Hindawi
Complexity
Volume 2018, Article ID 9280787, 12 pages
https://doi.org/10.1155/2018/9280787

Research Article

Case-Based Reasoning: The Search for Similar Solutions and Identification of Outliers

P. S. Szczepaniak and A. Duraj

Institute of Information Technology, Lodz University of Technology, Ul. Wólczanska 215, 90-924 Lodz, Poland

Correspondence should be addressed to A. Duraj; agnieszka.duraj@p.lodz.pl

Received 20 April 2018; Revised 24 June 2018; Accepted 8 July 2018; Published 26 August 2018

Academic Editor: David Gil

The present paper applies the case-based reasoning (CBR) technique to the problem of outlier detection. Although CBR is a widely investigated method with a variety of successful applications in the academic domain, so far, it has not been explored from an outlier detection perspective. This study seeks to address this research gap by defining the outlier case and the underlining specificity of the outlier detection process within the CBR approach. Moreover, the case-based classification (CBC) method is discussed as a task type of CBR. This is followed by the computational illustration of the approach using selected classification methods, that is, linear regression, distance-based classifier, and the Bayes classifier.

1. Introduction

Case-based reasoning (CBR) is a computational problem-solving method that can be effectively applied to a variety of problems [1–10]. Broadly construed, CBR is the process of solving newly encountered problems by adapting previously effective solutions to similar problems (cases). Very important results concerning the equivalence of the learning power of symbolic and case-based methods were presented by Globig and Wess [7]. The authors introduced a case-based classification (CBC) as a variant of the CBR approach and integrated it with basic learning techniques. In particular, they presented the relationship between the case base, the measure of distance, and the target concept of the learning process, while constructing a number of algorithms of great practical significance. Those results justify the validity of the approach to outlier detection proposed in this paper.

In a negative scenario of the CBR cycle execution, the assessment of the nearest neighbour case or other proposed similar cases is negative, which implies that probably no neighbouring cases are useful. In this situation, the current case under consideration is a new one and becomes a candidate to be called an outlier. In such case, the solution must be determined in a different way, but after the solution has been positively revised, the case should be included into the case base of the CBR system. Moreover, some of the cases already included in the case base which were never or hardly ever invoked and adapted by a large number of CBR system uses can be considered outliers. Interesting works that are worth mentioning in the context of outlier processing are those by Smyth and Keane [8, 9] and Richter et al. [10].

A considerable amount of literature has been published on outlier detection and analysis. These studies deal with diverse problem domains involving various types of data, including numeric, textual, categorical, and mixed-attribute records [11–18]. However, to the best of the authors' knowledge, no previous study has investigated case-based reasoning (CBR) from an outlier detection perspective. In this paper, outliers are defined more generally than they used to be to date, that is, as cases in the sense of CBR.

The paper is organized as follows: in Section 2, the principles of the case-based reasoning technique are presented. In Section 3, new definitions of case outlier in relation to CBR are given. Next, case-based classification is described, and computational illustration of outlier detection is given.

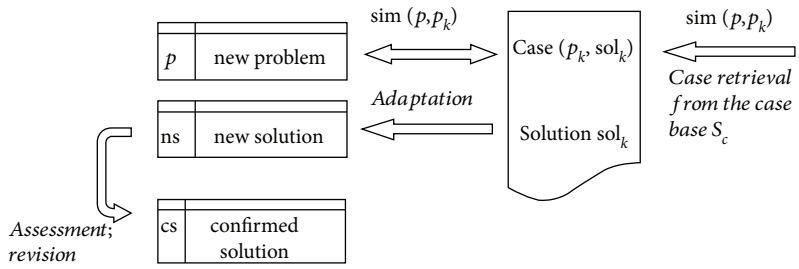

Figure 1: The CBR principle.

Finally, the last section gives a brief summary and critique of the findings.

2. Case-Based Reasoning (CBR)

Case-based reasoning (CBR) is considered a method for problem solving [1–3, 10], and *case-based classification* (CBC) is a task type of CBR. A detailed explanation of CBC is provided in Sections 4 and 5.

In the simplest definition within the CBR methodology, the *case* is understood as an ordered pair:

$$c_i = (p, \mathrm{sol})_i, \tag{1}$$

or, in an extended form, as a triple

$$c_i = (p, \mathrm{sol}, \mathrm{eff})_i. \tag{2}$$

In (1), the previously examined situation (problem) is stored with its solution, and implicitly, the solution was a success. The effect manifested in (2) describes the results obtained through the implementation of the solution.

Medicine-related terms are the following: p—set of symptoms; sol—diagnosis or diagnosis with treatment; and eff—prognosis.

Some relevant data structures need to be used for the proper representation of both: problems and solutions. Here, the *attribute-value representation* is of practical importance. The possible attributes are name, set of values assigned to the name, or a variable. Another term frequently used to describe an attribute is feature.

The concept of *similarity* and its proper application is crucial for the implementation of the CBR system. In general, there are two ways for the computationally applicable similarity representation—relation or function. To reduce the theoretical considerations, the following assumptions are made, which immediately refer to the concept of case:

$$\begin{aligned} 0 \le \mathrm{sim}\left(c_i, c_j\right) \le 1, \\ \mathrm{sim}(c_i, c_i) = 1. \end{aligned} \tag{3}$$

The solution of a new problem p starts with the retrieval of the most similar case (according to the selection criterion); say that this is the kth case, from the base of previously solved cases (S_c). The search for the nearest neighbour comes from the hypothesis that *similar problems have similar solutions*, although the nearest neighbour is not the only reasonable approach. Two situations are possible: (a) the features of both entire cases—the query and the candidate ones—may be compared, or (b) relevant, significant portions of cases can be considered. Then, one considers the associated solution s ol_k, which is either accepted in the given form or must be modified to be useful for the given new problem. This process is referred to as *case adaptation* (Figure 1). It is recommended that the *internal assessment* of the proposed solution of the current problem is performed within the CBR system. An external validation, called *revision*, is the definitive proof for correctness or practical usefulness of the proposed solution—*confirmed solution*. A case is added to the case base if it is recognized as a new one.

As an alternative to the concept of similarity, the concept of distance can be applied to the implementation of the CBR system. However, from the theoretical point of view, these two notions not only reflect different aspects of interpretation but also differ in terms of computational implementation.

