Mind the Gap: Addressing Incompleteness Challenge in Case-Based Reasoning Applications

Fateh Boulmaiz, Patrick Reignier, Stephane Ploix

MIND THE GAP: ADDRESSING INCOMPLETENESS CHALLENGE IN CASE-BASED REASONING APPLICATIONS

Fateh Boulmaiz[1], Patrick Reignier[1], and Stephane Ploix[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France
[*] Corresponding author: fateh.boulmaiz@univ-grenoble-alpes.fr

Abstract. Data quality is a crucial aspect of case-based reasoning (CBR), and incomplete data is a ubiquitous challenge that can significantly affect the accuracy and effectiveness of CBR systems. Incompleteness arises when a case lacks relevant information needed to solve a problem. Existing CBR systems often struggle to handle such cases, leading to suboptimal solutions, and making it challenging to apply CBR in real-world settings. This paper highlights the importance of data quality in CBR and emphasizes the need for systems to handle incomplete data effectively. The authors provide for the first time a framework for addressing the issue of incompleteness under the open-world assumption. The proposed approach leverages a combination of data-driven and knowledge-based techniques to detect incompleteness. The approach offers a promising solution to the incompleteness dimension of data quality in CBR and has the potential to improve the practical utility of CBR systems in various domains as illustrated by the results of a real data-based evaluation.

Keywords: Case based reasoning · Data quality· Data completeness.

1 Introduction

In the era of extensive digitization and ubiquitous computing, ensuring the quality of data manipulation has emerged as a critical challenge for companies and academic research across various fields such as database, artificial intelligence, image processing, information systems, and more. Numerous studies have highlighted the significant impact of data quality on the handling process. For instance, research has shown that the quality of data utilized in machine learning algorithms directly influences their performance [21,26]. A survey conducted in [1] has also shed light on the adverse effects of poor data quality on a country's economy. It estimates that the US economy alone loses over $3 trillion annually due to poor data quality, and this financial cost is still on the rise [2].

As organizations face increasingly complex data issues that can impact their profitability, and as research proposes more data quality-sensitive algorithms, the importance of accurate and trustworthy data has never been more critical. However, because of the diverse objectives and ways of using data, different data

quality dimensions (requirements) exist, which characterize quality properties such as accuracy, completeness, and consistency. Despite the extensive literature devoted to data quality (see [11] for an overview), it is worth noting that: 1) There is no consensus on the properties that should be considered when defining a data quality standard, despite ongoing research in this area [18]. For example, while the authors in [28] identify 179 dimensions for data quality, a more recent study [15] describes more than 300 properties that should be considered for defining data quality; 2) Although some requirements have been universally identified as important, there is no agreement on their precise definitions. The same requirement name may have different meanings in different studies; 3) Because of the diversity of data sources, the multitude of quality dimensions, and the specificity of the application domain, data quality assessment is a domain-specific process. Therefore, it is not possible to propose a generic data quality assessment approach that can be applied to all data-intensive applications.

Although data quality has been extensively studied by the database and data mining communities, it has been overlooked in the machine learning domain, where the focus is on developing learning algorithms and reasoning approaches that assume high-quality data. This paper aims to address this gap by exploring the issue of data quality in the context of machine learning, with the goal of improving the robustness of learning algorithms and reducing the impact of poor quality data on overall results. However, due to the broad scope of both data quality and machine learning, certain limitations were necessary to make this work feasible. Specifically, we restrict our research to approaches based on the case-based reasoning paradigm, as the machine learning domain is too vast to provide a single data quality assessment method that is valid for all approaches. Furthermore, we only investigate the data completeness dimension, which is still largely unexplored in the context of case-based reasoning. In this paper, we propose a heuristic based on the change point detection method to address the issue of data incompleteness in the CBR approach.

The remainder of the paper is organized along the following lines: Section 2 reviews the existing literature on the topic. Section 3 describes the background of the research. Section 4 outlines the problem statement through an motivating example followed by a formulation of the problem. Section 5 details the proposed approach to address the problem under consideration. Section 6 evaluates the proposed approach through a real case study and discusses the results. Section 7 concludes the paper and presents future work.

2 Related Work

Data incompleteness is a common problem in various domains, and many methods have been proposed to address this issue. However, traditional methods for handling missing data assume a closed-world assumption, which means that any unobserved value is assumed to be missing at random from the same distribution as the observed data. This assumption can be problematic in some cases where data is incompletely observed under an open-world assumption, where

unobserved values can be missing because they are not present in the data-generating process. One of the early works in this area is the paper [24], which introduced the concept of the open world assumption in the context of incomplete data. This challenge has been studied in several domains, and various approaches have been proposed to handle data incompleteness under the OWA. One approach is to use probabilistic models to reason about missing values, such as Bayesian networks and Markov logic networks (MLNs). For example, in the domain of natural language processing (NLP), there has been extensive work on using probabilistic models to handle data incompleteness in text corpora [17]. MLNs have also been used to handle incomplete data in other domains, such as bioinformatics [20] and image analysis [4].

Recently, deep learning methods have also been proposed for handling data incompleteness under the open world assumption. . For instance, DeepProbLog [19] is a probabilistic programming language that combines deep learning with logic programming to handle data incompleteness in relational domains. Another example of such technique is DeepImpute method [3], which is a deep learning-based imputation method that has been applied in various domains, such as genomics and biomedical data analysis. Other deep learning approaches for handling incomplete data include generative models [6] and adversarial training [29]. Works [9,23] demonstrate the effectiveness of deep learning models for imputing missing data.

