Mathematical Problems of Computer Science 52, 21–29, 2019.

UDC 519.8

Information - Theoretic Methods for Anomaly Detection

Mariam E. Haroutunian and Tigran S. Badasyan

Institute for Informatics and Automation Problems of NAS RA
e-mail: armar@ipia.sci.am; tigranbadasyan@gmail.com

Abstract

Maintaining the security of digital systems with a huge amount of data is one of the main concerns of IT specialists in these times. Anomaly detection in systems is one of the solutions to overcome this challenge. Anomaly detection means finding patterns that are not normal or deviate from normal behavior in a system. Anomaly detection has various applications in bio-informatics, image processing, cyber security, security for databases, etc. There are many groups of methods that are used for anomaly detection including statistical methods, neural network methods and information theoretic methods. In this paper we survey pros and cons of anomaly detection based on information theoretic techniques.

Keywords: Anomaly / outlier detection, Entropy, Relative entropy, Local-search heuristic algorithm LSA.

1. Introduction

Anomaly or outlier is an important concept of the data analysis. Anomaly is defined as a deviation from normal data. It means that the data object is not similar to the other observations in the data set. It is very important to detect these objects during the data analysis to treat them differently from the other data. The anomaly detection methods are widely used for the following purposes:

- Credit card (and mobile phone) fraud detection;
- Suspicious Web site detection;
- Whole-genome DNA matching;
- ECG-signal filtering;
- Suspicious transaction detection and so on.

The investigation of anomaly detection techniques is very important in such fields as

- Medical and public health;
- Decision making;
- Business intelligence;
- Industrial Damage Detection

and others.

The anomaly detection problem has become a recognized rapidly-developing topic of the data analysis. A significant number of methods have recently been proposed. Many surveys and studies are devoted to this problem, see for example [1] - [4]. They concern different aspects of anomaly detection. [1] reviews traditional outlier detection algorithms, [2] overviews the existing research for the problem of detecting anomalies in discrete sequences, [3] is focused on ensemble learning ones, while [4] only involves the latest and popular anomaly detection methods for the data with high dimensionality and mixed types, on which the classical detection methods cannot operate very well.

There is no widely acceptable formal definition of the anomaly. The precise definition of the outlier depends on the specific problem and its data representation. At an abstract level, an anomaly is defined as a pattern that does not conform to expected normal behavior. A straightforward anomaly detection approach, therefore, is to define a region representing normal behavior and declare any observation in the data that does not belong to this normal region as an anomaly. But several factors make this approach very challenging.

Specific formulation of the problem is determined by several different factors such as the nature of the input data, the availability of labels, the constraints and requirements induced by the application domain. This justifies the need for the broad spectrum of anomaly detection techniques. Based on the research area the majority of anomaly detection techniques can be categorized into classification, nearest neighbor, clustering, statistical and information theoretical techniques.

Each of the large number of anomaly detection techniques has it's own unique strengths and weaknesses. It is important to know which anomaly detection technique is best suited for a given problem. Given the complexity of the problem space, it is not feasible to provide such an understanding for every anomaly detection problem.

None of the previous surveys, however, focuses on information theoretical detection methods. In this paper we survey the state-of-the-art techniques of anomaly detection based on information theory. The aim is to explore the use of tools/proposals of Information Theory, which is capable of being used to solve the problem of anomaly detection from new perspectives.

2. Some Information Theoretic Measures

Entropy is an important concept in information theory and communication theory [5]. It measures the uncertainty of a collection of data items:

$$H(X) = -\sum_{x} p(x) \log p(x),$$

for a data set X of items x with probabilities $p(x)$. Entropy specifies the number of bits required to encode and transmit the classification of data item.

For anomaly detection entropy can be used as a measure of the regularity of audit data. The smaller the entropy, the more regular is the data set. High regularity data can predict future events because events repeated many times in the current data set are likely to appear in the future.

Conditional entropy of X given Y is

$$H(X|Y) = -\sum_{x,y} p(x,y) \log p(x|y),$$

where $p(x, y)$ is the joint probability of x and y and $p(x|y)$ is the conditional probability of x given y. This measure can be used for anomaly detection as the value of dependence regularity. As in case of entropy, the smaller the conditional entropy, the better.

