Conditional Anomaly Detection with Soft Harmonic Functions

Michal Valko*[0], Branislav Kveton†, Hamed Valizadegan‡, Gregory F. Cooper§ and Milos Hauskrecht¶
*INRIA Lille - Nord Europe, SequeL project, 40 avenue Halley, Villeneuve d'Ascq, France, e-mail: michal.valko@inria.fr
†Technicolor Labs, Palo Alto, California, USA, e-mail: branislav.kveton@technicolor.com
‡Computer Science Department, University of Pittsburgh, Pennsylvania, USA, e-mail: hamed@cs.pitt.edu
§Department of Biomedical Informatics, University of Pittsburgh, Pennsylvania, USA, e-mail: gfc@pitt.edu
¶Computer Science Department, University of Pittsburgh, Pennsylvania, USA, e-mail: milos@pitt.edu

***Abstract*—In this paper, we consider the problem of conditional anomaly detection that aims to identify data instances with an unusual response or a class label. 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 on several synthetic and UCI ML datasets in detecting unusual labels when compared to several baseline approaches. We also evaluate the performance of our method on a real-world electronic health record dataset where we seek to identify unusual patient-management decisions.**

***Keywords*-conditional anomaly detection, outlier and anomaly detection, graph methods, harmonic solution, backbone graph, random walks, health care informatics**

I. Introduction

Anomaly detection is the task of finding unusual elements in a set of observations. Most existing anomaly detection methods in data analysis are unconditional and look for outliers with respect to all data attributes [1], [2]. Conditional anomaly detection (CAD) [3], [4], [5] is the problem of detecting unusual values for a subset of variables given the values of the remaining variables. In other words, one set of variables defines the context in which the other set is examined for anomalous values.

CAD can be extremely useful for detecting unusual behaviors, outcomes, or unusual attribute pairings in many domains [6]. Examples of such problems are the detection of unusual actions or outcomes in medicine [4], [5], [7], investments [8], law [9], social networks [10], politics [11] and other fields [6]. In all these domains, the outcome strongly depends on the context (patient conditions, economy and market, case circumstances, etc.), hence the outcome is unusual only if it is compared to the examples with the same context.

In this work, we study a special case of CAD that tries to identify the unusual values for just one target variable given the values of the remaining variables (attributes). The target variable is assumed to take on a finite set of values which we also refer to as labels, because of its similarity to the classification problems.

In general, the concept of anomaly in data in the existing literature is somewhat ambiguous and several definitions have been proposed in the past [1], [2]. Typically, an example is considered anomalous when it is not expected from some underlying model. For the practical purposes of this paper, we define the conditional anomaly detection as follows:

> **Problem statement** (★): Given a set of n past examples $(\mathbf{x}_i, y_i)_{i=1}^n$ (with possible label noise), identify and rank instances i in recent m examples $(\mathbf{x}_i, y_i)_{i=n+1}^{n+m}$ that are unusual.

In this statement, we do not assume that the labels $\{y_i\}_{i=1}^n$ are perfect; they may also be subject to the label noise.

In order to assess the anomalousness of an example, we typically output an anomaly score. One way to define the score is to use a probabilistic model M and calculate the anomaly score as the probability of a different label: $P(y \neq y_i | \mathbf{x}_i, M)$. However, the probabilistic model M is not known in advance and must be estimated from available data. This may lead to two major complications that are illustrated in Figure 1: First, an instance may be far from the past observed data points. Because of the lack of the support for alternative responses, it is difficult to assess the anomalousness of these instances. We refer to these instances as *isolated points*. Second, the examples on the boundary of the class distribution support may look anomalous due to their low likelihood. These boundary examples are also known as *fringe points* [12].

One approach to the CAD task is to construct a classification model on the past data $(\mathbf{x}_i, y_i)_{i=1}^n$ and apply it to $(\mathbf{x}_i, y_i)_{i=n+1}^{n+m}$ to check if the assigned labels $(y_i)_{i=n+1}^{n+m}$ are correct. However, in that way we would disregard the labels $(y_i)_{i=n+1}^{n+m}$ which are available and can be utilized to improve the performance of the CAD task by leveraging the interaction between the labels of past examples and the new observed examples.

Because the underlying conditional distribution of the data is unknown, a non-parametric approach that looks for the

[0]The research work presented in this paper was done when Michal Valko was a PhD student at the University of Pittsburgh and was supported by funds from NIH awarded to University of Pittsburgh.