Similarity and distance are considered objective notions. From the practical point of view, one looks for useful tools. Usefulness is considered a subjective notion, which can be stated a posteriori. Yet, it can be in some sense expressed by the notion of acceptance interpreted on the basis of the *preference relation* ([1] Chapter 2, [19]):

Given the query q, $c_i >_q c_j$ means that case c_i is preferable to case c_j.

Both similarity and distance can determine preference relations:

$$c_i >_q c_j, \quad \text{if} \quad \mathrm{sim}(q, c_i) \ge \mathrm{sim}\left(q, c_j\right), \tag{4}$$

$$c_i >_q c_j, \quad \text{if} \quad \mathrm{dis}(q, c_i) \le \mathrm{dis}\left(q, c_j\right). \tag{5}$$

The intuitive explanation is that in (4) we look for inclusive arguments, whereas in (5) we prefer rejection or being out of the cluster. The usability of the CBR system is an important feature, which depends strongly on the size and growth of the case base. In larger case bases, the retrieval stages are more expensive. To keep the size of the case base within the limits ensuring the efficiency and proper performance of the system, it is necessary to apply appropriate deletion policies.

In [8, 10], the authors described how the competence of a CBR system can be modelled and how deletion policies can exploit this model to guard against competence depletion while controlling the size of the case base in a manner that guards against the swamping problem. For this reason, the authors found it useful to consider four basic competence

categories of cases: *auxiliary*, *spanning*, *support*, and *pivotal*. They are defined using the concepts of coverage and reachability which are formulated as follows:

Definition 1. Coverage. Given a case base $S_c = \{c_i\}$, $i \in I$. For $c_i \in S_c$,

$$\text{coverage}(c_i) = \{c_j \in S_c : \text{adaptable}(c_i, c_j)\}. \quad (6)$$

Definition 2. Reachability. Given a case base $S_c = \{c_i\}$, $i \in I$. For $c_i \in S_c$,

$$\text{reachable}(c_i) = \{c_j \in S_c : \text{adaptable}(c_j, c_i)\}. \quad (7)$$

The coverage of a case is the set of target problems that can be used to solve. The reachability of a target problem is the set of cases that can be used to provide a solution for the target.

A case is an *auxiliary case* if the coverage it provides is subsumed by the coverage of one of its reachable cases. The cases of this category end to lie within clusters of cases and they do not affect competence at all. Their deletion only reduces the efficiency of the CBR system. Competence is not reduced because if one case is deleted then a nearby case can be used to solve any target that the deleted auxiliary could solve.

The coverage spaces of *spanning cases* span regions of the problem space that are independently covered by other cases. If cases from these linked regions are deleted, then the spanning case may be necessary. In general, they do not directly affect the competence of the system.

Support cases are a special class of spanning cases. They exist in groups, each support providing coverage similar to the others in a group. They also do not affect competence directly. While the deletion of any case (or any proper subset) of a support group does not reduce competence, the removal of the group as a whole is analogous to deleting a pivot and does reduce competence.

A case is called a *pivotal case* if its deletion directly reduces the competence of the system (irrespective of the other cases in the case base). Using the above estimates of coverage and reachability, a case is pivotal if it is reachable by no other case but itself.

The above-mentioned case categories provide a means of ordering cases for deletion in terms of their competence contributions. The auxiliary cases are the least important as they make no direct contribution to competence; next are the support cases, then the spanning cases and, finally, the pivotal cases. The following sections of the paper focus on the last-mentioned of these categories.

The implementation of CBR phases is determined by the *domain* of application, for example, engineering, medicine, or business. For example, medical diagnosis may be considered a simple classification task or a complicated reasoning problem in which one deals with in complete information that requires to be supplemented by redefinition (e.g., extension) of the cases during the repetition of CBR cycles.

The CBR-based approach may be useful in several *task types*, such as information retrieval, planning, design, and classification. The discussion presented in Section 4 focuses on the latter type.

3. Outlier Case

In the literature, there is no single, universally applicable definition of the term outlier, since the formulation of such a definition depends largely on a particular area of application. Thus, the term *outlier* is used to refer to a multitude of concepts, as reflected by the following definitions:

(i) An outlier is an observation which deviates so much from the other observations to arouse suspicions that it was generated by a different mechanism [15].

(ii) Outliers are noise points lying outside the set which defines the clusters, or alternatively, outliers can be defined as points lying outside the set of clusters but are separated from the noise [11].

(iii) An outlier is an observation which deviates so much from the other observations to arouse suspicions that it was generated by a different mechanism [12].

(iv) An observation (or subset of observations) appears to be inconsistent with the remainder of that set of data [14].

(v) A point p in a data set is an outlier with respect to the parameters k and λ, if no more than k points in the data set are at a distance λ or less from p [13].

(vi) Let $X = \{x_1, x_2, \ldots, x_N\}$ for $N \in N$ be a finite, nonempty set of objects. Let S be a finite, nonempty set of attributes (features) of the set of objects X: $S = \{s_1, s_2, \ldots, s_n\}$. Then a subset of objects $X_{\text{out}} \in X$ will be called outliers in the set X if and only if for any subset of attributes $s_i \in S$. The cardinality of subset X_{out} is determined by the linguistic quantifier Q, that is, "little," "few," "very few," "very little," "almost no," and the like [20].

The last ten years have seen increasingly rapid advances in the field of outlier detection, and a variety of outlier detections methods have been proposed, for example, [17, 21–27]. In general, two main approaches to this problem may be distinguished. One way seeks to develop innovative outlier detection algorithms assuming the general definition of an outlier. The other approach, regardless of the application domain, is to employ similar or even the same algorithms, while considering different definitions of the outlier.

For the CBR technique, the following three definitions of outlier case $c_{\text{out}} = (p, \text{sol})_{\text{out}}$ or $c_{\text{out}} = (p, \text{sol}, \text{eff})_{\text{out}}$ can be considered:

(1). An *outlier* can be in general understood as a *pivotal case* (cf. Section 3). Formally, it is defined as follows [8–10]:

$$c_{\text{out}} = \text{pivot}(c), \quad \text{iff} \quad \text{reachable}(c) - \{c\} = \varnothing. \quad (8)$$

Outliers are too isolated to be solved by any other case.

The inconsistency criterion leads to the following definition:

(2). Outlier case c_{out} is understood as the case that appears to be inconsistent with the other cases of S_c. The inconsistency is due to the process of internal assessment or final revision.

