Nonparametric Anomaly Detection

Final Project Report Team Members: Changbai Liu, Chen Sun
EECS 545-001 Yiyang Wang, Haonan Zhu
Prof. Mert Pilanci Due Date: 12/17/2017

I Introduction

Anomaly detection intents to identify new incoming events that deviate from the given set of normal events, and it has wide range of applications including data security and system monitoring where failure to identify potential anomaly behavior such as intrusion or system failures will lead to catastrophic outcomes. This leads to two major performance measures about an anomaly detector: False Alarm Probability (The probability that a detector declared an event to be anomaly but the event is normal) and Detection Probability (The probability of correctly identify an anomaly event). In typical anomaly detection setting, all the events from normal set are treated as sample points from an unknown distribution, and anomaly events are considered to come from uniform random variable. However, in real application, there are seldom known information about the nominal distribution of interest which most of parametric method replies heavily upon. Therefore, nonparametric method has been of high interests in the past decades. The detector is completely characterized by its corresponding decision region, but the exact construction of optimal decision rule with unknown distribution is known to be intractable. The specification of optimality will be discussed in the background section. The main goal of this project is to apply efficient and accurate nonparametric anomaly detection methods [2], [10], [13] to approximate the optimum decision region for real world dataset.

In this Project, our group focused on the recent work by Lei [6], where a novel approach combining the idea of conformal prediction [11] with density estimation is proposed. We implemented the proposed method in[6], and we tested it on a bench-mark anomaly detection data set namely KDD Cup 1999 [1][7]. We compared our method with one-class support vector machine (SVM) mentioned in one of the midterm practice problem, where we are able to demonstrate the supreme performance of the algorithm proposed in [6] both quantitatively and qualitatively.

This report is organized as follows. Section II introduced the background of anomaly detection where we formally introduced the problem mathematically, and related work is being discussed. Section III provided a detailed explanation of the method proposed in [6]. Then section IV presented numerical results where we implemented the method in [6] and compared it with one-class SVM. Section V reported our conclusion about this work, and finally individual contribution to the project is reported in Section VI.

II Background and Related Work

This section introduces the formal problem statement of anomaly detection and provides an overview of related work.

Anomaly Detection

Denote the event space as Ω, 1 stands for anomaly event while 0 stands for normal event, the objective of anomaly detection is to derive a decision rule $f : \Omega \mapsto \{0, 1\}$ from a given set of normal events (data) $\{x_1, x_2...x_n\} \subseteq \Omega$. The decision rule can be completely characterized by its decision region, which is defined as $\Omega_f \subseteq \Omega$ such that:

$$f(x) = \left\{ \begin{array}{ll} 1, & x \notin \Omega_f \\ 0, & o.w. \end{array} \right\}$$

For typical numerical data that has continuous value, our event space will be Euclidean Space $\mathbb{R}^d$ or namely the feature space. For many of the application of interests, we can assume the observed normal data $\{x_1, x_2...x_n\}$ are realizations of independent i.i.d random variables from an unknown distribution $\mathbf{P}$. Then the anomaly detection problem can be interpreted as a supervised binary classification problem, where only one class of data (normal data) is provided. To make this connection explicitly, we formulate the problem as a sequential decision problem as follows. Given a data set $\{x_1, x_2...x_n\}$ corresponding to label 0 (nominal data), and assumed to be i.i.d draws from an unknown distribution $\mathbf{P}$. Then we want to classify incoming data points $\{x_{n+1}, x_{n+2}, ...\}$ into either label 0 (from the distribution $\mathbf{P}$ or label 1(not from the distribution).

