Proceedings of Machine Learning Research 71:33–42, 2017 KDD 2017 Workshop on Anomaly Detection in Finance

Spotlighting Anomalies using Frequent Patterns

Jaroslav Kuchař * JAROSLAV.KUCHAR@FIT.CVUT.CZ
Web Intelligence Research Group, Faculty of Information Technology, Czech Technical University in Prague,Thákurova 9, 160 00 Prague 6, Czech Republic

Vojtěch Svátek SVATEK@VSE.CZ
Department of Information and Knowledge Engineering, Faculty of Informatics and Statistics, University of Economics, W. Churchill Sq. 4, 130 67 Prague 3,Czech Republic

Abstract

Approaches for anomaly detection based on frequent pattern mining follow the paradigm: if an instance contains more frequent patterns, it means that this data instance is unlikely to be an anomaly. This concept can be used in financial industry to reveal contextual anomalies. The main contribution of this paper is an approach that includes a novel formula for computation of anomaly scores. We evaluated the proposed approach on baseline datasets and present a use case on a real world financial dataset. We also propose a way how to explain the anomaly to the users. Implementations of the evaluated algorithms and experiments are available online in R.

Keywords: frequent pattern mining, anomaly detection, financial data

1. Introduction

The anomaly detection task has become popular in many domains and encompasses several techniques of revealing instance in data that deviate from others. The application domain ranges from medicine, security or engineering to fraud detection. There are many techniques that are mainly built on top of statistical or machine learning approaches and are primarily divided into supervised and unsupervised (Chandola et al., 2009; Hodge and Austin, 2004; Aggarwal, 2013). Unsupervised anomaly detection is usually conceived either as extreme value analysis, which usually fits only to univariate data, or as proximity-based approaches employing clustering or density-based algorithms (Goldstein and Uchida, 2016). In this paper we however focus on unsupervised approaches based on frequent pattern mining (FPM). The main idea behind the approaches based on frequent pattern mining is that if an instance contains more frequent patterns, it is unlikely to be an anomaly. The presence or absence of the frequent patterns is then used to compute an overall anomaly score for each instance.

We present our work as framed by the analysis of financial data (e.g. budgets, EU funding etc.) within an EU-funded project called *OpenBudgets.eu*. Anomaly detection in financial data mostly utilizes approaches that require numeric values and rely on statistical or clustering techniques (Leung et al., 2006). Since the financial data includes multiple data types, we focused on algorithms that exploit the mining of frequent patterns and their

* also affiliated with Department of Information and Knowledge Engineering, Faculty of Informatics and Statistics, University of Economics, W. Churchill Sq. 4, 130 67 Prague 3,Czech Republic

availability or unavailability for the detection of anomalies. As an anomaly we can thus, for instance, consider an amount of money that was spent by an institution in an unusual category. We are therefore interested not only in unusual amounts, but also in contextual approaches that handle combinations with other descriptors of the data instance. The following main motivations lead us to algorithms based on the frequent pattern mining: 1) Since the data in the financial industry usually does not contain any ground truth information or it is generally difficult to have prior feedback from domain experts, we focus on unsupervised techniques (assuming posterior expert evaluation of the results top-rated by the algorithm). 2) We cannot rely only on statistical methods that analyze extreme values. Namely, the presence of extreme values may strongly depend on associated contextual information such as the category of the spending. 3) From the data type point of view, there is a requirement of processing a mixture of different data types. 4) Last but not least, most existing approaches only focus on identification of anomalies. In our approach we also emphasize the explanations of reasons why an instance is considered as an anomaly.

The paper is organized as follows: Section 2 presents details about the proposed algorithm and an experimental evaluation, Section 3 describes the use case on a real world financial data, Section 4 presents related work, while Section 5 concludes the paper.

