Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Multivariate Conditional Outlier Detection and Its Clinical Application

Charmgil Hong and **Milos Hauskrecht**
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
{charmgil, milos}@cs.pitt.edu

Abstract

This paper overviews and discusses our recent work on a multivariate conditional outlier detection framework for clinical applications.

Background

Over the past decades, the quality of healthcare and its improvement have been the center pieces of many public programs and initiatives. Recent studies on patient safety, however, revealed that preventable medical errors are more widespread than initially thought, which are now estimated to be one of the leading causes of death (James 2013). While computer-based tools aimed to prevent the occurrence of errors exist, they require substantial human expert input; hence they are expensive to build and their coverage is small.

One way to alleviate the issue is to use the data mining and machine learning techniques, and identify unusual patient care patterns based on the clinical data stored in electronic medical record (EMR) systems. The motivation is that *the cases requiring medical attention for reconsideration could be identified by detecting statistical outliers in patient care patterns* (Hauskrecht et al. 2007; 2013). Existing data-driven outlier detection methods, however, are not suitable for clinical applications, because they merely attempt to identify unusual data instances that deviate much from the majority in the dataset. On the other hand, clinical data often show the conditional relations that the responses (*i.e.* diagnoses, orders, administrations, *etc.*) are strongly based on the observation (*context*) of the patient. In addition, those responses are usually *correlated* to each other (*e.g.* a set of medications that are usually ordered together). However, such correlations have not been fully explored yet for the purpose of outlier detection.

Our research aims at developing automated methods of *multivariate conditional outlier detection*, which consider both context and correlations in identifying outliers, and applying the methods to support clinical decision making. More specifically, we study *model-based* outlier detection techniques that we represent a data distribution using a statistical model and identify outliers based on the deviation from the model. Accordingly, our research interests are twofold: (1) how to build a multivariate conditional model from data, and (2) how to apply the model to estimate the *outlier scores* of individual data instances. In the following, we highlight the objectives and outcomes of each fold.

Our Approach

Modeling of Multivariate Responses

In the first fold, our key objective is *to accurately and efficiently learn a compact representation of complex clinical records*. For clinical data, this is particularly challenging because each record may contain hundreds to thousands of observations and responses. The conventional approach is to build an oversimplified model, for the sake of computational efficiency, by disregarding any correlations among the responses. However, studies have suggested that those correlations often contain important modeling information, which could improve the accuracy of the model to a great extent (Read et al. 2009; Batal, Hong, and Hauskrecht 2013).

We formulate the modeling of EMRs in the multi-dimensional learning framework (van der Gaag and de Waal 2006; Batal, Hong, and Hauskrecht 2013), which allows us to efficiently find and exploit the correlation structure of data. Namely, we assume that each patient is associated with d discrete-valued variables representing clinical responses. 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 response assignment $\mathbf{y} = \{y_1, ..., y_d\}$. 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 function h^* assigns to an instance the maximum a posteriori (MAP) assignment of the responses.

$$h^*(\mathbf{x}) = \arg\max_{y_1,...,y_d} P(Y_1 = y_1, ..., Y_d = y_d|\mathbf{X} = \mathbf{x}) \quad (1)$$

A challenge in modeling $P(\mathbf{Y}|\mathbf{X})$ is that the number of all possible responses is exponential in d. We develop ways to improve the modeling accuracy without a heavy burden of computational complexity. In (Batal, Hong, and Hauskrecht 2013), we present efficient structure learning and exact inference algorithms by assuming the correlations among response variables form a tree. In (Hong, Batal, and Hauskrecht 2014; 2015), we extend this tree-structured model and develop theoretically sound multi-

dimensional ensemble frameworks. Compared to existing multi-dimensional ensemble approaches (Read et al. 2009), our methods produce more accurate and consistent results.