In addition to these methods, there have been several studies focused on understanding the causes and consequences of data incompleteness. For instance, The authors in [30] analyzed the impact of missing data on the medical domain and concluded that the missingness mechanism (i.e., the reason why data is missing) plays a critical role in determining the appropriate analysis strategy.

Overall, the literature on data incompleteness is vast, and a wide range of techniques have been developed to address this challenge. However, there is no one-size-fits-all solution. Each approach has its strengths and weaknesses, and the appropriate method depends on the characteristics of the data and the objectives being addressed by the application.

Surprisingly, despite the pervasiveness of incomplete data issue in real-world applications and the importance of addressing this issue in various fields, there has been a lack of attention given to incomplete data in case-based reasoning. Thus, there is a need for further research in CBR to address the issue of incomplete data and develop approaches that can detect, handle, and reason with incomplete information effectively.

3 Background

3.1 Case-based reasoning and data completeness

Case-based reasoning (CBR) is a reasoning paradigm based on a case base $\mathbb{CB}$ representing a collection of source cases. A case $\mathbb{C}$ represents an experience of problem-solving, usually defined by a couple $(\mathfrak{p}, \mathfrak{s})$ wherein $\mathfrak{p}$ is a problem in

the considered application domain and $\mathtt{s}$ is its solution. In the following, we assume a finer representation of the case $\mathbb{C}$ as a triplet $(\mathtt{C}^{\mathbb{C}}, \mathtt{A}^{\mathbb{C}}, \mathtt{E}^{\mathbb{C}})$ [8]. Let $\mathbb{C}^S$, $\mathbb{A}^S$, and $\mathbb{E}^S$ be three sets. The context $\mathtt{C}^{\mathbb{C}}$ is an element of $\mathbb{C}^S$ representing the phenomena undergone by the application domain. The actions $\mathtt{A}^{\mathbb{C}}$ is an element of $\mathbb{A}^S$ modeling controllable phenomena of the application domain. The effects $\mathtt{E}^{\mathbb{C}}$ is an element of $\mathbb{E}^S$ describing the consequence of the application of the actions $\mathtt{A}^{\mathbb{C}}$ to the context $\mathtt{A}^{\mathbb{C}}$. The intuition underlying the CBR paradigm is formulated by Hypothesis 1. The process of solving a target case $\mathbb{C}_{tg}$, which is formed initially from the context only, consists in calculating the relevant actions, which once applied will produce effects, generating a new source case in the case base.

Assumption 1 (Consistency) *The effects of applying similar actions to similar contexts are similar.*

Precisely, the reasoning strategy starts by looking for the set $\mathtt{SIM}^{\mathbb{C}_{tg}}$ of source cases $\mathbb{C}_{sr}$ similar to the target case $\mathbb{C}_{tg}$ (retrieval stage), followed by the modification of the actions of the cases $\mathbb{C}_{sr}$ to match the specificity of the context of the case $\mathbb{C}_{tg}$, generating thus the actions $\mathtt{A}^{\mathbb{C}_{\mathtt{tg}}}$ (adaptation stage). According to the adopted validation stage, the effects $\mathtt{E}^{\mathbb{C}_{\mathtt{tg}}}$ of the application of $\mathtt{A}^{\mathbb{C}_{\mathtt{tg}}}$ to the context $\mathtt{C}^{\mathbb{C}_{\mathtt{tg}}}$ are generated, and thus the new target case $\mathbb{C}_{tg}(\mathtt{C}^{\mathbb{C}_{\mathtt{tg}}}, \mathtt{A}^{\mathbb{C}_{\mathtt{tg}}}, \mathtt{E}^{\mathbb{C}_{\mathtt{tg}}})$, if approved, is integrated into the case base $\mathbb{CB}$ (memorization stage). This can be formalized as follows:

$$\mathtt{CBR\ system} : \mathtt{Memorization} \circ \mathtt{Validation} \circ \mathtt{Adaptation} \circ \mathtt{Retrieval}$$

$$\mathtt{Retrieval\ function} : \mathtt{C}^{\mathbb{C}_{\mathtt{tg}}} \longmapsto \mathtt{SIM}^{\mathbb{C}_{tg}} = \{\mathbb{C}_{sr}\} \subseteq \mathbb{CB}$$

$$\mathtt{Adaptation\ function} : \mathtt{SIM}^{\mathbb{C}_{tg}} \cup \mathtt{C}^{\mathbb{C}_{\mathtt{tg}}} \longmapsto \mathtt{A}^{\mathbb{C}_{\mathtt{tg}}} \cup \{\mathtt{failure}\}$$

$$\mathtt{Validation\ function} : \mathtt{A}^{\mathbb{C}_{\mathtt{tg}}} \longmapsto \mathbb{C}_{tg}(\mathtt{C}^{\mathbb{C}_{\mathtt{tg}}}, \mathtt{A}^{\mathbb{C}_{\mathtt{tg}}}, \mathtt{E}^{\mathbb{C}_{\mathtt{tg}}})$$

$$\mathtt{Memorization\ function} : (\mathbb{CB}, \mathtt{C}^{\mathbb{C}_{\mathtt{tg}}}) \longmapsto \mathbb{CB} \cup \mathbb{C}_{tg}(\mathtt{C}^{\mathbb{C}_{\mathtt{tg}}}, \mathtt{A}^{\mathbb{C}_{\mathtt{tg}}}, \mathtt{E}^{\mathbb{C}_{\mathtt{tg}}})$$

To conduct the different steps of the reasoning process, a CBR system draws on a set of knowledge spread over four containers: domain, case, similarity, and adaptation knowledge [25]. Usually, each stage of the reasoning process is supported by several knowledge containers because of the close connections existing between them.

