Methods to detect different types of outliers

Conference Paper · March 2016

DOI: 10.1109/SAPIENCE.2016.7684114

CITATIONS
0

READS
3,200

2 authors, including:

Dr SUVANAM Sasidhar Babu
Sree Narayana Gurukulam College of Engineering
51 PUBLICATIONS **99** CITATIONS

SEE PROFILE

METHODS TO DETECT DIFFERENT TYPES OF OUTLIERS

[1]Divya. D, [2] Dr. Suvanam Sasidhar Babu

[1] Department of Computer Science and Engineering, Adi Shankara Institute of Engineering & Technology, Kalady, Kerala, India

[2] Department of Computer Science and Engineering, Sree Narayana Gurukulam College of Engineering, Kadayiruppu, Kerala, India

[1]a.divya.d@gmail.com [2] sasidharmails@gmail.com

Abstract

Outliers are those data that deviates significantly from the remaining data. Outliers has emerging applications in irregular credit card transactions, used to find credit card fraud, or identifying patients who shows abnormal symptoms due to suffering from a particular type of disease. This paper gives an idea about the various approaches and techniques used in outlier detection and the areas in which outlier detection is used and also about how outlier detection is handled in higher dimensional data.

Keywords Outliers*, Outlier Mining, Tuples*

1. Introduction

There exist data objects that do not comply with the general behavior of the data. These are called outliers. Hawkins defines' an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism'[1]. Outliers are different from noise data. Noise data may be due to some random error. They need to be removed from the dataset. But outliers may contain some relevant information. This is due to the fact that" This is because of the fact that "one person's noise is another person's signal". Outliers may be result of variability that is inherent in the data. The salary of the manager of a company could naturally stand out as an outlier among the salary of the other employees in the firm. .i.e. salary of the manager may be high compared to that of other employees. This data appears like an outlier. But this data need not be removed.

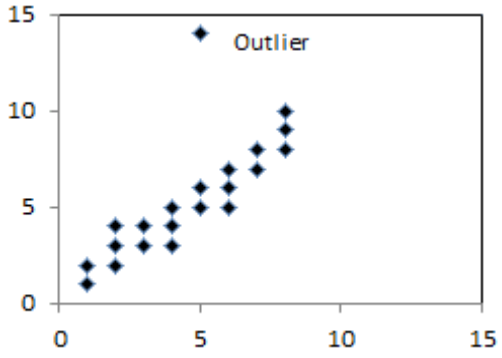


Fig.1 Outlier Detection

Applications of outlier detecion [2]

Fraud detection • Purchasing behavior of a credit card owner usually changes when the Purchasing behavior of a credit card owner usually changes when the card is stolen

Medicine

• Unusual symptoms or test results may indicate potential health problems of a patient

Sports statistics

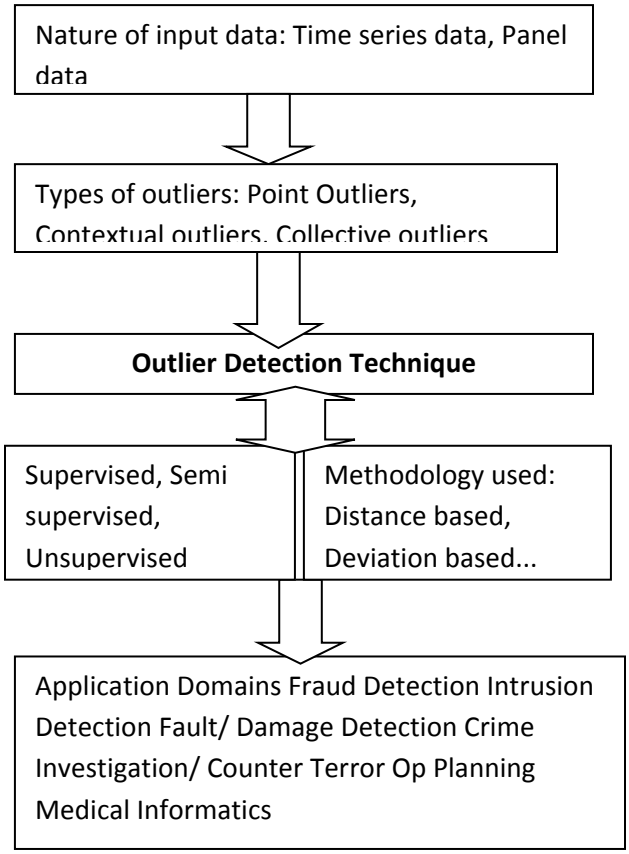
• In many sports various parameters are recorded for players in order to In many sports, various parameters are recorded for players in order to evaluate the players' performances• Outstanding (in a positive as well as a negative sense) players may be identified as having p abnormal parameter values

