The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

Multivariate Conditional Outlier Detection: Identifying Unusual Input-Output Associations in Data

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

Abstract

We study *multivariate conditional outlier detection*, a special type of the conditional outlier detection problem, where data instances consist of continuous input (context) and binary output (responses) vectors. We present a novel outlier detection framework that identifies abnormal input-output associations in data using a decomposable conditional probabilistic model. Since the components of this model can vary in their quality, we combine them with the help of weights reflecting their reliability in assessment of outliers. We propose two ways of calculating the component weights: *global* that relies on all data and *local* that relies only on the instances similar to the target instance. Experimental results on data from various domains demonstrate the ability of our framework to successfully identify multivariate conditional outliers.

Introduction

Multivariate conditional outlier detection (MCOD) is an outlier detection problem that analyzes instances in data $\mathcal{D} = \{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}\}_{n=1}^{N}$, where each instance consists of an m-dimensional *continuous* input vector (context attributes) $\mathbf{x}^{(n)} = (x_1^{(n)}, ..., x_m^{(n)})$ and a d-dimensional *binary* output vector (responses attributes) $\mathbf{y}^{(n)} = (y_1^{(n)}, ..., y_d^{(n)})$. Its goal is to precisely identify abnormal response patterns in $\mathbf{Y}$ given context $\mathbf{X}$; *i.e.*, to detect the instances with unusual input-output associations. MCOD fits well various practical outlier detection problems that require contextual understanding of data. For example, recent social media services allow users to tag their content (*e.g.*, online documents, photos, or videos) with keywords and thereby permit keyword-based retrieval. These annotations sometimes include irrelevant tags (entered by mistake) that could be effectively pinpointed if the conditional relations between content and tags are considered. Likewise, evidence-based expert decisions (*e.g.*, functional categorization of genes, medical diagnosis and treatment decisions for patients) occasionally involve errors that could lead to critical failures. Such erroneous decisions would be adequately identified via contextual analysis of evidence-decision pairs.

Despite its importance and usefulness, MCOD has received much less attention in the literature than *unconditional* outlier detection (Chandola, Banerjee, and Kumar 2009; Kriegel, Kröger, and Zimek 2010). Briefly, unconditional outliers are expressed in the joint space of all data attributes and do not consider any context that may help to differentiate the observed data and their unusualness. As a result, the application of unconditional outlier methods to an MCOD problem may lead to incorrect results. Take for example the problem of identification of mistaken image tags in a collection of annotated images. The application of unconditional outlier detection methods to the joint space of both images and tags may return images with rare themes instead of images with mistaken tags (false positives) due to the scarcity of the themes in the dataset. Similarly, unusual annotations on images with frequent themes may not be detected due to the abundance of the similar themes in the dataset (false negative).

The MCOD problem is challenging because both the contextual- and inter-dependences of data instances should be taken into account when identifying outliers. We tackle this by building a probabilistic model of $P(\mathbf{Y}|\mathbf{X})$. The model is learned from all available data, hence summarizing key dependences among data components and their strength. Conditional outliers are then identified with the help of this model: A conditional outlier corresponds to a data instance that is assigned a low probability by the model. We note that the meaning of 'low probability' should not be interpreted in absolute terms, but relative to probabilities associated with other outcomes. For example, the probability of 0.1 for a binary outcome is low relative to its opposite outcome, 0.9. However, if there are 10 possible outcomes and four of these are assigned probability 0.02, 0.1 cannot be considered low.

To convert the above idea into a workable MCOD framework, multiple issues need to be resolved. First, it is unclear how the probabilistic model $P(\mathbf{Y}|\mathbf{X})$ should be represented and parameterized. To address this issue, we use structured probabilistic data models that provide an efficient representation of input-output relations by decomposing the model into a product of univariate probabilistic components. Second, the quality of the probabilistic models trained on finite size data and inaccuracies in probability estimates may negatively affect their outlier detection performance. To overcome this, we propose new outlier scoring methods that combine probability estimates with the help of weights, reflecting their reliability in assessment of outliers. In par-

ticular, we present two ways of calculating the component weights: *global* that relies on all data, and *local* that relies only on the instances similar to the target instance.

Conditional outliers in varied application contexts may manifest themselves differently across the different output dimensions – in some applications, outliers are manifested in one or just a few output dimensions (*e.g.*, mistaken image tags or expert decisions); in others, abnormal output signals may occur across many output dimensions simultaneously (*e.g.*, mass surveillance for disease outbreaks). We experiment with our MCOD approach and demonstrate its usefulness across the different application contexts.