When the distance has been defined, the following definition may also be employed:

(3). A case c_i in a case base is called outlier c_{out} with respect to the parameters k and λ, if no more than k cases in that base are at a distance λ or less from c_i. It is assumed that values k and λ confirm the claim about outlierness.

The practical result of finding an outlier is that no useful modification of the solution of the nearest neighbour is possible within the CBR system, which means that the system will not generate an effective, satisfying, and useful solution. In this case, it is necessary to find a solution outside the system and then add it as a new case to the S_c case base.

It is also possible to verify if

(4). Outliers are some of the cases included in S_c which were never or hardly ever adapted by a large number of uses of the CBR system.

However, the above definition (4) is just an observation concerning the work of the CBR system and gives no insight into the nature of the cases under examination, which in fact do not have to represent outliers (items possessing anomalous features).

4. Case-Based Classification (CBC)

A *classifier* is a function which transforms S into K, where K denotes the number of subsets S_k identified in S; that is, $S_k \in S, k \in K$. In other words, K can be understood as the number of labels which can be assigned to objects in S. For a case-based classifier, the following notation is used:

$$(S_c, \text{sim}), \tag{9}$$

where $S_c \subset S$, while sim is defined on $S \times S_c$.

The class of a new object c_i is determined using the defined form of sim assuming that other objects used for comparison are already labelled. Usually, the nearest labelled neighbour is sought or another similar approach is applied; for example, k most similar cases are found and voting for the choice of a proper neighbour is performed. A new case in *case-based classification* is given by the description of an object (problem), and the goal is to assign the correct label (solution) to this object. In CBC, case $c_i = (p, \text{sol})_i$ as defined in (1) is determined entirely by the problem, because the label (class) is uniquely assigned to the object (multilabel classification is assumed to be beyond the scope of the present discussion). In other words, if k is identified as the label (class) of case c_i, then $c_i \in S_k$.

Within the CBC, it is assumed that if for two cases $c_i = (p, sol)_i$ and $c_j = (p, sol)_j$ the problems' descriptions p_i and p_j are similar, then both cases c_i and c_j can be assigned the same class or similar classes. However, the notion *similarity of the classes* must be defined.

It should be mentioned that within the CBR (CBC), learning can be performed, and thus, an initially approximate classifier function can be adjusted. The learning can be performed by modifying the similarity (or distance) measure or by supplementing the case base with new instances.

The characteristic stage of any CBR system is *case adaptation*. With the CBC, the procedure tends to be simpler. For example, if the retrieved similar case is the nearest labelled neighbour, then its solution is the best-known one, and the new solution can be proposed only by performing an external validation, called *revision*. In some situations, this can lead to the introduction of a new label k and consequently extension of set K.

When we seek for the cases that lie at the greatest distance from the already labelled numerous and dominating group of cases, then two situations are possible:

(i) The case is a single one, and an introduction of a new outlier class is necessary.

(ii) The case belongs to one of the existing outlier clusters.

However, when working with preference relation (5), one looks for objects which do not belong to the dominating class of many similar cases. Such objects are called outliers, and in some situations, they may require the introduction of a new label k.

As an example, let us consider the CBC-supported medical diagnosis. The term "supported" needs to be emphasized, because the CBC system only suggests the possible diagnosis, and the final statement falls exclusively within the competence of the physician.

Let the case examined by the system be of the following form:

$$c_i = (p, \text{sol})_i, \tag{10}$$

where p and sol denote the sets of symptoms and diagnosis, respectively. The solution sol_p proposed by the supporting system as the nearest neighbour must be revised by the physician. It may be the case that another solution sol_{co} is indicated by the expert as the correct one. The next step is to verify if the case $c = (p, \text{sol}_{\text{co}})$ already exists. If not, it needs to be included into the case base of the CBC system. If it lies at a great distance from the already classified cases, it may be referred to as an outlier.

5. Selected Computational Approaches

Classification as a method of supervised learning uses labelled observations, with the labels taking nominal values. The purpose of learning is to create a classifier that will assign objects to classes.

The model of the classifier is created according to the pattern of decision classes, most often prepared by an expert. Outliers are deviations from this model. There are many classification methods available, such as decision trees, probabilistic models such as the Bayesian classifier, k-nearest neighbour algorithm, support vector machine, and neural networks, to name a few. Let us examine the outliers in the data set shown in Figure 2. It can be easily noticed that the data belong to the two classes highlighted in green and blue. Note that points A, B, and C lie at a great distance from the rest of the objects. The classification of this data set using the k-NN classifier may result in assigning object A to the blue class, object B to the green one, while object C, depending on the number of neighbours of the k-NN algorithm, may be assigned to the green or to the blue one. In any case, this results in an increased classification error.

A crucial issue related to classification-based outlier detection is the selection of an appropriate classification model. This is a difficult task due to the rarity and atypical character of the feature vector which describes the outlier. While building a classification model, the expert determines decision functions for a particular set of features (a set which is known and often occurring). According to the definitions proposed by Hawkins [15] or Aggarwal [12], the outlier is a vector of atypical and rare characteristics that are not foreseeable for an expert. Therefore, there is a problem with class imbalance or lack of indication of the class with features that point to the existence of outliers.

Another problem associated with the classification of outlying objects is the lack of possibility to balance classes. The layered sampling technique does not provide equivalence of classes. It does not perform well in the case of outliers. There are few outlying objects in the whole set. In the process of layered sampling, the records are first separated according to their classes, and the classes with a small number of objects are selected. In the next step, the objects of the dominant class are randomly selected (the class is regarded as dominant if the majority of objects in the analyzed set belong to it). Yet, an object may appear that has not been assigned to any class.

The characteristic stage of the CBR working cycle, as described in Section 2, is case adaptation. Within CBC, the situation is simple. The assignment to the class occurs as a result of the defined similarity or distance function. For example, if a similar case is found and it is the closest labelled neighbour, its solution is known, and the only way to propose a new solution is by performing external validation. In some situations, this may lead to the introduction of a new label k and, consequently, the extension of the set of classes.

