A User Friendly Framework for Anomalous Pattern Detection

Feng Chen
Department of Computer Science
University at Albany, SUNY

July 12, 2014

1 Introduction

Anomalous pattern detection refers to the detection of any interesting or anomalous patterns in the data, in which each pattern is characterized as the subset of data records affected. The anomalous pattern detection task arises in a variety of applications, such as disease surveillance, where we must detect emerging outbreaks of disease in the very early stages; credit card fraud detection, where we attempt to detect patterns of suspicious credit card transactions; and road network congestion detection, where we attempt to detect non-recurrent congested road links. A popular hypothesis-testing-based framework for anomalous pattern detection assumes that, under the alternative hypothesis ($H_1(S)$), the majority of the data is generated from the same distribution representing the normal behavior of the system, and the subset anomalous data records (S) are generated from a different distribution as the characterization of anomalous patterns [4]. The null hypothesis (H_0) assumes no anomalous records. The problem of anomalous pattern detection can be formalized as the maximization of the log-likelihood ratio statistic function ($F(S)$) over all possible subsets ($S \subseteq$ data), where

$$F(S) = \log\left[\frac{\text{prob(data}|H_1(S))}{\text{prob(data}|H_0)}\right]. \tag{1}$$

Depending on the distributions assumed for normal and anomalous data records, many methods have been proposed, including expectation-based Poisson statistic [1], Kulldorff statistic [2], fast subset scan [3], fast generalized subset scan [4], and various others. For example, in disease surveillance, suppose data $\equiv \{d_1, \cdots, d_N\}$, where $d_i \equiv (c_i^t, b_i^t)$, c_i^t refers to the number of reported respiratory cases in a county s_i on day t, and b_i^t refers to the expected count calculated based on the historical data. We assume a Possion distribution $c_i^t \sim \text{Possion}(b_i^t)$ for normal data records, and a different Possion distribution $c_i^t \sim \text{Possion}(qb_i^t)$ for anomalous data records, where q, $q > 1$, is a unknown parameter that can be estimated via maximum likelihood estimation (MLE). Then we obtain expectation-based Poisson statistic, and the log-likelihood ratio can be derived as $F(S) = C\log(C/B) + B - C$, if $C > B$, and $F(S) = 0$ otherwise, where C and B are respectively the aggregate count $\sum_{i\in S} c_i^t$ and aggregate baseline $\sum_{i\in S} b_i^t$.

Although previous methods have been shown effective in a number of applicaitons [5, 6, 7], these methods have very limited capability to integrate users' prior knowledge and special requirements, except that they allow users to specify the distributions and their hyper parameters using Bayesian techniques [8, 9]. For example, there are potential costs $c(S)$ (e.g., for manual validation) associated with the identified anomalous nodes, and users may want to simultaneously minimize the overall cost. As a second example, users may have manually labeled some normal and anomalous nodes, and want to identify anomalous subsets that are consistent with their labels. As a third example, when data is a graph, users may want to add special structure constraints, such as the connectivity constraint, and to simultaneously maximize (or minimize) some statistic function of the sub-graph induced by S, such as density (or cut) function.

2 A User Friendly Framework based on First Order Logic

In order to design a user friendly framework that supports the integration of user's prior knowledge and special requirements, we present a regularization framework for anomalous pattern detection as follows:

