ECONOMETRICS
EKONOMETRIA
Advances in Applied Data Analysis

Econometrics. Ekonometria.
Advances in Applied Data Analysis
Year 2024, Vol. 28, No. 4
ISSN 2449-9994
journals.ue.wroc.pl/eada

Comparison of Machine Learning and Statistical Approaches of Detecting Anomalies Using a Simulation Study

Klaudia Lenart

University of Economics in Katowice, Doctoral School

e-mail: klaudia.lenart@edu.uekat.pl

ORCID: 0000-0001-8135-9362

Quote as: Lenart, K. (2024). Comparison of Machine Learning and Statistical Approaches of Detecting Anomalies Using a Simulation Study. *Econometrics. Ekonometria. Advances in Applied Data Analysis*, *28*(4), 23-31.
DOI: 10.15611/eada.2024.4.02

JEL: C15, C18

Abstract

Aim: An anomaly is an observation or a group of observations that is unusual for a given dataset. Anomaly detection has many applications, not only as a step of data preparation but also, for example, as a way of identifying credit card fraud detection, network intrusions and much more. There are diverse methods of anomaly detection. In particular two groups of methods have been developed independently – statistical methods and machine learning algorithms. Those methods are rarely compared. While statistical methods focus on formulating a measure of the abnormality of the observations, supervised machine learning makes it possible to use data about typical observations and previously identified anomalies. The aim of this paper was to compare the two approaches by conducting a simulation study.

Methodology: A simulation study was conducted, during which the data was generated using copula functions. For the purpose of generating different types of anomalies, marginal distributions of the variables were manipulated. The effectiveness of each method was evaluated based on measures of classification model performance.

Results: While the accuracy of the statistical methods was dependent on the precise prediction of the percentage of the anomalies that would occur in the data, the machine learning algorithms' recall was significantly lower when the change in the marginal distribution of the value parameters was smaller.

Implications and recommendations: For the statistical methods included in the study, knowledge about the distribution of the variables was crucial while the supervised machine learning algorithms required acquiring a training dataset. Unlike machine learning algorithms, the statistical methods performed with similar accuracy even when the change in the marginal distribution parameters' value was smaller.

Originality/value: The two approaches to anomaly detection presented in the paper are not often compared, usually used by two separate groups of researchers – statisticians and machine learning or data science specialists.

Keywords: anomaly detection, simulation study, machine learning

1. Introduction

Methods of anomaly detection have been developed in several different fields of study and therefore various approaches can be found in the literature. Most notably two fields of study: statistics and data science, gaining more and more popularity in recent years, have independently developed methods of anomaly detection. Although the most popular methods used by data scientists, for example machine learning algorithms, are unquestionably based on statistics and econometrics, researchers who focus on this subject matter are often not interested in statistical methods. Thus there are not many studies comparing the more traditional statistical methods with machine learning algorithms. The aim of this paper was to compare the accuracy of supervised learning machine learning algorithms with two chosen statistical methods of anomaly detection. Hence a simulation study was performed to compare the performance of the methods under different circumstances.

Statistical methods of anomaly detection often focus on developing a measure of the abnormality of the observations. The most common and simplest approach is to measure the distance between a given observation and the rest of the observations from a given dataset (Mehrotra et al., 2017). When an econometric model is estimated on the data, it is possible to measure the influence of observations on the values of model parameters estimators. Although they were not included in the conducted study, it is worth noting that some approaches rooted in statistics are based on the probability of a given value (or values) being the realisation of a variable with a given distribution. An example of this can be control charts used in quality control. Machine learning algorithms are also often used for anomaly detection. In cases of supervised learning, identification of anomalies can be treated as a classification task with two classes – anomalies and typical observations.

