ISSN:0975-9646

Oulier Analysis Using Frequent Pattern Mining – A Review

Yagnik Ankur N.[#], Dr .Ajay Shanker Singh[*]

[#]*M. Tech(CE) Research Scholar ,RK University,Rajkot,Gujarat India*
[*]*Asst.Professor, Dept. of Computer Engg,R.K.University,Rajkot,Gujarat,India*

Abstract. An outlier in a dataset is an observation or a point that is considerably dissimilar to or inconsistent with the remainder of the data. Detection of such outliers is important for many applications and has recently attracted much attention in the data mining research community. In this paper, we present a new method to detect outliers by discovering frequent patterns (or frequent item sets) from the data set. The outliers are defined as the data transactions that contain less frequent patterns in their item sets. We define a measure called FPOF (Frequent Pattern Outlier Factor) to detect the outlier transactions and propose the Find FPOF algorithm to discover outliers. The experimental results have shown that our approach outperformed the existing methods on identifying interesting outliers.

Keywords :Frequent pattern mining ,Association rules ,Data mining research, Applications, FPOF, abnormalities, discordant, deviants, anomalies.

1.INTRODUCTION

An outlier is a data point which is significantly different from the remaining data. Hawkins formally defined [205] the concept of an outlier as follows:

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.”

Outliers are also referred to as *abnormalities*, *discordant*, *deviants*, or *anomalies* in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves in an unusual way, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems and entities, which impact the data generation process. The recognition of such unusual characteristics provides useful application-specific insihts.

Some examples are as follows:

Intrusion Detection Systems: In many host-based or networked computer systems, different kinds of data are collected about the operating system calls, network traffic, or other activity in the system.

This data may show unusual behavior because of malicious 2 *OUTLIER ANALYSIS* activity. The detection of such activity is referred to as intrusion detection.

Credit Card Fraud: Credit card fraud is quite prevalent, because of the ease with which sensitive information such as a credit card number may be compromised. This typically leads to unauthorized use of the credit card. In many cases, unauthorized use may show different patterns, such as a buying spree from geographically obscure locations. Such patterns can be used to detect outliers in credit card transaction data.

Interesting Sensor Events: Sensors are often used to track various environmental and location parameters in many real applications. The sudden changes in the underlying patterns may represent events of interest. Event detection is one of the primary motivating applications in the field of sensor networks.

Medical Diagnosis: In many medical applications the data is collected from a variety of devices such as MRI scans, PET scans or ECG time-series. Unusual patterns in such data typically reflect disease conditions.

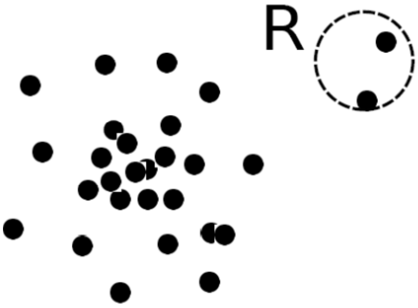
Law Enforcement: Outlier detection finds numerous applications to law enforcement, especially in cases, where unusual patterns can only be discovered over time through multiple actions of an entity. Determining fraud in financial transactions, trading activity, or insurance claims typically requires the determination of unusual patterns in the data generated by the actions of the criminal entity.

Earth Science: A significant amount of spatiotemporal data about weather patterns, climate changes, or land cover patterns is collected through a variety of mechanisms such as satellites or remote sensing. Anomalies in such data provide significant insights about hidden human or environmental trends, which may have caused such anomalies. In all these applications, the data has a “normal” model, and anomalies are recognized as deviations from this normal model. In many cases such as intrusion or fraud detection, the outliers can only be discovered as a sequence of multiple data points, rather than as an individual data point. For example, a fraud event may often reflect the actions of an individual in a particular sequence. The specificity of the sequence is relevant to identifying the anomalous event. Such anomalies are also referred to as *collective anomalies*, because they can only be inferred collectively from a *set or sequence* of data points. Such collective anomalies typically represent unusual *events*, which need to be discovered from the data.

The output of an outlier detection algorithm can be one of two types:

Most outlier detection algorithm output a score about the level of “outlierness” of a data point. This can be used in order to determine a ranking of the data points in terms of their outlier tendency. This is a very general form of output, which retains all the information provided by a particular algorithm, but does not provide a concise summary of the small number of data points which should be considered outliers. A second kind of output is a binary label indicating whether a data point is an outlier or not. While some algorithms may directly return binary labels, the outlier scores can also be converted into binary labels. This is typically done by imposing thresholds on outlier scores, based on their statistical distribution. A binary labeling contains less information than a scoring mechanism, but it is the final result which is often needed for decision making in practical applications.

