Conditional Anomaly Detection Using Soft Harmonic Functions: An Application to Clinical Alerting

Michal Valko MICHAL@CS.PITT.EDU
Computer Science Department, University of Pittsburgh, PA

Hamed Valizadegan HAMED@CS.PITT.EDU
Computer Science Department, University of Pittsburgh, PA

Branislav Kveton BRANISLAV.KVETON@TECHNICOLOR.COM
Technicolor, Palo Alto, PA

Gregory F. Cooper GFC@PITT.EDU
Department of Biomedical Informatics, University of Pittsburgh, PA

Milos Hauskrecht MILOS@CS.PITT.EDU
Computer Science Department, University of Pittsburgh, PA

Abstract

Timely detection of concerning events is an important problem in clinical practice. In this paper, we consider the problem of conditional anomaly detection that aims to identify data instances with an unusual response, such as the omission of an important lab test. We develop a new non-parametric approach for conditional anomaly detection based on the soft harmonic solution, with which we estimate the confidence of the label to detect anomalous mislabeling. We further regularize the solution to avoid the detection of isolated examples and examples on the boundary of the distribution support. We demonstrate the efficacy of the proposed method in detecting unusual labels on a real-world electronic health record dataset and compare it to several baseline approaches.

1. Introduction

With the advances in health-care and with more data being handled and stored electronically, the opportunities increase for machine learning to improve the health care. Despite continuous improvement in medical practice, medical errors remain a very serious problem. According to recent data, medical errors are the 8-th leading cause of death in the US population (Kohn et al., 2000). We aim to identify medical errors that correspond to unusual patient-management decisions, such as ordering a medication. Our hypothesis is that patient-management decisions that are unusual with respect to past patients may be due to errors and that it is worthwhile to raise an alert if such a condition is encountered. Typical systems for medical error detection rely on the clinical knowledge, such as expert-derived rules. Extracting such knowledge is costly, time-consuming, and very difficult with medical practice constantly changing. Machine learning can offer a viable alternative: using past medical records to detect anomalies. Detecting anomalies in the patient-management decisions has the potential to help avoid medical errors, which could lead to improved quality of care and decreased costs.

Traditional anomaly detection methods used in data analysis are unconditional and look for outliers with respect to all data attributes (Markou & Singh, 2003). The conditional anomaly detection (CAD) problem (Hauskrecht et al., 2007; Song et al., 2007) seeks to detect unusual values for a subset of variables given the values of the remaining variables. In the special case when the target variable is a class label, the problem is often called mislabeling detection. While we focus on this special case of anomalies in the class label, the main objective is different: we are interested in constructing a system that can raise an alert when an anomalous label of a new example is observed. Formally, we want to solve the following problem:

> **Problem statement** (★): Given a set of n past observed examples $(\mathbf{x}_i, y_i)_{i=1}^{n}$ (with possible label noise), check if any instance i in recent m examples $(\mathbf{x}_i, y_i)_{i=n+1}^{n+m}$ is unusual.

Appearing in the *ICML 2011 Workshop on Machine Learning for Global Challenges*, Bellevue, WA, USA, 2011.

In general, we seek to reliably identify anomalies on the response (decision or class) variables for all possible values of the context (input or feature) variables. Not knowing the underlying model, that generates the (attributes, label) pairs, may lead to two major complications. First, a given instance may be far from the past observed data points (e.g. patient cases). Because of the lack of the support for alternative responses, it is difficult to asses the anomalousness of these instances. We refer to these instances as *isolated points*. Second, the examples on the boundary of the class distribution support, also known as *fringe points*, may look anomalous due to their low likelihood.

Because the underlying conditional distribution of the data is unknown, a non-parametric approach that looks for the label consistency of the instances on their neighborhood (e.g. k-nearest neighbor or k-NN) can be very useful (Papadimitriou & Faloutsos, 2003). The problem with relying on models such as k-NN is that they fail to detect clusters of anomalous instances. Our approach differs from typical local neighborhood approaches in two important aspects. First, it respects the structure of the manifold and accounts for more complex interactions in the data. Second, it solves the problem of isolated and fringe points by decreasing the confidence in predicting an opposite label for such points through regularization.