– Detecting measurement errors

• Data derived from sensors (e.g. in a given scientific experiment) may contain measurement errors

ISBN: 978-1-4673-8592-3

Outlier detection method summarized as follows.



2. Types of Outliers

An important aspect of an outlier detection technique is the nature of the desired outlier. Outliers can be classified into following three categories: 1) Point Outliers 2) Contextual Outliers 3) Collective Outliers. [3]

1) Point Outliers

If an individual data instance can be considered as anomalous with respect to the rest of data, then the Instance is termed as a point outlier. This is the simplest type of outlier and is the focus of majority of research on outlier detection

2) Contextual Outliers

If a data instance is anomalous in a specific context (but not otherwise), then it is termed as a contextual outlier.
in the credit card fraud detection with contextual as time of purchase. Suppose an individual usually has a weekly shopping bill of $100 except during the Christmas week, when it reaches $1000.

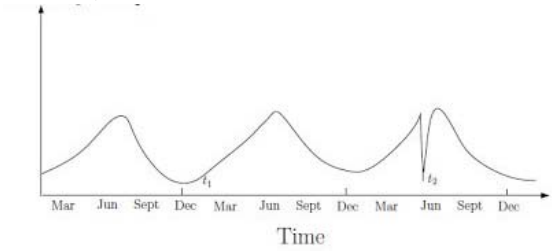


Fig 2: Contextual outliers

A new purchase of $1000 in a week in July will be considered a contextual outlier, since it does not conform to the normal behavior of the individual in the context of time.

3) Collective Outliers

If a collection of related data instances is anomalous with respect to the entire data set, it is termed as a collective outlier. The individual data instances in a collective outlier may not be outliers by themselves, but their occurrence together as a collection is anomalous. Collective outliers can occur only in data sets in which data instances are related.

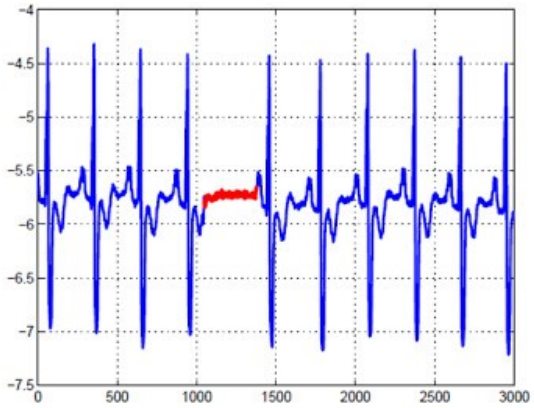


Fig 3: Collective outliers

ISBN: 978-1-4673-8592-3 Copyright notice: 978-1-4673-8594-7/16/$31.00 ©2016 IEEE

3. Types of supervision in outlier detection

Outlier detection methods can be divided into supervised, unsupervised and semi supervised outlier detection methods. [3]

Supervised outlier detection

Supervised methods make use of the training dataset which contains labels for normal data objects as well as for the outlier objects.ie. Whenever a test data come that data object will be divided into two classes. Those data that comply with properties of the normal data set in the training is classified as normal and those data that has a deviation is treated as outliers. For classification those data that fit with the model are considered as normal data objects and the data that has a deviation from the model is an outlier or we can say that is there are two labels normal and outliers and test data is divided into these two categories.