There are many different ways to construct CBC classes. For example, one can

(i) consider the distance between objects;

(ii) determine the number of neighbours that should be used for prediction;

(iii) use an additional weight or include all variables as equally important in the classification;

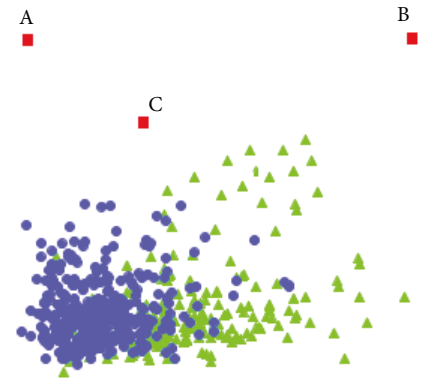

Figure 2: Example of objects A, B, and C being distance outliers.

(iv) apply a specific kind of standardization.

One way to approach outlier identification is to employ a binary statement about whether an object is an outlier or not. This method relies on the subjective opinion of the expert. Another way is to estimate and determine the degree to which the indicated object is an outlier. According to Aggarwal [12], the most interesting observations are those for which the degree of dissimilarity is the highest.

In classification tasks, two approaches to outlier detection can be distinguished, namely, the statistical approach and the approach based on the distance measure between objects. Relating this to the CBC approach, solution sol of problem p can be determined either by the use of a properly defined density function or by the use of a chosen similarity measure. In other words, the search is performed for objects that are at the maximum distance from the already labelled objects, that is, those that are least probable.

The following two approaches are possible:

(A) Determination of outliers without preliminary analysis of the considered set of cases

(B) Two-stage procedure:

Stage 1: dividing the analyzed set of cases into subgroups on the basis of an additional classification criterion (e.g., the medical criterion: healthy or ill);

Stage 2: determination of outliers

In both (A) and (B), the chosen classification method is applied, that is, distance-based outlier detection, Bayesian classifier, and linear regression by calculating Cook's measure.

The statistical approach is directly related to the probability distribution. It assumes that the values of objects in the analyzed set have a specified probability distribution. The objects for which the values of attributes deviate from the distribution are referred to as outliers. In this case, one can specify

(i) nonconformity tests for different probability distributions;

(ii) tests for known or unknown values of probability distribution parameters (distribution characteristics such as mean and standard deviation);

(iii) others

5.1. Regression for Outlier CBR Search. The statistical approach has certain limitations. The tests conducted pertain to a single attribute and, therefore, are not very useful or appropriate for multidimensional data. An additional difficulty may be the complexity and cost of the performed calculations related to the estimation of unknown parameters of polynomial probability distributions. For a more profound discussion, the reader is referred to [12, 14, 15, 28].

In the statistical approach, the so-called loss function is introduced, which enables the calculation of the cost of the classifier's error (mistake). An example of a loss function might be in the form of (11), where 1 means an incorrect decision, while 0 denotes a correct one.

$$L(r,s) = \begin{cases} 1, & r \neq s, \\ 0, & r = s. \end{cases} \tag{11}$$

The main idea of regression is to determine the vector of weights of each independent variable in order to minimize errors. In the regression model, the original independent variables are transformed into independent weighed variables.

Given the notation introduced in Sections 2 and 4, the CBR for the classification of data using regression should be considered as follows: the notation given in Section 3 can be defined as follows:

(i) p (problem)—to find the best distinction between objects of different classes

(ii) sol (solution)—to determine the values of regression parameters in such a way to enable the best adjustment of the values to a given data set

(iii) eff (effect)—the influence of the object on the regression model

The comparison of the two cases of c_k and c_i in the regression model consists in predicting the difference in the output attributes between them according to

$$y_i - y_k = a_1(x_{i1} - x_{k1}) + a_2(x_{i2} - x_{k2}) + \cdots + a_n(x_{\text{in}} - x_{kn}). \tag{12}$$

We can distinguish the following types of exceptional cases:

(i) Case of c_k is an exceptional case c_{out} if in a matrix of differences between two successive values of attributes for the compared cases, there are values different from 0 or greater than the set threshold pr.

(ii) Case of c_k is an exceptional case c_{out} if there has been a significant change in the regression model coefficient and the estimated measures DFFITS, DFBETAS, and Cook (or even one of them) take values above the determined threshold eff.

The regression model is influenced by the so-called high leverage points, which do not necessarily correspond to outliers. In the case of a regression-based classification, outliers are detected on the basis of measures which determine the impact of a given object on the regression model used. For example, Cook's measure determines the level of influence of an object on the model by calculating the squares of the difference between the predicted values of the response variable across the whole sample (the whole set) and the values in the model where the ith observation (ith object) was omitted. eff can be defined as Cook's measure based on Cook's equation (13).

$$D_i = \frac{(y_i - \widehat{y})^2}{ps^2} \frac{h_i}{(1-h_i)^2} = \frac{e_i^2}{p\text{MSE}} \frac{h_i}{(1-h_i)^2}, \tag{13}$$

where D_i is the residual of ith observation, p is the number of parameters in the model, h_i is the influential value of this observation, s is the standard error of the estimator, and MSE is the average square error. Factors $e_i^2/p\text{MSE}$ and $h_i/(1-h_i)^2$ are called the measure of variability and the measure of the leverage of a given observation, respectively.

The high eff value, which is Cook's measure as defined by (13) (value of $D_i > 1$ is considered high), indicates that the deletion of the ith observation from the population has a strong influence on the regression model and, thus, that observation is considered to be influential. Other popular measures that determine the impact of an outlying object on a regression model are DFFITS (difference in FITS) and DFBETAS (difference in betas).

An object is an outlier if for a small sample the value $\text{eff} = |\text{DFFITS}i|$ or $\text{eff} = |\text{DFBETAS}_j(i)|$, $|\text{DFFITS}_i|$, and $|\text{DFBETAS}_j(i)|$ is greater than 1.0. For a large sample, an object is an outlier if the value $|\text{DFFITS}i|$ exceeds $2\sqrt{p/n}$ and $|\text{DFBETAS}_j(i)|$ exceeds $2/\sqrt{n}$. More details can be found in [6].

Depending on the measure adopted to determine the influence of a given object on the regression equation, the eff value must be greater than 1 or, for large samples, greater than $2\sqrt{p/n}$ or $2/\sqrt{n}$.

5.2. Bayes Outlier Case-Based Model. The Bayesian CBR model defines cases according to (1) and (2), as introduced in Section 2. The case is defined by problem p and solution sol. The problem p is a description of objects with characteristic features, for example, a collection of dishes or food products. The solution sol is an allocation of an object to a class, a quintessential observation that it is the best representation of the class. The sim function defining the similarity of the objects is defined by the density function. The new object to be classified belongs to the ith case if the density function is the largest. For the outlier case, the probability function obtained does not indicate the maximum value, but the smallest probability.