2. Anomaly Score Computation

Our algorithm utilizes the principles of well known approaches to (associative) frequent pattern mining such as Apriori or FP-Growth (Agrawal and Srikant, 1994; Han et al., 2000). The main limitation of these methods is that they do not fully support numeric values and a discretization step is required while preprocessing the data. Although there are other existing approaches that can handle multiple data types, contextual anomalies or numerical values (Chandola et al., 2009), we focused on FPM as it allows to build on the top of existing algorithms whose output in the form of associations is well interpretable (Fürnkranz and Kliegr, 2015).

2.1. Frequent Pattern Isolation

Although several formulas for computing the anomaly score exist, we propose an amended method inspired by an existing algorithm called Isolation Forests (IF) (Liu et al., 2008, 2012). The original IF algorithm is based on a set of decision trees and the anomaly score is computed on top of the concept of separating an instance from the rest of the instances. Shorter paths in decision trees indicate better isolation and higher anomaly scores.

We call the proposed formula *Frequent Pattern Isolation* (FPI). It is defined as:

$$FPI = mean(FPI_{pContrib} \cup FPI_{pen}) \tag{1}$$

$$FPI_{pContrib} = \bigcup_{P \in MP} \frac{1}{support(P) * length(P)} \tag{2}$$

$$FPI_{pen} = \bigcup_{\#penalizations} size(data) \tag{3}$$

Algorithm 1: Anomaly score computation.

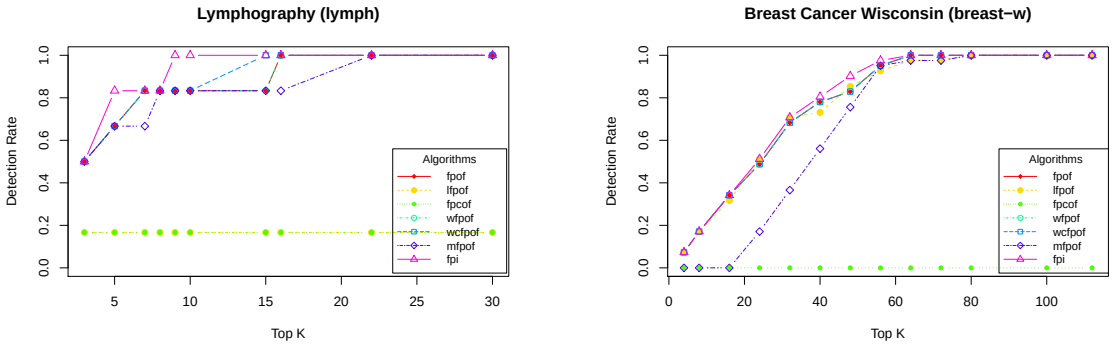
input : Input data: *data*
Minimum relative support of a pattern: *minSupp*
Maximum length of a pattern: *maxLen* (≥ 1)
output: Anomaly score for each input instance: *scores*

begin
  // Frequent Pattern Mining using Apriori, FP-Growth, ...
  frequentPatterns = FPM(*data, minSupp, maxLen*)
  scores = {}
  foreach *instance in data* **do**
    // Find all frequent patterns that match the instance
    matchingPatterns = $\{P \mid P \in frequentPatterns \wedge P\, matches\ \ instance\}$
    // Compute contributions of matching patterns using their support and length
    pContrib = $\{1/(support(P) * length(P)) \mid P \in matchingPatterns\}$
    // Number of penalizations as a number of values of the instance that are not part of any matching pattern
    penalizations = $\mid items(instance) \mid - \mid unique(items(matchingPatterns)) \mid$
    // Compute mean value of pattern contrib. and penalizations (size(data) corresponds to the contribution of a pattern with the lowest possible support)
    score = $mean(pContrib \cup \{a_i = size(data) \mid 1 \leq i \leq penalizations\})$
    $scores = scores \cup score$
  end
  return *scores*
end

The interpretation is that the anomaly score is computed as a mean value of (the multiset of) specific values representing contributions of matching patterns and so-called *penalizations* (Equation (1)). Contributions of matching patterns *MP* (Equation (2)) are proportional to their support and length. Penalizations (Equation (3)) serve as compensation in situations when the matching patterns only match a subset of the item descriptors and the contributions cannot be thus properly determined. More details are available in the following detailed description of the whole algorithm.