2. Background

Label propagation on the graph is widely used for semi-supervised learning (SSL). The general idea is to assume the consistency of labels among the data which are 1) close to each other and 2) lie on the structure (manifold/cluster). The two examples are the *Consistency Method* of Zhou et al. (Zhou et al., 2004) and the *Harmonic Solution* of Zhu et al. (Zhu et al., 2003), which are the instances of unconstrained regularization (Cortes et al., 2008). Let G be the similarity graph with the nodes corresponding to $\{\mathbf{x}_i\}_{i=1}^{n+m}$ and with the weighted edges W encoding pairwise similarities between the nodes. We denote by $\mathcal{L}(W)$ the (unnormalized) graph Laplacian defined as $\mathcal{L}(W) = D - W$ where D is a diagonal matrix whose entries are given by $d_{ii} = \sum_j w_{ij}$. In the transductive setting, the unconstrained regularization searches for soft (continuous) label assignment such that it maximizes fit to the labeled data and penalizes for not following the manifold structure:

$$\boldsymbol{\ell}^\star = \min_{\boldsymbol{\ell}\in\mathbb{R}^n} (\boldsymbol{\ell}-\mathbf{y})^\intercal C(\boldsymbol{\ell}-\mathbf{y}) + \boldsymbol{\ell}^\intercal K \boldsymbol{\ell}, \qquad (1)$$

where K is a symmetric regularization matrix and C is a symmetric matrix of empirical weights. C is usually diagonal and the diagonal entries often equal to some fixed constant c_l for the labeled data and c_u for the unlabeled. In a SSL setting, $\mathbf{y}$ is a vector of pseudo-targets such that $y_i \in \{\pm 1\}$ is the label of the i-th example when the example is labeled, and $y_i = 0$ otherwise. The appealing property of (1) is that it can be computed by the following closed form solution:

$$\boldsymbol{\ell}^\star = (C^{-1}K + I)^{-1}\mathbf{y} \qquad (2)$$

3. Methodology

We now propose a way to compute the anomaly score from (2). The output $\ell^\star$ of (1) for the example i can be rewritten (sgn refers to the sign function) as:

$$\ell_i^\star = |\ell_i^\star| \times \mathrm{sgn}(\ell_i^\star) \qquad (3)$$

SSL methods use $\mathrm{sgn}(\ell_i^\star)$ in (3) as the predicted label for i. For an unlabeled example, the closer the value of ℓ_i is to ± 1, the more consistent labeling information was propagated to it. The key observation, which we exploit in this paper, is that we can interpret $|\ell_i^\star|$ as a confidence of the label. We define the *anomaly score* s_i as the absolute difference between the actual label y_i and the inferred soft label ℓ_i:

$$s_i = |\ell_i^\star - y_i|. \qquad (4)$$

We will now address the problems of isolated examples by setting $K = \mathcal{L}(W) + \gamma_g I$, where we diagonally regularize the graph Laplacian. Intuitively, such a regularization lowers the confidence value $|\ell_i^\star|$ of all examples; however it reduces the confidence score of outlier points relatively more. In the fully labeled setting, the *hard* harmonic solution (Zhu et al., 2003) degenerates to the weighted k-NN. To alleviate this problem, we allow labels to spread on the graph by using soft constraints in the unconstrained regularization problem (1). In particular, instead of $c_l = \infty$ we set c_l to a finite constant and we set $C = c_l I$. With such a setting we can solve (1) using (2):

$$\boldsymbol{\ell}^\star = \left(c_l^{-1}\mathcal{L}(W) + \left(1 + \frac{\gamma_g}{c_l}\right) I \right)^{-1} \mathbf{y}. \qquad (5)$$

To avoid computation of the inverse, we may calculate (5) by solving a system of linear equations. We then plug the output of (5) into (4) to get the anomaly score. We will refer to this score as SoftHAD score. Intuitively, when the confidence is high but $\mathrm{sgn}(\ell_i^\star) \neq y_i$, we will consider the label y_i of the case $(\mathbf{x}_i, y_i)$ anomalous.

Backbone Graph The computation of the system of linear equations (5) scales with cubic[1] time complexity. This is not feasible for a graph with more than several thousands of nodes. To address the problem, we use *data quantization* (Gray & Neuhoff, 1998) and sample a set of nodes from the training data to create G. We then substitute the nodes in the graph with a smaller set of $k \ll n + m$ distinct centroids which results in $O(k^3)$.