Detection of Outliers in Clinical Responses

Our interest in the second fold lies in *measuring the degree of outlier for the clinical data, based on the model obtained from the first fold*. An important advantage of our multi-dimensional models is that they give a well-defined estimate of the posterior response probabilities. That is, using the decomposable structure of the multi-dimensional models (Read et al. 2009; Batal, Hong, and Hauskrecht 2013), we can easily estimate not only the posterior joint $P(\mathbf{y}|\mathbf{x})$ but also the posterior of individual responses $P(y_i|\mathbf{x}) : i = 1, ..., d$ for any $(\mathbf{x}, \mathbf{y})$ pair. This is an extremely useful property in many clinical applications, such as diagnosing diseases and allergic drug reactions. We leverage this advantage for solving the clinical outlier detection problem.

In (Hong and Hauskrecht 2015), we formalize the transformation of data from their original space to the d-dimensional conditional probability space of $P(y_i|\mathbf{x})$, and define the methods to estimate the outlier score for each data instance in the new space. Compared to the existing multivariate outlier detection methods, our approach produces more robust results and is able to identify outliers when they are either sparse (manifested in one or only few dimensions) or dense (affecting multiple dimensions).

Evaluation and Discussion

We present experimental results on a clinical dataset obtained from Cincinnati Children's Hospital Medical Center (Pestian et al. 2007). The dataset contains 978 instances; each consists of 1,449 features ($\mathbf{x}$) extracted from clinical progress notes and 45 binary response variables ($\mathbf{y}$) representing the diseases diagnosed. We compare our *Multivariate Conditional Outlier DEtection* method (MCODE) (Hong and Hauskrecht 2015) with two state-of-the-art multivariate outlier detection methods: *Local Outlier Factor* (LOF) (Breunig et al. 2000) and *One-class SVM* (OS) (Amer, Goldstein, and Abdennadher 2013). We performed 10-fold cross validation; on each round, we perturbed 0.5% of the data by randomly flipping 1 to 5 response variables (hence, the outliers represent misdiagnoses), and evaluated how the methods identify the outliers. Figure 1 shows the results in terms of the area under the precision-recall curve (AUCPR). We can verify that our proposed approach consistently outperforms the existing multivariate outlier detection methods.

To conclude, we summarized our research on multivariate conditional outlier detection in the context of clinical application. We described the problems and objectives of the research, and highlighted our model-based outlier detection approach. The wide deployment of the EMR systems and the growing availability of large clinical datasets open up new opportunities for understanding the dynamics of diseases, patient conditions, and efficacy of various treatments. Accordingly, we believe the findings from our research will not only contribute to the quality of healthcare by boosting the utilization of EMRs and providing advanced clinical decision support, but also enable a more comprehensive analysis of complex clinical data and their underlying structures.

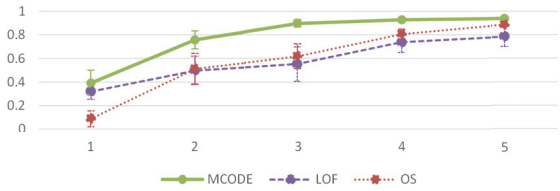

Figure 1: Performance comparison in AUCPR.

Acknowledgments

This work was supported by grant R01GM088224 from the NIH. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Amer, M.; Goldstein, M.; and Abdennadher, S. 2013. Enhancing one-class support vector machines for unsupervised anomaly detection. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*.

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 Information and Knowledge Management*.

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, 93–104. ACM.

Hauskrecht, M.; Valko, M.; Kveton, B.; Visweswaran, S.; and Cooper, G. F. 2007. Evidence-based anomaly detection in clinical domains. In *AMIA Annual Symposium Proceedings*, 319-323.

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. Conditional outlier detection in multivariate responses. In *(pending)*.

Hong, C.; Batal, I.; and Hauskrecht, M. 2014. A mixtures-of-trees framework for multi-label classification. In *Proceedings of the 23nd ACM International Conference on Information and Knowledge Management*, CIKM '14. ACM.

Hong, C.; Batal, I.; and Hauskrecht, M. 2015. A generalized mixture framework for multi-label classification. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM.

James, J. T. 2013. A new, evidence-based estimate of patient harms associated with hospital care. *Journal of patient safety* 9(3):122–128.

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*, 97–104.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2009. Classifier chains for multi-label classification. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer-Verlag.

van der Gaag, L. C., and de Waal, P. R. 2006. Multi-dimensional bayesian network classifiers. In *Probabilistic Graphical Models*.