Kullback-Leibler divergence also known as **relative entropy** between two probability distributions on the same X is

$$D(P \parallel Q) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

For anomaly detection KL divergence can be used to measure the distance of regularities between the training data set and the test data set. Again, the smaller the relative entropy, the better.

3. Information Theory Based Anomaly Detection Techniques

Information theoretic techniques analyze the information content of a data set using different information theoretic measures such as entropy, relative entropy, and so on. Such techniques are based on the following key assumption: anomalies in data induce irregularities in the information content of the data set.

In [6] the information-theoretic measures are proposed for anomaly detection in the following general approach:

- Measure the regularity of the data and perform the appropriate data transformation. Iterate this step if necessary so that the data set used for modeling has high regularity.
- Determine how the model should be built, i.e. how to achieve the best performance or the optimal performance/cost trade-off, according to the regularity measure.
- Use relative entropy to determine whether a model is suitable for a new data set (e.g., from a new environment).

Different sequence lengths are used to show the relationship between regularity and detection performance. In practice, one can simply compute the regularity of a given data set and determine how to build a model, because computing regularity, in general, is much more efficient than computing a model. This is one of the advantages of this approach in comparison with the others, where there is no quideline for building a model and explaining its performance. However, the relationship between regularity and detection performance was shown for the classifier model, while there are other probabilistic algorithms, that can be used for anomaly detection. How the information-theoretic measures can be used for these algorithms is an open question.

Some real-life applications contain *categorical data*, however, most of the outlier algorithms are focused on numerical data and do not perform well when applied to categorical data. The problem of outlier detection in categorical data is considered in [7]. Here entropy is used to measure the degree of disorder of a dataset. The optimization problem is described as follows: finding a subset of k objects such that the expected entropy of the resultant dataset after the removal of this subset is minimized. The greedy optimization scheme that uses local search heuristic is studied for quality - time tradeoff. As a result the **Local search algorithm (LSA)** was presented, which has been shown to be the best in detecting outliers both in terms of accuracy and speed when compared with other techniques.

Working of LSA can be defined as follows.

- For a dataset DS, k outliers are to be detected using entropy,
- The number of outliers to be generated (k) is user defined.

- The initial set of outliers (SO) is empty and all the dataset's records are marked as nonoutliers.

- k scans are carried out to select k records as outliers.

- During each scan, each record tagged as a non-outlier is temporarily removed from the dataset and the change in entropy is calculated.

- The record that achieves the maximum decrease in entropy by removing that record, is selected as an outlier and added to SO.

- This continues for each scan until the size of OS reaches the defined value of k.

As the optimal number of outliers varies from dataset to dataset, it is not possible to predict the number of outliers or to define a standard or fixed number of outliers that can be applicable for every dataset. This is the disadvantage of LSA.

In [8] an outlier detection technique for categorical data is proposed which is based on entropy and called Automated Entropy Value Frequency (AEVF). It requires no user input and will always generate the optimal number of outliers. This is the extended version of the LSA, which introduces the new terms: entropy difference gap and max entropy gap. The **entropy difference gap** is the difference in the values of change of entropy between one record and the next. The **max entropy gap** is the maximum entropy difference gap that can exist as defined by the user. If the entropy difference gap becomes larger than the max entropy gap, the algorithm terminates.

The algorithm is a two-step process:

- Generating entropy change values. The change in entropy value for each record in a data set is generated and stored in a table. Once all records have been processed, the table is updated so that the record with the maximum entropy change value is at the top, followed by the other records in descending order of entropy change values.

- Generating outliers. The entropy difference gap is determined and then compared with the max entropy gap for each value from the top of the table downwards. If the entropy difference gap is less than or equal to the max entropy gap then the algorithm continues carrying out comparisons down the table, otherwise the algorithm terminates and all the records up to that point are added to OS, which is then displayed as the output.

The disadvantage of this approach is that it is not possible to define a standard max entropy gap which could be applied to every data set. In [8] an automated algorithm is proposed by combining the two above algorithms, which requires only the dataset as input and does not require either the number of outliers or the max entropy gap as user input.

