Discovering Reliable Evidence of Data Misuse by Exploiting Rule Redundancy

Laura Genga*, Nicola Zannone

Eindhoven University of Technology

Anna Squicciarini

Pennsylvania State University

Abstract

Big Data offers opportunities for in-depth data analytics and advanced personalized services. Yet, while valuable, data analytics might rely on data that should not have been used due to, e.g., privacy constraints from the data subject or regulations. As decision makers and data controllers often act outside any control mechanism and with no requirement of transparency, it is challenging to verify whether constraints on data usage are actually satisfied. In this work, we relate the problem of finding evidence of data misuse to the identification of unique decision rules, i.e. rules that have likely been used for decision making. Accordingly, we propose an approach to find reliable evidence of data misuse in the context of classification problems using association rule mining, along with novel metrics to assess the level of redundancy among decision rules. Our proposed approach is able to identify the use of sensitive information in decisional processes along with their context. We evaluated our approach through both controlled experiments and two case studies using real-life event data. The results show that our approach finds more reliable evidence of data misuse compared to previous work.

*Corresponding author

Email addresses: l.genga@tue.nl (Laura Genga), n.zannone@tue.nl (Nicola Zannone), a.squicciarini@edu.it (Anna Squicciarini)

1. Introduction

Personal data are increasingly used to drive a large number of decisions, concerning individuals' professional or personal life, for example granting a loan or offering a job, societal issues, like smart mobility and urbanization, or to provide users with customized services like web search, recommendations, personalized medicine, and so on. These decisions and how they are reached often lack transparency in the era of Big Data, as they are the result of analytical processes outside the direct control of the data subjects [1]. Henceforth, in the aftermath, it is often challenging to verify whether data usage restrictions (i.e., restriction on the use of sensitive attributes) were held.

This phenomenon is typically known as *data misuse*. Here, we refer to data misuse as any situation in which the processing of personal data does not comply with the intended use of data as defined by the data subject and/or violates the requirements and constraints imposed by privacy and data protection regulations. In particular, data protection regulations regard certain personal information as sensitive (e.g., gender, race, political opinions, medical information) and allows the processing of such information only under certain circumstances.

Data misuse is considered one of the main issues underlying Big Data Analytics [2]. While an increasing body of work has proposed approaches to deal with this issue, several challenges are still unaddressed [3]. One of them regards the existence of possible correlations existing among (values of) sensitive and (values of) non-sensitive attributes. In presence of these correlations, the analysis of observational data might be biased, since one might determine unreliable associations between sensitive data and the decision made, which were actually due to the presence of some other data correlated with the decision and sensitive data. However, most of the previous work focused on detecting data misuse (e.g., [4, 5]) ignores the effects of these correlations

when discovering evidences of misuse.

In this paper, we aim to extract *reliable* evidence of data misuse in the context of data analytics, wherein analytical processes are based on data classification, taking into account possible correlations among sensitive and non-sensitive attributes. Our objectives are multi-fold: *i)* determining whether sensitive attributes were used for classification purposes and, thus, were in fact used inappropriately, *ii)* determining the values of these attributes and their combination with other possibly sensitive (or non-sensitive) attributes, to uncover the *context* of the misuse. The intuition is that sensitive attributes are often used to target certain demographics or groups of users (e.g., underserved communities). Therefore, it is of utmost importance to determine under which circumstances sensitive attributes are used for decision making.

The detection of practices in which data usage does not comply with privacy requirements can be performed by analyzing the underlying decision process. However, such a process is often not available. To this end, we model analytical processes as a classification problem and employ association rule mining to identify correlation among itemsets. We adopt a relaxed notion of correlation, similar to the one adopted in [6], intended as a trend according to which when the antecedent of a rule holds, then also the consequent of the rule holds (i.e., it is possible to devise a certain degree of dependence between attribute values in the antecedent and attribute values in the consequent of the rule). In this setting, we consider a rule "reliable" if it captures correlations between sensitive attribute values and a given decision that are not captured by any other rule including only non-sensitive attributes. Accordingly, we reduce the problem of finding reliable evidence of data misuse to the problem of redundancy reduction. In particular, we model the effects of possible correlations among dataset attributes in terms of *rule redundancy*. We argue that whenever a non-sensitive and a sensitive attribute are (at least partially) correlated, then one can expect to detect in the ruleset rules describing a significantly overlapping set of records, possibly with

both sensitive and non-sensitive attributes. We posit that these redundant rules affect the reliability of the results and, thus, should be removed.

