible because prescription claims are issued by the pharmacy and contain the medication information, while medical claims which record the patient diagnosis are issued by the prescriber. These two sets of claims could even be issued by different insurance programs. The correspondence between the prescription and medical claims data requires accurately linking patient information between the two sets of claims, and even if they could be accurately linked, a perfect direct correspondence between these two different claims data could be hard to find in a model. In other words, there exists some latent variables which have no direct connections among them, therefore we need to use some model smoothing skills for the Bayesian network model learning and inference, and the backing off to a model without latent variables is one frequently used smoothing skill. As paper [17]and paper [18] show, a simple example using a back-off model as in n-gram language models is to fill in one portion of the probability space of the trigram with bigram and unigram probability estimates which are more frequently seen. This smoothing could be realized by backing off or by interpolating against the weaker models. Good smoothing would greatly improve the final estimated results.

The paper [15] is mainly concerned with expense auditing for corporate travel and entertainment, and we adopt very similar approach and methodology proposed in [15]. For example, [15] examined expense claims submitted by employees in various focus areas including ground transportation, restaurant tip, etc., and evaluated the expense claims of different entities such as individuals or entire departments in these areas. A normalized baseline model for each focus area was defined, and for each entity, any abnormal behavior was detected if there were significant departures from the expected behavior. This approach assumes that any abnormal behavior only takes a small fraction of the overall set of expense claims so that normalized baseline models can be reliably estimated. [15] used the likelihood ratio score to identify the entities with abnormal behavior, which can be computed as the likelihood ratio of the actual behavior over the set of all related variables for each entity to the predictions of the normalized baseline model for this same set of related variables. [15] also used Monte Carlo methods, which are similar to those used in scan statistics [16] to evaluate the relevant p-values of the likelihood ratios. Although the general methods proposed in [15] and in our paper are similar, applying that method to the healthcare prescription fraud is significantly more challenging due to the complexity of the data and the related model. Therefore, we adopt the graphical model, the Bayesian network model in particular, to accommodate the sparse, high- dimensional data.

To summarize, we propose a methodology for the off-line application of fraud detection for prescription claims data. The proposed approach uses Bayesian network to incorporate all available data including diagnosis codes, procedure codes, medication history, and other prescriber and patient attributes which characterize the interaction between the prescriber and the patient. The baseline model that we obtain captures the relationship among all the above attributes under the Bayesian network framework, and any anomalous prescription would be detected based on the corresponding small posterior probability inferred by the the baseline model.

References

[1] D. J. Spiegelhalter, Probabilistic Expert Systems in Medicine: Practical Issues in Handling Uncertainty, Statistical Science, 2(1), pp. 25-30, 1987.

[2] J. Pearl, Causality: Models, Reasoning and Inference, 29, Cambridge: MIT press, 2000.

[3] F. V. Jensen and T. D. Nielsen, Bayesian networks and decision graphs Springer, 2007.

[4] P. J. F. Lucas, L. C. van der Gaag, A. Abu-Hanna, Bayesian networks in biomedicine and health-care, Artificial Intelligence in medicine, 30(3), pp. 201-214, 2004.

[5] M. Verduijn, N. Peek, P. M. J. Rosseel, E. de Jonge, B. A. J. M. de Mol, Prognostic Bayesian Networks: I: Rationale, Learning Procedure, and Clinical Use, Journal of Biomedical Informatics, 40(6), pp. 609-618, 2007.

[6] J. Donkers, K. Tuyls Computational Intelligence in Bioinformatics, Studies in Computational Intelligence, Springer, pp. 75-111, 2008.

[7] A.Meloni, A.Ripoli, V.Positano, L.Landini, Improved Learning of Bayesian Networks in Biomedicine, ISDA '09, pp. 624-628, 2009.

[8] G. F. Cooper, D. H. Dash, J. D. Levander, W. K. Wong, W. R. Hogan and M. M. Wagner, Bayesian biosurveillance of disease outbreaks, Proceedings of the 20th conference on Uncertainty in artificial intelligence, pp. 94-103. AUAI Press, 2004.

[9] D. J. Wilkinson Bayesian methods in bioinformatics and computational systems biology, Briefings in Bioinformatics 8(2), pp. 109-116, 2007.

[10] Travaille P, Muller RM, Thornton D, van Hillegersberg J (2011) Electronic fraud detection in the U.S. medicaid healthcare program: lessons learned from other industries. In: Proceedings of the seventeenth americas conference on information systems.

[11] United States General Accounting Office (2000) Health care fraud. schemes to defraud, medicare, medicaid and private health insurers. GAO/T-OSI-00-15

[12] Aral KD, Guvenir HA, Sabuncuoglu I, Akar AR (2012) A prescription fraud detection model. Comput Methods Prog Biomed 106:37-46

[13] Hand DJ (2010) Fraud detection in telecommunications and banking: discussion of Becker, Volinsky and Wilks (2010) and Sudjianto et al. (2010), 52(1), 34?38

[14] Sudjianto A, Nair S, Yuan M, Zhang A, Kern D, Cela-Diaz F (2010) Stat Methods Fighting Financ Crimes 52(1):5-19

[15] Iyengar V, Boier I, Kelley K, Curatolo R (2007) Analytics for audit and business controls in corporate travel and entertainment. In: Proceedings sixth australasian data mining conference (AusDM 2007) CRPIT, vol 70. Gold Coast, pp 3-12.

[16] Kulldorff M (1997) A spatial scan statistic. Commun Statist Theor Meth 26(6):1481-1496

[17] J. J. Carlson, S. D. Sullivan, L. P. Garrison, P. J. Neu- mann, D. L. Veenstra, Linking payment to health out- comes: A taxonomy and examination of performance- based reimbursement schemes between healthcare payers and manufacturers, Health Policy, 96(3), pp. 179-190, 2010.

[18] S. F. Chen and J. Goodman, An Empirical Study of Smoothing Techniques for Language Modeling, Association for Computational Linguistics, pp. 310-318, 1996.