Our Approach[1]

Our approach works by analyzing data instances corresponding to input-output pairs with a statistical model representing the conditional joint distribution $P(\mathbf{Y}|\mathbf{X})$. To build the model we first decompose the conditional joint into a product of conditional univariate distributions using the chain rule of probability: $P(Y_1, ..., Y_d|\mathbf{X}) = \prod_{i=1}^{d} P(Y_i|\mathbf{X}, \boldsymbol{\pi}(Y_i))$, where $\boldsymbol{\pi}(Y_i)$ denotes the parents of Y_i; *i.e.*, all the output variables preceding Y_i (Read et al. 2009). That is, the decomposition lets us represent $P(\mathbf{Y}|\mathbf{X})$ in terms of d univariate conditional factors, $P(Y_i|\mathbf{X}, \boldsymbol{\pi}(Y_i))$, each factor representing one output dimension. Multiple probabilistic models (*e.g.*, logistic regression, naïve Bayes, or support vector machine with probabilistic output (Platt 1999)) can be used to represent these factors and learn them from data. In this paper, we use a logistic regression model to represent each of these factors. This choice of base model allows us to effectively regularize and handle high-dimensional feature spaces, defined by a mixture of continuous and discrete variables (Ng 2004).

Once the model of $P(\mathbf{Y}|\mathbf{X})$ is learned from data, it can be applied to estimate conditional probability for any data instance $<\mathbf{x}, \mathbf{y}>$. Outliers are the instances that have a low probability estimation $\widetilde{P}(\mathbf{y}|\mathbf{x}; \mathcal{M})$, where $\mathcal{M}$ denotes a trained model. For computational convenience and to match the definition of the outlier score (higher score implies stronger outlier), we define our multivariate conditional outlier score as the negative logsum of d univariate probability estimates, one per output dimension:

$$Score_{\text{MCOD}}(\mathbf{y}|\mathbf{x}) = -\log \widetilde{P}(\mathbf{y}|\mathbf{x}; \mathcal{M}) \quad (1)$$

$$= \sum_{i=1}^{d} -\log \widetilde{P}(y_i|\mathbf{x}, \boldsymbol{\pi}(y_i); \mathcal{M}) \quad (2)$$

Decomposable Data Model with Circular Dependences In theory, the decomposed conditional joint in the above MCOD score (Equation 2) should be invariant regardless of the chain order (order of Y_i). Nevertheless, in practice, different chain orders produce different conditional joint distributions as they draw in models learned from different data (Dembczynski, Cheng, and Hüllermeier 2010; Hong, Batal, and Hauskrecht 2015). For this reason, several structure learning methods determining the optimal set of parents have been proposed (Zhang and Zhang 2010; Hong, Batal, and Hauskrecht 2015). However, such methods require at least $O(d^2 t_m)$ of time, where t_m denotes the time of learning a base statistical model (*e.g.*, logistic regression). This may prohibit many MCOD applications whose output dimensionality d is high.

We address the issue by relaxing the chain rule and by permitting *circular dependences* among the output variables. Specifically, we let $\boldsymbol{\pi}(Y_i)$, the parents of Y_i, be all the remaining output variables and approximate Equation 2:

$$Score_{\text{MCOD}}(\mathbf{y}|\mathbf{x}) \simeq \sum_{i=1}^{d} -\log \widetilde{P}(y_i|\mathbf{x}, \mathbf{y}_{-i}; \mathcal{M}) \quad (3)$$

where $\mathbf{y}_{-i}$ denotes the values of all other output variables except y_i. This approximation allows us to capture the interactions among the output variables, as well as the input-output relations, without expensive learning time. Although the new conditioning set for each output dimension always includes all other outputs, the outputs not contributing to the prediction can be regularized out when learning the model from data, and hence the complexity of the individual models can be controlled.

Outlier Scoring with Reliability Weights The above MCOD score implicitly assumes that all our probability estimates and the models generating them are of high quality. However, in practice, the models that produce the probability estimates may not be all equally reliable as they are trained from a finite number of samples (especially when the number of input and output variables is high, and the sample size is small). Also, some dimensions of $Y_i|\mathbf{X}, \boldsymbol{\pi}(Y_i)$ may not fit well the base statistical assumption (which in this work is a logistic curve) and result in miscalibrated estimations. Consequently, if we treat $P(Y_i|\mathbf{X}, \boldsymbol{\pi}(Y_i))$ for all $i = 1, ..., d$ equally and merely search for the regions with low probabilities, the resulting scores degenerate to a noisy vector, which makes the detection of true irregularities hard.