4. Experiments

To evaluate our SoftHAD method, we compare it to the following baselines: (1) 1-class SVM approach in which we cover each class by a separate 1-class SVM (Schölkopf et al., 1999), (2) Quadratic discriminant analysis (QDA) model (Hastie et al., 2001), (3) SVM classification model (Vapnik, 1995) with RBF kernel, and (4) Weighted k-NN approach (Hastie et al., 2001) that uses the same weight metric W.

4.1. UCI ML Datasets

We first evaluated our method on the three UCI ML datasets (Frank & Asuncion, 2010) for which an ordinal response variable was available to calculate the true anomaly score. In particular, we selected 1) *Wine Quality* dataset with the response variable *quality* 2) *Housing* dataset with the response variable *median value of owner-occupied homes* and 3) *Auto MPG* dataset the response variable *miles per gallon*. In each of the dataset we scaled the response variable y_r to the $[-1, +1]$ interval and set the class label as $y := y_r \geq 0$. We randomly switched the class labels for three percent[2] of examples. The true anomaly score was computed as the absolute difference between the original response variable y_r and the (possibly switched) label. Table 1 compares the agreement scores to the true score for all methods on (2/3, 1/3) train-test split. We see that SoftHAD either performed the best or was close to the best method.

	Wine Quality	**Housing**	**Auto MPG**
QDA	75.1% (1.3)	56.7% (1.5)	65.9% (2.9)
SVM	75.0% (9.3)	58.5% (4.4)	37.1% (8.6)
1-class SVM	44.2% (1.9)	27.2% (0.5)	50.1% (3.5)
wk–NN	67.6% (1.4)	44.4% (2.0)	61.4% (2.3)
SoftHAD	74.5% (1.5)	71.3% (3.2)	72.6% (1.7)

Table 1. Mean anomaly agreement score and variance (over 100 runs) for CAD methods on the 3 UCI ML datasets.

4.2. Medical data

In this real-world experiment, we evaluated CAD on data extracted from electronic health records (EHR) of 4,486 patients. Our goal was to detect unusual lab test orders or medication administrations. We divided EHRs into two groups: a training set (2646 patients), and a test set (1840 patients). For each patient, we segmented the data according to the length of the patient stay where we considered all the patient data available at 8:00am each day. These patient instances were then converted into: (1) 9,282 features and (2) 749 labels/tasks – reflecting whether a particular lab was ordered or a particular medication was given within a 24-hour period. This segmentation led to 51,492 patient-state instances, such that 30,828 were used for training and 20,664 for testing. More details can be found in (Hauskrecht et al., 2010).

Parameters for the graph-based algorithms To construct G, we computed the similarity weights as:

$$w_{ij} = \exp\left[-\left(||\mathbf{x}_i - \mathbf{x}_j||^2_{2,\psi}\right)/\sigma^2\right],$$

where ψ is a weighing of the features (we used Wilcoxon score (Hanley & McNeil, 1982)) and σ is a length scale parameter. We chose σ as 10% of the empirical variance of the Euclidean distances. For each label, we sampled an equal number of positive and negative instances to construct a k-NN graph. We set $k = 75$, $c_l = 1$ and varied γ_g and the graph size.

Scaling for multi-task anomaly detection In this dataset, we have 749 binary labels. We want to output an anomaly score which is comparable among the different tasks/labels so we can, for example, set a unified threshold when the system is deployed in practice. To achieve this score comparability, we propose a simple approach where we take the minimum and the maximum score obtained for the training set and scale all scores for the same task linearly so that the score after the scaling ranges from 0 to 1.

4.3. Results and Conclusion

For the dataset described above, we computed the SoftHAD anomaly scores according to (5). We asked the panel of 15 clinical experts to evaluate the 222 patient case-label pairs (selected from $749 \times 20,664$ test case-label pairs), such that every case-label was evaluated by 3 experts who decided whether the alert was clinically relevant. We finally evaluated the performance of the CAD methods using the area under the ROC curve.

[1] The complexity can be further improved to $O(n_u^{2.376})$ with the Coppersmith-Winograd algorithm.

[2] We also performed the experiments with 1% to 10% of switched labels with the same trends.