2. Literature Review

It is important to note that many authors use the terms ‘anomaly’ and ‘outlier’ interchangeably (Aggarwal, 2017; Chandola et al., 2009; Mehrotra et al., 2017). An anomaly can be defined as an observation that is unusual for a given dataset. In his definition of an outlier, Hawkins (1980) indicated that the level of deviation from other observations is significant enough to arouse suspicion that it was generated by a different mechanism. This is a key factor for most of the applications of anomaly detection. An observation being generated by a different mechanism can be attributed to an error in the data or an occurrence of a rare event. In the case of the former, the values do not reflect any observed realisations of the variables, while the occurrence of the latter disrupts the relations between the variables that can be observed under normal circumstances. It is difficult to define the level of abnormality that can allow the researcher to assume that an observation is an anomaly. It is often necessary to use expert knowledge about a given phenomenon. As noted by Green (1976), some distributions are more prone to occurrence of outliers, making the identification of observations generated by a different mechanism much more difficult. In practice the level of abnormality that an observation has to reach in order to be flagged as an anomaly is also influenced by the consequences of falsely identifying it as a normal observation and vice versa. This will depend on the application of anomaly detection, for example in medicine incorrectly identifying a normal observation as an anomaly can mean additional tests being requested for a given patient, while not identifying an anomaly can be much more dangerous.

This paper focuses on detecting point anomalies, meaning a singular observation that is unusual when compared to the entire dataset. It is worth noting that other types of anomalies exist:

- a contextual anomaly is an observation unusual in a given context, for example when seasonality can be observed, an observation unusual in one period may be normal in another;
- a group anomaly is a set of observations that, although separately can be considered normal, are unusual when appearing in a sequence (Chandola et al., 2009).

When the existence of an anomaly is caused by an error in the data, it is important to identify and either correct or delete the erroneous observation, thus anomaly detection can be used as a step in data preparation process that should precede other research. Nevertheless, it is important to note that in that case, incorrectly identifying observations as anomalies and excluding them from a dataset may result in a skewed perception of the variables' distribution.

When dealing with large amounts of data, anomaly detection methods may help identify the most interesting parts of the data. An example of this can be found in astrology, where researchers often cannot analyse all of the collected data (Das et al., 2015; Baron & Poznanski, 2017; Faaique, 2024). Anomaly detection is also used in cyber-security as part of an IDS (Intrusion Detection System), in particular neural networks are often used in those systems (Maddireddy, 2024). A change in the typical behaviour of a user, for example unusual login time or location, can be a sign of an attack (Jabez & Muthukumar, 2015). Similarly, in credit card fraud detection, purchases atypical for a given credit card owner can be flagged as potential fraudulent transactions (Prarthana & Gangadhar, 2017; Thimonier et al., 2024). An investor whose behaviour differs from the rest, may be in possession of knowledge that is not available to the wider public. This allows for the anomaly detection methods to be used in identifying insider trading (Kulkarni et al., 2017). It was also proposed by Serrano-Cinca et al. (2019) to search for accounting indicators as indicators of bankruptcy. Machine learning algorithms can also be used to detect anomalies in less conventional data, such as images (Liu et al., 2024).

3. Methodology

3.1. Anomaly Detection Methods Included in the Study

During the simulation study, two statistical methods (one distance based and one based on identifying influential observations) were compared with three supervised machine learning algorithms: *k* nearest neighbours, random forest and support vector machine.

The logic behind the use of the distance between observations in anomaly detection is quite intuitive, yet calculating the distance between each pair of observations can be very time consuming, especially for larger datasets. A more practical approach is to use the average or sum of distance to *k* nearest neighbours (*k* is unchanging as these two approaches are equivalent). For each *i*-th datapoint *k,* the nearest neighbours are identified. Near(*p,j*) is defined as *j*-th nearest neighbour of point *p,* and *d*(*a,b*) as a function of distance between to datapoints *a* and *b*. The statistic used in this method is calculated as:

$$\alpha(p) = \sum_{j=1}^{k} d(p, Near(p, j)). \tag{1}$$

The observations for which the value of $\alpha(p)$ is greater than a set value can be identified as anomalies. If there is an expected number of anomalies that will occur in the data, then a quantile of the vector of $\alpha(p)$ values can be used as the cut-off point (Mehrotra et al., 2017).