Completeness. In keeping with existing literature on Knowledge bases [12] and databases [14], we consider completeness through an ideal reference domain knowledge container $\mathtt{K}^{\mathtt{D}}_{\mathtt{R}}$, which captures all the real-world aspects of the application domain. The domain knowledge $\mathtt{K}^{\mathtt{D}}$ of CBR system is complete if the application of any actions (defined in $\mathtt{K}^{\mathtt{D}}$) to any context (likewise defined in $\mathtt{K}^{\mathtt{D}}$) generates the same effects on $\mathtt{K}^{\mathtt{D}}$ as on $\mathtt{K}^{\mathtt{D}}_{\mathtt{R}}$.

Definition 1. *[Completeness] Completeness refers to the ability of the domain knowledge container of a CBR system to describe every relevant state of the domain application environment.*

The principal barrier to assessing and achieving completeness, as stated in Definition 1, is the Open World Assumption. The latter states that if a given

piece of real-world knowledge is not represented in the Knowledge domain $\mathtt{K}^{\mathtt{D}}$, then that knowledge is not necessarily false, it may be real-world true but not included in the $\mathtt{K}^{\mathtt{D}}$.

A plethora of work has been done on data quality assessment, which continues to be an intense research domain in such diverse fields as relational databases, big data, machine learning, data mining, etc. Data quality verification remains a challenging process for several reasons:

- *Data quality verification is a permanent process.* This is due to the data nature (particularly, their velocity) on one side and the different processings performed on the data (e.g., data cleaning) on the other side.
- *Data quality verification is strongly dependent on the application-task domain.* The different dimensions of data quality are evaluated by metrics whose specification strongly depends on the needs of the user/expert, the application domain (the aeronautics domain does not have the same requirements in terms of data quality as the education domain, for example) but also on the task (in the health domain, there are different requirements for the diagnostic phase and the treatment phase).

3.2 Change point analysis

Change points in a data set modeling a system are defined by abrupt shifts in the data. These change points can represent transitions that occur between states of the modeled system due to hidden changes in the properties of the data set. Determining the change points in a data set is the objective of the change point analysis approaches, which have sparked an increasing work in statistics [27] as well as in several application domains such as climate [13], medical [31], finance [10].

More formally, consider a system characterized by non-stationary random phenomena and modeled by a multivariate vector $\Omega = \{\omega_1, \ldots, \omega_m\}$ whose values are defined in $\mathbb{R}^{d \geq 1}$ and consisting of m samples. It is further supposed that the vector Ω is piecewise stationary, i.e., certain phenomena of the system change abruptly at unknown instants $t_1, t_2, \ldots, t_m$. The detection of the change points consists in solving a model detection problem whose objective is to determine the optimal segmentation S based on a quantitative criterion to be minimized. Specifically, it consists in identifying the number m of changes and finding the indices $t_{i_{(1 \leq i \leq m)}}$.

4 Problem setting

4.1 Motivating example

We motivate the need to guarantee the completeness of data in a CBR system through a concrete scenario. Let's consider the scenario of a CBR-based energy management system (EMS) that monitors a building equipped with an air-conditioning (AC) system, but the EMS designer has not envisaged any means

to discover the AC system function. On two days with a similar context (e.g., the same weather conditions) and the same actions, if the AC system was turned on one day but not on the other (this phenomenon cannot be detected by the system), the two days would have different effects (e.g., different indoor temperatures), which calls into question the founding assumption of the CBR technique.

4.2 Problem statement

Existing CBR systems exploit directly the case base to carry out the different steps of the CBR cycle, assuming that the domain knowledge is consistent. Indeed, by adopting the consistency assumption (Assumption 1), it is implicitly admitted that the completeness hypothesis is valid. However, it is arguably not warranted, especially considering the modeling of a complex domain with many dependent variables. The violation of the completeness assumption poses some substantive issues:

- the system has no guarantee that the principle of the CBR approach (Assumption 1) is respected.
- the CBR system cannot identify incomplete data and therefore cannot determine which data reflects reality for use in the reasoning process.
- as a consequence of the previous statements, the performance of the reasoning process may degrade as the case base includes cases that are wrongly judged as similar.

The failure of one of four knowledge containers to be adequately defined (incomplete) can be overwhelming to the whole CBR system unless any of the remaining knowledge containers can fill the missing knowledge. As a result, either the CBR system will fail to respond or provide inaccurate solutions. In particular, it was established that incomplete domain knowledge generates such a critical dysfunction of a CBR system [5]. Incomplete domain knowledge in a CBR system most likely leads to the generation of incomplete cases. Moreover, the retrieval process is burdened by the absence of missing data since the similarity evaluation is biased by the incompleteness of the data. Furthermore, incomplete cases can also degrade the adaptation process when the adaptation knowledge is acquired automatically from the case base.

It is obvious that the problem of incompleteness verification can be reformulated as a hidden variable detection problem. Indeed, an incompleteness situation occurs in a case base when a group of similar cases produces different effects, which is necessarily a consequence of the existence of context and/or action variables that are not considered in the similarity evaluation process.