$$\max_{S\in\{0,1\}^N} F(S) - \Phi_\lambda(S), \tag{2}$$

where $\Phi : \{0,1\}^N \to \mathbb{R}$ is a penalty component, $\Phi(S) = \sum_{j=1}^m \lambda_j \phi_j(S)$, $\lambda = (\lambda_1, \cdots, \lambda_m)$, N refers to the total number of data records, and m refers to the number of penalty functions. Each penalty function $\phi_j(S)$ relates to a specific first order logic (FOL) rule ($w : r_{body} \to r_{head}$) that is composed of a conjunctive body ($r_{body} \equiv l_1 \wedge \cdots, \wedge\ l_n$), a single head ($r_{head} \equiv l$), and a weight of the rule $w \in \mathbb{R}$. Each atom $l_i \in \{1(\text{True}), 0(\text{False})\}$. The distance from satisfaction of a FOL rule is defined as "$\max\{0, r_{body} - r_{head}\}$". For example, in disease surveillance, suppose users want to simultaneously minimize the additive cost function of the subset S: $c(S) = \sum_{d\in S} c(d)$. This can be realized by defining the FOL rule (Rule 1): "$c(d) : isAnomaly(d) \to$ False", where d is a data record. Given that there are N different data records, this FOL rule has totally N instances: "$c(d_i) : isAnomaly(d_i) \to 0$ (False)", $i = 1, \cdots, N$. The penalty function ($\phi_1(S)$) related to this FOL rule is then defined as the sum of weighted distances of all its instances: $\phi(S) \equiv \sum_{i=1}^N c(d_i) \cdot \max\{0, isAnomaly(d_i) - 0\} = \sum_i^N c(d_i) \cdot S_i$, where $S_i \in \{0, 1\}$. If this FOL rule is the only FOL rule defined, then Problem 2 can be reformulated as:

$$\max_{S\in\{0,1\}^N} F(S) - \lambda_1 \cdot \sum_{i=1}^N c(d_i) \cdot S_i, \tag{3}$$

where λ_1 is the trade-off parameter. As a second example, suppose we have manually labeled a set of normal data records ($\mathcal{N}$) and a set of anomalous data records ($\mathcal{A}$). We can define two FOL rules (Rule 2 and Rule 3) as follows: "$+\infty : isAnomaly(d) \to$ 0 (False), $d \in \mathcal{N}$" and "$+\infty :\ \neg isAnomaly(d) \to$ 0 (False), $d \in \mathcal{A}$". The penalty functions related to these two FOL rules can be readily derived as "$+\infty \cdot \sum_{i\in\mathcal{N}} S_i$" and "$+\infty \cdot \sum_{j\in\mathcal{A}}(1 - S_j)$", respectively. As a third example, suppose the data is a graph ($\mathcal{G}(\{d_1, \cdots, d_N\}, \mathcal{E})$) and we want to simultaneously maximize the cut function of the subset S, where $\mathcal{E} \subseteq \{1, \cdots, N\} \times \{1, \cdots, N\}$. We can define two FOL rules (Rule 4 and Rule 5): "$1.0 :\ isAnomaly(d_i) \wedge neighbor(d_i, d_j) \to isAnomaly(d_j)$" and "$1.0 :\ \neg isAnomaly(d_i) \wedge neighbor(d_i, d_j) \to \neg isAnomaly(d_j)$". The penalty functions related to Rule 4 and Rule 5 can be derived as "$\sum_{(i,j)\in\mathcal{E}, S_i=1}(S_i - S_j)$ " and "$\sum_{(i,j)\in\mathcal{E}, S_i=0}(S_j - S_i)$ ", respectively. If we consider all the five FOL rules together, the resulting problem can be reformulated as:

$$\max_{S\in\{0,1\}^N} F(S) - \lambda_1 \cdot \sum_i^N c(d_i) \cdot S_i - \lambda_2 \cdot +\infty \cdot \sum_{i\in\mathcal{N}} S_i - \lambda_3 \cdot +\infty \cdot \sum_{j\in\mathcal{A}}(1 - S_j) - \lambda_4 \sum_{(i,j)\in\mathcal{E}, S_i=1} (S_i - S_j) - \lambda_5 \sum_{(i,j)\in\mathcal{E}, S_i=0} (S_j - S_i). \tag{4}$$

Note that, if we set $\lambda_4 = \lambda_5$, then the sum of the last two components is identical to cut function [10].

3 Inference

Given the testing data, predefined FOL rules, and trade-off parameters (λ), the inference task is to calculate the optimal subsets S that maximize the objective function (2). The trade-off parameters can be selected through cross validation. There are two directions to design efficient inference algorithms. First, we can directly solve the problem in the discrete space. In general, the penalty function $\Phi_\lambda(S)$ is neither sub-modular nor super-modular, and it is difficult to design efficient algorithms with theoretical guarantees. However, we can still apply greedy algorithms [11, 12] that were originally proposed for sub-modular or super-modular functions, and it can be readily proved that these algorithms guarantee to find a local optimum of Problem (2), although it is not clear about the closeness between the local optimum and global optimum.