To this end, we investigate possible cases of redundancy and study the effects of redundant rules on the reliability of the discovered evidence of data misuse. Our metrics for assessing redundancy *(i)* take into account the presence/absence of sensitive itemsets; *(ii)* measure redundancy between both overlapping and partially overlapping rules, as well as between rules involved in inclusion relationships; *(iii)* evaluate the level of redundancy between a rule and other rules in a given ruleset. Based on these metrics, we propose a framework to infer reliable evidence of data misuse, i.e. rules involving sensitive itemsets that are not redundant with respect to rules not involving sensitive itemsets. We carried out a set of experiments on both synthetic and real-world datasets, and showed how the proposed metrics for the evaluation of rule redundancy and relevance can improve the overall accuracy of the discovered evidence of misuse and better understand its context (i.e., which additional attributes and attribute values were used to reach a given decision).

Although we present our approach in the context of data misuse, it can be applied in other contexts that require unveiling systematic biases in decisional processes with respect to selected attributes. An example of application in the security domain is vulnerability analysis, in which risk labeling usually tends to be biased towards technical characteristics of software bugs, resulting in a huge number of alerts when there is actually little or no risk of attack [7]. In this context, our approach could be applied to uncover biases in observational data, thus allowing the security analyst to focus on those alerts that have more definite causes. Another domain of application is the monitoring of industrial control systems. A widely-used strategy to prevent malfunctioning in those systems is to define so-called "invariants", i.e. rules describing casual relations between certain (set of) variables (e.g., sensor values) and certain states of the system. These invariants are often manually defined by human system

designers. However, this is a time-consuming and error-prone procedure, which can easily lead to miss critical relations. To address this issue, several authors have lately started employing association rule mining to support the automatic extractions of these relations [8, 9]. Those approaches, however, typically lead to a large number of invariants to be monitored [10]. Our approach can complement these techniques to find those invariant that provide reliable justifications for malfunctioning.

The paper is organized as follows. Section 2 provides preliminaries on association rule mining. Section 3 introduces the problem of data misuse in decisional processes. Section 4 discusses the issue of data misuse detection in the presence of redundant rules, showing limitations of previous approaches. Section 5 presents our metrics to evaluate redundancy, together with an approach to address redundancy reduction. Section 6 discusses the results obtained from a set of experiments on synthetic and real-world datasets. Finally, Section 7 discusses related work and Section 8 concludes the paper and provides directions for future work.

2. Preliminaries

This section provides general notions related to association rule mining. Table 1 summarizes the formal notation used thorough the paper.

A *dataset* D is a set of *records*, each describing a given subject (e.g., a person applying for a job offer). More formally, given a set of *attributes* (aka *features*) $\mathcal{A} = \{A_1, \ldots, A_n\}$ each representing a characteristic of the subject, a record is a set $\{A_1 = v_{1_l}, \ldots, A_n = v_{n_k} \mid \forall_{i:1\ldots n} v_{i_j} \in dom(A_i)\}$, where $dom(A_i)$ denotes the domain of attribute A_i. Every pair $A_i = v_{i_j}$ is called *item*, and a set of items *itemset*. Hereafter, Λ denotes the set of all possible items. We use $cov(X)$ to indicate the set of records covered by itemset X in the given dataset.

An *association rule* r is an implication of the form $r : X \rightarrow Y$, where X and Y

are two itemsets (i.e., $X, Y \subseteq \Lambda$), respectively called *antecedent* and *consequent* of the rule [11]. Intuitively, an association rule indicates that if an itemset X occurs in a record, then itemset Y also occurs in the same record. A record is *covered* by a rule r if it matches the itemsets defined in both the antecedent and consequent of the rule, i.e. $cov(r) = cov(X \cup Y)$.

Since the number of association rules that can be generated from a dataset is usually huge, a common practice in association rule mining is to employ metrics that evaluate the *relevance* of each rule and filter out all rules whose relevance is below a predefined threshold. Two very well-known metrics are rule *support*, which represents the percentage of samples in the dataset covered by a rule, and *confidence*, which represents the percentage of samples fulfilling the rules among those fulfilling its antecedent. Formally, given an itemset X and a dataset D, $supp(X) = \frac{|cov(X)|}{|D|}$. The support and confidence of an association rule $r : X \rightarrow Y$ with respect to D are then defined as $supp(r) = \frac{|cov(r)|}{|D|}$ and $conf(r) = \frac{supp(r)}{supp(X)}$. It is worth noting that confidence represents the conditional probability of having the consequent of the rule given its antecedent. A third measure that is often used is *lift*, defined as $lift(r) = \frac{supp(r)}{supp(X) \times supp(Y)}$. The lift measures how different is the actual joint probability of $X \cup Y$ with respect to the probability that we would expect if they are independent. If its value is close to 1, then the antecedent and the consequent are likely independent; otherwise, they are considered as positive/negative correlated. A plethora of other metrics of relevance have been proposed in literature; interested readers can refer, for instance, to [12] for an overview.