An influential observation can be defined as an observation that either individually or together with several other observations has a demonstrably larger impact on the calculated values of various estimates than most other observations (Belsley et al., 1980). Based on this definition, an influence based method was developed that identifies observations whose presence in the dataset has the most influence on the values of parameter estimates of a linear regression function with form defined as:

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_1 x_{1i} + \ldots + \alpha_k x_{ki} + \xi_i. \tag{2}$$

This influence is measured based on *T* statistic that is calculated in the following way:

1. Estimation of the parameters of the linear regression function based on the entire dataset. The estimates are denoted as $a_{00},\ldots,a_{0k}$.
2. For each $i \in [1,2,\ldots,n]$ a subset of the data containing $n-1$ elements is created by excluding *i*-th observation.
3. Estimation of the parameters of the linear regression function based on the created subsets of data. The estimates are denoted as $a_{i0},\ldots,a_{ik}$.
4. Calculation of the differences between the values of the parameters estimated based on the entire dataset and *i*-th subset

$$R_{ij} = a_{i\,j} - a_{0\,j}. \tag{3}$$

5. Standardising the differences for each $j \in [0,1,\ldots,k]$.

$$RS_{ij} = \frac{R_{ij} - \bar{R}_{\iota}}{S_{R_j}} \tag{4}$$

where S_{Rj} is a standard deviation of *j*-th parameter estimates calculated on the subsets of data.

6. Calculation of the *T* statistic values for each $i \in [1,2,\ldots,n]$ using the formula

$$T_i = \sum_{j=0}^{k} |RS_{i\,j}|. \tag{5}$$

The author defined *Tp* as the quantile of *T* statistic's vector corresponding to the percentage of the observations identified as anomalies which equals *p*. The *i*-th observation is classified as an anomaly if the following condition is met:

$$T_i \geq T_p. \tag{6}$$

Unlike the statistical methods described above, to utilise the machine learning algorithms included in the study, a training dataset was needed for which the correct classification of each observation is known. The *k* nearest neighbours algorithm classifies a given observation based on the class of the majority of *k* observations closest to it. Random forest is an ensemble-based method that requires constructing multiple decision trees which use a random subset of the dataset's independent variables. This was defined by Breiman (2001) as a classifier consisting of a tree-structures classifiers $\{h(x,\Theta_k), k = 1,\ldots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input *x.* The classification is decided based on which class was most often the result returned by individual decision trees. The support vector machine algorithm is based on identifying a hyperplane or a set of hyperplanes that can accurately separate the different classes.

3.2. Simulation Study

A simulation study was conducted to test the accuracy of the methods described above. The data were generated using the copula functions. The data generation process was implemented in the R program, using the copula package (Hofert et al., 2024; Yan, 2007).

A *m*-dimensional copula is function *C* with domain $[0,1]^m$ when the following conditions are met (Nelsen, 1998):

- $C(1,\ldots,1,a_n,1,\ldots,1) = a_n$
- $C(a_1,\ldots,a_m) = 0$ if $a_i = 0$ for every $i \leq m$
- C is m-increasing.

The foundation of the theory of copulas, as well as its applications in statistics, can be found in Sklar's theorem (Sklar, 1959).

Let *H* be a joint distribution function with margins *F* and *G*. Then there exists copula C such that for all x,y in $\bar{R}$,

$$H(x, y) = C(F(x), G(y)). \tag{7}$$

If *F* and *G* are continuous, then *C* is unique; otherwise, *C* is uniquely determined on RanF x RanG. Conversely, if *C* is a copula and *F* and *G* are distribution functions, then function *H* is a joint distribution function with margins *F* and *G*.

Throughout the years, many copulas have been proposed differing in dependence structure and having unique properties. In the performed simulation study, a normal copula was used because it allows correlation parameter Θ to have a positive or negative value; it is also known as Gaussian copula and was first described by Lee (1983).

Using copulas allowed for the generation of multidimensional data with set marginal distributions and a correlation matrix. In every generated dataset a small percentage of observations was added, generated with a different marginal distribution specification. The marginal distributions of the variables for the typical observations is shown in Table 1.