Let us consider problem p described using the attributes. $A = \{A_1, A_2, \ldots, A_m\}$, $m \in N$. The Bayesian case-based classifier assigns the label k for the case of $c_i (c_i \in S_k)$. The case c_i is a "prototype" representation of the class (in a sense) of similar observations and is encoded as the vector:

$$c_i = (P_i(a_{11}), \ldots, P_i(a_{1m}), \ldots, P_k(a_{m1}), \ldots, P_k(a_{mn})), \quad (14)$$

where $P_i(a_{ij})$ expresses the probability that the A_i attribute has the value of a_{ij} in the class k.

Of course, the c case database consists of t cases $c_l, \ldots, c_t$, each of which is provided with a unique c_k label. Initially, cases are defined by an expert (alternatively, they may come from a large observation database using statistical clustering methods).

The designated conditional probability (a posteriori) $P(c_K \mid X)$ means that the object $x_i \in X$ is classified into the case $c_{kj} \in K$.

Let $X = \{x_1, x_2, \ldots, x_N\}$ be the set of objects and $k = \{1, \ldots, p\}$. Let the distribution of objects be a discrete probability distribution or probability density $P(x \mid k) \equiv f_k(x)$. Let us introduce the following designations:

(i) $P(K)$—unconditional probability (a priori) of the occurrence of the case K

(ii) $P(X \mid K)$—conditional probability, where the object X belongs to the case K

(iii) $P(X)$—unconditional probability of the occurrence of the object x_i.

$$P(K|X) = \frac{P(K) * P(X|K)}{P(X)}. \quad (15)$$

An object belongs to the case c_K if it fulfils the maximum likelihood principle or the maximum a posteriori principle. The maximum likelihood principle (ML) selects the case $c_{kj} \in K$, which maximizes the conditional probability of the given objects $o \in$ OT (OT objects used as training data).

$$K_{\mathrm{ML}} = \arg\max P(O|k). \quad (16)$$

The maximum a posteriori principle (MAP) consists in selecting the case $c_{kj} \in K$ with the maximum probability a posteriori:

$$K_{\mathrm{MAP}} = \arg\max P(k|O). \quad (17)$$

The case receives a new label (the outlier label) if its $c_{\mathrm{out}} = (p, \mathrm{sol})_{\mathrm{out}}$ is the same as for at least two other different cases. The maximum likelihood principle or the maximum a posteriori principle is not met.

The case receives a new label (the outlier label) if the probability for each previously defined case is much smaller than the threshold assumed by the expert; for example, the value is smaller than 25% of the value of the smallest probability determining the given case.

For the classifiers based on the probability theory (especially for the Bayes classifiers), it is possible to introduce classification weights, which have an impact on the a priori probability value of the decision classes. The other estimates remain unchanged [12]. A special case occurs when the highest probability a posteriori is obtained for several classes. In this situation, it is not possible to unambiguously state to which class the object should be classified. In addition, according to [12, 13, 29, 30], it is not in any case justified to assign an object to the class with the highest probability a posteriori. The authors then propose a classification threshold.

Example 5.1. Let o_i denote objects $O = \{o_1, o_2, \ldots, o_n\}$ for $n \in N$ and r, s be classes to which we assign new objects. If for object o_i the estimated probability for class r is 0.6 and for class s is 0.4, according to Bayer's rules of the classifier, the o_i object is assigned to class r. However, if the threshold for class s is 0.35, then the object o_i is assigned to class s.

5.3. Distance-Based Outlier Detection. Another way to detect outliers is to calculate the distance between objects according to a selected measure.

Taking into account the denotations introduced in Section 3 and definitions for detecting outliers using the distance-based algorithm, we have the following:

(i) p—a problem, that is, the division of objects into c_i classes

(ii) sol—a solution that assigns an object to a class

(iii) dis—the distance between two objects

Assigning a new object to a given case takes place after the distance of that case to the labelled cases is determined. Case c_i gets the c_{out} outlier label if the distance of this case to the other cases exceeds the designated dtc threshold.

In most data classification tasks (similar to those described above), the detection of outlying objects is based on the distance threshold criterion (dtc). If the distance of object o_i, defined by the expert, to object o_k is greater than the specified threshold, then object o_i is considered as an outlier.

$$\mathrm{dis}(o_i, o_k) > \mathrm{dtc}. \quad (18)$$

These objects may represent unusual and previously unknown behaviours or operations. They have a small number of neighbours. They are not removed from the set and still participate in the data analysis but are considered outliers.

We can also say that object o_i is a distance-based outlier in the data set $O = \{o_1, o_2, \ldots, o_n\}$, $n \in N$ if and only if the distance of at most proc percentage of objects from set O is smaller than the distance *dis* an equation (19) is true where $d(o_k, o_i)$ is the measure of the distance between objects o_k and o_i.

$$\frac{||o_i| d(o_k, o_i) \leq \mathrm{dis}|}{|O|} \leq \mathrm{proc}. \quad (19)$$

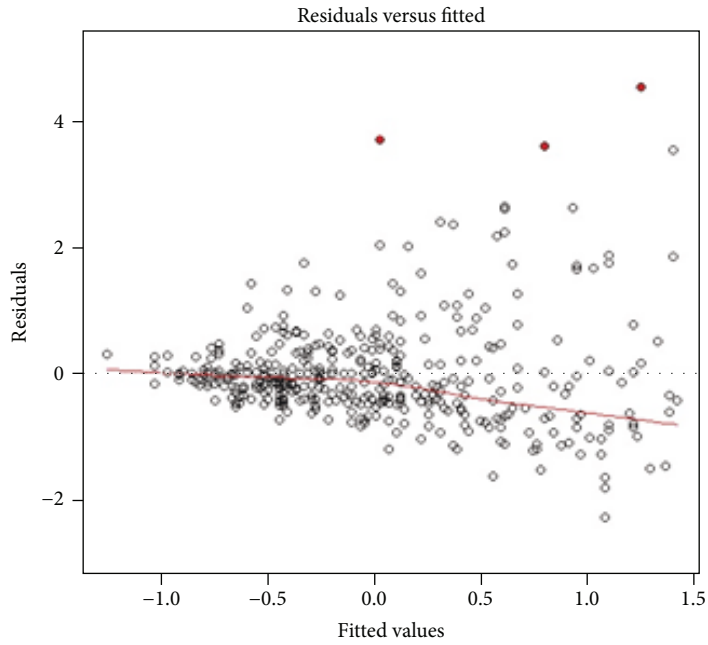

Figure 3: Graph of dependencies between the regression model for the original data set from [32] and residual values with three outliers market.