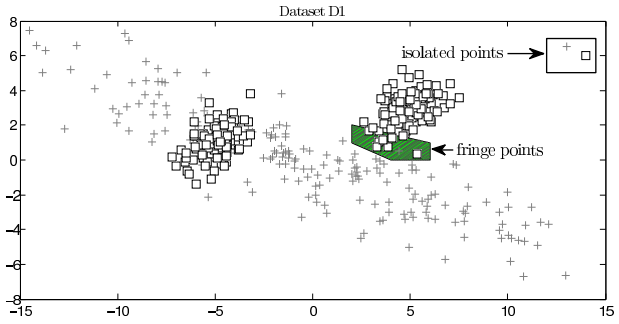

Figure 1. Challenges for CAD — the squares and the plus signs represent the examples from the two different classes: 1) **Fringe points** are the examples on the boundary of the class distribution support 2) **Isolated points** are the examples far from the majority but not within a different class.

label consistency of the instances on their neighborhood (e.g., k-nearest neighbor or k-NN) can be very useful [12]. One problem with relying on models such as k-NN is that they fail to detect clusters of anomalous instances.

In this paper, we develop and present a new non-parametric method to tackle CAD and its challenges. Our method relies on the similarity graph of instances and attempts to assess whether the response to an input variable agrees with responses of data points in its neighborhood using propagation of labeling information through the graph. Our method differs from typical local neighborhood methods 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.

Similar to other graph-based approaches for propagation of information (e.g., semi-supervised learning), the solution to our approach demands the computation of the inverse of the similarity matrix that can be challenging for large number of instances. We address this problem by proposing a method for building a smaller backbone graph that approximates the original graph.

In summary, the main contributions of this paper are:

- We introduce a label propagation approach on the data similarity graph for conditional anomaly detection to estimate the confidence of the labels.
- We propose a specific regularization that avoids unconditional outliers and fringe points (Figure 1).
- We present a compact computation of unconstrained regularization to account for the approximate backbone graph with different node weights.
- We propose a scaling approach that adjusts for multi-task predictions and makes anomaly scores comparable.
- We verify the efficacy of the proposed algorithm on some synthetic datasets, several UCI datasets and a challenging real-world dataset of patients' health records.

In the following, we first review the related work in Section II-A, and label propagation on the data similarity graph in Section II-C. In Section III-A, we adopt the label propagation on the data similarity graph for CAD problem and propose a regularization that addresses the isolated and fringe points problems. Next, in Section III-B, we show how to create a smaller backbone graph to deal with more than a few thousand examples. We report the results of our approach on the synthetic datasets, UCI ML datasets, and a real-world medical dataset in Section IV.

II. Background

A. Related work

While traditional anomaly detection has been studied for a long time [1], [2], the methods for CAD are relatively new and the research in this area is only emerging [3], [4], [13]. Typically, the existing CAD methods adopt and use classification methods, such as generative models [4], [13], [14], or max-margin classifiers [5], [7]. In Section IV, we compare our method to these approaches.

The work on CAD, when the target variable is restricted to discrete values only, is closely related to mislabeling detection [15] and cross-outlier detection [12]. The objective of mislabeling detection is to 1) make a yes/no decision on whether the examples are mislabeled, and 2) improve the classification accuracy by removing the mislabeled examples.

Brodley et al. [15] used different classification approaches to remove mislabeled samples including single and ensemble classifiers. Bagging and boosting are applied in [16]. Jiang et al. [17] used an ensemble of neural nets to enhance the performance of a k-NN classifier. Sanchez et al. [18] introduced several other k-NN based approaches including *depuration*, *nearest centroid neighborhood* (NCN) and *iterative* k-NCN. Finally, Valizadegan and Tan [19] introduced an objective function based on the weighted nearest neighbor approach and solved it with Newton method.

The main difference between mislabeling detection and CAD is that mislabeling detection identifies and removes the mislabeled examples in order to learn better classifiers, while CAD is interested in ranking examples according to the severity of conditional anomalies in data. This is the main reason our evaluations in Section IV measure the rankings of the cases being anomalous and not the improved classification accuracy when we remove them. Nevertheless, we do compare (Sections IV-A and IV-B) to the methods typically used in mislabeling detection.

Papadimitriou and Faloutsos [12] define cross-outliers as examples that seem normal when considering the distribution of examples from the assigned class, but are abnormal when considering the examples from the other class. For

each sample $(\mathbf{x}, y)$, they compute two statistics based on the the similarity of $\mathbf{x}$ to its neighborhood from the samples belonging to class y and samples not belonging to class y. An example is considered anomalous if the first statistic is significantly smaller than the second statistic. However, their method is not very robust to fringe points (Figure 1) [12]. The conditional anomaly approach developed in this paper addresses this problem.

B. Notation

We use the following notation throughout the paper. Let $(\mathbf{x}_i, y_i)_{i=1}^{n+m}$ be the collection of n past and m recent observed examples. Without loss of generality, we limit y to binary class labels, i.e., $y \in \{\pm 1\}$. Let G be the similarity graph constructed on the nodes $\{\mathbf{x}_i\}_{i=1}^{n+m}$ with weighted edges W. The entries w_{ij} of W encode the pairwise similarities between $\mathbf{x}_i$ and $\mathbf{x}_j$. 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 $d_{ii} = \sum_j w_{ij}$.