Table 1. Marginal distributions of the typical observations in generated data where N(μ, σ) denotes the normal distribution with expected values equal to μ and standard deviation equal σ

Variable	Distribution
Y	N(2,2)
x1	N(4,4)
x2	N(10,4)
x3	N(10,7)
x4	N(3,1.2)

Source: author's own work.

4. Results

To test how the compared methods perform in different situations, the performed simulation study included four different variants of generating anomalous observations as shown in Table 2. In the first variant, changes in the mean parameter of the marginal distribution of *y* variable were introduced. Additionally, to test the methods' sensitivity to more subtle changes in the distribution, different percentages of the parameter were changed. The second variant is similar to the first, but with a different percentage of the anomalies added. Variants 3 and 4 introduced different types of anomalies added at the same time, with a new type of anomaly added only to the test dataset in variant 4.

Table 2. Variants of methods of generating anomalies used in the simulation study

Variant	Percent of anomalies	Method of introducing anomalies
1	5%	Change of the mean parameter of the marginal distribution of *y* variable by 25%, 50% or 100%
2	3%	Change of the mean parameter of the marginal distribution of *y* variable by 25%, 50% or 100%
3	5%	Change of the mean parameter of the marginal distribution of *y* variable to 4 and -4 or changing the distribution to exponential distribution with $\lambda = 2$
4	5%	Change of the mean parameter of the marginal distribution of *y* variable to 4 and -4 (second value in test dataset only)

Source: author's own work.

As can be observed in Figure 1, the machine learning algorithms are less accurate the smaller the change in the mean of the marginal distribution of *y* variable. The accuracy of the method identifying the influential observations and the method based on farthest distance to *k* nearest neighbours (knn) were

not significantly impacted by the size of the change in the mean of the marginal distribution of *y* variable. Comparing the values of accuracy and its recall, reveals that for the method based on farthest distance to *k* nearest neighbours, *k* nearest neighbours (knn ML) and random forest, a relatively high value of accuracy was achieved by correctly identifying the typical observations and not the anomalies. This emphasises the importance of comparing different measures instead of just their accuracy. All of these methods in most cases presented in Figure 1, correctly identified fewer than 50% of the anomalies.

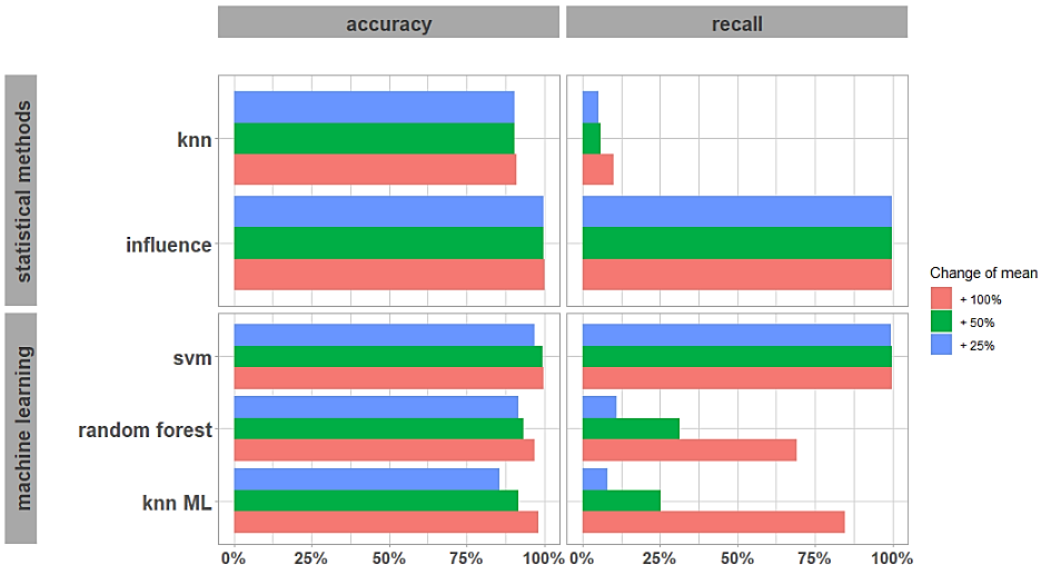

Fig. 1. Accuracy and recall of the presented methods – variant 1

Source: author's own work in the R program.