In terms of the CBC approach, the new case c_i is assigned to a known class if dis is the smallest or receives an outlier label if $\text{dis}(o_i, o_k) > \text{dtc}$ or (19) is true.

The classification-based distance outlier detection may be affected by difficulties due to a high number of dimensions. As the dimensionality increases, all objects are situated at a similar distance to each other. It may be the case that the distance between the object and its nearest neighbour approaches the distance to the furthest neighbour. Therefore, all parameters must be carefully selected. The essential advantage of using distance measures in outlier detection is the fact that it does not require a priori knowledge of the probability distributions.

Figure 2 also highlights the division of distance outliers into global distance outliers and local distance outliers. Objects A and B are global distance outliers because their distance from objects in the whole set is great. Object C can be considered in terms of its isolation degree relative to the nearest neighbourhood, that is, the object from the blue class which is closest to object C and the object from the green class which is closest to object C. Then, the local outlier factor (LOF) is determined. More details on this can be found in [31].

5.4. Evaluation. The evaluation of the performance of both methods was based on a mean square error and a matrix of errors. Cases of correct classification, that is, TP (true positive) and TN (true negative) as well as cases of incorrect classification, that is, FP (false positive) and FN (false positive), were taken into consideration in the matrices of errors. Sensitivity (SE), specificity (SP), and the accuracy were calculated, according to (20). The detection error was determined as the ratio of the number of misclassifications to the sum of all detections.

$$\begin{aligned} \text{SE} &= \frac{\text{TP}}{(\text{TP} + \text{FN})}, \\ \text{SP} &= \frac{\text{TN}}{(\text{TN} + \text{FN})}, \\ \text{ACC} &= \frac{(\text{FP} + \text{FN})}{(\text{TN} + \text{FP} + \text{FN} + \text{TP})}. \end{aligned} \tag{20}$$

It should be noted that a large number of FP or FN contribute to an increase in the classification error. The consequence of FP detection is the detection of outliers. This may also be the reason for creating a class with a new pattern.

6. Practical Example

6.1. Classic Methods. The experimental research was carried out using the benchmark (repository) database [32], which originally contained 868 records. The data collected included information concerning blood glucose, glucose (plas), blood pressure- (pres-) diastolic blood pressure, skin thickness (skin)—thickness of skin on triceps (mm), age, weight, BMI, pregnancies (preg)—the attribute stating the number of pregnancies of the patient, and inheritance risk ratio—the factor of the risk of inheriting diabetes. Over 2000 records were taken into account. The data set was examined for the presence of outliers using the classic regression method (cf. Figure 3). Three cases of outliers were detected in the set under examination using Cook's measure.

To make the experiments more reliable, the data set was extended by 1200 new records in which 9 known outliers

Table 1: Best results of outlier detection using classic methods.

Classic	Number of detected outliers	Percentage of correct detections (%)
Regression	7	58
Bayes	6	50
k-NN	5	42

Table 2: Measures of rating classic methods.

Method	SE sensitivity	SP specificity	ACC accuracy	Classification error
Bayes	0.69	0.52	0.31	0.35
k-NN	0.67	0.41	0.3	0.42

were incorporated. The resulting total number of data records was 2077, in which 12 known outliers were hidden.

The results obtained by three classic methods used for comparison are collected in Table 1. Measures of rating are shown in Table 2.

In the literature [33] referring to the k-NN method, the following formula is recommended for determining the optimal value of parameter k: $k^* = \sqrt{N}$, where N is the number of cases chosen to learn.

However, this recommended value $k^* = 21$ did not work for the database examined and only 1 or 2 outliers were detected. In general, each value bigger than 10 was unsatisfactory. The best results were obtained for $k = 5$ and $k = 6$, which were determined experimentally.

In the case of the k-nearest neighbour classifier, outliers are the objects whose distance from the nearest neighbour is much greater than that for the other objects. Thus, it is possible to specify that the distance between objects cannot be greater than the distance given by the expert. Figure 4 shows an example data dispersion where the circles are healthy persons, pluses (crosses) indicate the class of healthy people, and the rhombuses represent the outliers. The distance of the nearest neighbour in the case of 5 points is much higher than that in the other cases. Therefore, these points are likely to be classified as false positives or false negatives. This leads to an increased classification error.

6.2. Regression CBC: Cook's Measure. In the case of the CBR method with the use of linear regression, Cook's measure was applied to estimate the level of influence of the object on the model.

The outlying objects indicated by Cook's measure are shown in Figure 5. Cook's values above the line indicate the existence of outliers in the analyzed set. The graph of dependencies between the CBC regression model and residual values for 2077 cases with 10 outlier cases marked is shown in Figure 6.

6.3. Naive Bayesian Classifier CBC. The naive Bayesian classifier, as a probabilistic classifier, estimates the frequency of occurrence of objects with specified parameters for each class. In our case, outliers occur very rarely. Thus, it is

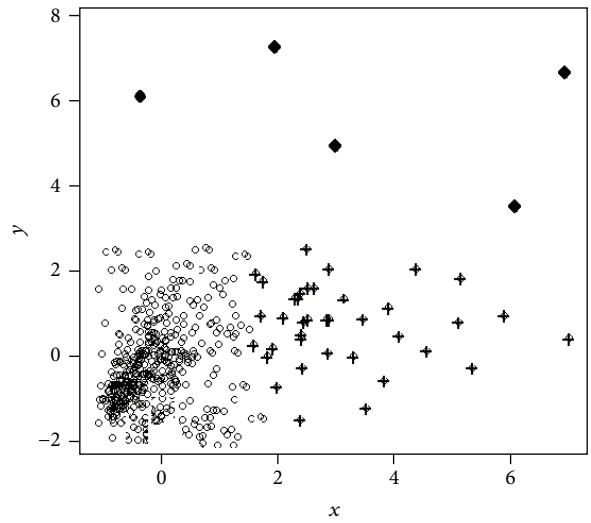

Figure 4: Outliers clearly distancing from the other objects—the result obtained by k-NN ($k = 5$).

