Journal of COMPUTER AND FORENSIC SCIENCES

CFS 2023, Vol. 2, Issue 1, pp. 53–61
https://doi.org/10.5937/jcfs2-45169
Original research paper

Application of Machine Learning Methods for Anomaly Detection in Internet Advertising

Marko Živanović[1*], Svetlana Štrbac-Savić[2], and Zlatogor Minchev[3]

[1] Academy of Technical and Art Applied Studies, School of Electrical and Computer Engineering, Belgrade, Serbia; markozivanovic998@gmail.com

[2] Academy of Technical and Art Applied Studies, School of Electrical and Computer Engineering, Belgrade, Serbia; svetlana.strbac@viser.edu.rs

[3] Joint Training Simulation and Analysis Center, Institute of ICT, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences; zlatogor@bas.bg

* Corresponding author: markozivanovic998@gmail.com

Received: June 24, 2023 • Accepted: July 3, 2023 • Published: July 7, 2023

Abstract: This research deals certain with issues regarding downloading data from the Internet, i.e., Internet page advertising, and certain mechanisms to take care of the integrity of the data that is put into the dedicated processing context afterwards. The work also relates to e-commerce, as some advertising scenarios provide high error rates with pricing, which may be unacceptable in various scenarios, such as renting or selling a home. This paper presents a brief overview of the outlier detection methods and machine learning-based classifiers that are used to determine the number of anomalies in the analyzed dataset. This work contributes to the operation of organizations that deal with data accuracy and integrity, such as home rental or selling agencies.

Keywords: data retrieval; machine learning; anomaly detection; e-commerce.

1. INTRODUCTION

The development of the Internet and Internet services provided significant improvements in electronic business and marketing. Sellers who sell goods in such applications often strive to ensure as much profit as possible or to have their product highlighted in the best possible way in order to sell and market the product faster. Through the Internet portal, customers have a detailed overview of products in a certain part of the city or country. However, when entering data by staff, it often happens that the data is not entered correctly, that is, the type of data entered is appropriate, but the values entered for that data in the system are not expected and are inadequate for end users. For example, in the information system for the sale of apartments in the city center, the price per square meter is

Citation: M. Živanović, S. Štrbac-Savić, and Z. Minchev "Application of Machine Learning Methods for Anomaly Detection in Internet Advertising," *Journal of Computer and Forensic Sciences*, vol. 2, no. 1, pp. 53–61, 2023, https://doi.org/10.5937/jcfs2-45169.

extremely higher than on the outskirts of the city, in another residential building, or in a building within the same residential unit. Such information systems do not have great accuracy and seriousness, so users usually avoid them and they become less frequented over time. However, a data tracker cannot independently keep track of tens or thousands of records in a database. So, it is obvious that it is necessary to provide automated puzzles of software that employ certain machine learning methods as well as other artificial intelligence algorithms in order to distinguish the data in the ads that is classified as anomaly from those classified as regular data. Various commercial websites that are the most popular on the real estate sales market in the Republic of Serbia were used to check the classifier. Although there are numerous data classification techniques, it is necessary to use the best possible classifiers for these types of data. In machine learning and artificial intelligence, the rule is that one solution is not fully usable in another system. Models for classifying anomalies in data related to real estate prices are not the same as models for classifying anomalies in information system failures per unit of time. Therefore, it is necessary to find a solution that is good enough or acceptable with a certain degree of accuracy.

This paper presents research on machine learning-based anomaly detection classifiers, which are a very good basis for further investigation in the field of anomaly detection in both commercial and non-commercial data sets. Future information systems are expected to have exceptional accuracy and precision, as well as avoid errors that occur during data entry. Precision and accuracy can be achieved by detecting data errors and increasing the severity of various portals and services for the sale and placement of products.

