Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Multivariate Conditional Anomaly Detection and Its Clinical Application

Charmgil Hong and **Milos Hauskrecht**
Computer Science Department
University of Pittsburgh
Pittsburgh, PA 15260

Abstract

This paper overviews the background, goals, past achievements and future directions of our research that aims to build a multivariate conditional anomaly detection framework for the clinical application.

Background and Goals

We humans are prone to error. Despite startling advances in medicine, the occurrence of medical errors remains a persistent and critical problem. Although various computer-aided monitoring devices support medical practices to prevent errors, because those tools are primarily knowledge-based built by clinical experts, they are expensive and their clinical coverage is incomplete.

We develop a new detection framework that identifies statistically anomalous patient care patterns based on past clinical information stored in an electronic health record (EHR) systems. Our hypothesis is that the detection of anomalies in patient care patterns corresponds to identifying cases that need medical attention for reconsideration. Typical anomaly detection methods, however, simply attempt to identify unusual data instances that do not conform with the majority of examples in the dataset, and are not suitable in the clinical context. This is because clinical decisions on patient care are strongly based on the condition of the patient (Hauskrecht et al. 2013). In addition, patient care generally consists of multiple clinical actions which often show correlations between the individual actions (e.g., a set of medications that are usually ordered together). However, such correlations have not been vigorously exploited in the context of anomaly detection. Our framework aims to improve the anomaly detection performance by identifying *multivariate conditional anomalies* where we are interested in the patterns exhibit dependencies among individual clinical actions conditioned on the patient condition.

Approaches

Our approach to identify multivariate conditional anomalies consists of the following two phases: (1) We first build a predictive probabilistic model from EHRs using the *multi-dimensional learning* methods. Then, (2) we apply the model to estimate an *anomaly score* that measures how unlikely a care pattern for the patient is. Below we further describe each of these phases.

Multi-dimensional Modeling of Clinical Data

Multi-dimensional classification (MDC) (Zhang and Zhou 2013) has received much attention in recent years, due to its wide applications. For example, an image can be annotated with multiple tags (Boutell et al. 2004); and a patient may be diagnosed with multiple diseases (Pestian et al. 2007).

We formulate the modeling of EHRs in the MDC framework by assuming each patient is associated with d discrete-valued class variables that represent patient care patterns. The objective is to learn a function that assigns to each patient, represented by its feature vector $\mathbf{x} = \{x_1, ..., x_m\}$, the most probable assignment of the clinical actions $\mathbf{y} = \{y_1, ..., y_d\}$. One approach to this task is to model the conditional joint distribution $P(\mathbf{Y}|\mathbf{X})$. Assuming the 0-1 loss function, the optimal classifier h^* assigns to an instance the maximum a posteriori (MAP) assignment of class variables:

$$h^*(\mathbf{x}) = \arg\max_{\mathbf{y}} P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) \quad (1)$$

$$= \arg\max_{y_1,...,y_d} P(Y_1 = y_1, ..., Y_d = y_d|\mathbf{X} = \mathbf{x}) \quad (2)$$

A challenge in modeling $P(\mathbf{Y}|\mathbf{X})$ is that the number of all possible class assignments is exponential in d. One may tackle the issue by assuming all class variables are conditionally independent of each other and learn d models for each class variable separately (Boutell et al. 2004). However, this approach often fails because it does not take advantage of the multivariate dependencies among the class variables, which is the key to facilitate the learning of MDC.

Our goal in the first phase is therefore to develop efficient multi-dimensional patient care models and methods that overcome the above mentioned difficulties. We start from the decomposition assumption has been introduced with Classifier Chains (Read et al. 2009). The method directly models the class posterior distribution $P(\mathbf{Y}|\mathbf{X})$ by decomposing the relations among class variables using the chain rule:

$$P(Y_1, ..., Y_d|\mathbf{X}) = \prod_{i=1}^{d} P(Y_i|\mathbf{X}, Y_1, ..., Y_{i-1}), \quad (3)$$

where each factor $P(Y_i|\mathbf{X}, Y_1, ..., Y_{i-1})$ in the chain is a classifier that is learned separately by incorporating the predictions of preceding classifiers as additional features.

In (Batal, Hong, and Hauskrecht 2013), we have proposed to restrict the dependency structure to a tree instead of a chain. By having the tree-structure assumption, we presented an efficient structure learning method that finds the optimal dependency relations among class variables, and a linear-time exact MAP inference algorithm based on belief propagation (Koller and Friedman 2009). Later, we extended the tree-structured model and developed statistically sound multi-dimensional ensemble frameworks (Hong, Batal, and Hauskrecht 2014; 2015). Compared to existing multi-dimensional ensemble approaches (Read et al. 2009), our methods learn models from data in more principled ways and produce more accurate and consistent results.