AEVF algorithm is:

- Finding the optimal max entropy gap. The change in entropy value for each record in the dataset is generated and stored in a table. The average entropy change is determined, which is set as the max entropy gap. Once all records have been processed, the table is updated to show the record with the maximum entropy change at the top of the table, followed by descending values of entropy change.

- Finding the optimal number of outliers. The entropy difference gap is determined and then compared with the max entropy gap. If the entropy difference gap is less than or equal to the max entropy gap then the algorithm continues working down the table carrying out the comparison. Otherwise the algorithm terminates and all the records up to that point are added to OS, which is then displayed as the output.

A novel, parameter-free method, **COMPREX**, for identifying anomalies using pattern-based **compression** is introduced in [9]. Compression-based techniques have been explored mostly in communications theory, for reduced transmission cost and increased throughput and in databases, for reduced storage cost and increased query performance. In the mentioned paper the proposed method finds a collection of dictionaries that describe the norm of a database succinctly, and subsequently flags those points dissimilar to the norm with a high compression cost as anomalies. This approach exhibits four key features:

1) it is parameter free; it builds dictionaries directly from data, and requires no user - specified parameters such as distance functions or density and similarity thresholds,

2) it is general; it works for a broad range of complex databases, including graph, image and relational databases that may contain both categorical and numerical features,

3) it is scalable; its running time grows linearly with respect to both database size as well as number of dimensions, and

4) it is effective; experiments on a broad range of datasets show large improvements in both compression, as well as precision in anomaly detection, outperforming its state-of-the-art competitors.

Techniques focused on anomaly detection in *graph-based data* have recently been the subject of attention . A general, comprehensive survey of state-of-the-art methods for anomaly detection in data represented as graphs is provided in [10].

Here we focus only on information -theoretic approaches. In [11] the authors introduced two techniques for graph-based anomaly detection. The first, **anomalous substructure detection**, looks for specific, unusual substructures within a graph. In the second method, **anomalous subgraph detection**, the graph is partitioned into distinct sets of vertices (subgraphs), each of which is tested against the others for unusual patterns. Using the concept of conditional entropy a measure of graph regularity called **conditional substructure entropy** has been introduced to define the number of bits needed to describe an arbitrary substructures surroundings. A substructure is a connected subgraph of the overall graph. By surroundings, authors are referring to the edges and vertices adjacent to the substructure. The surroundings can be thought of as a set of extensions to the substructure; an extension of a substructure is defined to be the addition of either a single vertex (along with the edge connecting it to the substructure), or a single edge within the substructure.

Let Y be defined to contain all n-vertex substructures within the graph, and X contain all extensions of the substructures in Y. For a given substructure $y \in Y, P(y)$ is defined as the number of instances of y in G, divided by the total number of instances of all n-vertex substructures. For particular substructures $x \in X, y \in Y$, $P(x|y)$ is defined to represent the percentage of instances of y that extend to an instance of x. By analogy of the conditional entropy for string data, the **conditional substructure entropy** is defined as

$$H(X|Y) = -\sum_{x,y} p(y)[p(x|y)\log p(x|y) + (1-p(x|y))\log(1-p(x|y))].$$

Entropy and KL divergence methods have been regarded as effective methods for detecting *abnormal traffic* based on IP address-distribution statistics or packet size-distribution statistics [12] - [14].

In [15] two new and effective anomaly-based detection metrics (generilized entropy and information distance) were proposed, which identify DDoS attacks early and accurately. An effective IP traceback scheme was proposed based on an information distance metric that can trace all attacks back to their own local area networks (LANs) in a short time.

The **Renyi or generalized entropy** of order α is defined as follows:

$$H_\alpha(x) = \frac{1}{1-\alpha}\log_2(\sum_i p_i^\alpha), \ \ \alpha \geq 0, \ \ \alpha \neq 1.$$

The **Renyi divergence** between probability distributions P and Q is:

$$D_\alpha(P||Q) = \frac{1}{\alpha-1}\log_2(\sum_i p_i^\alpha q_i^{1-\alpha}), \ \ \alpha \geq 0.$$

The Renyi divergence, as the generalized divergence, can deduce many concrete divergence forms according to different values of order α.