C. Label Propagation

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 some structure (a manifold or a cluster). The two examples are the *consistency method* of Zhou et al. [20] and the *harmonic solution* of Zhu et al. [21]. The inference of the labels by the approach of Zhu et al. [21] can be interpreted as a random walk on G with the transition matrix $P = D^{-1}W$. The harmonic solution satisfies the *harmonic* property $\ell_i = \frac{1}{d_{ii}} \sum_{j \sim i} w_{ij} \ell_j$.[1]

Harmonic solution and consistency method are the instances of a bigger class of the optimization problems called the unconstrained regularization [22]. 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:

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

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 is the label of the i-th example when the example is labeled, and $y_i = 0$ otherwise. Many methods can be derived from (1). For example, for the (hard) harmonic solution $K = \mathcal{L}(W)$, $c_l = \infty$, and $c_u = 0$. Consistency method has K equal to the normalized graph Laplacian $K = I - D^{-1/2} W D^{-1/2}$ and $c_u = c_l$ is set to a non-zero constant. The appealing property of (1) is that its solution can be computed in closed form as follows [22]:

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

[1] $j \sim i$ denotes that j and i are neighbors in G

III. Methodology

In this section we show how to solve the CAD problem ($\bigstar$) using label propagation on a data similarity graph and how to compute the anomaly score. In particular, we will build on the harmonic solution approach (Section II-C) and adopt it for CAD in the following ways: 1) show how to compute the confidence of mislabeling, 2) add a regularizer to address the problem of isolated and fringe points, 3) use soft constraints to account for a fully labeled setting, and 4) describe a compact computation of the solution from a quantized backbone graph.

A. Conditional Anomaly Detection

The label propagation method described in Section II-C can be applied to CAD by considering all observed data as labeled examples with no unlabeled examples. The setting for matrix C is dependent on the quality of the past observed data. If the labels of the past observed data (or any example from the recent sample) are guaranteed to be correct, we set the corresponding diagonal elements of C to a large value to make their labels fixed. Notice that specific domain techniques can be used to make sure that the collected examples from the past observed data have correct labels. In this paper, we assume that we do not have access to such prior knowledge and therefore, the observed data are also subject to label noise.

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

$$\ell_i^\star = |\ell_i^\star| \times \operatorname{sgn}(\ell_i^\star) \tag{3}$$

SSL methods use $\operatorname{sgn}(\ell_i^\star)$ in (3) as the predicted label for i. For an unlabeled example, when the value of ℓ_i is close to ± 1, then the labeling information that was propagated to it is more consistent. Typically, that means that the example is close to the labeled examples of the respective class. The key observation, which we exploit in this paper, is that we can interpret $|\ell_i^\star|$ as a confidence in the label. Our situation differs from SSL, as all our examples are labeled and we aim to assess the confidence of *already labeled* example. Therefore, we define the *anomaly score* as the absolute difference between the actual label y_i and the inferred soft label ℓ_i:

$$s_i = |\ell_i^\star - y_i|. \tag{4}$$

We will now address the problems illustrated in Figure 1. Recall that the isolated points are the examples that are (with respect to some metric) far from the majority of the data. Consequently, they are surrounded by few or no nearby

points. Therefore, no matter what their label is, we do not want to report them as conditional anomalies. In other words, we want CAD methods to assign them a low anomaly score. Even when the isolated points are far from the majority data, they still can be orders of magnitudes closer to the data points with the opposite label. This can make a label propagation approach falsely confident about that example being a conditional anomaly. In the same way we do not want to assign a high anomaly score to fringe points just because they lie on a distribution boundary. To tackle these problems we set $K = \mathcal{L}(W)+\gamma_g I$, where we diagonally regularize the graph Laplacian. Intuitively, such a regularization lowers the confidence value $|\boldsymbol{\ell}^\star|$ of all examples; however it reduces the confidence score of far outlier points relatively more. To see this, notice (Section IV-C) that the similarity weight metric is an exponentially decreasing function of the Euclidean distance. In other words, such a regularization can be interpreted as a label propagation on the graph with an extra sink. The sink is an extra node in G with label 0 and every other node connected to it with the same small weight γ_g. The edge weight of γ_g affects the isolated points more than other points because their connections to other nodes are small.

In the fully labeled setting, the *hard* harmonic solution degenerates to the weighted k-NN. In particular, the hard constraints of the harmonic solution do not allow the labels spread beyond other labeled examples. However, despite the fully labeled case, we still want to take the advantage of the manifold structure. 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 also set $C = c_l I$. With such a setting of K and C, we can solve (1) using (2) to get:

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

$$= \left(c_l^{-1}\mathcal{L}(W) + \left(1 + \frac{\gamma_g}{c_l}\right) I\right)^{-1}\mathbf{y}. \tag{6}$$

To avoid computation of the inverse,[2] we calculate (6) using the following system of linear equations:

$$\left(c_l^{-1}\mathcal{L}(W) + \left(1 + \frac{\gamma_g}{c_l}\right) I\right)\boldsymbol{\ell}^\star = \mathbf{y} \tag{7}$$

We then plug the output of (7) into (4) to get the anomaly score. We will refer to this score as the SoftHAD score. Intuitively, when the confidence is high but $\operatorname{sign}(\ell_i^\star) \neq y_i$, we will consider the label y_i of the case $(\mathbf{x}_i, y_i)$ conditionally anomalous.