A comparison of the values shown in Figure 1 and Figure 2 revealed the significance of the correct prognosis of the number of anomalies that can be expected to appear for the accuracy of the statistical methods. The change of the percentage of anomalies added into the data did not significantly impact on the performance of the machine learning algorithms. The performance of the machine learning algorithms was again impacted by the size of the change in the mean of the marginal distribution of y variable, although in all the cases the accuracy and recall that can be observed for the svm was close or equal to 100%.

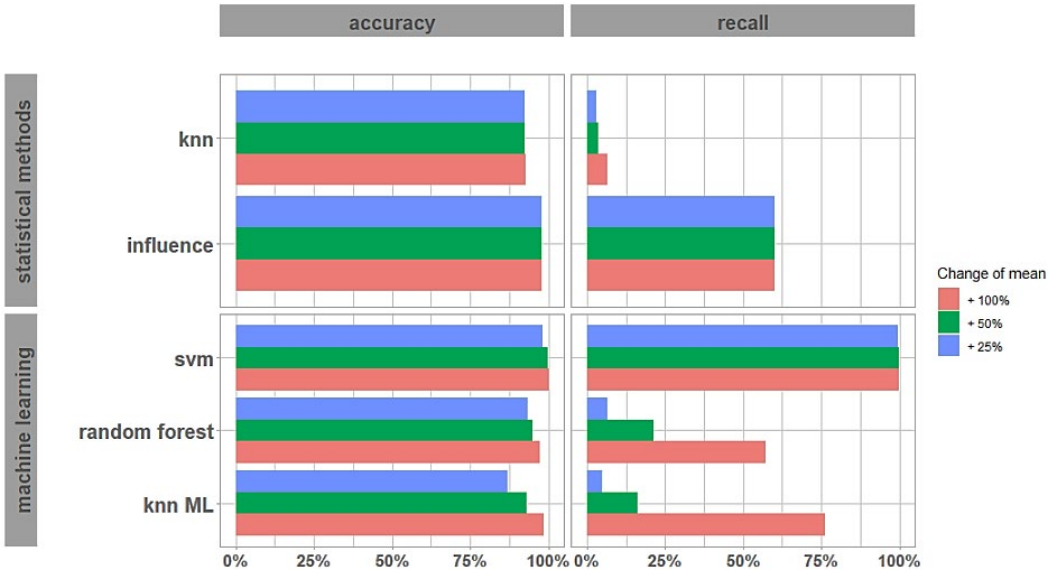

Fig. 2. Accuracy and recall of the presented methods – variant 2

Source: author's own work in the R program.