For a given incoming data point x, there will be four possible of outcomes summarized in the table below:

Decision Truth	Declared Anomaly (Label 1, H_1)	Declared Norminal (Label 0,H_0)
x is anomaly (H_1)	Detection	Miss
x is nominal (H_0)	False Alarm	Rejection

With the above definition, there will be four probability of interests to evaluate an anomaly detection as $P_F = P(H_1|H_0)$ (False Alert Probability), $P_R = P(H_0|H_0)$ (Rejection Probability), $P_D = P(H_1|H_1)$ (Detection Probability), $P_M = P(H_0|H_1)$ (Miss Probability). Notice that $P_F + P_R = 1$ and $P_D + P_M = 1$, so there are only two degrees of freedom P_F and P_D for evaluating a hypothesis test. A describable anomaly detector would provide a high detection probability with a low false alarm probability, but there is implicit trade-off between the two, which can be shown mathematically [5]. Due to the trade-off, in anomaly detection problem, we are interested in derive an optimum detector (Highest detection probability) with a controlled false alert probability. A well known lemma from statistics community below will provide what we need:

Lemma II.1 (Neyman-Pearson Lemma) *When Performing a hypothesis test between two simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, the likelihood-ratio test which rejects H_0 in favor of H_1 when*

$$\Lambda(x) = \frac{L(x|\theta_0)}{L(x|\theta_1)} \leq \eta$$

where

$$P(\Lambda(x) \leq \eta | H_0) = \alpha$$

is the most powerful test at significance level α for a threshold η.
Note: Most powerful test means it has the highest detection probability, and significant level here is equivalent to the desired upon bound of false alarm probability. If we assumed the second distribution $L(x|\theta_1)$ is uniform, this lemma reduced to find the minimum volume set where the likelihood function integrated to α.

If we assume d=1, and $\mathbf{P}$ is a normal distribution with mean μ and variance σ^2. Then this problem will be similar to Linear Discriminative Analysis (LDA), with the difference be we are assuming the second distribution to be uniform instead of another Gaussian distribution.

Neyman-Pearson Lemma indeed provided critical insights for an optimum anomaly detector, but it cannot be extended to general setting because the distribution $\mathbf{P}$ is unknown.

Related Work

As stated in the Neyman-Pearson Lemma, one of the common parametric method of anomaly detection will be based on generalized likelihood ratio test. But these class of method would require knowledge of distribution $\mathbf{P}$, and even with given distribution, the search for optimum decision region could be intractable if we are working with high dimensional data.

The one-class SVM method mentioned in one of the practice problem for midterm exam is one of the popular nonparametric method to perform anomaly detection, whose performance will be compared with conformal prediction in Section IV. However, like general SVM methods, it has draw backs because there is no standard rule in terms of kernel function selection.

Current state-of-art nonparametric anomaly detection includes, .In [3][2][13] utilized graphical approach such as k-nearest neighbor graph and minimum spanning tree, and they can be interpreted as conformal prediction with various conformal score as stated in [6]. The conformal prediction method [9][11] is a general method that provides nonparametric predictions using exchangability. In [6], this method is combined with kernel density estimator to construct conformal prediction set that satisfy two optimum property. This proposed method can alternatively be seen as an anomaly detector where The optimum properties can be interpreted as controlled false alert rate and asymptotically converges to the uniform-most-powerful test under the assumption that anomaly is uniformly distributed. We will describe this general method and its connection to anomaly detection in Section III. Section IV reports numerical results.

III Conformal Prediction

Exchangeability

Formal Definition: A finite sequence of random variables $X_1, X_2, ..., X_n$ is Exchangeable if for any finite permutation σ of the indices 1,2,...,n, the joint probability distribution of the permuted sequence $X_{\sigma(1)}, X_{\sigma(2)}, ..., X_{\sigma(n)}$ is the same as the joint probability distribution of the original sequence:

$$\mathbb{P}(X_1, X_2, ...X_n) = \mathbb{P}(X_{\sigma_{(1)}}, X_{\sigma_{(2)}}, ..., X_{\sigma_{(n)}}), \forall\sigma$$

One of the important implication of the exchangeability is that the marginal probability density function is also exchangeable. In particular, for any given event E, the marginal probability $\mathbb{P}(X_i \in E) = \mathbb{P}(X_j \in E)\ \forall i, j$

For example, Given A finite sequence of independent Bernoulli trials $Z_1, Z_2, ...Z_n$ with probability p. Then by exchangeability the probability of the sequence start with 01 is the same as 10.