[2]due to numerical instability

B. Backbone Graph

The computation of the system of linear equations (7) scales with complexity[3] $O(n^3)$. This is not feasible for a graph with more than several thousand nodes. To address the problem, we use *data quantization* [23] 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 $k \ll n$ distinct centroids which results in $O(k^3)$ computation.

We improve the approximation of the original graph with the backbone graph, by assigning different weights to the centroids. We do it by computing the multiplicities (i.e., how many nodes each centroid represents). In the following we will describe how to modify (7) to allow for the computation with multiplicities.

Let V be the diagonal matrix of multiplicities with v_{ii} being the number of nodes that centroid $\mathbf{x}_i$ represents. We will set the multiplicities according to the empirical prior. Let W^V be the compact representation of the matrix W on G, where each node $\mathbf{x}_i$ is replicated v_{ii} times. Let L^V and K^V be the graph Laplacian and regularized graph Laplacian of W^V. Finally, let C^V be the C in (1) with the adjustment for the multiplicities. C^V accounts for the fact that we care about "fitting" to train data according to the train data multiplicities. Then:

$$\begin{aligned} W^V &= VWV \\ L^V &= \mathcal{L}(W^V) \\ K^V &= L^V + \gamma_g V \\ C^V &= V^{1/2} C V^{1/2} \end{aligned}$$

The unconstrained regularization (1) now becomes:

$$\boldsymbol{\ell}^{V\star} = \min_{\boldsymbol{\ell}\in\mathbb{R}^n} (\boldsymbol{\ell}-\mathbf{y})^\intercal C^V (\boldsymbol{\ell}-\mathbf{y}) + \boldsymbol{\ell}^\intercal K^V \boldsymbol{\ell} \tag{8}$$

and subsequently (6) becomes:

$$\begin{aligned} \boldsymbol{\ell}^{V\star} &= \left(\left(C^V\right)^{-1} K^V + I\right)^{-1}\mathbf{y} \\ &= \left(V^{-1/2} C^{-1} V^{-1/2}(L^V + \gamma_g V) + I\right)^{-1}\mathbf{y} \\ &= \left((c_l V)^{-1}\left(L^V + \gamma_g V\right) + I\right)^{-1}\mathbf{y} \\ &= \left(1/c_l V^{-1} L^V + c_l\gamma_g + I\right)^{-1}\mathbf{y} \end{aligned}$$

With these adjustments the anomaly score that accounts for the multiplicities is equal to $|\boldsymbol{\ell}^{V\star} - \mathbf{y}|$.