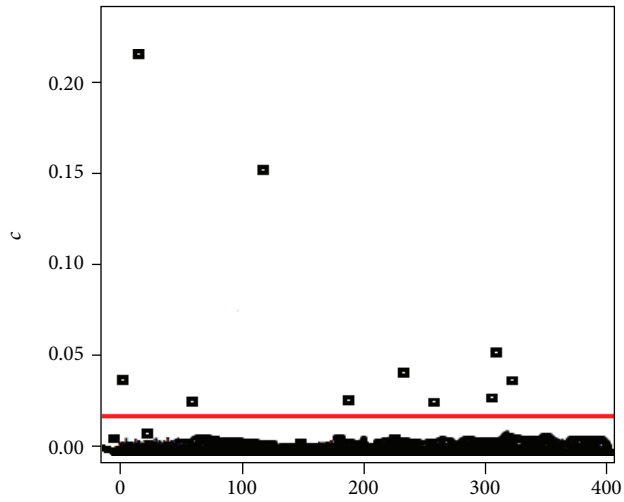

Figure 5: Outliers indicated using Cook's measure.

difficult to speak of the determination of occurrence frequency. It is not always helpful to use Laplace's expansion. In addition, the naive Bayesian classifier assumes that the total density of objects is a product of boundary densities. The testing of the CBR method with the Bayes classifier consisted of two stages. The classification was performed for all cases under consideration, that is, the whole given data set. Due to the fact that there were outliers in the analyzed database, the value of the classification error obtained was 0.27 (see Table 3).

The classification error decreased after using the Bayes outlier case-based reasoning method, in which a separate class of outlier cases was initially found, without preliminary classification. The results obtained using four evaluation measures are summarized in Table 3.

The classification error decreased after using the Bayes outlier case-based reasoning method, in which a separate class of outlier cases was initially found, without preliminary

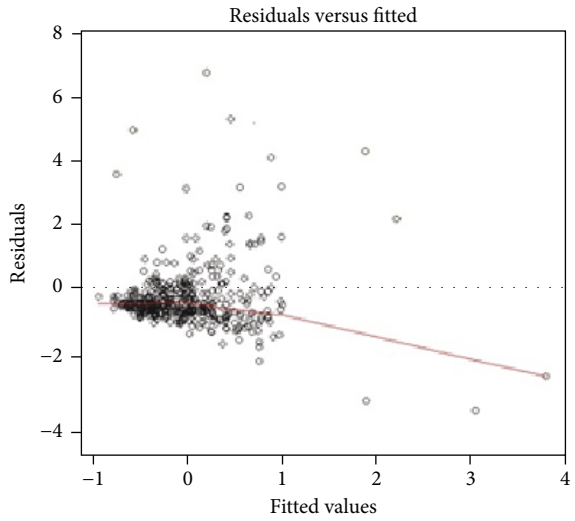

Figure 6: Position of ten outliers detected in the extended database.

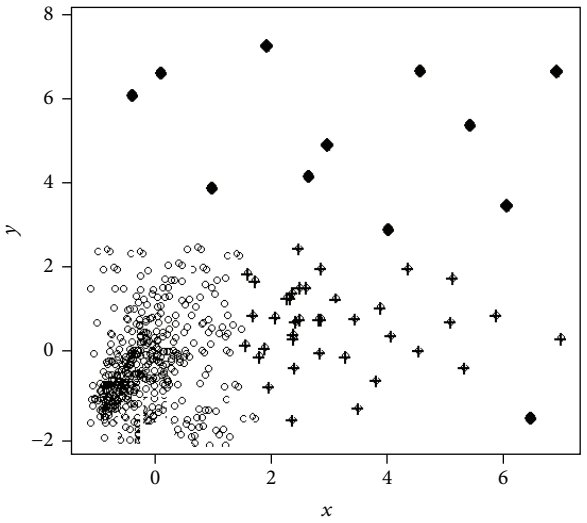

Figure 7: Outliers detected for distance-based CBC.

Table 3: Measures of rating classifiers for Bayes case-based reasoning for two approaches (A) and (B).

Bayes	SE sensitivity	SP specificity	ACC accuracy	Classification error
A	0.75	0.66	0.27	0.27
B	0.8	0.7	0.23	0.21

Table 4: Measures of rating distance-based reasoning.

Distance-based CBC	SE sensitivity	SP specificity	ACC accuracy	Classification error
A	0.81	0.79	0.26	0.6
B	0.77	0.56	0.30	0.28

classification. The results using four evaluation measures are summarized in Table 3.

6.4. Distance-Based CBC. The distane-based classification, like the Bayesian classification, was used in two stages (A) and (B). The results are given in Table 4 and Figure 7.

7. Summary

The paper has presented the application of case-based reasoning (CBR) and case-based classification (CBC) to the problem of outlier detection. The formal definition of case outlier has been introduced. The study has demonstrated a CBC framework for the interpretation of several classification approaches. The method proposed here was validated using a practical example from the field of medicine.

The results obtained using the CBR approach, which are described in detail in Section 5, were significantly better than classic methods. For example, the graph of dependencies between the designated CBC regression model and residual values for 2077 cases with 10 outliers marked is shown in Figure 6. Better results were obtained also using the Bayes CBC method and distance-based CBC (Figure 7). The complete set of results is collected in Table 5.

Table 5: Best results of outlier detection using CBR (CBC) approach.

CBR (CBC)	Number of detected outliers	Percentage of correct detections (%)
Regression	10	83
Bayes	11	92
Distance-based CBC	12	100

As can be seen from Tables 3, 4, and 5, the presence of outliers in the data set makes the classification much more difficult. The implementation of CBR without creating a subgroup of cases resulted in the decrease in sensitivity and accuracy, while increasing the classification error.

The application of the case-based reasoning approach resulted in the classification error decreasing by 0.06 and 0.32 for the Bayes classifier and the distance-based CBC, (Tables 3 and 4), respectively. It should be emphasized that the strongest resistance to the occurrence of outliers was demonstrated by the Bayes outlier case-based classification. However, the results produced of the Bayes and distance-based classification method are of similar quality.

Nomenclature

S: The universe of all objects
S_c: The case base
c_i: The ith case, $i \in I$
sim: Similarity (formal definition is the appropriate approach)
dis: Distance
p: Problem
sol: Solution
eff: Effect.