Unsupervised outlier detection

In unsupervised clustering outlier is not identified by the model instead the data objects are clustered and those data which are far away from each of these clusters are considered as outliers. Whenever the test data comes grouping is performed and outliers are detected by using this mechanism.

Semi supervised clustering

Such techniques assume the availability of labeled instances for only one class. It is often difficult to collect labels for other class. Techniques that assume availability of only the outlier instances for training are not very popular. Normal behavior is typically well-defined and hence it is easier to construct representative models for normal behavior from the training data.

4. Different Methods of outlier detection

4.1 Distance based Method

Normal data objects have a dense neighborhood Outliers are far apart from their neighbors, i.e., has a less dense neighborhood. Given a radius ε and a percentage π, this is the required density. A point p is considered an outlier if at most π percent of all other points have a distance to p less than ε.

```
Input: k, t, b, D
Output: Ok, the set of top t outliers in D
Initialization: c ← 0, O ← ∅
        while B ←get next block(D, b)<> ∅ do
        forall the b ∈ B do Nk(b) ← ∅;
        forall the i = 1 to |D| do
        x = getFromfile(i, D);
        forall the b ∈ B, b 6= x do
        if |Nk(b)| < k or dist(b, x) <
        maxdist(b, Nk(b)) then
        Nk(b) ← Update nbors(Nk(b), x, k);
        δk(b) ← max{kb − yk : y ∈ Nk(b)};
        if δk(b) < c then B ← B \ b;
        end
        end
        end
for b=1 to B do
newO ← newO S
[b; δk(b); Nk(b)];
end
Ok ← Find Top t(newO ∪ Ok, t);
c ← min{δk(y, D) : y ∈ O};
end
```

Algorithm 1: Distance based outlier detection

There are different types of distance based outlier methods [4].

Indexed-based Method

In index based method Index is used to search for neighbors of each object O within radius D around that object.. Once K (K = N(1-p)) neighbors of object O are found, O is not an outlier.

Nested-loop Method

In Nested loop method the buffer space is divided into two halves (first and second arrays). Break data into blocks and then feed two blocks into the arrays. Then directly computes the distance between each pair of objects, inside the array or between arrays Decide the outlier.

4.2 Distribution Method

Given a certain kind of statistical distribution like a Gaussian distribution [5] compute the parameters

ISBN: 978-1-4673-8592-3 Copyright notice: 978-1-4673-8594-7/16/$31.00 ©2016 IEEE

assuming all data points have been generated by such a statistical distribution. Outliers are points that have a low probability to be generated by the overall distribution. Normal data objects follow a known distribution and occur in a high probability region of this model. Outliers deviate strongly from this distribution. A key drawback of this category of tests is that most of the distributions used are uni variate. [6]But for many Knowledge discovery (KDD) applications, the underlying distribution is unknown. Fitting the data with standard distributions is costly, and may not produce satisfactory results.

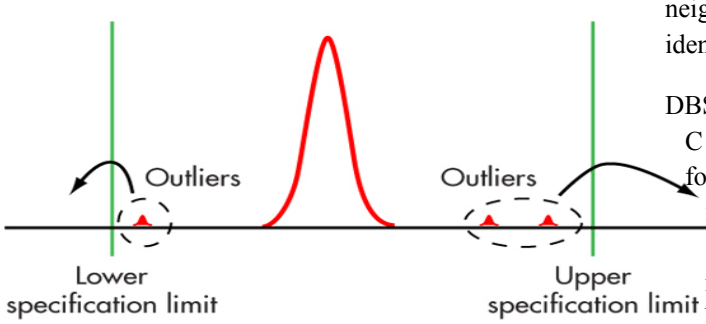


Fig.4 Distribution Based Outlier Detection

4.2 Depth based methods

Search for outliers at the border of the data space but independent of statistical distributions. This method [7] organizes data objects in convex hull layers. Outliers are objects on outer layers. In this each data object is represented as a point in a *k*-d space, and is assigned a depth. With respect to outlier detection, outliers are more likely to be data objects with smaller depths. Those data with higher depth are considered to be normal data points.