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

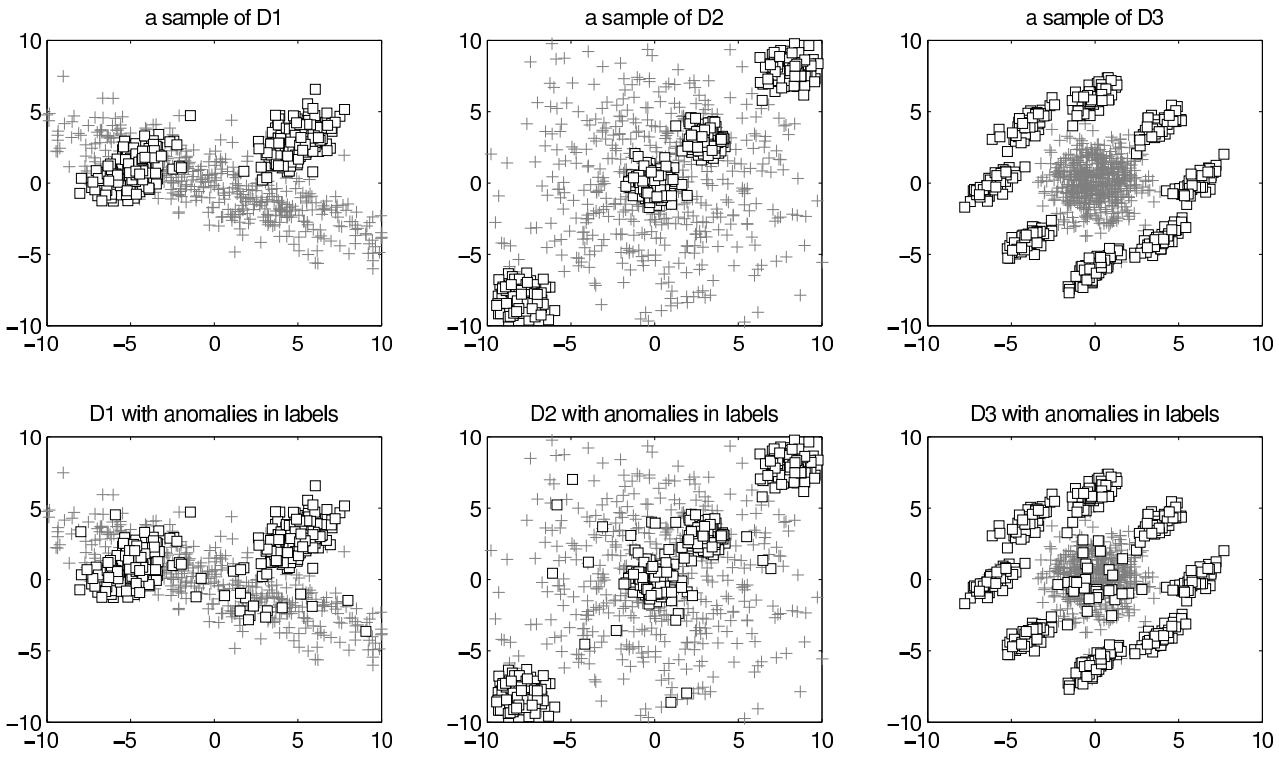

Figure 2. Synthetic Data. Top: A sample of datasets D1, D2, and D3. Bottom: Synthetic datasets after changing the labels of 3% of the examples.

IV. Experiments

To evaluate the performance of our SoftHAD method, we compare it to the following baselines:

- 1-class SVM approach in which we cover each class by a separate 1-class SVM [24] with RBF kernel and the anomaly score equals to the distance of the example from the learned boundary of its own class. This method is an example of the traditional anomaly detection method adopted for CAD.
- Quadratic discriminant analysis (QDA) model [25], where we model each class by a multivariate Gaussian, and the anomaly score is the class posterior of the opposite class.
- SVM classification model [26] with RBF kernel where we consider an example anomalous if it falls far on the opposite side of the decision boundary. This method is an example of the classification method adopted for CAD and was used by Valko et al. [5].
- Weighted k-NN approach [25] that uses the same weight metric W as SoftHAD, but relies only on the labels in the local neighborhood and does not account for the manifold structure.

A. Synthetic data

The evaluation of a CAD is a very challenging task when the true model is not known. Therefore, we first evaluate and compare the results of different CAD methods on three synthetic datasets (D1, D2, and D3) with known underlying models that let us compute the true anomaly scores.

We show the three datasets we used in our experiments in Figure 2. All datasets are modeled as the mixtures of multivariate Gaussians and the class densities we used to generate these datasets vary in locations, shapes and mutual overlaps. Dataset D1 is similar to XOR type of data with one of the classes modeled by a single elongated Gaussian. In D2, the classes overlap and the form of D3 is close to the concentric circles but the clusters are non-overlapping. For each dataset, we sampled 500 examples from the class $+1$ and 500 examples from the class -1 for the training set and the same number of the examples for the testing set. For each experiment we sample the datasets 100 times. After the sampling, we randomly switch the class labels for three percent of examples for both training and testing set.

We then calculate the true anomaly score as

$$P(y \neq y_i|\mathbf{x}_i) = 1 - P(y = y_i|\mathbf{x}_i),$$

reflecting how anomalous the label of the example is with respect to the true model. Each of the methods outputs a score which orders the examples according to the belief of the anomalous labeling. For each of the CAD methods, we assess how much this ordering is consistent with the ordering of the true anomaly score. We did it by counting the number

of swapped pairs between the true and predicted ordering which is equivalent to the Wilcoxon score or AUC score.[4] The less number of swapped pairs, the higher the agreement score.

Table I compares the agreement scores (or equivalently, the AUCs) of the experiment for all methods. The results demonstrate that our method outperforms all other baselines and comes closest to the order induced by the true model. We also evaluated the linear versions of SVM and 1-class SVM, but the results were inferior to the ones with the RBF kernel.

	Dataset **D1**	Dataset **D2**	Dataset **D3**
QDA	73.4% (2.4)	48.0% (1.4)	54.5% (1.2)
SVM	59.8% (4.8)	58.8% (5.7)	50.8% (2.2)
1-class SVM	50.7% (1.3)	52.3% (1.2)	59.4% (1.6)
wk–NN	66.6% (1.4)	64.6% (1.3)	60.5% (1.5)
SoftHAD	81.3% (1.8)	82.4% (1.6)	63.0% (2.9)

Table I
MEAN ANOMALY AGREEMENT SCORE AND VARIANCE (OVER 100 RUNS) FOR CAD METHODS ON THE 3 SYNTHETIC DATASETS.