Formally, consider a case base $\mathbb{CB} = \{\mathbb{C}_i\}_{1 \leq i \leq n}$ consisting of a finite number n of cases $\mathbb{C}_i$. Each element of the latter is described by a set of features. The context $\mathtt{C}^{\mathbb{C}_i}$ of case $\mathbb{C}_i$ is specified by $\mathtt{C}^{\mathbb{C}_i} = \{O^{\mathbb{C}_i}_{\mathtt{C}j}\}_{1 \leq j \leq n_1}$, where the observed features $O_{\mathtt{C}}$ are defined on the knowledge domain $\mathtt{K}^{\mathtt{D}}_{\mathtt{C}}$. The actions $\mathtt{A}^{\mathbb{C}_i}$ are modeled by the features $\{O^{\mathbb{C}_i}_{\mathtt{A}j}\}_{1 \leq j \leq n_2}$ which are defined on the knowledge domain $\mathtt{K}^{\mathtt{D}}_{\mathtt{A}}$, and the effects are specified on the knowledge domain $\mathtt{K}^{\mathtt{D}}_{\mathtt{E}}$ by the features $\{O^{\mathbb{C}_i}_{\mathtt{E}j}\}_{1 \leq j \leq n_3}$. The knowledge domain $\mathtt{K}^{\mathtt{D}}$ of the CBR system is defined by $\mathtt{K}^{\mathtt{D}} = \mathtt{K}^{\mathtt{D}}_{\mathtt{C}} \cup \mathtt{K}^{\mathtt{D}}_{\mathtt{A}} \cup$

$\mathtt{K}^{\mathtt{D}}_{\mathtt{E}}$. Let's also assume, $\{H_{\mathtt{C}j}\}_{1\leq j\leq m_1}$, $\{H_{\mathtt{A}j}\}_{1\leq j\leq m_2}$, and $\{H_{\mathtt{E}j}\}_{1\leq j\leq m_3}$ are the hidden features of the context, action, and effect elements respectively. We denote the reference knowledge domain by $\mathtt{K}^{\mathtt{D}}_{\mathtt{R}} = \mathtt{K}^{\mathtt{D}} \cup \{H_{\mathtt{C}j}\}_{1\leq j\leq m_1} \cup \{H_{\mathtt{A}j}\}_{1\leq j\leq m_2} \cup \{H_{\mathtt{E}j}\}_{1\leq j\leq m_3}$.

The completeness evaluation problem of a CBR system against $\mathtt{K}^{\mathtt{D}}_{\mathtt{R}}$ consists in identifying eventual incompleteness situations in the case base. An incompleteness situation is formalized as:

$$\mathtt{Incompleteness\ situation} \Leftrightarrow \exists\ \mathbb{C}_1, \mathbb{C}_2 \in \mathbb{CB}/$$
$$(\{O^{\mathbb{C}_1}_{\mathtt{C}j}\} = \{O^{\mathbb{C}_2}_{\mathtt{C}j}\})_{1\leq j\leq n_1} \wedge (\{O^{\mathbb{C}_1}_{\mathtt{A}j}\} = \{O^{\mathbb{C}_2}_{\mathtt{A}j}\})_{1\leq j\leq n_2} \wedge (\{O^{\mathbb{C}_1}_{\mathtt{E}j}\} \neq \{O^{\mathbb{C}_2}_{\mathtt{E}j}\})_{1\leq j\leq n_3}$$
$$\Longrightarrow\ \exists f \in \{H_{\mathtt{C}j}\}_{1\leq j\leq m_1} \cup \{H_{\mathtt{A}j}\}_{1\leq j\leq m_2}$$

For effectiveness reasons, we argue that is a prerequisite to check the completeness of the data as early as possible in the problem-solving process, i.e., before starting the reasoning cycle. Furthermore, the incompleteness assessment process must be launched whenever the case base is updated.

5 Incompleteness checking in the CBR system

In this section, we detail the workflow of our I2CCBR (InCompleteness Checking CBR) algorithm to evaluate data incompleteness in a CBR system. This section is divided according to the global architecture of the I2CCBR algorithm into two parts. In this workflow, starting from splitting the case base into the best possible segmentation by grouping the cases according to their effects, we exploit the resulting partitions to search for possible incomplete situations by relying on an effective method based on context and action knowledge.

5.1 Case base partitioning

The process of partitioning the case base aims to identify possible patterns in the cases' effects, i.e., detecting and estimating changes in the statistical properties of the effects, so that cases having similar effects can be grouped into the same cluster. In this section, a hybrid method based on the change point detection approach is proposed to achieve this objective. This is a mixture of two techniques: the cumulative sum (CUMSUM) technique proposed in [22] and the bootstrapping mechanism introduced in [16]. In short, the detection of change points in the case effects model is an iterative process involving the following two steps:

Step 1: Cumulative sums. Considering the notation introduced in Section 3.1, cumulative sums $\mathcal{CS}_i$ of the effect variables are calculated by the recursive formula described in Equation (1). Note that the cumulative sums do not represent the cumulative sums of the effect variables but rather they represent the cumulative sums of the differences between the values and the average $\overline{\mu}$. Consequently, the last cumulative sum ($\mathcal{CS}_n$) is always null.

By plotting the chart of cumulative sums $\mathcal{CS}_i$, potentials change points in the effect variables could be identified as changes in the direction of the diagram.

However, the cumulative sums chart cannot determine with certainty either the existence of these change points or the indices of the cases corresponding to these changes. These two problems are the focus of the second step. For the sake of the second step, it will be necessary to estimate the change magnitude $\mathcal{CS}_M$ of cumulative sums $\mathcal{CS}_i$. One way to do so is to apply Formula (2).

$$\forall\ i \leq n,\ \mathcal{CS}_i = \begin{cases} \mathcal{CS}_{i-1} + (\mathtt{E}^{\mathbb{C}_i} - \overline{\mu}) \\ 0, \qquad \text{if } i = 0 \end{cases} \tag{1}$$

With n – the number of cases, $\overline{\mu}$ – the average of the effect variable given as $\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathtt{E}^{\mathbb{C}_i}$.

$$\mathcal{CS}_M = \max_{1 \leq i \leq n}\ \mathcal{CS}_i - \min_{1 \leq i \leq n}\ \mathcal{CS}_i \tag{2}$$

Step 2: Bootstrapping. The first objective of the this step is to determine a confidence level for the observed change points. In the following, this issue is addressed with a bootstrapping approach. The rationale underlying the bootstrapping is to imitate the behavior of the cumulative sums $\mathcal{CS}_i^b$ in the case where there is no change in the patterns of effects. The resulting cumulative sums will provide a baseline for comparing the cumulative sums $\mathcal{CS}_i$ of the effects of the cases in their original order (as calculated in step 1). The bootstrapping process consists in applying the same process from step 1 to the randomly reorganized case base, which produces the cumulative sums $\mathcal{CS}_i^b$ and the change magnitude $\mathcal{CS}_M^b$.