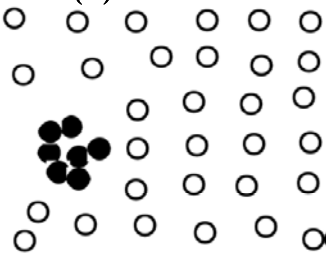
2.TYPES OF OUTLIERS (I)



Global Outlier

- Three kinds: *global, contextual* and *collective* outliers
- **Global outlier** (or point anomaly)
  - Object is O_g if it significantly deviates from the rest of the data set
  - Ex. Intrusion detection in computer networks
  - Issue: Find an appropriate measurement of deviation
- **Contextual outlier** (or *conditional outlier*)
  - Object is O_c if it deviates significantly based on a selected context
  - Ex. 80^o F in Urbana: outlier? (depending on summer or winter?)
  - Attributes of data objects should be divided into two groups
    - Contextual attributes: defines the context, e.g., time & location
    - Behavioral attributes: characteristics of the object, used in outlier evaluation, e.g., temperature
  - Can be viewed as a generalization of *local outliers*—whose density significantly deviates from its local area

2.1 Types of Outliers (II)



Collective Outlier

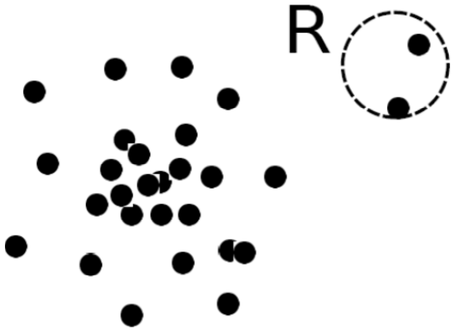
- **Collective Outliers**
  - A subset of data objects *collectively* deviate significantly from the whole data set, even if the individual data objects may not be outliers
  - Applications: E.g., *intrusion detection*:
    - When a number of computers keep sending denial-of-service packages to each other
  - Detection of collective outliers
    - Consider not only behavior of individual objects, but also that of groups of objects
    - Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects.
- A data set may have multiple types of outlier
- One object may belong to more than one type of outlier

3.CHALLENGES OF OUTLIER DETECTION

- Modeling normal objects and outliers properly
  - Hard to enumerate all possible normal behaviors in an application
  - The border between normal and outlier objects is often a gray area
- Application-specific outlier detection
  - Choice of distance measure among objects and the model of relationship among objects are often application-dependent
  - E.g., clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations
- Handling noise in outlier detection
  - Noise may distort the normal objects and blur the distinction between normal objects and outliers. It may help hide outliers and reduce the effectiveness of outlier detection
- Understandability
  - Understand why these are outliers: Justification of the detection

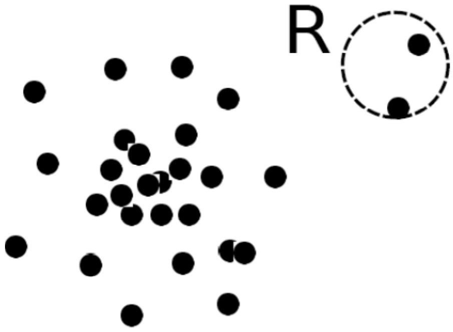
Specify the degree of an outlier: the unlikelihood of the object being generated by a normal mechanism.

4.Outlier Detection (1): Statistical Methods



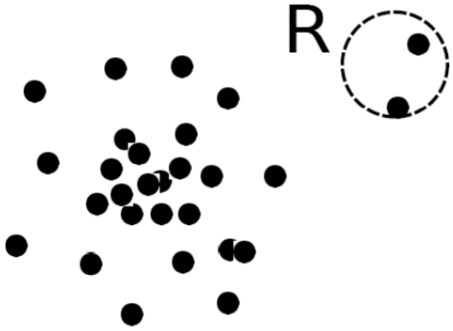
- Statistical methods (also known as model-based methods) assume that the normal data follow some statistical model (a stochastic model)
  - The data not following the model are outliers.
- Example (right figure): First use Gaussian distribution to model the normal data
  - For each object y in region R, estimate $g_D(y)$, the probability of y fits the Gaussian distribution
  - If $g_D(y)$ is very low, y is unlikely generated by the Gaussian model, thus an outlier
- Effectiveness of statistical methods: highly depends on whether the assumption of statistical model holds in the real data
- There are rich alternatives to use various statistical models
  - E.g., parametric vs. non-parametric

4.1 Outlier Detection (2): Proximity-Based Methods



- An object is an outlier if the nearest neighbors of the object are far away, i.e., the **proximity** of the object is **significantly deviates** from the proximity of most of the other objects in the same data set
- Example (right figure): Model the proximity of an object using its 3 nearest neighbors
  - Objects in region R are substantially different from other objects in the data set.
  - Thus the objects in R are outliers
- The effectiveness of proximity-based methods highly relies on the proximity measure.
- In some applications, proximity or distance measures cannot be obtained easily.
- Often have a difficulty in finding a group of outliers which stay close to each other
- Two major types of proximity-based outlier detection
  - Distance-based vs. density-based