To alleviate the issues, we propose to consider the reliability of each estimated conditional probability and incorporate it into the outlier score. For notational convenience, let ρ_i denote a conditional probability estimate for a data point $<\mathbf{x}, \mathbf{y}>$ on output dimension i, and let $\boldsymbol{\rho} = (\rho_1, \cdots, \rho_d)$. The MCOD score (either Equation 2 or 3) is rewritten as:

$$Score_{\text{MCOD}}(\mathbf{y}|\mathbf{x}) = -\sum_{i=1}^{d} \log \rho_i \quad (4)$$

One way to incorporate the reliability of each probability estimate and combine it with conditional probabilities is to define a weighted score:

$$Score_{\text{MCOD-RW}}(\mathbf{y}|\mathbf{x}) = -\sum_{i=1}^{d} w_i \log \rho_i \quad (5)$$

where w_i denotes the reliability weight of the model used to score the i-th output dimension. Trivially, when $w_i = 1$ for all $i = 1, ..., d$, the score becomes equivalent to Equation 4.

[1] **Notation:** For notational convenience, we will omit the index superscript $^{(n)}$ when it is not necessary. We may also omit variable names when they are clear; e.g., $P(Y_1 = y_1|\mathbf{X} = \mathbf{x}) = P(y_1|\mathbf{x})$.

Reliability Weights The Brier score (Brier 1950) measures the quality of the model based on model's probability outputs. It is defined as mean squared error between the predicted probabilities and observed outcomes. For our weighting purpose (Equation 5), however, direct application of the Brier score to the assessment of model quality would not be appropriate as it imposes different penalties for different errors and varies the distribution of errors (the mean squared error penalizes larger errors more than smaller errors (Willmott and Matsuura 2005)). Therefore we compute the reliability without squaring the error (*i.e.*, mean estimation error), which lets us estimate the quality of each estimate dimension ρ_i without distorting the distribution of errors. We finally define the reliability weight w_i by taking the inverse of this reliability measure. More formally, let $\epsilon_i^{(n)} = 1-\rho_i^{(n)}$ be the estimation error in probability on the dimension i for the n-th data instance. The reliability weight w_i (Equation 5) is defined as: $w_i = N/\sum_{n=1}^{N} \epsilon_i^{(n)}$. This effectively prioritizes the components of the outlier score, such that the contribution of outlier scores for more reliable partial models and their output dimensions increases, whereas that of noisy (unreliable) models and their dimensions decreases.

Local Reliability Weights The above weighting scheme assumes that the reliability of probability estimates (*i.e.*, the quality of a model) is invariant across all data regions. However, the assumption often does not hold because in most practical problems, especially in high-dimensional data spaces, data is not uniformly distributed in its attribute space. As a result, modeling and estimation of $P(Y_i|\mathbf{X}, \boldsymbol{\pi}(Y_i))$ cannot be achieved properly in the regions where data are sparse. We tackle such a sparsity issue by evaluating the reliability of each dimension of $\boldsymbol{\rho}^{(n)}$ locally in the region around the instance that we want to test:

$$Score_{\text{MCOD-LRW}}(\boldsymbol{\rho}^{(n)}) = -\sum_{i=1}^{d} w_i^{(n)} \log \rho_i^{(n)} \qquad (6)$$

where $w_i^{(n)} = |N_k(n)|/\sum_{n\in N_k(n)} \epsilon_i^{(n)}$ and $N_k(n)$ denotes k-nearest neighbors of the n-th instance in the input space.

Experiments

Through the empirical analysis below, we would like to demonstrate the advantages of (1) adopting the conditional outlier detection approach, (2) considering the dependence relations among outputs, (3) applying reliability weights and local reliability weights to outlier scores. Specifically, we compare the performance of our proposed outlier scores (MCOD, MCOD-RW, and MCOD-LRW; Equations 4-6), computed with the models that permit circular dependences, against two baseline methods:

- *Local outlier factor* (LOF) (Breunig et al. 2000) is one of the most widely used unconditional outlier detection method that identifies outliers using relative local densities. We apply LOF to the joint space of all data attributes.
- *Conditional outlier detection with d independent models* (COD) solves the problem by considering d independent conditional probability models $P(Y_i|\mathbf{X})$ (hence, the dependences among the output variables are not considered) and by computing Equation 4 with these models.

To obtain data models in COD, MCOD, MCOD-RW, and MCOD-LRW, we use L_2-penalized logistic regression as the base statistical model and choose their regularization parameters by cross validation. In LOF and MCOD-LRW, we use the Mahalanobis distance to find nearest neighbors and set the number of neighbors $k = 100$.