Data Availability

In the research, public repositories were used, and they are available on the website http://archive.ics.uci.edu/ml/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. Lenz, B. Bartsch-Spörl, H. D. Burkhard, and S. Wess, *Case-Based Reasoning Technology: From Foundations to Applications. Volume 1400*, Springer, 2003.

[2] S. K. Pal, T. S. Dillon, and D. S. Yeung, *Soft Computing in Case Based Reasoning*, Springer Science & Business Media, 2012.

[3] P. Perner, *Case-Based Reasoning on Images and Signals. Volume 73*, Springer, 2008.

[4] K. D. Althoff, R. Bergmann, S. Wess et al., "Case-based reasoning for medical decision support tasks: the inreca approach," *Artificial Intelligence in Medicine*, vol. 12, no. 1, pp. 25–41, 1998.

[5] A. S. Ochi-Okorie, "Disease diagnosis validation in TROPIX using CBR," *Artificial Intelligence in Medicine*, vol. 12, no. 1, pp. 43–60, 1998.

[6] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Volume 571*, John Wiley & Sons, 2005.

[7] C. Globig and S. Wess, "Learning in case-based classification algorithms," in *Algorithmic Learning for Knowledge-Based Systems. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 961*, K. P. Jantke and S. Lange, Eds., pp. 340–362, Springer, Berlin, Heidelberg, 1995.

[8] B. Smyth and M. T. Keane, "Remembering to forget," in *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence*, pp. 377–382, Montreal, Quebec, Canada, August 1995.

[9] B. Smyth and M. T. Keane, "Footprint based retrieval," in *Case-Based Reasoning Research and Development. ICCBR 1999. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, K. D. Althoff, R. Bergmann, and L. Branting, Eds., pp. 134–148, Springer, Berlin, Heidelberg, 1999.

[10] M. M. Richter, R. O. Weber, and C. B. Reasoning, *A Textbook. Organic Chemistry*, John Wiley & Son, Inc, New York, NY, USA, 2013.

[11] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *SIGMOD '01 Proceedings of the 2001 ACM SIGMOD International Conference on Management of data*, pp. 37–46, Santa Barbara, CA, USA, May 2001.

[12] C. C. Aggarwal, *Outlier Analysis*, Springer Science & Business Media, 2013.

[13] E. M. Knorr and R. T. Ng, "A unified notion of outliers: properties and computation," in *KDD'97 Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pp. 219–222, Newport Beach, CA, USA, August 1997.

[14] V. Barnett and T. Lewis, *Outliers in Statistical Data. Volume 3*, Wiley, New York, NY, USA, 1994.

[15] D. M. Hawkins, *Identification of Outliers. Volume 11*, Springer, 1980.

[16] A. Duraj, P. S. Szczepaniak, and J. Ochelska-Mierzejewska, "Detection of outlier information using linguistic summarization," in *Flexible Query Answering Systems 2015. Advances in Intelligent Systems and Computing*, pp. 101–113, Springer, 2016.

[17] A. Duraj and P. S. Szczepaniak, "Information outliers and their detection," in *Information Studies and the Quest for Transdisciplinarity. Volume 9, Chapter 15*, M. Burgin and W. Hofkirchner, Eds., pp. 413–437, World Scientific Publishing Company, 2017.

[18] I. Jurisica, J. Mylopoulos, J. Glasgow, H. Shapiro, and R. F. Casper, "Case-based reasoning in IVF: prediction and knowledge mining," *Artificial Intelligence in Medicine*, vol. 12, no. 1, pp. 1–24, 1998.

[19] J. N. Mordeson, D. S. Malik, and S. C. Cheng, *Fuzzy Mathematics in Medicine*, Springer-Verlag New York, Inc, 2000.

[20] A. Duraj, A. Niewiadomski, and P. S. Szczepaniak, "Outlier detection using linguistically quantified statements," *International Journal of Intelligent Systems*, vol. 33, no. 8, pp. 1590–1601, 2018.

[21] A. Duraj and L. Chomatek, "Outlier detection using the multiobjective genetic algorithm," *Journal of Applied Computer Science*, vol. 25, no. 1, pp. 29–42, 2017.

[22] A. Duraj and D. Zakrzewska, "Effective outlier detection technique with adaptive choice of input parameters," in *Intelligent Systems'2014. Advances in Intelligent Systems and Computing*, pp. 535–546, Springer, 2015.

[23] A. Duraj and L. Chomatek, "Supporting breast cancer diagnosis with multi-objective genetic algorithm for outlier detection," in *Advanced Solutions in Diagnostics and Fault Tolerant Control. DPS 2017. Advances in Intelligent Systems and Computing*, J. Kościelny, M. Syfert, and A. Sztyber, Eds., pp. 304–315, Springer, 2017.

[24] M. Kalisch, M. Michalak, M. Sikora, Ł. Wróbel, and P. Przystałka, "Data intensive vs sliding window outlier detection in the stream data an experimental approach," in *Artificial Intelligence and Soft Computing. ICAISC 2016. Lecture Notes in Computer Science*, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. Zadeh, and J. Zurada, Eds., pp. 73–87, Springer, 2016.

[25] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Reverse nearest neighbors in unsupervised distance-based outlier detection," *IEEE transactions on knowledge and data engineering*, vol. 27, no. 5, pp. 1369–1382, 2015.

[26] G. O. Campos, A. Zimek, J. Sander et al., "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 891–927, 2016.

[27] C. Titouna, M. Aliouat, and M. Gueroui, "Outlier detection approach using bayes classifiers in wireless sensor networks," *Wireless Personal Communications*, vol. 85, no. 3, pp. 1009–1023, 2015.

[28] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," in *Data Warehousing and Knowledge Discovery. DaWaK 2002. Lecture Notes in Computer Science*, Y. Kambayashi, W. Winiwarter, and M. Arikawa, Eds., pp. 170–180, Springer, Berlin, Heidelberg, 2002.

[29] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Procedia Computer Science*, vol. 60, pp. 708–713, 2015.

[30] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB Journal*, vol. 8, no. 3-4, pp. 237–253, 2000.

[31] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, Dallas, TX, USA, May 2000.

[32] C. J. Merz and P. M. Murphy, *{UCI} Repository of Machine Learning Databases*, 1998.

[33] G. G. Enas and S. C. Choi, "Choice of the smoothing parameter and efficiency of *k*-nearest neighbor classification," in *Statistical Methods of Discrimination and Classification*, pp. 235–244, Elsevier, 1986.