4.2 Outlier Detection (3): Clustering-Based Methods



Normal data belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters

Example (right figure): two clusters

- All points not in R form a large cluster
- The two points in R form a tiny cluster, thus are outliers

Since there are many clustering methods, there are many clustering-based outlier detection methods as well

Clustering is expensive: straightforward adaption of a clustering method for outlier detection can be costly and does not scale up well for large data sets

5.Challenges for Outlier Detection in High-Dimensional Data

Interpretation of outliers

- Detecting outliers without saying why they are outliers is not very useful in high-D due to many features (or dimensions) are involved in a high-dimensional data set
- E.g., which subspaces that manifest the outliers or an assessment regarding the “outlier-ness” of the objects

Data sparsity

- Data in high-D spaces are often sparse
- The distance between objects becomes heavily dominated by noise as the dimensionality increases

Data subspaces

- Adaptive to the subspaces signifying the outliers
- Capturing the local behavior of data

Scalable with respect to dimensionality

- # of subspaces increases exponentially

6.VARIOUS OUTLIER DETECTION METHOD COMPARISON

Issues	Outlier Analysis Using FP Approach	Term Based Approach	Phrase Based Approach	Emerging Pattern Approach
Polysemy	Not found	Found	Found	Partially found
synonymy	Not found	Found	Found	Partially found
Info. filtering	Easy	Complex	Complex	Complex
Object noise	Very less	Less	Medium	More
Noise handling	Complex	Easy	Easy	Complex
PDS	Post processing required	Not required	Feature extraction	Concatenation required
Ambiguity	Very less	Less	More	More
Object divergence	More	n/a	n/a	n/a

7.CONCLUSION

Frequent pattern mining and outlier detection are two integral parts of data mining and have attracted attentions in their own fields. Based on frequent patterns, this paper has proposed a new outlier detection method. The effectiveness of the method was verified by the experimental results.

Using the same process and functionality to solve both frequent pattern mining and outlier discovery is highly desirable. Such integration will be a great benefit to business users because they do not need to worry about the selection of different data mining algorithms. Instead, they can focus on data and business solution. More importantly, some commercial data mining software do not provide the functionality of outlier discovery, hence it is easier to discover outliers directly using the frequent pattern mining results (since most commercial data mining software provide association mining module).

8.REFERENCES

[1] Gerd Stumme, Ra_k Taouil, Yves Bastide, Nicolas Pasquier, and Lot_ Lakhal. Computing iceberg con-cept lattices with t. Data & Knowledge Engineering,42(2):189{222, 2002.

[2] Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura. An e_cient algorithm for enumerating closed patterns in transaction databases. In Discovery Sci-ence, pages 16{31, 2004.

[3] J. Han, J. Pei, Y. Yin and R. Mao, “Mining frequent patterns without candidate generation: a frequent-pattern tree approach”, Data Mining and Knowledge Discovery, (2004), Vol. 8, No.1, pp. 53–87.

[4]J. Pei, J. Han, andW.Wang. Constraint-based sequential pattern mining in large databases.

[5] X. Yan, J. Han, and R. Afshar. CloSpan: Mining closed sequential patterns in large datasets.

[6] M. Zaki. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 40:31–60,2001.

[7] G. M. Landau and J. P. Schmidt. An algorithm for approximate tandem repeats.

[8] F. Masseglia, F. Cathala, and P. Poncelet. The PSP approach formining sequential patterns.

[9] A.M. Carvalho, A.T. Freitas, A.L. Oliveira, and M.-F. Sagot, “A Highly Scalable Algorithm for the Extraction of Cis-Regulatory Regions,” Proc. Asia-Pacific Bioinformatics Conf. (APBC), pp. 273-282, 2005.

[10] Y. Zhang and M.J. Zaki, “EXMOTIF: Efficient Structured Motif Extraction,” Algorithms for Molecular Biology, vol. 1, pp. 21-38, 2006.

[11] F. Fassetti, G. Greco, and G. Terracina, “Mining Loosely Structured Motifs from Biological Data,” IEEE Trans. Knowledge and Data Eng., vol. 20, no. 11, pp. 1472-1489, Nov. 2008.

[12] L. DS, “Transcription Factors: An Overview,” Int’l J. Biochemistry and Cell Biology, Vol. 29, no. 12, pp. 1305-1312, 2007.

[13]Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993ACM-SIGMODinternational conference on management of data (SIGMOD’93), Washington, DC, pp 207–216

[14]Agrawal R, Shafer JC (1996) Parallel mining of association rules: design, implementation, and experience. IEEE Trans Knowl Data Eng 8:962–969

[15]Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 1994 international conference on very large data bases (VLDB’94), Santiago, Chile, pp 487–499

structured data.