Algorithm 1 presents the algorithm of computing the anomaly scores. It starts with the mining of all frequent patterns that meet the standard predefined criteria: minimum relative support and maximal length of the pattern (the number of items in the pattern with no repetitions). It continues with the computation of the anomaly score for each instance. Each matching pattern contributes to the final score, where the isolation principle is included as follows: the matching patterns that are more frequent (with higher support) and contain more items (higher length) produce a significantly lower score than less frequent (lower support) and shorter patterns. If the data instance contains short but infrequent patterns, it follows the isolation principle and the instance is likely to be an anomaly. The mining of all frequent patterns is a complex task, since the matching of all patterns with all data instances is computationally intensive. The FPM algorithms thus mostly focus on extraction of the most frequent patterns, while the patterns for the least frequent items are unavailable. Patterns with low frequency are at the same time longer and composed of more frequent sub-patterns. FPI therefore includes penalization for situations when only a limited amount of patterns is available and the data instance is not completely covered with the existing set of frequent patterns. Specifically, FPI uses the highest possible contribution (equal to the number of instances in the data) as penalization for each unmatched attribute of the

instance. It corresponds to the isolation using a pattern of length one with the lowest possible support (equal to $1/size(data)$). The penalization significantly increases the final score of such an instance and isolates it from others. The final score is computed as a mean value of all contributions from the matching patterns and penalizations (An example is in Table 2).



(a) Lymph dataset (b) Breast-w dataset

Figure 1: Visualization of the *Detection Rate*.

2.2. Experimental Evaluation

We evaluated the proposed algorithm on two standard UCI[1] datasets that are well-established in research and evaluation of anomaly detection based on frequent patterns: Lymphography(*lymph*) and Breast Cancer Wisconsin(*breast-w*):

- *lymph*: 148 instances, 18 attributes and 4 classes, where instances associated with the *normal* or *fibrosis* class are considered as anomalies ($\approx 4\%$ of instances).
- *breast-w*: 699 instances, 10 attributes and 2 classes, where instances associated with the *malignant* class are considered as anomalies. We have redone a preprocessing of the dataset so that it is highly imbalanced – only one in every six malignant records was chosen (resulting in $\approx 8\%$ of anomaly instances) (Hawkins et al., 2002).

We compared our approach with the baseline algorithm FPOF and its modifications or successors: LFPOF, FPCOF, WFPOF, WCFPOF and MFPOF (See Section 4 for more details). The setting of the algorithms is the same as in specific evaluations of individual algorithms including the baseline FPOF: $minimumSupport = 0.1$ and $maximumLength = 5$.

Since the algorithms are unsupervised, we use the class attribute only for the evaluation of results. The class attribute is thus not available as input during the runs of anomaly detection. As a metric to measure the quality of the detection we use the *Detection Rate@Top-K*: the instances are sorted according to the computed anomaly score and we measure the proportion of anomalies in the Top-K lists of instances. It is defined as a

1. https://archive.ics.uci.edu/ml/

number of detected true anomalies divided by the number of all true anomalies (Said et al., 2013).

Table 1: AUC values for all algorithms

	fpof	lfpof	fpcof	wfpof	wcfpof	mfpof	fpi
lymph	0.986	0.5	0.5	0.986	0.991	0.984	**0.996**
breast-w	0.99	0.99	0.5	0.99	0.99	0.979	**0.992**

Figures 1(*a*),1(*b*) present the detection rates for selected lists of Top-K anomalies. The K values were selected according to existing research and evaluations (He et al., 2005; Said et al., 2013). Table 1 presents the AUC values that are highly influenced by the nature of such imbalanced datasets and specific algorithmic approaches. The results for the FPOF algorithm confirm the results in the original paper. Algorithm FPCOF does not perform well for neither dataset, LFPOF performs well for *breast-w* but nut for *lymph*. Their low performance is caused by the general setting of algorithms and the domains of the specific datasets. Other algorithms are able to detect significantly more anomalies at lower values of the K parameter. Our modification of the anomaly score computation slightly improves the detection rate when compared to the other provided algorithms. The likely main reason for the improvement is that it takes into account the coverage of frequent patterns and the penalization mechanism. All experiments are reproducible and are available on-line in form of a notebook for R[2].