Figure 3 shows the top 5 anomalies identified by each method on D3. We see that only the soft harmonic method was able to identify the top conditional anomalies, which correspond to the examples with switched labels in the middle region that carries a lot of counter-support and hence leads to the highest anomaly score.

B. UCI ML Datasets

We also evaluated our method on the three UCI ML datasets [27], 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 with 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 $y := y_r \geq 0$. As with the synthetic datasets, we randomly switched the class labels for three percent 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 II compares the agreement scores to the true score for all methods on (2/3, 1/3) train-test split. Again, we see that SoftHAD either performed the best or was close to the best method.

C. Medical data

In this experiment, we evaluated our method on the problem of detecting unusual patient-management actions [7].

[4]AUC commonly used for classification is a special case with the true score being ± 1.

	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 II
MEAN ANOMALY AGREEMENT SCORE AND VARIANCE (OVER 100 RUNS) FOR CAD METHODS ON THE 3 UCI ML DATASETS.

We asked a panel of clinical experts to judge the outputs of the CAD methods for the clinical relevance.

1) Data: We used the data extracted from electronic health records (EHRs) of 4,486 patients as described in [7]. The patients were divided into a train set (2646 patients) and a test set (1840 patients). Patient records were segmented in time (every day of a patient's visit at 8:00am) to obtain 51,492 patient-state instances, such that 30,828 were train and 20,664 test instances. The data in EHRs for these instances were then converted into 9,282 features – a vector representation of the patient state. For every patient-state instance we had 749 decision labels (or tasks) which were possible lab-order and medication decisions with true/false values, reflecting whether a particular lab was ordered or a particular medication was given within a 24-hour period.

2) Evaluation: We evaluated our CAD method on 222 patient-instance/action pairs. We selected these 222 cases such that they represented a wide range of low, medium and high anomaly scores according to the baseline SVM method [5], [7]. Each instance/action pair was evaluated by three different clinical experts determining whether the action is anomalous and whether this anomaly is clinically relevant. To assess the example, we used the majority rule (two out of three experts). We then evaluated the quality of CAD methods using the area under the ROC (AUC) metric. We compared SoftHAD method to the three baselines 1) weighted k–NN on the same graph 2) SVM with RBF kernel 3) 1-class SVM with RBF kernel described at the beginning of this section.

3) 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 and σ is a length scale parameter. The reason for the different feature weights is the high dimensionality of the data. Without any feature scaling, a distance based on 9K features would make any two points almost equidistant and thus meaningless. Therefore, we weighted the features based on their discriminative power according to the univariate Wilcoxon score [28]. Next, σ is chosen so that the graph is reasonably sparse [29]. We followed [19] and chose σ as 10% of the empirical variance

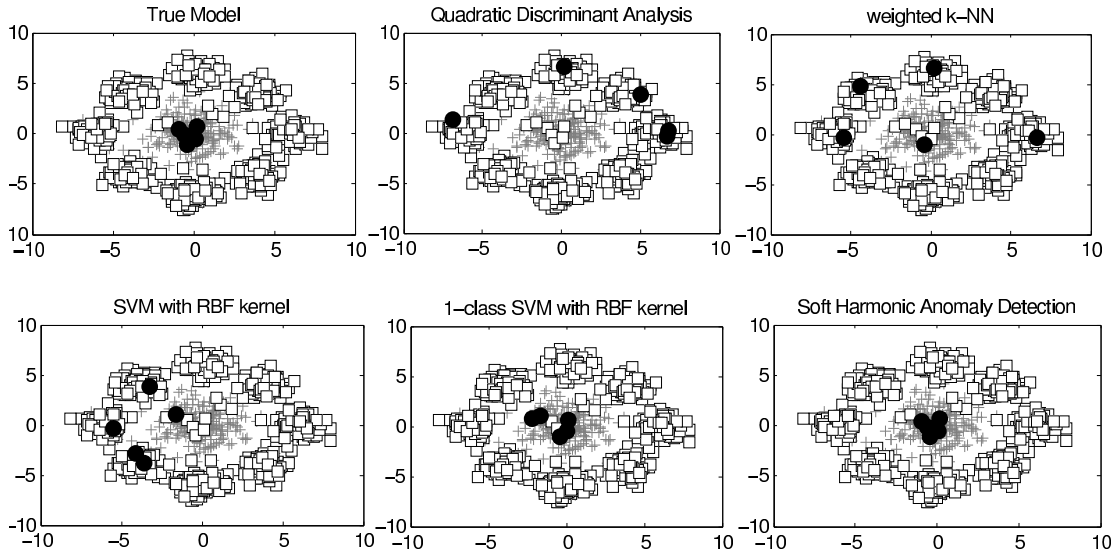

Figure 3. Black dots depict the top five conditional anomalies based on the score for each of the methods on D3. The top five conditional anomalies according to the true model are in the middle (top left).

of the Euclidean distances. Based on the experiments, our algorithm is not sensitive to the small perturbations of σ; what is important is that the graph does not become disconnected by having all edges of several nodes with weights close to zero. 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.