Second, we may consider the relaxed version of Problem (2) in the numerical space, and find a local optimum using convex optimization and rounding techniques. Specifically, we consider Lukasiewicz t-norm

and its corresponding co-norm as the relaxation of the logical AND and OR, respectively. The logical conjunction ($\wedge$), disjunction ($\vee$), and negation ($\neg$) are relaxed as follows: $l_1 \wedge l_2 = \max\{0, l_1 + l_2 - 1\}$, $l_1 \vee l_2 = \min\{l_1 + l_2, 1\}$, and $\neg l_1 = 1 - l_1$. Then the resulting penalty function $\Phi_\lambda(S)$ becomes a convex function [15]. For the log-likelihood ratio function ($F(S)$), we consider its convex surrogate function as the relaxed version. Taking the expectation-based Possion statistic as an example, the convex surrogate function of its log-likelihood ratio function is the same form itself. By further relaxing $S \in \{0, 1\}^N$ as $S \in [0, 1]^N$, the resulting relaxed problem becomes a convex optimization problem, and efficient techniques such as alternating-direction method of multipliers (ADMM) [13, 14] can be applied to find a global optimum of the relaxed problem. By rounding of the numerical solution, we obtain a discrete solution of the original problem.

4 Conclusion

Traditional methods for anomalous pattern detection have been shown effective in a number of applications. However, these methods have very restrictive capability to integrate users' prior knowledge and special requirements that are required in many emerging applications. This paper presents a novel user friendly framework to address this challenge. For future work, we plan to evaluate the effectiveness and efficiency of this framework in a variety of real world applications, such as disease outbreak detection, road network congestion detection, and event detection in social media. In addition, we plan to consider higher order logic languages in our framework, such that more complicated constraints can be supported.

References

[1] Neill, D. B., Moore, A. W., Sabhnani, M. R. and Daniel, K. Detection of emerging space-time clusters. In *KDD*, pp. 218-227, 2005.

[2] Kulldorff, M. A spatial scan statistic. *Communs Statist. Theor. Meth.*, vol. 26, pp. 1481–1496, 1997.

[3] Neill, D. B. Fast subset scan for spatial pattern detection. *J. R. Statist. Soc. B*, vol. 74, pp. 337-360, 2012.

[4] McFowland III, E., Speakman, S., and Neill, D.B. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research*, vol. 14, pp. 1533–1561, 2013.

[5] Neill, D.B. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics*, pp. 8–20, 2009.

[6] Oliveira, D., Neill, D.B., Garrett Jr., J.H., and Soibelman, L. Detection of patterns in water distribution pipe breakage using spatial scan statistics for point events in a physical network. *Journal of Computing in Civil Engineering*, vol. 25(1), pp. 21–30, 2011.

[7] Chen, F. and Neill, D.B. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *KDD*, 2014 (To Appear).

[8] Neill, D.B., Moore, A.W., and Cooper, G.F. A Bayesian spatial scan statistic. In *NIPS*, vol. 18, pp. 1003–1010, 2006.

[9] Neill, D.B. Fast Bayesian scan statistics for multivariate event detection and visualization. *Statistics in Medicine*, vol. 30(5), pp. 455–469, 2011.

[10] Boykov, Y. and Kolmogorov, V. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *PAMI*, vol. 26(9), pp. 1124 - 1137, 2004.

[11] Buchbinder, N., Feldman, M., Naor, J., and Schwartz, R. A tight (1/2) linear-time approximation to unconstrained submodular maximization. *FOCS*, 2012.

[12] Iyer, R., Jegelka, S., and Bilmes, J. Fast Semidifferential-based Submodular Function Optimization. *ICML*, 2013.

[13] Bach, S., Broecheler, M. , Getoor, L. , and O'Leary, D. Scaling MPE inference for constrained continuous Markov random fields with consensus optimization. In *NIPS*, 2012.

[14] Bach, S., Huang, B., London, B., and Getoor, L. Hinge-loss Markov Random Fields: Convex Inference for Structured Prediction. In *UAI*, 2013.

[15] Kimmig, A., Bach, S., Broecheler, M., Huang, B., and Getoor, L. Hinge-loss Markov Random Fields: Convex Inference for Structured Prediction. In *NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 2012.