2.3. Explanations of Anomalies

Existing studies are mainly focused on the identification of anomalies themselves, while the presentations of reasons and explanations of anomalies are still limited. We propose the following complementary approaches providing more insights into the identified anomalies. The main idea is based on our experiments and discussions with domain experts: it is important to be able to properly explain which value contributes to the overall anomaly score most strongly (see examples in Section 3):

Attribute explanations This approach provides numeric proportions explaining the overall overview of contributions of each individual instance descriptor. It is computed similarly to the main FPI anomaly score. Since we already know the partial contribution scores of frequent patterns for an instance, we identified attributes that are members of those patterns and proportionally divide the contribution to all their members. If the attribute is not covered with any frequent pattern, the penalization score is assigned in the same fashion as for the FPI score. This allows to get brief insights into the contributions of individual attributes.

Visual explanations As visual explanation we decided for a set of bar plots. They allow to summarize the frequencies together with the cardinality of values for all available descriptors separately. This kind of visualization provides a brief overview about the differences of selected instances.

2. https://gist.github.com/jaroslav-kuchar/16155a8c431898866808eda5d4693593

Table 2: Overview of outputs for ESF-CZ-2007-2013 use case.

Anomaly Instance (amount=3.34B):
- Matching: 1 pattern (1 of 3 attributes = 33.3%)
- Patterns (support):
  - {partnerTypeBroader= Educational and research Institution}(0.028)
- Anomaly score: **71552.57**
- Computed as: $mean(\{1/(0.028 * 1), 107311, 107311\})$

Instance with the highest amount (amount=7.45B):
- Matching: 2 patterns (2 of 3 attributes = 66.6%)
- Patterns (support):
  - {partnerTypeBroader=Business subject}(0.22)
  - {operationalProgrammeBroader=1-5}(0.0001)
- Anomaly score: **38007.53**
- Computed as: $mean(\{1/(0.22 * 1), 1/(0.0001 * 1), 107311\})$

Regular Instance:
- Matching: 7 patterns (3 of 3 attributes = 100%)
- Patterns (support):
  - {amount=[0.00e+00,7.45e+06)}(0.85)
  - {partnerTypeBroader=Other}(0.21)
  - {partnerTypeBroader=Other, amount=[0.00e+00,7.45e+06)}(0.2)
  - {operationalProgrammeBroader=7-1}(0.18)
  - {operationalProgrammeBroader=7-1,amount=[0.00e+00,7.45e+06)}(0.17)
  - {partnerTypeBroader=Other, operationalProgrammeBroader=7-1}(0.15)
  - {partnerTypeBroader=Other, operationalProgrammeBroader=7-1,amount=[0.00e+00,7.45e+06)}(0.15)
- Anomaly score: **3.2**
- Computed as: $mean(\{1/(0.85 * 1), 1/(0.21 * 1), 1/(0.2 * 2), 1/(0.18 * 1), 1/(0.17 * 2), 1/(0.15 * 2, 1/(0.15 * 3))\})$

3. Financial Data Use Case

For the real-world use case we selected data about the Czech segment of the European Social Funds (ESF-CZ-2007-2013)[3]. It contains information about projects with additional attributes (e.g. partner, partner type or operational program) and the amounts of assigned money. The dataset contains 107,311 instances in total and it is also focused on the ability of our approach to work with larger datasets. The goal is to reveal instances in data that deviate from others.