4) Scaling for multi-task anomaly detection: So far, we have described CAD only for a single task (anomaly in a single label). In this dataset, we have 749 binary tasks (or labels) that correspond to 749 different possible orders of lab tests or medications. In our experiments, we compute the CAD score for each task independently. Figure 4 shows the CAD scores for two of them. CAD scores close to 1 indicate that the order should be done, while the scores close to 0 indicate the opposite. The ranges for the anomaly scores can vary among the different labels/tasks, as one can notice in Figure 4. However, we want to output an anomaly score which is comparable among the different tasks/labels so we can 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.

5) Results: In Figure 5, we fixed $\gamma_g = 1$ and vary the number of examples we sample from the training set to construct the similarity graph, and also compare 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. This is due to the multiplicity

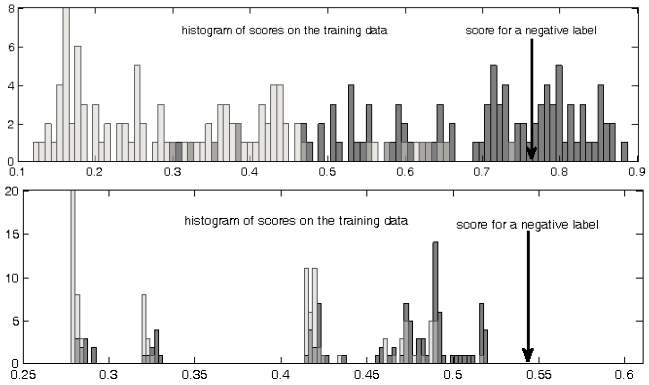

Figure 4. Histogram of anomaly scores for 2 different tasks. The scores for the top and bottom task range from 0.1 to 0.9 and from 0.25 and 0.61, respectively. The arrow in both cases points to the scores of the evaluated examples, both with negative labels. Despite the score is lower for the bottom task, we may believe that it is more anomalous because it is more extreme within the scores for the same task.

adjustment for the backbone graph (Section III-B). Since we use the same graph both for SoftHAD and weighted k–NN, we anticipate that we are able to outperform weighted k–NN due to the label propagation over the data manifold and not only within the immediate neighborhood. In Figure 6, we compare SoftHAD to the CAD using SVM with an RBF kernel for different regularization settings. We sample 200 examples to construct a graph (or train an SVM) and vary the γ_g regularizer (or cost c for SVM). We outperform the SVM approach over the range of regularizers. The AUC for the 1-class SVM with an RBF was consistently below 55%, so we do not show it in the figure. We also compared

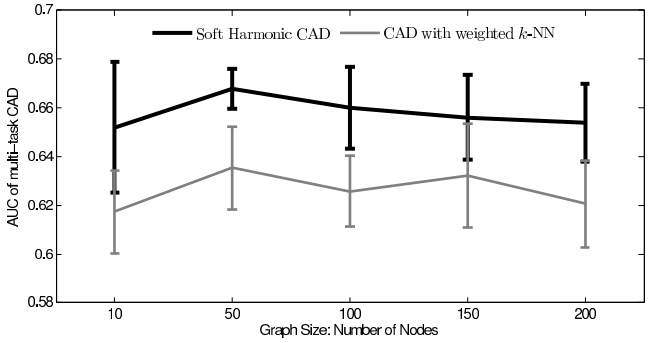

Figure 5. Medical Dataset: Varying graph size. Comparison of 1) SoftHAD and 2) weighted k-NN on the same graph.

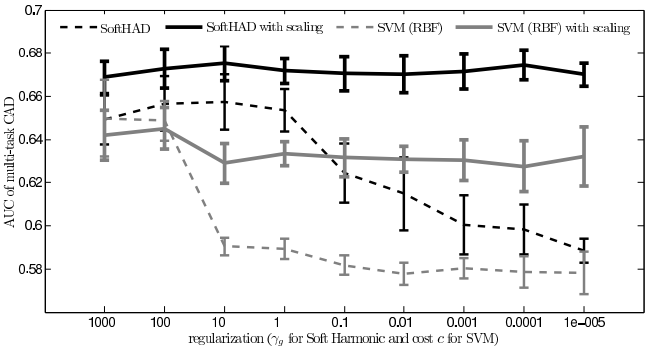

Figure 6. Medical Dataset: Varying regularizer 1) γ_g for SoftHAD 2) cost c for SVM with RBF kernel.

the two methods with scaling adjustment for this multi-task problem (Figure 6). The scaling of anomaly scores improved the performance of both methods and makes the methods less sensitive to the regularization settings.