When the plot of the cumulative sums $\mathcal{CS}_i^b$ is likely to remain closer to zero than the chart of the original cumulative sums $\mathcal{CS}_i$, a change has probably taken place. The estimation of the index of confidence in the existence of a change point includes conducting a significant k number of bootstraps and determining the number (let l be this number) of situations for which the magnitude of change $\mathcal{CS}_M^b$ is smaller than the change magnitude $\mathcal{CS}_M$ of the original case base. The confidence index $\mathtt{CI}$ that a shift in the pattern of effects is given by Formula (3).

$$\mathtt{CI} = \frac{l}{k} \times 100 \tag{3}$$

Change point position. If the confidence level is high enough (typically around 90%) to confirm the existence of a change point, one way to estimate the case index corresponding to the change in the model of effects is to use the mean square error (MSE) metric. The case base is divided into two parts containing z and $n - z$ cases, where z is the index of the last case preceding the change in effect model. The estimation of the index z consists in solving an optimization problem whose objective is to minimize Function 4.

$$\mathtt{MSE} = \sum_{i=1}^{z} (\mathtt{E}^{\mathbb{C}_i} - \overline{\mu}_1)^2 + \sum_{i=z+1}^{n} (\mathtt{E}^{\mathbb{C}_i} - \overline{\mu}_2)^2 \tag{4}$$

With $\overline{\mu}_1 = \frac{1}{z}\sum_{i=1}^{z} \mathtt{E}^{\mathbb{C}_i}$, $\overline{\mu}_2 = \frac{1}{n-z}\sum_{i=z+1}^{n} \mathtt{E}^{\mathbb{C}_i}$.

Once a change point is identified, the case base is divided into two case bases, a first case base including cases from 1 to z and the remaining cases composing the second case base. The process described in steps 1 and 2 is then iteratively applied on each of the case bases until there are no other change points in the case bases. as a result, more changes, if existing, are detected.

5.2 Incompleteness detection

Let $m \neq 0$ be the number of change points detected in the case base $\mathbb{CB}$. Let $\mathcal{I}$ denotes the set of indices of the cases whose effects represent a change in the model, such as $|\mathcal{I}| = m$ and $\mathcal{I} = \{\mathtt{I}_j\}_{1 \leq j \leq m}$. Precisely, index $\mathtt{I}_j$ corresponds to the index of the case preceding the j^{th} change in the model of the effect variables. Then, the case base $\mathbb{CB}$ can be broken into $m+1$ groups $\mathbb{G}_{1 \leq j \leq m+1}$ Such that constraints (5) are satisfied.

$$\begin{aligned} \mathbb{CB} &= \bigcup_{j=1}^{m+1} \mathbb{G}_j \\ \mathbb{G}_j &= \begin{cases} \{\mathbb{C}_t\}_{\mathtt{I}_{j-1} < t \leq \mathtt{I}_j}, & \text{if } 2 \leq j \\ \{\mathbb{C}_t\}_{1 \leq t \leq \mathtt{I}_j}, & \text{if } j = 1 \end{cases} \end{aligned} \tag{5}$$

The idea behind the completeness evaluation is to detect situations where two cases with similar actions and similar contexts but different effects. Specifically, the investigation of possible incompleteness situations is performed as follows.

1. given the set of groups $\{\mathbb{G}_j\}$, for each group $\mathbb{G}_j$, which represents the set of cases with similar effects, compute the maximum context-action distance $\mathtt{D}^{max}_{\mathtt{CA},j}$ and minimum one $\mathtt{D}^{min}_{\mathtt{CA},j}$ between cases. Let $\mathtt{S}^{\mathtt{CA}}_j$ denote the interval $\left[\mathtt{D}^{min}_{\mathtt{CA},j}, \mathtt{D}^{max}_{\mathtt{CA},j}\right]$. Let $\mathtt{S}^{\mathtt{E}}_j = \left[\mathtt{E}^{min}_j, \mathtt{E}^{max}_j\right]$ be the effect variable interval of group $\mathbb{G}_j$.
2. a situation of incompleteness is reliably identified if there exist two cases $\mathbb{C}_1$ and $\mathbb{C}_2$, located respectively in two different groups $\mathbb{G}_1$ and $\mathbb{G}_2$ whose effect models differ, such that the context-action distance between $\mathbb{C}_1$ and $\mathbb{C}_2$ belongs to one of the intervals $\mathtt{S}^{\mathtt{CA}}_1$, $\mathtt{S}^{\mathtt{CA}}_2$. Formally:

$$\begin{aligned} &\exists\ \mathbb{C}_{i'} \in \mathbb{G}_i, \mathbb{C}_{j'} \in \mathbb{G}_j, k \in \{i, j\} / \\ &\mathtt{D}_{\mathtt{CA}}(\mathbb{C}_{i'}, \mathbb{C}_{j'}) \in \mathtt{S}^{\mathtt{CA}}_k \implies \mathtt{S}^{\mathtt{E}}_i \cap \mathtt{S}^{\mathtt{E}}_j = \emptyset \ \vee (\mathtt{E}^{\mathbb{C}_1} \notin \mathtt{S}^{\mathtt{E}}_i \cap \mathtt{S}^{\mathtt{E}}_j \wedge \mathtt{E}^{\mathbb{C}_2} \notin \mathtt{S}^{\mathtt{E}}_i \cap \mathtt{S}^{\mathtt{E}}_j) \\ &\vee (\mathtt{E}^{\mathbb{C}_m} \notin \mathtt{S}^{\mathtt{E}}_i \cap \mathtt{S}^{\mathtt{E}}_j, m \in \{i', j'\} \wedge \mathbb{C}_m \notin \mathbb{G}_k) \end{aligned}$$