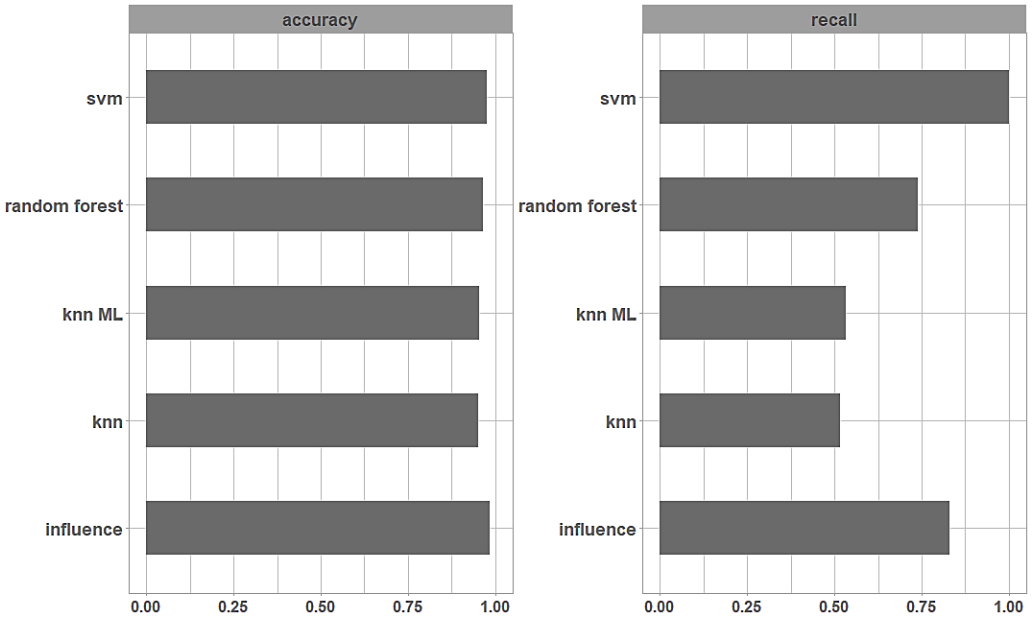

Fig. 3. Accuracy and recall of the presented methods – variant 3

Source: author's own work in the R program.

As shown in Figure 3, introducing different types of anomalies at once, negatively impacts recall achieved by both the statistical methods and the machine learning algorithms, with the exception of the method based on farthest distance to *k* nearest neighbours and the svm. Out of the machine learning algorithms, the svm was both most accurate and achieved the highest value of recall. The method based on the influence achieved the highest accuracy, but the value of recall observed for this method was significantly lower. It is important to note that for svm, the value of recall was 100% and higher than the accuracy of this algorithm. This means that all the anomalies were correctly identified, however some typical observations were incorrectly classified as anomalies.

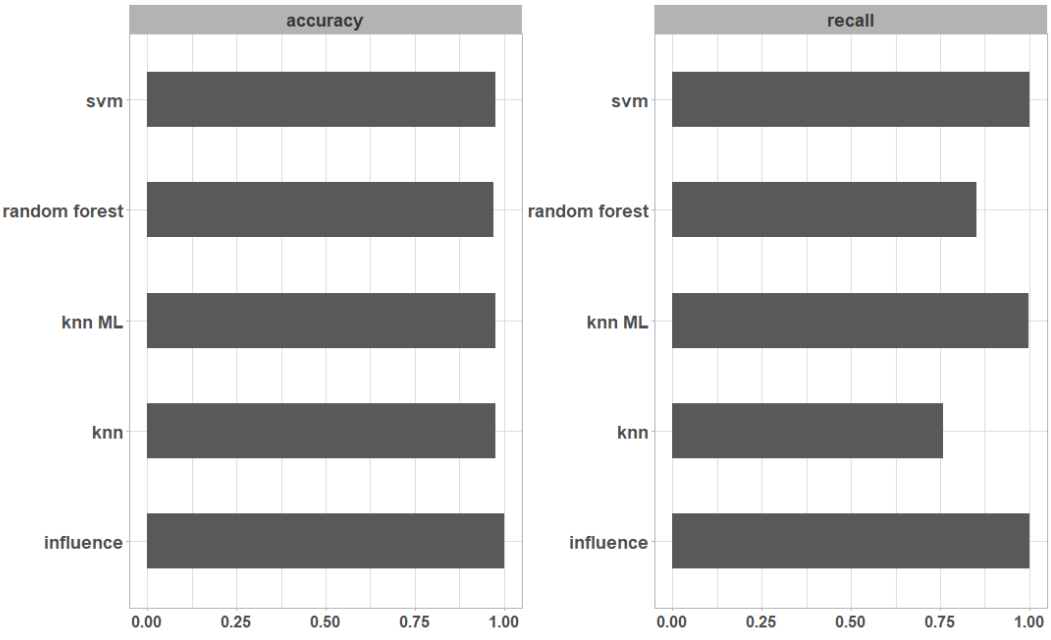

Fig. 4. Accuracy and recall of the presented methods – variant 4

Source: author's own work in the R program.