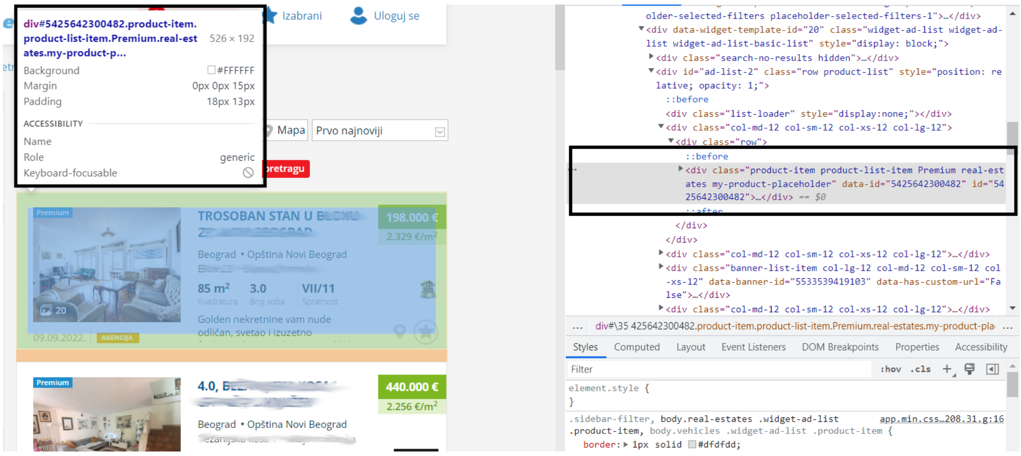
2. RELATED WORK

According to a vast amount of scientific papers published in the previous twenty years, anomaly detection has become one of the major issues in intrusion detection systems. These systems try to seek out deviations from normal behavior that indicate various attack scenarios (both ones on purpose and ones unintended), faults, system and network defects, resource overloading of any suspicious kind (such as providing huge traffic inside the company after business hours, for example), and other anomalous behavior, e.g., dropping a support for email or data service. This paper provides the reader with a brief review of possible (un)supervised learning algorithms for anomaly detection. The references cited within will cover some basic theoretical issues, guiding the researcher further [1]. Vignotto and Engelke [2] proposed two new algorithms for anomaly detection that rely on approximations from the theory of extreme values, which are more robust in such cases. Algorithms employ hypotheses that test points that are extremely far away from the points belonging to the training classes that are more likely to be anomalous objects. Although it may be correct in theory, in practice it depends on how clearly the normal behavior is absorbed – how sterile was the normal behavior training environment acquisition, i.e., how distant were the objects belonging to the normal class from anomalies, and how successful was the noise removal. However, the authors provided the results motivated by the univariate theory of extreme values that make aforementioned theory relatively precise. Rao et al. [3] demonstrated the effectiveness of classifiers in simulations and on real datasets. In this paper, the authors presented cascade grouping of K-means clustering and decision tree ID3 methods to seek out computer network anomalies. One of the IoT's main tasks is to seek out collected data alternations automatically, which includes data pattern deviation techniques. One specific set of algorithms, nowadays known as called deep learning (which actually originated some time ago, but there was no commercial hardware at that time to support it), can search for a specific relationship in billions of corporate IoT data, analyze them, classify them, and provide the right decisions [4]. Azavedo and Hoegner [5] examined the predictability of 299 capital market anomalies enhanced by 30 machine learning approaches and over 250 models in a dataset with more than 500 million firm-month anomaly observations.

3. MATERIALS AND METHODS

Downloading data from Internet ads is a process that involves the use of automata, or "robots" for extracting content from Web pages. Unlike the so-called "scraping" (Web scraping – downloading data from Internet services using automated scripts) of the screen that extracts the pixels of the image and saves them, Web scraping refers to downloading data from a web page and therefore also downloads data that has been generated from a database. All content is mapped from a database that is independent of the organization (relational or non-relational) to another database without the need for manual rewriting of data. Web scraping is used by various e-commerce services and product marketing sites that rely on permitted data collection from Internet sites that support Internet data scraping [6].