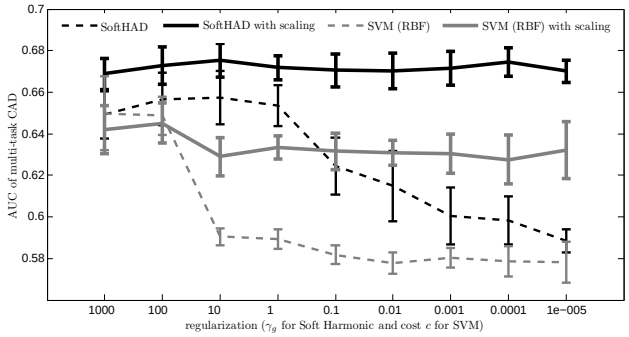

Figure 1. Medical Dataset: Varying regularizer 1) γ_g for SoftHAD 2) cost c for SVM with RBF kernel.

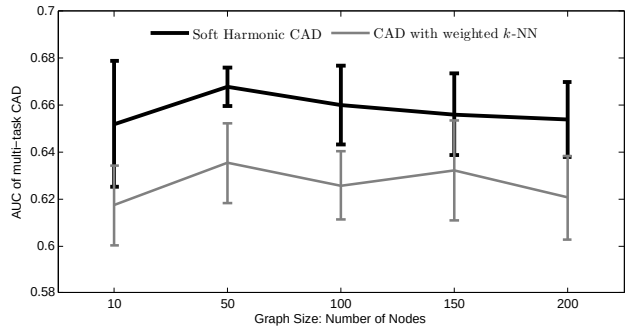

Figure 2. Medical Dataset: Varying graph size. Comparison of 1) SoftHAD and 2) weighted k-NN on the same graph.

In Figure 1, we compared SoftHAD vs. CAD using SVM with RBF kernel for different regularization settings. We sampled 200 examples to construct G (or train an SVM) and varied the γ_g regularizer (or cost c for SVM). Scaling anomaly scores to the same range improved the performance of both methods and makes the methods less sensitive to the regularization settings. We outperformed SVM approach over the range of regularizers. In Figure 2, we fixed $\gamma_g = 1$ and varied the number of examples we sampled from the training set to construct the similarity graph and compared it to the weighted k–NN. The error bars show the variances over 10 runs. Notice that the both of the methods are not too sensitive to the graph size. In future, we plan to extend the approach to the online anomaly detection. This can be beneficial for the deployment of our SoftHAD method in hospitals.

This research work was supported by grants R21LM009102, R01LM010019, and 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

Cortes, Corinna, Mohri, Mehryar, Pechyony, Dmitry, and Rastogi, Ashish. Stability of transductive regression algorithms. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

Frank, A. and Asuncion, A. UCI ML repository, 2010. URL http://archive.ics.uci.edu/ml.

Gray, Robert and Neuhoff, David. Quantization. *IEEE Transactions on Information Theory*, 44(6), 1998.

Hanley, J. A. and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, April 1982.

Hastie, T., Tibshirani, R., and Friedman, J. H. *The Elements of Statistical Learning*. Springer, 2001.

Hauskrecht, M., Valko, M., Kveton, B., Visweswaram, S., and Cooper, G. Evidence-based anomaly detection. In *Annual American Medical Informatics Association Symposium*, pp. 319–324, November 2007.

Hauskrecht, M., Valko, M., Batal, I., Clermont, G., Visweswaram, S., and Cooper, G. Conditional outlier detection for clinical alerting. *Annual American Medical Informatics Association Symposium*, 2010.

Kohn, L., Corrigan, J., and Donaldson, M. *To Err Is Human: Building a Safer Health System*. National Academy Press, Washington DC, 2000.

Markou, Markos and Singh, Sameer. Novelty detection: a review, part 1: statistical approaches. *Signal Process.*, 83(12):2481–2497, 2003. ISSN 0165-1684.

Papadimitriou, Spiros and Faloutsos, Christos. Cross-outlier detection. In *Advances in Spatial and Temporal Databases, 8th International Symposium, SSTD 2003*, volume 2750, pp. 199–213, 2003.

Schölkopf, Bernhard, Platt, John C., Shawe-taylor, John, Smola, Alex J., and Williamson, Robert C. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:2001, 1999.

Song, Xiuyao, Wu, Mingxi, and Jermaine, Christopher. Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 19(5): 631–645, 2007. ISSN 1041-4347.

Vapnik, Vladimir N. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Scholkopf, B. Learning with local and global consistency. *Advances in NIPS*, 16:321–328, 2004.

Zhu, Xiaojin, Ghahramani, Zoubin, and Lafferty, John. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th ICML*, pp. 912–919, 2003.