We experimentally selected the following subset of (three) attributes: *partnerTypeBroader* (textual representation of the partner type, e.g., national institution), *operationalProgrammeBroader* (identifier of an operational programme), and *certifiedEu* (amount of money certified by the EU).

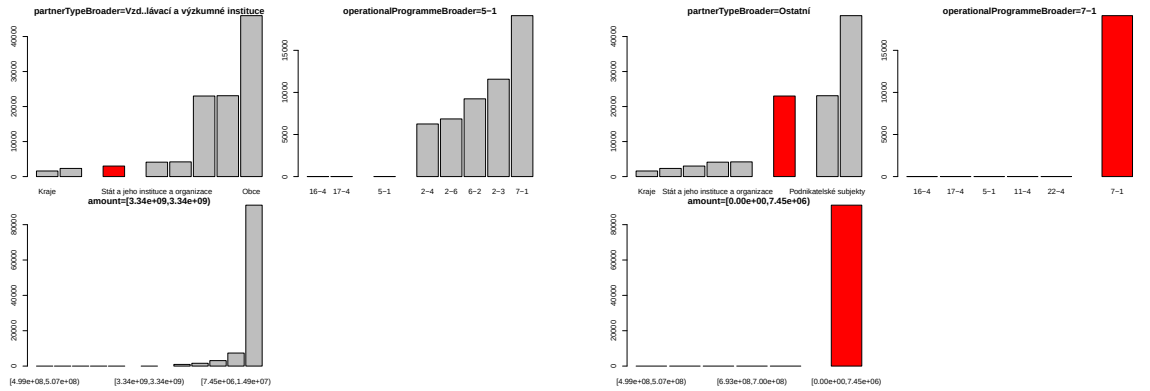
The data contains only one numeric attribute for the certified amount. If we only relied on the analysis of extreme numeric values, the highest amount (and the most deviated value from others) would be around 7.45 billion (in the given currency unit). We can consider the instance with this amount as a baseline kind of anomaly.

To compare it with anomalies detected by FPI, we used the following parameter setting: *minimumSupport*=0.0001 and *maximumLength* of patterns was unlimited. The amount values were discretized to 1000 equal-length intervals. With the selected setting we can extract patterns that are available for at least 10 instances of the data. It allows to utilize the penalization, as a part of the FPI, for values with lower frequencies. FPI produced anomaly scores from 3.2 to 71 553 (with 1245 frequent patterns).

The details for the three selected instances are in Table 2. The instance with the lowest score is matched with 7 frequent patterns that cover all attributes of the instance and there is thus no need for penalization. The instance with the highest anomaly score is matched with only one frequent pattern, which covers only one attribute, with significantly lower support. The two remaining attributes (infrequent operational program together with the infrequent amount of about 3.34 billion) are penalised. For the instance with the extreme amount value, the FPI algorithm computes an anomaly score of 38 007 and matches two attributes with two patterns. Only one attribute (corresponding to an infrequent amount)

3. Provided by the OpenBudgets.eu project: https://github.com/openbudgets/datasets

is penalized. Since there are other instances that contain more infrequent values that are also relevant, the instance with the highest amount is not assigned the highest anomaly score.



(*a*) Anomaly instance (*b*) Regular instance

Figure 2: Visual explanations.

Figure 2(*a*) and Figure 2(*b*) demonstrate the visual explanations for the anomaly and the regular instance, via comparing the frequency of the value associated to the instance (red bar) with frequencies of other values. The anomaly contains less frequent values than the regular instance.