6 Evaluation

The objective of the experiment is to investigate the reliability and efficiency of the I2CCBR algorithm to discover incompleteness in a case base. First, we describe the dataset used in the experimentation, then we present the experimental setup, and finally, we report the results.

Dataset. To investigate the effectiveness of the I2CCBR approach, we conducted experiments using real word dataset. We used the real dataset from [7] that resulted from the experiment of the motivation example (see Section 4.1). More precisely, the authors proposed a CBR-based approach to improve the energy efficiency of buildings considering the comfort of the occupants. The approach is evaluated through a case study where data are collected from numerous sensors deployed in an academic research office.

Collected data are classified into three categories according to the casec structure presented in Section 3.1. The context data, which besides the meteorological data, includes the number of occupants. The action data model opening/closing of doors/windows. The effect data concern the temperature and the concentration of CO_2 in the office. A case corresponds to one-day measurements. The case base used in the present evaluation consists of 98 cases ordered by their measurements' dates. In this experiment, we are restricted to the incompleteness evaluation regarding the indoor temperature as the only effect variable.

Experimental setup. To avoid biasing the results of the similarity assessment due to the dominant influence of variables with large values, the context and action data are rescaled between 0 and 1 using the MinMax strategy.

In this experiment, the weighted Euclidean distance is used as a similarity function to evaluate the context-action-based similarity between two cases. It is beyond the scope of this work to detail the process of weighting context and action variables. We adopted the approach developed in [8] to estimate these weights.

The experiments were performed on a 13" MacBook Pro laptop equipped with an Intel® Core™ i7-8559U CPU 2.70 GHz, 16 GB of RAM, powered by Windows 10 pro 64 bit. The I2CCBR algorithm is implemented in Python 3.9. The code was ran in Jupyter Notebook 6.4.

Case base with random incompleteness. At this stage, we do not know whether the case base is complete since the modeling of the building environment is difficult due to the high number and the complex interactions between the phenomena influencing the energetic behavior of a building. To check the efficiency of the I2CCBR algorithm, we need a baseline for which it certainly presents incompleteness situations.

We constructed an incomplete case base $\mathbb{CB}^I$ from the case base $\mathbb{CB}$. The process of generating the incompleteness situations is described as follows:

- we introduced incompleteness in $\mathbb{CB}$ by randomly choosing and modifying 5% of the cases (5 cases).
- as the office where the experimentation took place was not equipped with an air-conditioning system, the modification of the chosen cases consists in integrating a new action variable modeling the air conditioning in the office. This variable simulates the presence of a hidden variable in the CBR system. Values of this variable are sampled from a discrete uniform distribution between 18° C and 23° C.
- the effects following the application of the new actions (turn on the air conditioner) to the context of the chosen cases are generated using the physical

model of the office. For consistency, the real effects of the other cases are simulated by the physical model too.

Empirical results. Figure 1 plots the average of the real effect variable (temperature) of the 98 cases (green curve) and the corresponding simulated values (red curve). Note that the simulated effects of cases $\mathbb{C}_6$, $\mathbb{C}_{19}$, $\mathbb{C}_{25}$, $\mathbb{C}_{59}$, and $\mathbb{C}_{71}$ are far from their original ones. The significant discrepancy between the original values and the simulated ones is due to the influence of the hidden variable (modeling air conditioning) on the effects of these cases, i.e., these cases correspond to the five randomly chosen and modified cases. Note also that the two curves overlap almost all along the plot. The Mean Absolute Percentage Error (MAPE) analysis, excluding the modified cases, indicates that the variation of the simulated data from the real data is less than 2.50%, showing the robustness of the physical model used in the simulation of the effect variables.

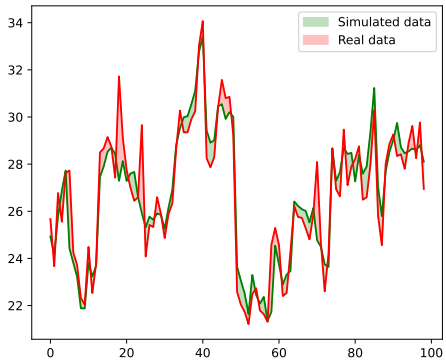

Fig. 1: Real and simulated effect.

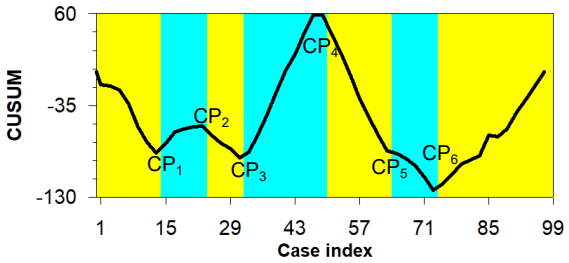

Fig. 2: Change point detection

The results of the first step of the C2CBR algorithm, which consists in detecting changes in the model of effects (temperature), are depicted in Figures 2 and 3. In Figure 2 plotting the cumulative sums of the effect variable, each color

change in the background corresponds to an abrupt change in the direction of the chart indicating the occurrence of a pattern change, i.e., each swap between the yellow and turquoise colors models a change point. It emerges that there are 6 permutations of background colors, which correspond to 6 change points.