4.3 Deviation based methods

In deviation based method [8], given a set of data points Outliers are points that do not fit to the general characteristics of that set, i.e., the variance of the set is minimized when removing the outliers. Deviation based technique identify outliers by inspecting the characteristics of objects and consider an objects that deviates these features as an outlier deviation based outlier detection does not use statistical tests or distance based measures to identify exceptional objects, Instead, it identifies outliers by examining the main characteristics of objects in a group. [9]Objects that "deviate" significantly from this description are considered as outliers. Hence in this approach the term deviation is typically used to refer to outliers.

4.4 Density Based Methods

It relies on the local outlier factor (LOF) of each point [11], which depends on the local density of its neighborhood. The neighborhood is defined by the distance to the MinPts-th nearest neighbor. In typical use, objects with a high LOF are flagged as outliers. Tang introduced a connectivity based outlier factor (COF) scheme that improves the effectiveness of LOF scheme. The density around an outlier is considerably different to the density around its neighbors. By using this property the outliers are identified.

```
DBSCAN(D, eps, MinPts) {
  C = 0
  for each point P in dataset D {
    if P is visited
      continue next point
    mark P as visited
    NeighborPts = regionQuery(P, eps)
    if sizeof(NeighborPts) < MinPts
      mark P as NOISE
    else {
      C = next cluster
      expandCluster(P, NeighborPts, C, eps, MinPts)
    }
  }
}
```

Algorithm 2: Density based outlier detection

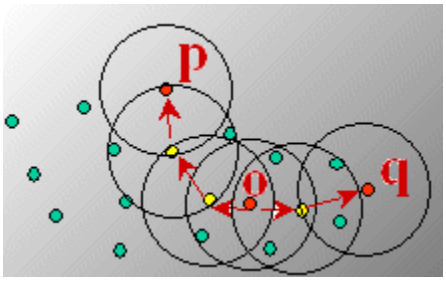


Fig.5. Density Based Outlier Detection

Distance-based outlier detection models have problems with different densities the main problem is how to compare the neighborhood of points from areas of different densities. If clusters of different densities are not clearly separated, LOF will have problems

ISBN: 978-1-4673-8592-3 Copyright notice: 978-1-4673-8594-7/16/$31.00 ©2016 IEEE

4.5 Clustering based methods

Outlier detection based methods on the fact that one person's noise is another person's signal. For clustering algorithms [12] noise data or rare events are outliers.

Input:
DATA, a dataset with k variables and n observations;
distance function d;
hierarchical algorithm h;
nc a number of clusters to use (entailing a level of cut of the hierarchy);
threshold t for the size of small clusters.
Output:
Out, a set of outlier observations.
Out ← φ
Obtain the distance matrix D by applying the distance function d to the observations in DATA
Use algorithm h to grow an hierarchy using the distance matrix D
Cut the hierarchy at the level l that leads to nc clusters
For each resulting cluster c Do
If sizeof(c) < t Then
Out ← Out ∪ {obs ∈ c}

Algorithm 3: Clustering based outlier detection

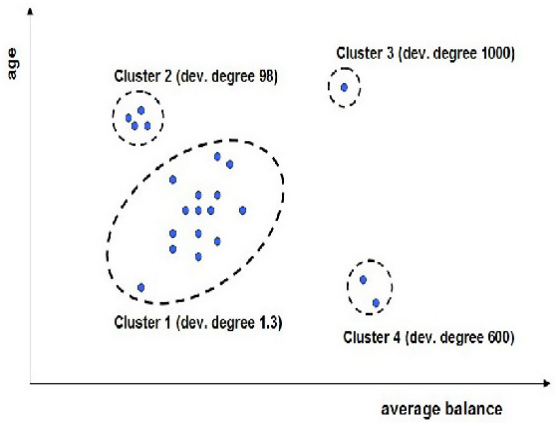


Fig.6. Clustering Based Outlier Detection

Most clustering algorithms are able to handle noise in the datasets. These refer to the returned noisy data as clustering-based outliers. Example clustering algorithms which also handle outliers are Birch, Clarans, Dbscan, Gdbscan, Optics and Proclus. However, the outliers are identified as by-products and are highly dependent on the algorithms used. In outlier analysis, we want to focus our efforts on outlier detections. In this case, finding outliers without the need of clustering operations is desirable.