As can be seen in Figure 4, the introduction of a new type of anomaly only in the test dataset did not significantly affect the accuracy and recall of the machine learning algorithms. Three methods – influence-based method, *k*-nearest neighbors algorithm and the svm – correctly identified all the anomalies. In the case of the two latter algorithms, achieving this result came at the cost of incorrectly classifying some of the typical observations as anomalies.

5. Discussion and Conclusions

The performed simulation study allowed to highlight some differences in the application of the described statistical methods and machine learning algorithms in anomaly detection. The scale of the changes of the marginal distribution parameters significantly impacted on the accuracy of the machine learning algorithms, which performed considerably better when a bigger change was introduced. Therefore, in the applications where a subtle difference in the distribution of the variables needs to be detected, for example quality control for industries like aircraft manufacturing, the use of statistical methods is recommended. It is also important to note that the supervised learning algorithms presented in this paper require a training data set that would accurately reflect reality. Nevertheless, it is worth noting that, as seen in variant 4 in the simulation study, introducing a new type of anomalies only in the test data set did not significantly affect the accuracy and recall of the machine learning algorithms.

For the statistical methods, knowledge about the distribution of the variables is important. In the case of the presented methods, it was crucial to accurately predict the percentage of anomalies than could be expected to appear in the dataset, alternatively a method of identifying the cut-off point could be used. The consistently high accuracy and recall of the svm algorithm implies that it is the most universal method out of those included in the simulation study, and as such it can be recommended to use if the researcher does not have comprehensive knowledge about the analysed phenomena.

Although the conducted simulation study allows to formulate some conclusions, further research using real world data is needed to confirm the findings and formulate more concrete recommendations. This study did not include unsupervised machine learning algorithms that are also often used in anomaly detection, in particular when a training dataset with labelled data is not available. Further research could examine the accuracy of those methods in comparison to statistical methods.

References

Aggarwal, C. C. (2017). *Outlier Analysis* (2nd ed. 2017). Springer International Publishing. https://doi.org/10.1007/978-3-319-47578-3

Baron, D., & Poznanski, D. (2017). The Weirdest SDSS Galaxies: Results from an Outlier Detection Algorithm. *Monthly Notices of the Royal Astronomical Society*, *465*(4), 4530-4555.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys (CSUR)*, *41*(3), 1-58.

Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of Artificial Intelligence in Machine Learning: Review and Prospect. *International Journal of Computer Applications*, *115*(9), 31-41.

Faaique, M. (2024). Overview of Big Data Analytics in Modern Astronomy. *International Journal of Mathematics, Statistics, and Computer Science*, *2*, 96-113.

Green, R. F. (1976). Outlier-prone and Outlier-resistant Distributions. *Journal of the American Statistical Association*, *71*(354), 502-505.

Hawkins, D. M. (1980). *Identification of Outliers* (Vol. 11). Springer.

Hofert, M., Kojadinovic, I., Maechler, M., & Yan, J. (2024). *copula: Multivariate Dependence with Copulas*. R package version 1.1-4. https://CRAN.R-project.org/package=copula

Jabez, J., & Muthukumar, B. (2015). Intrusion Detection System (IDS): Anomaly Detection Using Outlier Detection Approach. *Procedia Computer Science*, *48*, 338-346.

Kulkarni, A., Mani, P., & Domeniconi, C. (2017). Network-based Anomaly Detection for Insider Trading. *arXiv Preprint arXiv:1702.05809*.

Lee, L.-F. (1983). Generalized Econometric Models with Selectivity. *Econometrica: Journal of the Econometric Society*, *51*(2), 507-512.

Liu, J., Xie, G., Wang, J., Li, S., Wang, C., Zheng, F., & Jin, Y. (2024). Deep Industrial Image Anomaly Detection: A Survey. *Machine Intelligence Research*, *21*(1), 104-135.