The **information distance** is the symmetricized metric

$$D_\alpha(P,Q) = D_\alpha(P||Q) + D_\alpha(Q||P).$$

It is significant that this metrics can improve the systems' detection sensitivity by adjusting the value of order α of the generalized entropy and information distance metrics. As the proposed metrics can increase the information distance between the attack traffic and the legitimate traffic, they can effectively detect low-rate DDoS attacks early and reduce the false positive rate clearly. As a result [15] the proposed information metrics improve the performance of low-rate DDoS attacks detection and IP traceback over the traditional approaches.

As in [9], universal source coding has been used for anomaly detection in various papers mostly based on comparing code length, see for example, [16] - [20]. The comparison is done by a measure of entropy rate. The problem is that there are many dissimilar sources that have the same entropy rate. To overcome this issue a new approach based on the notion of **atypicality** is suggested in [21].

Most of the value in the information in some applications is in the parts that deviate from the average, that are unusual, atypical. Atypicality is defined as data that can be encoded with fewer bits in itself rather than using the code for the typical data. In [21] the authors show that this definition has good theoretical properties and develop an implementation based on universal source coding. This idea of finding alternative explanations for data rather than measuring some kind of difference from typical data is what separates this method from usual approaches in outlier detection and anomaly detection. Atypicality is purely a property of data and therefore there are no misses or false alarms: data is atypical or not.

4. Conclusion

The present work studied main information theoretic approaches for anomaly detection applications. Information theory provides a universal approach instead of looking at specific statistics of data.

The advantages of information theoretic techniques are as follows: (1) They can operate in an unsupervised setting. (2) They do not make any assumptions about the underlying statistical distribution for the data.

References

[1] V.Chandola, A.Banerjee and V. Kumar, "Anomaly detection: a survey", *ACM Computing Surveys*, vol. 41, no. 3, pp. 158, 2009.

[2] V.Chandola, A.Banerjee and V.Kumar, "Anomaly detection for discrete sequences: a survey", *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 823 - 839, 2012.

[3] C. C. Aggarwal and S.Sathe, Outlier Ensembles. An introduction—, Springer, 2017.

[4] X. Xu, H.Liu and M.Yao, *Recent Progress of Anomaly Detection*, Hindawi, Complexity, 2019.

[5] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd edition, *A Wiley-Interscience Publication*, 2006.

[6] W. Lee and D.Xiang, "Information-theoretic measures for anomaly detection', *In Proceedings of the IEEE Symposium on Security and Privacy*, IEEE Computer Society,pp. 130 - 143, 2001.

[7] Z.He, S.Deng, Xu X., "An optimization model for outlier detection in categorical data", *Advances in Intelligent Computing.*, Lecture Notes in Computer Science, vol 3644, Springer, pp. 400 - 409, 2005.

[8] U. Qamar, "Automated entropy value frequency (AEVF) algorithm for outlier detection in categorical data", *12th WSEAS Intern. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases*, Cambridge, UK, 2013.

[9] L. Akoglu, H. Tong, J. Vreeken and C. Faloutsos, "Fast and reliable anomaly detection in categorical data, *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, pp. 415424, 2012.

[10] L. Akoglu, H. Tong and D. Koutra, "Graph based anomaly detection and description: a survey", *Data Min Knowl Disc*, vol. 29, pp. 626 - 688, 2015.

[11] C. C. Noble and D. J. Cook, "Graph-based anomaly detection", *Proc. of the 9th ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, pp. 631636, 2003.

[12] Y. Gu, A. McCallum and D.Towsley, "Detecting anomalies in network traffic using maximum entropy estimation", *proc. ACM SIGCOMM Conf. Internet Measurement*, pp. 345-350, 2005.

[13] G. Nychis, V. Sekar, D. G. Anderson, H. Kim, H. Zhang, "An empirical evaluation of entropy-based traffic anomaly detection", *Proc. of the 8th ACM SIGCOMM Conference on Internet Measurement*, Greece, 2008.

[14] A. Wagner and B. Plattner, "Entropy based worm and anomaly detection in fast IP networks", *Proc. of the Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, WET ICE*, 2005.