Web scraping is used illegally when it comes to stealing the copyright found on a page on the Internet. Also, it is used for the purposes of illegal price reductions on the stock market. Companies dealing with competitive businesses such as shopping facilities, supply chains, content distribution, and the like are particularly vulnerable.



div#5425642300482.product-item.product-list-item.Premium.real-estates.my-product-p... 526 × 192
Background □#FFFFFF
Margin 0px 0px 15px
Padding 18px 13px
ACCESSIBILITY
Name
Role generic
Keyboard-focusable ⃠

TROSOBAN STAN U [redacted]
198.000 €
2.329 €/m²
Beograd • Opština Novi Beograd
[redacted]
85 m² Kvadratura
3.0 Broj soba
VII/11 Spratnost
Golden nekretnine vam nude odličan, svetao i izuzetno
09.09.2022. AGENCIJA

4.0, [redacted]
440.000 €
2.256 €/m²
Beograd • Opština Novi Beograd
[redacted]

```
        older-selected-filters placeholder-selected-filters-1">…</div>
      ▼<div data-widget-template-id="20" class="widget-ad-list widget-ad-
        list widget-ad-list-basic-list" style="display: block;">
        ▶<div class="search-no-results hidden">…</div>
        ▼<div id="ad-list-2" class="row product-list" style="position: re
          lative; opacity: 1;">
            ::before
            <div class="list-loader" style="display:none;"></div>
          ▼<div class="col-md-12 col-sm-12 col-xs-12 col-lg-12">
            ▼<div class="row">
                ::before
              ▶<div class="product-item product-list-item Premium real-est
                ates my-product-placeholder" data-id="5425642300482" id="54
                25642300482">…</div> == $0
                ::after
              </div>
            </div>
          ▶<div class="col-md-12 col-sm-12 col-xs-12 col-lg-12">…</div>
          ▶<div class="banner-list-item col-lg-12 col-md-12 col-sm-12 col
            -xs-12" data-banner-id="5533539419103" data-has-custom-url="Fa
            lse">…</div>
          ▶<div class="col-md-12 col-sm-12 col-xs-12 col-lg-12">…</div>
…  div#\35 425642300482.product-item.product-list-item.Premium.real-estates.my-product-pla  …
```

```
element.style {
}
.sidebar-filter, body.real-estates .widget-ad-list        app.min.css…208.31.g:16
.product-item, body.vehicles .widget-ad-list .product-item {
  border: ▶ 1px solid □#dfdfdd;
```

Figure 1. *An example of an information system with HTML tags.*

There are three sets of data in the paper. The first set consists of 1067 data taken from the website. Each record has its name in text format, square footage as an integer, number of rooms as an integer, and apartment prices in decimal format.

This applies to other sets as well. The second one consists of 750 data taken from the website, while the third one consists of 1050 real data instances. Therefore, the data are not artificially generated but are real and taken from the system for selling apartments.

We have used the data from the Web sites of apartment sale or renting agencies with the sole purpose of preventing malicious financial gain for these agencies. Several Web sites have been scraped, and we have chosen three for this research. The main goal is to download the data on prices and potentially eliminate the competitors' adversarial behavior (leading to could-have-been-done social engineering-based attacks) in the labor market. The so-called attacks are taking place in an industry where price plays a major role in customers' decisions and where customers are spending money. Most often, these are travel agencies, computer equipment stores, etc. For example, a computer store monitors other competitors and analyzes the market. A permanent scraping system also tracks prices that change in real time. Then Internet browsers rank higher the Web sites that offer lower prices than other competitors.

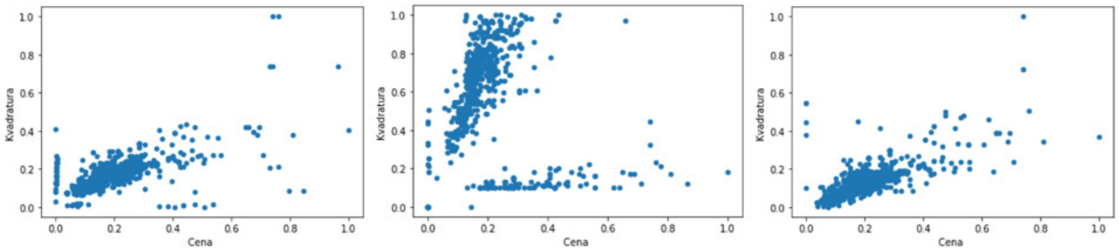


Figure 2. *Visual display of data from commercial Web systems. Label "Cena" on the x-axis on all three diagrams represents the price of the apartment, while label "Kvadratura" on the y-axis represents the size of the flat. Both images are exported from the Python programming language and are x- and y-axis scaled to the range [0, 1] with the purpose of better presentation to the reader.*