This fact is also supported by the explanation of attributes contributions. The total anomaly score is approximately the mean value of individual contributions. The anomaly instance has the highest score, since two attributes are rare (Figure 2(*a*)). The regular instance has the lowest score: all values are significantly more frequent than for the other instances (Figure 2(*b*)). Contributions of each attribute to the overall scores are as follows:

- *Anomaly instance (71553)* : partnerTypeBroader: 36, operationalProgrammeBroader: 107311, amount: 107311
- *Highest amount instance (38007)*:partnerTypeBroader: 4.6, operationalProgrammeBroader: 6707, amount: 107311
- *Regular instance (3.2)*:partnerTypeBroader: 3.2, operationalProgrammeBroader: 3.5, amount: 2.2

The FPI algorithm is able to detect instances of the data that are described by frequent patterns and associates them with the low anomaly scores. For instances that contain infrequent values the significantly higher anomaly score is assigned. The drawback of the method is that numerical values are treated as categorical values with the same effect as other values, and in this specific use case, amount plays a special role. As our future work we will focus on modifications of the formula in terms of a strengthening the influence for specific attributes. The complete report is available online in the form of a notebook for R[4].

4. https://gist.github.com/jaroslav-kuchar/0968328abaf7be7a2d34199e1d9cb571

4. Related Work

Our research is focused on unsupervised approaches that can handle multiple types of data a and can take into account all existing features as the relevant context. Many approaches are based on measuring deviations from standard distributions, using distance measures to indicate abnormal values or clustering algorithms to detect instances outside clusters (Chandola et al., 2009).

There are several existing approaches based on the FPM principles, which differ in application of the discovered frequent patterns. Algorithm FPOF (Frequent Pattern Outlier Factor) (He et al., 2005) is considered as a baseline approach that computes the anomaly score using the availability of pattern in the instance together with its frequency. Since FPOF uses conventional algorithms for the mining of all existing frequent patterns, one known limitation is that it uses, for the calculation, duplicates coming from subsets or supersets of a frequent pattern. Algorithm LFPOF/EFPOR (Zhang et al., 2010) decreases the influence of duplicates using only the longest frequent patterns; MFPOF (Lin et al., 2010) uses maximal frequent patterns (items having no frequent supersets), and WCFPOF (Ren et al., 2009) utilizes only closed frequent patterns (items having no superset with the same frequency). Algorithm FPCOF (Tang et al., 2009) measures how contradictory the existing patterns are, where a less contradictory pattern set means that the instance is more likely a normal instance. WFPOF (Xiao-Yun et al., 2007) extends the computation of the FPOF score by the influence of the length of the pattern in contrast to the size of the data instance.

Recent research also focused on optimization of the algorithmic complexity and computation time. The mining of all frequent patterns and their application is generally computationally intensive. There are also approaches that approximate the computation (Giacometti and Soulet, 2016) resulting in better computational time but featuring an approximation error. To be able to process large volumes of evolving data, the streaming approaches can be used (Said et al., 2015). Another solution is to invert the task and focus on the mining of infrequent patterns (Rahman et al., 2010). The availability of well performing approaches is however still limited.

5. Conclusions

In this paper we present our work in the domain of anomaly detection using frequent pattern mining. The main contribution is an innovated formula to compute the anomaly score. We first evaluated the proposed approach on two standard UCI datasets. The results show that our formula can provide better results when compared with existing approaches. We also present a use case that is focused on detection of unusual situations in financial data, where we demonstrate an application of the method in a real-world use case. Several questions remain open and will be addressed in our future work. We will focus on experiments and evaluations of results on OpenBudgets.eu datasets with the help of domain experts. Modification of the anomaly score computation in terms of specification of preferences for specific attributes will be also our research direction. The evaluations, experiments and

implementations of all algorithms[5] are publicly available to ensure reproducibility. The algorithm has been also recently integrated into the EasyMiner[6].

Acknowledgments

This research was supported by the European Union's H2020 EU research and innovation program via the OpenBudgets.eu project (under grant agreement No 645833). We thank Jindřich Mynář and our domain expert Lucie Sedmihradská who provided insight and expertise that greatly assisted the research.

References

Charu C. Aggarwal. *Outlier Analysis*. Springer Publishing Company, Incorporated, 2013. ISBN 978-1-4614-6395-5.