[15] Y. Xiang, K. Li and W. Zhou, "Low-rate DDoS attacks detection and traceback by using new information metrics", *IEEE Trans. on information, forensics and security*, vol. 6, no. 2, pp. 426 - 437, 2011.

[16] F. Pan and W. Wang, "Anomaly detection based-on the regularity of normal behaviors, *Proc. 1st Int. Symp. Syst. Control Aerosp. Astronaut.*, pp. 1046-11046-6, 2006.

[17] E. E. Eiland and L. M. Liebrock, "An application of information theory to intrusion detection, *Proc. Fourth IEEE Int. Workshop Inf. Assurance*, 2006.

[18] C.-K. Han and H.-K. Choi, "Effective discovery of attacks using entropy of packet dynamics, *IEEE Netw.*, vol. 23, no. 5, pp. 412, 2009.

[19] N. Wang, J. Han and J.Fang, "An anomaly detection algorithm based on lossless compression, *Proc. IEEE 7th Int. Conf. Netw. Archit. Storage*, pp. 3138, 2012.

[20] H. Shahriar and M. Zulkernine, "Information-theoretic detection of SQL injection attacks, *Proc. IEEE 14th Int. Symp. High-Assurance Syst. Eng.*, pp. 4047, 2012.

[21] A. Host-Madsen, E. Sabeti and C. Walton, “Data discovery and anomaly detection using atypicality: theory”, *IEEE Trans. on Inform. Theory*, vol. 65, no. 9, pp. 5302 - 5322, 2019.

Submitted 10.06.2019, accepted 18.10.2019.

Անոմալիաների հայտնաբերման ինֆորմացիոն - տեսական մեթոդներ

Մարիամ Ե. Հարությունյան և Տիգրան Ս. Բադասյան

ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտ
e-mail: armar@sci.am, tigranbadasyan@gmail.com

Ամփոփում

Անոմալիաների հայտնաբերումը նշանակում է հազվագյուտ տվյալների նույնականացում, որոնք նորմալ չեն կամ շեղվում են համակարգում նորմալ պահվածքից: Անոմալիայի հայտնաբերումը ունի տարբեր կիրառություններ կենսաինֆորմատիկայի, պատկերների մշակման, կիբերանվտանգության, տվյալների բազայի անվտանգության ապահովման և այլ ոլորտներում: Կան բազմաթիվ մեթոդների խմբեր, որոնք օգտագործվում են անոմալիաների հայտնաբերման համար, ներառյալ վիճակագրական մեթոդները, նեյրոնային ցանցի մեթոդները և ինֆորմացիայի տեսության մեթոդները: Այս հոդվածում քննարկվում են հիմնական ինֆորմացիոն-տեսական մոտեցումներն անոմալիաների հայտնաբերման կիրառությունների համար: Ինֆորմացիայի տեսությունը տալիս է համընդհանուր մոտեցում՝ տվյալների վիճակագրությունը վերլուծելու փոխարեն:

Բանալի բառեր՝ անոմալիայի հայտնաբերում, էնտրոպիա, հարաբերական էնտրոպիա, տեղային որոնման էվրիստիկ ալգորիթմ LSA:

Информационно-теоретические методы для обнаружения аномалий

Мариам Е. Арутюнян и Тигран С. Бадася

Институт проблем информатики и автоматизации НАН РА
e-mail: armar@sci.am, tigranbadasyan@gmail.com

Аннотация

Обнаружение аномалий означает опознавание редких данных, которые не являются нормальными или отклоняются от нормального поведения в системе. Обнаружение аномалий имеет различные применения в биоинформатике, обработке изображений, кибербезопасности, безопасности для баз данных и т. д. Существует много групп методов, которые используются для обнаружения аномалий, включая статистические методы, методы нейронных сетей и теоретико-информационные методы. В настоящей работе рассматриваются основные теоретико-информационные подходы для приложений обнаружения аномалий. Теория информации предоставляет универсальный подход вместо того, чтобы смотреть на конкретные статистические данные.

Ключевые слова: Обнаружение аномалий, энтропия, относительная энтропия, эвристический алгоритм локального поиска LSA.