The methods that were used during the realization of the task are enlisted below:

- Angle-based Outlier Detector (ABOD) – a class for detecting deviations from the pattern based on the angle formed between different features. There are two versions of ABOD supported – fast, based on KNNs, and original one, with $O(n^3)$ time complexity [7];
- Cluster-based Local Outlier Factor (CBLOF), which is also a clustering-based algorithm that measures the distance to the nearest large cluster center (normal data cluster) [8];
- Histogram-based outlier detection (HBOS) – with static bin numbers and the Birge-Rosenblatz automated bin number method [9];
- The Isolation Forest detection (IF);
- Local Correlation Integral (LOCI);
- Minimum covariance determinant (MCD);
- Multiple Objective Generative Adversarial Active Learning (MO-GAAL);
- Single-Objective Generative Adversarial Active Learning (SO-GAAL);
- Principal Component Analysis (PCA);
- Rotation-based Outlier Detector (ROD);

- Auto-encoder (AE) – a ANN for unsupervised useful data representations learning;
- Unsupervised anomaly detection using Empirical Cumulative Distribution Functions (ECOD), which is parameter-free;
- Isolation-based anomaly detection using Nearest-Neighbor Ensembles (INNE);
- Lightweight On-line Detector of Anomalies (LODA) [10].

Readers may consult [7–10] for details on the methods and algorithms employed, including some others. Acronyms will be used sub-sequentially.

4. RESULTS

According to the results obtained for the first data set, we can conclude that the most anomalies are concentrated around 102 to 109, so for this data set, the algorithms marked with a symbol † (in Tables 1–3) provide the best accuracy, and if we compare the raw data, we will see that they are significantly close to the accuracy.

Table 1. *Results obtained from the data in the first dataset (1067 in total).*

Algorithm	Anomalies	Regular data
ABOD	0	1067
CBLOF	53	1014
HBOS	50	1017
IF	54	1013
KNN	45	1022
Average KNN	34	1033
LOCI †	105	962
MCD †	109	958
MO-GAAL †	106	961
One-class SVM detector †	107	960
PCA †	107	960
ROD †	107	960
SP †	107	960
SO-GAAL †	107	960
AE †	107	960
CBLOF †	102	956
ECOD †	107	960
GMM †	105	962
INNE †	107	960
KDE †	107	960
Lightweight on-line detector of anomalies	82	985

According to the obtained results for the second data set, we can conclude that the most anomalies are concentrated around 72 to 75, so for this data set, the colored algorithms have the best accuracy, and if we compare the raw data, we will see that they are significantly close to the accuracy.

Table 2. *Results obtained from the data in the second dataset (750 in total).*

Algorithm	Anomaly	Regular data
ABOD	0	750
CBLOF	38	712
HBOS	34	716
IF	38	718
KNN	29	721
Average KNN	29	721
LOCI †	75	675
MCD †	74	676
MO-GAAL †	75	675
One-class SVM detector †	75	675
PCA †	72	678
ROD †	75	675
SP †	75	675
SO-GAAL †	75	675
AE †	74	676
CBLOF †	75	675
ECOD †	75	675
GMM †	75	675
INNE †	74	676
KDE †	75	675
Lightweight on-line detector of anomalies	61	689

According to the results obtained for the last data set, we can conclude that the most anomalies are concentrated around 97 to 105, so for this data set, the colored algorithms have the best accuracy, and if we compare the raw data, we will see that they are significantly close to accuracy.