The 6 identified change points serve to split the case base into 7 disjointed groups according to the chronological recording of case effects, as presented in Figure 3. Each change point is displayed by a shift in the turquoise-colored background and corresponds to a case index in the case base. A turquoise-shaded segment depicts a group containing all cases based on the current effect variable model formed by two successive change points. Each row of Table 1 provides detailed information on each change point. The index of the change case from which a model change is detected is assigned a confidence interval estimated at a 95% probability of being accurate. For instance, with a probability of 95%, the 6^{th} change point is estimated to be between cases 75 and 77. Furthermore, using Formula 3, a confidence index in each detected change point is reported to qualify the quality of the analysis. For instance, the system is 99% confident that the 6^{th} change point took place. Further information is also provided as averages of the groups' effect variables before and after a change point. Table 2 gives the values of the $\mathtt{S}^{\mathtt{CA}}$ and $\mathtt{S}^{\mathtt{E}}$ metrics (defined in Section 5.2) for each of the seven groups.

After applying the proposed heuristic, The five incompleteness situations that were artificially generated in the previous step have been correctly identified, as shown in Table 3. Each probable incompleteness situation is described by the two cases $\mathbb{C}_1^I$ and $\mathbb{C}_2^I$ generating this situation. Note that the cases in column $\mathbb{C}_1^I$ of Table 3 correspond to the five modified cases. For instance, the incompleteness situation S5 is observed between the cases $\mathbb{C}_{71}$ and $\mathbb{C}_{79}$ since the distance $\mathtt{D}_{\mathtt{CA}}(\mathbb{C}_{71}, \mathbb{C}_{79})$ is lower than the distance $\mathtt{D}_{\mathtt{CA}}^{max}$ of the group $\mathbb{G}_7$ to which the case $\mathbb{C}_{79}$ belongs but the effect variable of the case $\mathbb{C}_{71}$ belongs to the group $\mathbb{G}_6$ knowing that $\mathtt{E}^{\mathbb{C}_6} = 24.77 \notin \mathtt{S}_6^{\mathtt{E}} \cap \mathtt{S}_7^{\mathtt{E}}$ and $\mathtt{E}^{\mathbb{C}_7} = 28.43 \notin \mathtt{S}_6^{\mathtt{E}} \cap \mathtt{S}_7^{\mathtt{E}}$.

Table 1: Change points details.

Index	Confidence interval	Confidence index	From	To	Level
15	(11,15)	96%	24.49	27.814	4
25	(25,25)	94%	27.814	25.725	5
33	(33,37)	99%	25.725	29.733	3
51	(51,51)	100%	29.733	22.745	2
65	(65,67)	98%	22.745	25.305	3
75	(75,77)	99%	25.305	28.355	1

Table 2: Groups' properties.

Group	$\mathtt{S}^{\mathtt{CA}}$	$\mathtt{S}^{\mathtt{E}}$
$\mathbb{G}_1$	$[0.138, 0.526]$	$[21.88, 27.25]$
$\mathbb{G}_2$	$[0.199, 0.550]$	$[26.61, 28.70]$
$\mathbb{G}_3$	$[0.156, 0.516]$	$[25.26, 26.14]$
$\mathbb{G}_4$	$[0.166, 0.390]$	$[23.65, 33.42]$
$\mathbb{G}_5$	$[0.188, 0.377]$	$[21.35, 24.54]$
$\mathbb{G}_6$	$[0.155, 0.440]$	$[23.64, 26.42]$
$\mathbb{G}_7$	$[0.126, 0.602]$	$[25.79, 31.23]$

7 Conclusion

.

This study introduces a novel approach, called I2CCBR, which aims to tackle the challenge of data incompleteness in a CBR system based on the open-world assumption. The authors employed a combination of a change point detection

Table 3: Change points details.

Situation	$\mathbb{C}_1^I$	$\mathbb{C}_2^I$	$\mathtt{D}_{\mathtt{CA}}$
S1	$\mathbb{C}_6 \in \mathbb{G}_1$	$\mathbb{C}_{85} \in \mathbb{G}_7$	0.531
S2	$\mathbb{C}_{19} \in \mathbb{G}_2$	$\mathbb{C}_{39} \in \mathbb{G}_4$	0.184
S3	$\mathbb{C}_{25} \in \mathbb{G}_3$	$\mathbb{C}_{18} \in \mathbb{G}_2$	0.275
S4	$\mathbb{C}_{59} \in \mathbb{G}_5$	$\mathbb{C}_{69} \in \mathbb{G}_6$	0.369
S5	$\mathbb{C}_{71} \in \mathbb{G}_6$	$\mathbb{C}_{79} \in \mathbb{G}_7$	0.545

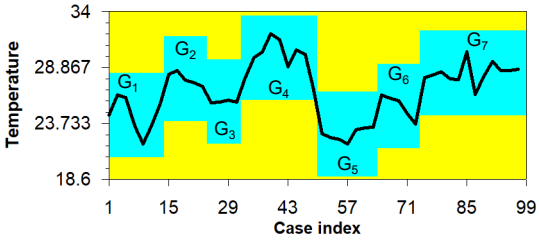

Fig. 3: Change points detection.

technique and a heuristic strategy to detect possible incompleteness in the case base. To the best of the authors' knowledge, this is the first attempt to address this issue. The effectiveness of the proposed approach was evaluated in a real-world experiment, and the results demonstrate its potential for practical implementation with promising outcomes.