Maddireddy, B. R. (2024). Neural Network Architectures in Cybersecurity: Optimizing Anomaly Detection and Prevention. *International Journal of Advanced Engineering Technologies and Innovations*, *1*(2), 238-266.

Mehrotra, K. G., Mohan, C. K., & Huang, H. (2017). *Anomaly Detection Principles and Algorithms*. Springer International Publishing. https://doi.org/10.1007/978-3-319-67526-8

Nelsen, R. B. (1998). *An Introduction to Copulas*. Springer science & business media.

Prarthana, T. S., & Gangadhar, N. D. (2017). User Behaviour Anomaly Detection in Multidimensional Data. *2017 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, 3-10.

Serrano-Cinca, C., Gutiérrez-Nieto, B., & Bernate-Valbuena, M. (2019). The Use of Accounting Anomalies Indicators to Predict Business Failure. *European Management Journal*, *37*(3), 353-375.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Annales de l'ISUP*, *8*(3), 229-231.

Thimonier, H., Popineau, F., Rimmel, A., Doan, B. L., & Daniel, F. (2024, February). Comparative Evaluation of Anomaly Detection Methods for Fraud Detection in Online Credit Card Payments. In *International Congress on Information and Communication Technology* (pp. 37-50).

Yan, J. (2007). Enjoy the Joy of Copulas: With a Package copula. *Journal of Statistical Software*, *21(4)*, 1-21. https://doi.org/10.18637/jss.v021.i04

Uczenie maszynowe i statystyczne metody wykrywania anomalii – porównawcza analiza symulacyjna

Streszczenie

Cel: Anomalia to obserwacja lub grupa obserwacji nietypowych dla danego zbioru danych. Wykrywanie anomalii ma wiele zastosowań, nie tylko jako etap przygotowania danych do dalszych analiz, lecz także jako sposób wykrywania oszustw z wykorzystaniem kart kredytowych, włamań do sieci i wielu innych. Istnieją różne metody wykrywania anomalii. Można wyróżnić dwie grupy metod, które rozwijane są niezależnie: metody statystyczne oraz algorytmy uczenia maszynowego. Grupy te nieczęsto są porównywane. Podczas gdy metody statystyczne oparte są na sformułowaniu miary nietypowości obserwacji, nadzorowane uczenie maszynowe umożliwia wykorzystanie danych zarówno o typowych obserwacjach, jak i wcześniej zidentyfikowanych anomaliach. Celem artykułu jest dokonanie porównania tych dwóch podejść na podstawie badań symulacyjnych.

Metodyka: W przeprowadzonych badaniach symulacyjnych wykorzystano dane wygenerowane przy użyciu funkcji kopula. W celu wygenerowania różnych rodzajów anomalii dokonano modyfikacji parametrów oraz postaci rozkładów brzegowymi zmiennych. Skuteczność każdej z metod została oceniona na podstawie miar dokładności klasyfikacji.

Wyniki: Podczas gdy skuteczność metod statystycznych zależna była od trafnego zaprognozowania procenta anomalii, jaki pojawi się w danych, metody uczenia maszynowego charakteryzowały się niższą czułością w przypadku wprowadzenia mniejszych zmian wartości parametrów.

Implikacje i rekomendacje: W przypadku metod statystycznych przedstawionych w ramach artykułu kluczowe było posiadanie wiedzy o rozkładzie zmiennych, podczas gdy do zastosowania algorytmów nadzorowanego uczenia maszynowego konieczne było posiadanie zbioru uczącego. W przeciwieństwie do uczenia maszynowego, metody statystyczne uzyskiwały podobną trafność w przypadku wprowadzenia mniejszych zmian wartości parametrów.

Oryginalność/wartość: Dwa podejścia do wykrywania anomalii zaprezentowane w artykule są nieczęsto porównywane. Zazwyczaj metody te są wykorzystywane przez dwie odrębne grupy badaczy – statystyków oraz specjalistów z zakresu uczenia maszynowego lub data science.

Słowa kluczowe: wykrywanie anomalii, badanie symulacyjne, uczenie maszynowe