Table 3. *Results obtained from the data in the third dataset (1050 in total).*

Algorithm	Anomaly	Regular data
ABOD	0	1050
CBLOF	51	999
HBOS	50	1000
IF	49	1001
KNN	47	1003
Average KNN	34	1006
LOCI †	105	945
MCD †	105	945
MO-GAAL †	104	946
One-class SVM detector †	104	946
PCA †	105	945
ROD †	104	946

SP †	104	946
SO-GAAL †	104	946
AE †	105	945
CBLOF †	95	953
ECOD †	105	945
GMM †	105	945
INNE †	105	945
KDE †	105	945
Lightweight on-line detector of anomalies	86	946

For the purpose of a brief visual presentation, we have put together a couple of graphics representing the results of LOCI, MCD, and MO-GAAL's results on all three datasets.

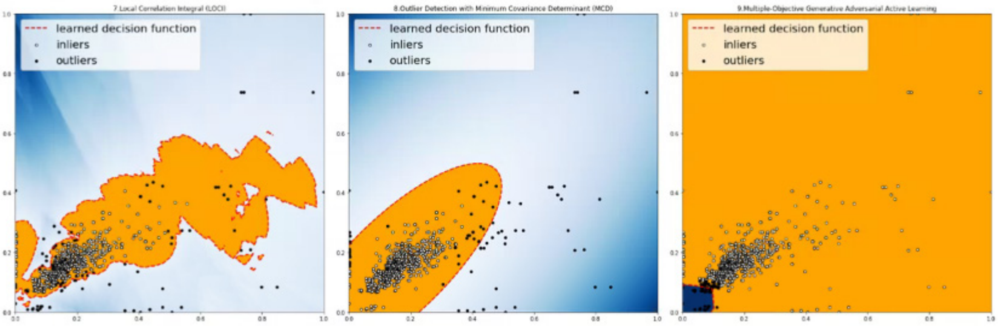


Figure 3. *LOCI, MCD, and MO-GAAL's results on the first dataset.*

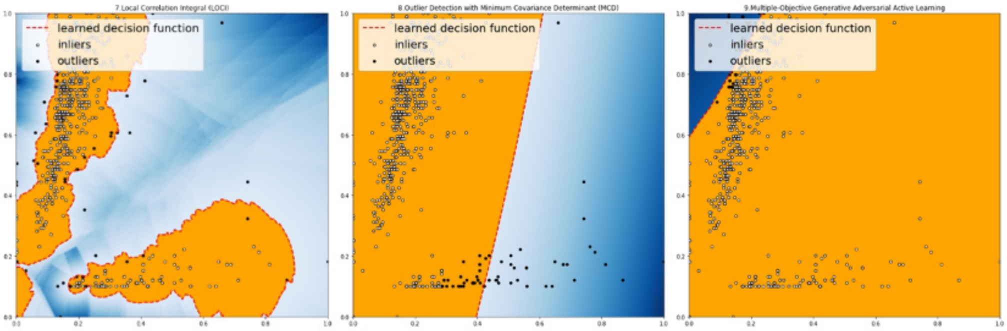


Figure 4. *LOCI, MCD, and MO-GAAL's results on the second dataset.*

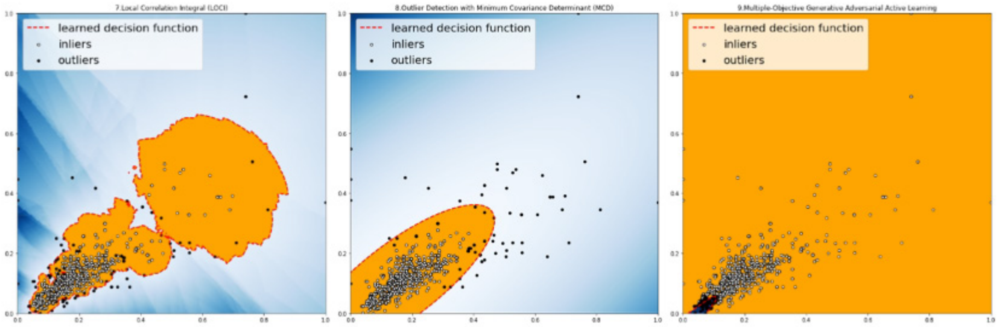


Figure 5. *LOCI, MCD, and MO-GAAL's results on the third dataset.*

7. DISCUSSION AND CONCLUSION

This paper presented the methods of collecting and visualizing the obtained data. Also, it was noted how the data was implemented in different ways within the Web pages. We chose the ratio of attributes between which the highest degree of data correlation occurred, as was the case with the ratio of apartment price and square footage. Using the principles of machine learning in anomaly theory and anomaly detection methods, we presented different types of anomalies and classifiers, as well as a graphical representation of the data. Data were successfully downloaded from three different sites for the purpose of better determination of anomalies. The idea was to prove that even if the data is different and downloaded from different websites, they have some similarities. We still performed the classification on the basis of two attributes (characteristics): of apartment price and square footage. Some anomalies are also obvious to humans, for example, there cannot be a physical quantity that is equal to 0, and its value is up to several thousand euros. However, not all anomalies are apparent in the system at first glance. Already, when we have more samples, we have a better overview of the exceptions in the data. A description of significant classifiers for anomaly detection and their implementation in different datasets is given. When detecting anomalies, exceptions and regular data are separately classified, and the number of them is stated. The number of exceptions is significant due to the assessment and evaluation of machine learning models. Systems that have