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1-55860-153-8.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009. ISSN 0360-0300.

Johannes Fürnkranz and Tomáš Kliegr. *A Brief Overview of Rule Learning*, pages 54–69. Springer International Publishing, Cham, 2015. ISBN 978-3-319-21542-6.

Arnaud Giacometti and Arnaud Soulet. *Frequent Pattern Outlier Detection Without Exhaustive Mining*, pages 196–207. Springer International Publishing, Cham, 2016. ISBN 978-3-319-31750-2.

Markus Goldstein and Seiichi Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE*, 11(4):1–31, 04 2016.

Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 1–12, New York, NY, USA, 2000. ACM. ISBN 1-58113-217-4.

Simon Hawkins, Hongxing He, Graham J. Williams, and Rohan A. Baxter. Outlier detection using replicator neural networks. In *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*, DaWaK 2000, pages 170–180, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-44123-9.

Zengyou He, Xiaofei Xu, Joshua Zhexue Huang, and Shengchun Deng. FP-outlier: Frequent pattern based outlier detection. *Computer Science and Information Systems/ComSIS*, 2 (1):103–118, 2005.

5. Available as a package for R: https://github.com/jaroslav-kuchar/fpmoutliers
6. EasyMiner is an open source web-based project for data mining based on associations - http://easyminer.eu/

Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004. ISSN 1573-7462.

Carson Kai-Sang Leung, Ruppa K. Thulasiram, and Dmitri A. Bondarenko. *An Efficient System for Detecting Outliers from Financial Time Series*, pages 190–198. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-35971-5.

Feng Lin, Wang Le, Jin Bo, Feng Lin, Wang Le, and Jin Bo. Research on maximal frequent pattern outlier factor for online high- dimensional time-series outlier detection. *Journal of Convergence Information Technology*, 10(5):66–71, 2010.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08)*, pages 413–422, 2008.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 6(1):3:1–3:39, March 2012. ISSN 1556-4681.

Ahmedur Rahman, Christie I. Ezeife, and Akshai K. Aggarwal. WiFi Miner: An Online Apriori-Infrequent Based Wireless Intrusion System. In Mohamed Gaber, Ranga Vatsavai, Olufemi Omitaomu, João Gama, Nitesh Chawla, and Auroop Ganguly, editors, *Knowledge Discovery from Sensor Data*, volume 5840 of *Lecture Notes in Computer Science*, chapter 5, pages 76–93. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-12518-8.

Jiadong Ren, Qunhui Wu, Changzhen Hu, and Kunsheng Wang. An approach for analyzing infrequent software faults based on outlier detection. In *Proceedings of the 2009 International Conference on Artificial Intelligence and Computational Intelligence - Volume 04*, AICI '09, pages 302–306, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3816-7.

Aiman Moyaid Said, Dhanapal Durai Dominic, and Brahim Belhaouari Samir. Outlier detection scoring measurements based on frequent pattern technique. *Research Journal of Applied Sciences, Engineering and Technology*, 6(8), 2013.

Aiman Moyaid Said, Dhanapal Durai Dominic, and Ibrahima Faye. Data stream outlier detection approach based on frequent pattern mining technique. *Int. J. Bus. Inf. Syst.*, 20(1):55–70, July 2015. ISSN 1746-0972.

Xianghong Tang, Guohui Li, and Gang Chen. Fast detecting outliers over online data streams. In *2009 International Conference on Information Engineering and Computer Science*, pages 1–4, Dec 2009.

Zhou Xiao-Yun, Sun Zhi-Hui, Zhang Bai-Li, and Yang Yi-Dong. A fast outlier detection algorithm for high dimensional categorical data streams. *Journal of Software*, 18(4), 2007.

Weiwei Zhang, Jianhua Wu, and Jie Yu. An improved method of outlier detection based on frequent pattern. In *2010 WASE International Conference on Information Engineering*, volume 2, pages 3–6, Aug 2010.