The next phase of this study involves expanding the experimental evaluation of the I2CCBR algorithm by testing it on more extensive datasets. This will enable the researchers to verify the findings obtained in the initial study.

References

1. Extracting business value from the 4 v's of big data. techreport, IBM, 2016. `http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data`, accessible on 2022/05/31.
2. 2019 Global data management research. Taking control in the digital age. Benchmarkreport, Experian UK&I, February 2019.
3. Cedric Arisdakessian, Olivier Bertrand Poirion, Breck Yunits, Xun Zhu, and Lana Garmire. Deepimpute: An accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. *Genome Biology*, 20, 10 2019.
4. S.A. Barker and P.J.W. Rayner. Unsupervised image segmentation using markov random field models. *Pattern Recognition*, 33(4):587–602, 2000.
5. Ralph Bergmann, Wolfgang Wilke, and Ivo Vollrath. Integrating general knowledge with object-oriented case representation and reasoning. In *4th German Workshop: Case-Based Reasoning - System Development and Evaluation*, 1996.
6. Edgar A. Bernal. Training deep generative models in highly incomplete data scenarios with prior regularization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2631–2641, 2021.
7. Fateh Boulmaiz, Stephane Ploix, and Patrick Reignier. A data-driven approach for guiding the occupant's actions to achieve better comfort in buildings. In *IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021.
8. Fateh Boulmaiz, Patrick Reignier, and Stephane Ploix. An occupant-centered approach to improve both his comfort and the energy efficiency of the building. *Knowledge-Based Systems*, 249:108970, 2022.
9. Wei Cao, Dong Wang, Jian Li, Hao Zhou, Yitan Li, and Lei Li. Brits: Bidirectional recurrent imputation for time series. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6776–6786, Red Hook, NY, USA, 2018. Curran Associates Inc.

10. Avraam Charakopoulos and Theodoros Karakasidis. Backward degree a new index for online and offline change point detection based on complex network analysis. *Physica A: Statistical Mechanics and its Applications*, 604:127929, 2022.
11. Corinna Cichy and Stefan Rass. An overview of data quality frameworks. *IEEE Access*, 7:24634–24648, 2019.
12. Luis Galárraga, Simon Razniewski, Antoine Amarilli, and Fabian M. Suchanek. Predicting completeness in knowledge bases. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, feb 2017.
13. Yitea Seneshaw Getahun, Ming-Hsu Li, and Iam-Fei Pun. Trend and change-point detection analyses of rainfall and temperature over the awash river basin of ethiopia. *Heliyon*, 7(9):e08024, 2021.
14. Martin Grohe and Peter Lindner. Probabilistic databases with an infinite open-world assumption. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - PODS '19*. ACM Press, 2019.
15. Anders Haug. Understanding the differences across data quality classifications: a literature review and guidelines for future research. *Industrial Management & Data Systems*, 121(12):2651–2671, aug 2021.
16. David V. Hinkley and Edna Schechtman. Conditional bootstrap methods in the mean-shift model. *Biometrika*, 74:85–93, 1987.
17. Tushar Khot, Niranjan Balasubramanian, Eric Gribkoff, Ashish Sabharwal, Peter Clark, and Oren Etzioni. Exploring Markov Logic Networks for question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 685–694, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
18. Siaw-Teng Liaw, Alireza Rahimi, Pradeep Ray, and Jane Taggart. Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *International Journal of Medical Informatics*, 82, 2013.
19. Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Neural probabilistic logic programming in deepproblog. *Artificial Intelligence*, 298:103504, 2021.
20. Adam Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla-Favera, and Andrea Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7 Suppl 1:S7, 02 2006.
21. Phuong T. Nguyen, Juri Di Rocco, Ludovico Iovino, Davide Di Ruscio, and Alfonso Pierantonio. Evaluation of a machine learning classifier for metamodels. *Software and Systems Modeling*, 20(6):1797–1821, sep 2021.
22. A. N. Pettitt. A simple cumulative sum type statistic for the change-point problem with zero-one observations. *Biometrika*, 67(1):79–84, 1980.
23. Son Phung, Ashnil Kumar, and Jinman Kim. A deep learning technique for imputing missing healthcare data. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019.
24. Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, 1987.
25. Michael M. Richter. The knowledge contained in similarity measures. International Conference on Case-Based Reasoning, ICCBR'95, Sesimbra, Portugal, 1995.
26. Valerie Sessions and Marco Valtorta. The effects of data quality on machine learning algorithms. In John R. Talburt, Elizabeth M. Pierce, Ningning Wu, and Traci Campbell, editors, *Proceedings of the 11th International Conference on Information Quality, MIT, Cambridge, MA, USA, November 10-12, 2006*. MIT, 2006.

27. Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, feb 2020.
28. Richard Wang and Diane Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 1996.
29. Wentao Wang, Tyler Derr, Yao Ma, Suhang Wang, Hui Liu, Zitao Liu, and Jiliang Tang. Learning from incomplete labeled data via adversarial data generation. In *2020 IEEE International Conference on Data Mining (ICDM)*, 2020.
30. Colin Wilcox, Soufiene Djahel, and Vasileios Giagos. Identifying the main causes of medical data incompleteness in the smart healthcare era. In *2021 International Symposium on Networks, Computers and Communications (ISNCC)*, 2021.
31. Seung-H. You, Eun J. Jang, Myo-S. Kim, Min-T. Lee, Ye-J. Kang, and Jae-E. Lee. Change point analysis for detecting vaccine safety signals. *Vaccines*, 9(3), 2021.