approximately similar or the same solutions give us a better value for the model. We have over twenty different classifiers, and each of them performs with almost similar accuracy on all three datasets. Each of the data sets also has its own visual representation, which is very significant and differs from the data set and the use case of the algorithm. Furthermore, space is left for the implementation of such a system in business intelligence and systems that will automatically detect anomalies, exclude such anomalies, and upload classified, accurate data to Web services.

FUNDING:

This research received no external funding.

INSTITUTIONAL REVIEW BOARD STATEMENT:

Not applicable.

INFORMED CONSENT STATEMENT:

Not applicable.

CONFLICTS OF INTEREST:

The authors declare no conflict of interest.

REFERENCES

[1] S. Omar, A. Ngadi, and H. H. Jebur, “Machine Learning Techniques for Anomaly Detection: An Overview,” *International Journal of Computer Applications*, vol. 79, no. 2, Oct., pp. 33–41, 2013.

[2] E. Vignotto and S. Engelke, “Extreme Value Theory for Anomaly Detection – The GPD Classifier,” *Extremes*, vol. 23, no 4, pp. 501–520, 2020.

[3] K. H. Rao, G. Srinivas, A. Damodhar, and M. V. Krishna, “Implementation of Anomaly Detection Technique Using Machine Learning Algorithms,” *International Journal of Computer Science and Telecommunications*, vol. 2, no. 3, June, pp. 25–31, 2011.

[4] T. Çavdar, N. Ebrahimpour, M. T. Kakız, and F. B. Günay, “Decision-making for the Anomalies in IIoTs Based on 1D Convolutional Neural Networks and Dempster-Shafer Theory (DS-1DCNN),” *The Journal of Supercomputing*, vol. 79, no. 2, pp. 1683–1704, 2023.

[5] V. Azevedo and C. Hoegnerm, “Enhancing Stock Market Anomalies with Machine Learning,” *Review of Quantitative Finance and Accounting*, vol. 60, no. 1, pp. 195–230, 2023.

[6] R. Mitchell, *Web scraping with Python: Collecting More Data from The Modern Web*. O’Reilly Media, Inc, 2018.

[7] N. Silva, J. Soares, V. Shah, M. Y. Santos, and H. Rodrigues, “Anomaly Detection in Roads with a Data Mining Approach,” *Procedia Computer Science*, vol. 121, pp. 415–422, 2017.

[8] Y. Zhao, Z. Nasrullah, and Z. Li, “PyOD: A Python Toolbox for Scalable Outlier Detection,” *Journal of Machine Learning Research*, vol. 20, no. 96, pp. 1–7, 2019.

[9] M. Ahmed, N. Choudhury, and S. Uddin, “Anomaly Detection on Big Data in Financial Markets,” In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2017, pp. 998–1001.

[10] pyod 1.0.5 documentation. [Online] Available: https://pyod.readthedocs.io/. [Accessed: May 29, 2023].