Dataset	*N/m/d*	Domain	Value Description	
			Input	Output
Mediamill	43,907/120/101	Video	Video frames	Concepts
Yahoo	11,214/21,924/30	Text	News articles	Topics
Yeast	2,417/103/14	Biology	Genes	Functionalities
Birds	645/276/19	Sound	Bird songs	Species

Table 1: Dataset characteristics. (N: number of instances, m: input dimensionality, d: output dimensionality)

Datasets We use *four* public datasets with multi-dimensional input and output (Table 1).[2] These are collected from various application domains, including semantic video/image annotation (*Mediamill*), text categorization (*Yahoo*), biology (*Yeast*), and sound recognition (*Birds*).

Simulating Outliers For the purpose of our comparative evaluation, we simulate multivariate conditional outliers by perturbing the output space of data. We take the following steps to simulate outliers. (1) In each simulation, select 1% of instances uniformly at random. (2) For each of the selected instances, perturb the values in $\{2.5, 5, 10, 20\}\%$ of the output dimensions uniformly at random ($y_{\text{outlier}} = |y_{\text{original}} - 1|$). The simulated outliers can be interpreted as contextually abnormal (erroneous) output signals in each application (see Table 1). For example, in *Mediamill* (video annotation), the outliers (perturbed output values) can be perceived as video frames with inaccurate concept tags. One important remark is that all methods (including both the model learning and outlier scoring stages) are run on data with simulated outliers. That is, we never learn a model on the unperturbed original data and detect outliers on the perturbed data. Such an experimental setting is impractical since in real applications we do not a priori know what data instances to remove to learn a model from outlier-free data.

Evaluation Metrics We evaluate the methods using the Average Precision-Alert Rate (APAR). Precision at Alert Rate r (P@r) measures precision at the top r-th percentile of outlier score (Hauskrecht et al. 2016). We average P@r over $r = [0.00, 0.01]$, which coincides with the ratio of simulated outliers in our experiments. Note that, in many real world applications, recall is considered no longer meaningful metric, as it can be computed only when true outliers are known as in our simulated study.

Results Table 2 shows the APAR of the five compared methods. All results are obtained from *ten* repeats. The num-

[2]Datasets are available at http://mulan.sourceforge.net/datasets-mlc.html (Tsoumakas, Katakis, and Vlahavas 2010).

$APAR_{[0.00,0.01]}$	Baselines		Ours			Baselines		Ours		
	LOF	COD	MCOD	MCOD-RW	MCOD-LRW	LOF	COD	MCOD	MCOD-RW	MCOD-LRW
	Outlier dimensionality = 2.5%					Outlier dimensionality = 5.0%				
Mediamill	0.14 ± 0.16	0.17 ± 0.09	0.26 ± 0.17	$\mathbf{0.61 \pm 0.12}$	$\mathbf{0.69 \pm 0.09}$	0.20 ± 0.17	0.06 ± 0.05	0.57 ± 0.14	$\mathbf{0.85 \pm 0.05}$	$\mathbf{0.90 \pm 0.04}$
Yahoo	0.01 ± 0.02	0.13 ± 0.06	$\mathbf{0.21 \pm 0.10}$	$\mathbf{0.36 \pm 0.09}$	$\mathbf{0.38 \pm 0.07}$	0.01 ± 0.03	0.25 ± 0.08	$\mathbf{0.43 \pm 0.11}$	$\mathbf{0.56 \pm 0.08}$	$\mathbf{0.58 \pm 0.07}$
Yeast	-	-	-	-	-	0.08 ± 0.07	0.04 ± 0.06	0.45 ± 0.12	$\mathbf{0.65 \pm 0.06}$	$\mathbf{0.65 \pm 0.05}$
Birds	-	-	-	-	-	0.04 ± 0.08	$\mathbf{0.34 \pm 0.22}$	$\mathbf{0.39 \pm 0.25}$	$\mathbf{0.45 \pm 0.21}$	$\mathbf{0.46 \pm 0.22}$
	Outlier dimensionality = 10.0%					Outlier dimensionality = 20.0%				
Mediamill	0.27 ± 0.16	$\mathbf{0.92 \pm 0.03}$	$\mathbf{0.91 \pm 0.04}$	$\mathbf{0.97 \pm 0.03}$	$\mathbf{0.98 \pm 0.03}$	0.30 ± 0.12	$\mathbf{0.99 \pm 0.02}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{1.00 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$
Yahoo	0.01 ± 0.01	0.32 ± 0.10	$\mathbf{0.42 \pm 0.13}$	$\mathbf{0.57 \pm 0.06}$	$\mathbf{0.57 \pm 0.07}$	0.01 ± 0.02	$\mathbf{0.36 \pm 0.13}$	0.25 ± 0.09	$\mathbf{0.39 \pm 0.05}$	$\mathbf{0.41 \pm 0.04}$
Yeast	0.08 ± 0.07	0.04 ± 0.06	0.45 ± 0.11	$\mathbf{0.64 \pm 0.06}$	$\mathbf{0.64 \pm 0.05}$	0.13 ± 0.09	0.17 ± 0.11	$\mathbf{0.52 \pm 0.08}$	$\mathbf{0.56 \pm 0.07}$	$\mathbf{0.55 \pm 0.08}$
Birds	0.07 ± 0.11	$\mathbf{0.42 \pm 0.31}$	$\mathbf{0.56 \pm 0.14}$	$\mathbf{0.66 \pm 0.18}$	$\mathbf{0.66 \pm 0.19}$	0.32 ± 0.22	$\mathbf{0.67 \pm 0.25}$	$\mathbf{0.78 \pm 0.19}$	$\mathbf{0.85 \pm 0.12}$	$\mathbf{0.84 \pm 0.13}$