Conformal Prediction

Given an i.i.d sample $\mathbf{X} = \{X_1, ..., X_n\}$ from $\mathbf{P}$. We are interested to quantify how likely a fixed point $x \in \mathbb{R}^d$ is a realization of X_{n+1}. A conformal score is a real-valued function that measures similarity between a given data point to the whole data set, and it is required to be symmetric in the entries of the whole data set (i.e the order of $x_1, ..., x_n$ does not matter. Formally, $\sigma_i = \sigma(\{X_1, ..., X_{n+1}\}, X_i)$ measures how similar X_i is to $\{X_1, ..., X_{n+1}\}$. Then by exchangability, under $H_0 : X_{n+1} = x$, the ranks (The order number $\sum_{j=1}^{n+1} \mathbf{1}[\delta_j \leq \delta_i]$) of σ_i is uniformly distributed among $\{1, ..., n+1\}$ with probability $\frac{1}{n+1}$. Thus we can compute the following P-value:

$$\pi_n(y) = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{1}[\delta_j \leq \delta_{n+1}]$$

And $\pi_n(y)$ is uniformly distributed among $\{\frac{1}{n+1}, \frac{2}{n+1}..., 1\}$, and hence for $\alpha \in (0,1)$, we have $\mathbb{P}(\pi_n(y) \leq \tilde{\alpha}) = \tilde{\alpha} \leq \alpha$, where $\tilde{\alpha} = \lfloor (n+1)\alpha \rfloor / (n+1) \approx \alpha$. We obtained our following decision rule with controlled false alert rate: declare H_0 if $\mathbb{P}(\pi_n(y) \geq \tilde{\alpha}$, and reject H_0 otherwise.

Kernel Density Estimator

IV Numerical Result

V Conclusion

VI Division of Work

Method Data Type	Changbai Liu	Chen Sun	Yiyang Wang	Haonan Zhu
Literature Study			✓	✓
Implementation of Algorithms	✓	✓	✓	✓
Validation	✓	✓		

References

[1] A. Asuncion and D.J. Newman (2007). UCI machine learning repository.

[2] K. Sricharan, R. Raich, and A. O. Hero III (2012b), Estimation of nonlinear functionals of densities with confidence, Information Theory, IEEE Transactions on, vol. 58, no. 7, pp. 4135–4159.

[3] A. O. Hero III (2006). Geometric entropy minimization (GEM) for anomaly detection and localization. In *Advances in Neural Information Processing Systems 19.*

[4] E.L. Lehmann and J.P. Romano (2005). Testing Statistical Hypotheses, 3rd Edition. Springer, New York.

[5] R. W. Keener (2010). Theoretical Statistics: Topics for a Core Course, 1st Edition. Springer, New York.

[6] J. Lei, J. Robins, and L. Wasserman (2013). Distribution-free prediction sets. *Journal of the American Statistical Association* **108**(501): 278-287.

[7] S. Rayana (2016). ODDS Library [http://odds.cs.stonybrook.edu]. Stony Brook, NY: Stony Brook University, Department of Computer Science.

[8] C. Scott and R. Nowak (2006). Learning minimum volume sets. In *Machine Learning Res* **7**: 665-704.

[9] G. Shafer and V. Vovk (2008). A tutorial on conformal prediction. In *JMLR.*

[10] K. Sricharan and A. O. Hero III (2011). Efficient anomaly detection using bipartite k-NN graphs. In *Advances in Neural Information Processing Systems 24.*

[11] V. Vovk, A. Gammerman, and G. Shafer (2005). Algorithmic learning in a random world. *Springer.*

[12] J. E. Yukich, Probability theory of classical Euclidean optimization problems. 1998.

[13] M. Zhao and V. Saligrama (2009). Anomaly detection with score functions based on nearest neighbor graphs. In *Advances in Neural Information Processing Systems 22.*