Multivariate Conditional Anomaly Detection

An important advantage of our multi-dimensional modeling approach compared to other MDC methods is that it gives a well-defined model of posterior class probability. That is, our model lets us estimate $P(\mathbf{y}|\mathbf{x})$ for any $(\mathbf{x}, \mathbf{y})$ input-output pair. In addition, by exploiting the decomposable structure of the model (Equation 3), we can easily estimate the likelihood of each individual decision made on a patient $P(y_i|\mathbf{x})$ – which in turn indicates how unlikely the decision is based on the observation.

Based on this probabilistic measure, we use multiple approaches to estimate an anomaly score, which allows rank-ordering of anomaly candidates. In our preliminary study on multivariate conditional anomaly detection, we showed the validity of the approach using a Mahalanobis distance-based anomaly detection method (Rousseeuw and Zomeren 1990) on the posterior class probability $P(\mathbf{y}|\mathbf{x})$ to identify anomalous clinical decisions. We currently investigate on more robust approaches to estimate the anomaly score that well reflects the conditional dependencies among clinical decisions. We also study on how to pinpoint the cause of anomalies to provide more informative feedback.

Experimental Results

To validate our approach and demonstrate its effectiveness, we present experimental results on a clinical dataset obtained from Cincinnati Childrens Hospital Medical Center (Pestian et al. 2007). The dataset has 978 instances; each consists of 1,449 features ($\mathbf{x}$) extracted from clinical progress notes and 45 binary class variables ($\mathbf{y}$) representing the diseases diagnosed. We compared two of our chain variations – *chain.mod1* (Batal, Hong, and Hauskrecht 2013) and *chain.mod2* (Hong and Hauskrecht 2015) – with the binary relevance (*BR*) model (Boutell et al. 2004), which ignores the relationships between individual clinical decisions. We performed 10-fold cross validation with 3 repeats. On each round, we perturbed 15% of test data by randomly flipping 1 to 5 class variables, and see whether the methods can correctly identify the anomalies. The anomaly score is evaluated by the Mahalanobis distance on the posterior class probability $P(\mathbf{y}|\mathbf{x})$.

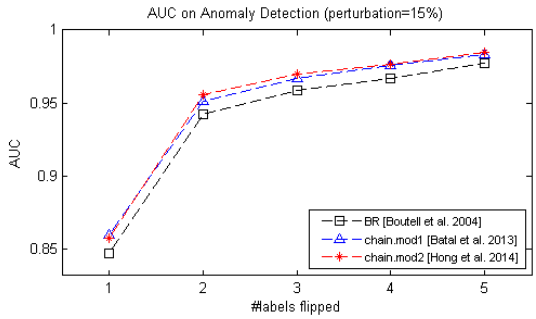

Figure 1: Performance comparison in AUC.

Figure 1 shows the results in terms of the area under receiver operating characteristic (AUC). We can clearly see that the anomaly detection performance has been consistently improved when the dependencies among clinical actions conditioned on patient condition are considered.

References

Batal, I.; Hong, C.; and Hauskrecht, M. 2013. An efficient probabilistic framework for multi-dimensional classification. In *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*, CIKM '13, 2417–2422. New York, NY, USA: ACM.

Boutell, M. R.; Luo, J.; Shen, X.; and Brown, C. M. 2004. Learning multi-label scene classification. *Pattern Recognition* 37(9):1757 – 1771.

Hauskrecht, M.; Batal, I.; Valko, M.; Visweswaran, S.; Cooper, G. F.; and Clermont, G. 2013. Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics* 46(1):47–55.

Hong, C., and Hauskrecht, M. 2015. Detecting multivariate conditional outliers using classifier chains. In *(pending)*.

Hong, C.; Batal, I.; and Hauskrecht, M. 2014. A mixtures-of-trees framework for multi-label classification. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, 211–220. New York, NY, USA: ACM.

Hong, C.; Batal, I.; and Hauskrecht, M. 2015. A generalized mixture framework for multi-label classification. In *(pending)*.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Pestian, J. P.; Brew, C.; Matykiewicz, P.; Hovermale, D. J.; Johnson, N.; Cohen, K. B.; and Duch, W. 2007. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, 97–104. Stroudsburg, PA, USA: Association for Computational Linguistics.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2009. Classifier chains for multi-label classification. In *Machine Learning and Knowledge Discovery in Databases*, volume 5782 of *Lecture Notes in Computer Science*, 254–269. Springer Berlin Heidelberg.

Rousseeuw, P. J., and Zomeren, B. C. v. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85(411):pp. 633–639.

Zhang, M., and Zhou, Z. 2013. A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on* PP(99):1.