V. Conclusion

We presented a non-parametric graph-based algorithm for conditional anomaly detection. Our algorithm goes beyond exploring just the local neighborhood (nearest neighbor approach) and uses label propagation on the data manifold structure to estimate the confidence of labeling. We evaluated our method on synthetic data, where the true model was known to confirm that the anomaly score from our method outperforms the others in ordering examples according to the true anomaly score. We also presented the evaluation of our method on the real-world data of patient health records, where the true model is not know, but when we used the experts in clinical care to evaluate the severity of our alerts.

In future, we plan to work on the structure anomalies where instead of computing the anomaly score independently for each label, we compute it *jointly*. With such a structured approach we can avoid the necessity of the score scaling.

VI. Acknowledgments

This research work was supported by the grants R21LM009102, R01LM010019, and R01GM088224 from the NIH and the grant IIS-0911032 from the NSF. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

References

[1] M. Markou and S. Singh, "Novelty detection: a review, part 1: statistical approaches," *Signal Process.*, vol. 83, no. 12, pp. 2481–2497, 2003.

[2] ——, "Novelty detection: a review, part 2: neural network based approaches," *Signal Process.*, vol. 83, no. 12, pp. 2499–2521, 2003.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, pp. 15:1–15:58, July 2009. [Online]. Available: http://doi.acm.org/10.1145/1541880.1541882

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

[5] M. Valko, G. Cooper, A. Seybert, S. Visweswaran, M. Saul, and M. Hauskrecht, "Conditional anomaly detection methods for patient-management alert systems," in *Workshop on Machine Learning in Health Care Applications in The 25th International Conference on Machine Learning*, 2008.

[6] K. Das, J. Schneider, and D. B. Neill, "Anomaly pattern detection in categorical datasets," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 169–176.

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

[8] S. Rubin, M. Christodorescu, V. Ganapathy, J. T. Giffin, L. Kruger, H. Wang, and N. Kidd, "An auctioning reputation system based on anomaly detection," in *Proceedings of the 12th ACM conference on Computer and communications security*, ser. CCS '05. New York, NY, USA: ACM, 2005, pp. 270–279.

[9] E. Aktolga, I. Ros, and Y. Assogba, "Detecting outlier sections in us congressional legislation," in *Proceedings of SIGIR*, 2010, IR.

[10] N. A. Heard, D. J. Weston, K. Platanioti, and D. J. Hand, "Bayesian anomaly detection methods for social networks," *Annals of Applied Statistics*, vol. 4, pp. 645–662, 2010.

[11] M. Kolar, L. Song, A. Ahmed, and E. P. Xing, "Estimating time-varying networks," *Annals of Applied Statistics*, vol. 4, pp. 94–123, 2010.

[12] S. Papadimitriou and C. Faloutsos, "Cross-outlier detection," in *Advances in Spatial and Temporal Databases, 8th International Symposium, SSTD 2003, Santorini Island, Greece, July 24-27, 2003, Proceedings*, T. Hadzilacos, Y. Manolopoulos, J. F. Roddick, and Y. Theodoridis, Eds., vol. 2750, 2003, pp. 199–213.

[13] X. Song, M. Wu, and C. Jermaine, "Conditional anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 631–645, 2007, fellow-Sanjay Ranka.

[14] M. Valko and M. Hauskrecht, "Distance metric learning for conditional anomaly detection," in *Twenty-First International Florida Artificial Intelligence Research Society Conference*. AAAI Press, 2008.

[15] C. E. Brodley and M. A. Friedl, "Identifying mislabeled training data," *J. Artif. Intell. Res. (JAIR)*, vol. 11, pp. 131–167, 1999.

[16] S. Verbaeten and A. V. Assche., "Ensemble methods for noise elimination in classification problems." in *Proceeding of 4th International Workshop on Multiple Classifier Systems*, 2003.

[17] Y. Jiang and Z.-H. Zhou, "Editing training data for knn classifiers with neural network ensemble." in *Lecture Notes in Computer Science 3173*, 2004, pp. 356–361.

[18] J. Sanchez, R. Barandela, A. I. Marques, R. Alejo, and B. J., "Analysis of new techniques to obtain quality training sets." *Pattern Recognition Letteres 24*, pp. 1015–1022, 2003.

[19] H. Valizadegan and P.-N. Tan, "Kernel based detection of mislabeled training examples," in *Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA*, 2007.

[20] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systems*, vol. 16, pp. 321–328, 2004.

[21] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the 20th International Conference on Machine Learning*, 2003, pp. 912–919.

[22] C. Cortes, M. Mohri, D. Pechyony, and A. Rastogi, "Stability of transductive regression algorithms," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 176–183.

[23] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[24] B. Scholkopf, J. C. Platt, J. Shawe-taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, p. 2001, 1999.

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

[26] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[27] A. Asuncion and D. Newman, "UCI machine learning repository," 2011. [Online]. Available: http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html

[28] J. A. Hanley and B. J. Mcneil, "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, April 1982.

[29] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.