4.6 Outlier detection in high Dimensional Data

Most of the outlier detection problems are in high dimensional domain which contains hundreds of dimensions. In high dimensional data sets only subspaces typically expose outliers. Many outlier mining algorithms use concepts of proximity in order to find outlier based on their relationship with rest of the data. However in high dimensional space[13], the data is sparse and notion of proximity fails to retain its meaningfulness. In fact, the sparcity of high dimensional data implies that every point is almost equally good outlier from perspective of proximity based definition.[14] When the data set is very sparse traditional concepts such as Euclidean distance between points, and nearest neighbor, become irrelevant Employing dissimilarity measures that can handle sparse data becomes imperative.[15] In addition, inspecting several, smaller views of the large, high-dimensional data can help uncover outliers, which would otherwise be masked by other outlier points if one were

6. Conclusion

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. There are a different variety of applications for outlier detection. Outliers are used in irregular credit card transactions, used to find credit card fraud, or identifying patients who shows abnormal symptoms due to suffering from a particular type of disease. Various approaches in outlier detection include local and global approaches.[16] There are different techniques used in outlier detection like distribution based distance based, deviation based, density based, depth based and clustering based methods. Outlier detection in high dimensional data uses different ways to handle different dimensions, like taking only numerical attributes first to find the outlier.

7. References

[1]. Kluwer Academic Publishers(2005),Outlier Detection, Knowledge Discovery Handbook,A Complete Guide for Practitioners and Researchers"

[2]. Hans-Peter Kriegel, Peer Kröger, Arthur Zimek:Outlier Detection Techniques.Tutorial at the

13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2009), Bangkok, Thailand, 2009

[3] Shin Ando, Clustering Needles in a Haystack: An Information Theoretic Analysis of Minority and Outlier Detection, Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, p.13-22, October 28-31, 2007

[4]. Knorr, E.M., Ng, R.T.: Finding Intensional Knowledge of Distance-based Outliers. In Proc. of VLDB, pp. 211-222. Scotland (1999)

[5]. Victoria Hodge and Jim Austin,(2004)A Survey of Outlier Detection Methodologies.

[6]. P. Fränti, O. Virmajoki, and V. Hautamäki. Graph-based agglomerative clustering. In Proceedings of The Third IEEE Int. Conf. on Data Mining, pages 525-528, Melbourne, Florida, November 2003.

[7].Yixin Chen, Xin Dang, HanxiangPeng,(2008) Outlier Detection: A Novel Depth Approach.

[8].Kluwer Academic Publishers(2005) Knowledge Discovery Handboo A Complete Guide for Practitioners and Researchers".

[9]. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, China Machine Press, Beijing, 2006

[10]. Jingke Xi , Outlier Detection Algorithms in Data Mining, Intelligent Information Technology Application, 2008. IITA '08. Second International Symposium (Volume:1), 2008

[11].M. Breunig, Hans-Peter Kriegel, Raymond T. Ng‡, Jörg Sander(2000) Local Outlier Factor:IdentifyingDensity-Based Local Outliers,Markus

[12]. Svetlana Cherednichenko(2005)Outlier detection in Clustering

[13].Charu C. Aggarwal, Philip S. Yu(2001) May 21-24,"Outlier Detection for High Dimensional Data", International conference, ACM SIGMOD, Santa Barbara, California USA.

[14]. Anna Koufakou · Michael Georgiopoulos,(2010) 20:259–289,A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes,Data Min Knowl Disc

[15]. FabrizioAngiulli and Clara Pizzuti, February (2005) vol. 17, no. 2, "Outlier Mining in Large High-Dimensional Data Sets", IEEE Transactions on knowledge and data engineering.

[16].Biao Huang a, Peng Yang a,b(2011),Finding key knowledge attribute subspace of outliers in high-dimensional dataset,Expert Systems with Applications 38.

ISBN: 978-1-4673-8592-3 Copyright notice: 978-1-4673-8594-7/16/$31.00 ©2016 IEEE