Table 2: Average precision-alert rate (over alert rate $= [0.00, 0.01]$). Numbers shown in bold indicate the best results on each experiment set (by paired t-test at $\alpha = 0.05$). Dashes (-) indicate the sets that we cannot create due to low-dimensional output.

bers shown in boldface indicate the best results (by paired t-test at $\alpha = 0.05$) on each experiment set. In general, APAR increases as the outlier dimensionality gets larger, because larger perturbations are easier to detect.

Comparing the conditional outlier detection approaches (COD, MCOD, MCOD-RW, and MCOD-LRW) to the unconditional approach (LOF), the conditional approaches are the clear winners. MCOD, MCOD-RW, and MCOD-LRW always produce better APAR than LOF. Although COD sometimes underperforms LOF (*Mediamill* and *Yeast*), more frequently COD outperforms LOF. On the other hand, as expected, LOF hardly detects conditional outliers, because it seeks unusual data patterns in the joint space of all attributes.

Between MCOD and COD, our MCOD method outperforms COD in most cases across all datasets. Recalling that the key difference between two methods is in the type of data model they adopt, this verifies the advantages of considering the dependence relations among the output variables.

To validate our outlier scores with reliability weighting, we compare the performance of MCOD-RW and MCOD-LRW to that of MCOD. Recall that all three methods use the same data representation, and the only difference is in how they compute the outlier scores. The results show that MCOD-RW and MCOD-LRW always improve APAR over MCOD. We also point out that MCOD-RW and MCOD-LRW are not only capable of improving APAR, but are also able to make the performance more consistent (the standard deviations often decrease after reliability weighting). Lastly, although it is not statistically significant, our local approach, MCOD-LRW, seems capable to further improve the performance of MCOD-RW (see *Mediamill* and *Yahoo*).

Conclusions

We presented a probabilistic framework for the multivariate conditional outlier detection (MCOD) problem that relies on a decomposable model of conditional joint probability, where data instances that are assigned a low probability by the model are considered to be outliers. To efficiently obtain data representations, we proposed to use a collection of individually trained probabilistic functions with a relaxed conditional independence assumption. To cope with potentially different model qualities, we introduced new MCOD scores that incorporate with our global and local reliability weighting schemes. We presented experimental results on real world datasets with simulated outliers that support our proposed MCOD methods.

Acknowledgments

The work in this paper 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

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.

Brier, G. W. 1950. Verification of Forecasts expressed in terms of probability. *Monthly Weather Review* 78(1):1–3.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM Comput. Surv.* 41(3):15:1–15:58.

Dembczynski, K.; Cheng, W.; and Hüllermeier, E. 2010. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 279–286. Omnipress.

Hauskrecht, M.; Batal, I.; Hong, C.; Nguyen, Q.; Cooper, G. F.; Visweswaran, S.; and Clermont, G. 2016. Outlier-based detection of unusual patient-management actions: An icu study. *Journal of Biomedical Informatics* 64:211 – 221.

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.

Kriegel, H.-P.; Kröger, P.; and Zimek, A. 2010. Outlier detection techniques. In *Tutorial at 2010 SIAM Conference on Data Mining*.

Ng, A. Y. 2004. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*. ACM.

Platt, J. C. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, 61–74. MIT Press.

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.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2010. Mining multilabel data. In *Data Mining and Knowledge Discovery Handbook*. Springer US. 667–685.

Willmott, C. J., and Matsuura, K. 2005. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research* 30(1).

Zhang, M.-L., and Zhang, K. 2010. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, 999–1008. ACM.