Nonparametric Anomaly Detection

Project Proposal
EECS 545-001
Prof. Mert Pilanci

Team Members: Changbai Liu, Chen Sun
Yiyang Wang, Haonan Zhu
Due Date: 11/03/2017

Problem Statement

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. In typical anomaly detection setting, all the events from normal set are treated as sample points from an unknown distribution. In real application, there are seldom known information about the nominal distribution of interest. Therefore, nonparametric method has been heavily investigates in the past decades. To control the false alarm rate, the concept of level set is widely used as stated in well-known Neyman-Pearson Lemma to achieve optimum detection efficiency, and the detector is completely characterized by its corresponding decision region. However, the exact construction of optimal decision region with unknown distribution is known to be intractable. The main goal of this project is to investigate existing work on efficient and accurate nonparametric anomaly detection methods that approximate the optimum decision region.

Project Description

The main reference our group is going to focus on is recent work by Lei [2], where a novel approach combining the idea of conformal prediction [12] with density estimation is proposed. With the general frame work of conformal prediction, there are a couple of existing nonparametric methods [10][13] can be interpreted as graph-based conformal prediction.

In [2], the author investigated the proposed method for classification problem. Our group intents to implement the method specifically in the context of anomaly detection, and compare its performance with relevant methods as in [10][13].

We plan to perform experiments on two data set. One is KDD Cup 1999 Data from [1] which is a widely used dataset to test network intrusion detector. Second is Benchmark Dataset for Time Series Anomaly Detection [4][7] from Yahoo, which is a data set for detecting unusual traffic on Yahoo servers. The performance of algorithms will be evaluated based on both averaged AUC (Area under ROC curve) and processing time (a total of training and test time) as in [10].

Anticipated Division of Work

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

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

[4] N. Laptev and S. Amizadeh (2015). Online dataset for anomaly detection. http://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70.

[5] .T. Liu, K.M. Ting, and Z.H. Zhou (2008). Isolation forests. In *Proceedings of ICDM 2008.*

[6] M. Noshad, K. R. Moon, S. Y. Sekeh, and A. O. Hero III (2017). Direct estimation of information divergence using nearest neighbor ratios. In *IEEE International Symposium on Information Theory.*

[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. 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] K.M. Ting, G.T. Zhou, F.T. Liu, and S.C. Tan (2010). Mass estimation and its applications. In *Proceedings of 16th ACM SIGKDD.*

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

[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.*