In this work, we restrict our attention to a special type of association rules, namely *class association rules* [13]. Let C be a *class attribute* representing the set of possible decisions (e.g., the approving/rejecting of a candidate for a job offer). Class association rules are association rules whose consequent consists of a single class item $C = c_i$. We use Ω to denote the set of possible class items. For the sake of simplicity, hereafter

Symbol	Description
D	dataset describing a decisional process.
$\mathcal{A}$	set of attributes.
$\mathcal{S}$	set of sensitive attributes.
A_i	an attribute, representing an entity's characteristic.
S_i	sensitive attribute.
C	class attribute.
Λ	set of all possible items.
Ω	set of all possible class items.
Σ	set of all possible sensitive itemsets.

Table 1: Formal notation used through the paper.

we use the term 'rule' to indicate a class association rule.

3. Data misuse in decisional processes

In this work, the term "data misuse" refers to any exploitation of personal data in a way that violates the data subject's intended use and/or constraints imposed by privacy and data protection regulation. Data misuse in decisional processes occurs when the decision reached by these processes is based on data whose usage is prohibited.

Given a dataset D defined over a set of attributes $\mathcal{A}$, involving a set of records describing the decisions made for a set of subjects, we model the underlying decision process as a set of class association rules. We mark the attributes that should not be used for decision making as *sensitive*. Hereafter, we use $\mathcal{S}$ to denote the set of sensitive attributes, with $\mathcal{S} \subseteq \mathcal{A}$, and Σ the set of possible sensitive itemsets, i.e. $\Sigma = \{S_i = v_{i_k} \mid S_i \in \mathcal{S} \wedge v_{i_k} \in dom(S_i)\}$.

Evidence of data misuse in a decisional process can be expressed as a set of

potentially misused itemsets (PMI), i.e. the set of itemsets that *i)* involves at least a sensitive item and *ii)* shows a significant correlation[1] with the class itemset. Formally, given a set of class association rules $r_i : \Sigma_i, \Delta_i \rightarrow \Omega_i$ representing the decisional process, we define the set of potentially misused itemsets as the set $PMI = \bigcup_i \{\Sigma_i \cup \Delta_i\}$ such that:

- $\Sigma_i \neq \emptyset \wedge \Sigma_i \subseteq \Sigma$;
- $\Delta_i \subseteq \Lambda \setminus \Sigma$;
- $\Sigma_i \cup \Delta_i$ *is correlated* with a class itemset $\Omega_i \subseteq \Omega$ where this correlation is defined in terms of relevance metrics (see Section 2).

Intuitively, potentially misused itemsets point out the risks of data misuse in the decisional process, i.e. the usage of one or more sensitive attributes for assigning the class label to the records in the dataset.

Example 1. *Table 2 shows an example dataset describing an applicant selection process. Specifically, its records represent a set of applicants for a job position annotated with the decision. Each record is described by the attributes shown in the head of the table, namely* $\mathcal{A} = \{Instruction, SpeakLanguage, PrevRole, Age, Country, Gender, PersStatus\}$ *and* $C = Decision$.

Several countries have issued employment regulations that require employers to conduct fair recruitment processes and impose legal obligations to avoid the use of sensitive personal characteristics as criteria for the selection of candidates. In our scenario, we capture these legal requirements by modeling $\mathcal{S} = \{Age, Country, Gender\}$.

[1] Recall that we use the term *correlation* in a broad sense, i.e. to indicate the fact that when attributes itemsets Σ_i, Δ_i occur in a record, then it is likely to have Ω_i.

Instruction	SpeakLanguage	PrevRole	Age	Country	Gender	PrevAppl	Decision	$\#Samples$
Doctorate	Y	Employee	< 25	USA	M	N	Y	200
Master	Y	Employee	< 25	USA	F	N	N	300
Master	N	Manager	< 25	China	M	Y	N	30
Master	Y	Employee	< 25	China	M	N	N	150
Bachelor	Y	Self-Employee	25-50	SA	M	N	N	140
Bachelor	Y	Self-Employee	25-50	SA	M	N	Y	100
Master	N	Self-Employee	25-50	USA	F	N	N	180
Master	Y	Employee	25-50	SA	M	N	N	60

Table 2: Example dataset describing features evaluated for applicants for a job position, along with the result of the application.

Table 3 reports some rules describing the decisional process recorded in Table 2. Let us suppose that a domain expert sets as conditions for the relevance metrics $supp(r) > 0.01$ *and* $conf\,(r) > 0.6$*. We can observe that rules* r_2*,* r_3*,* r_4*,* r_6 *and* r_7 *show possible evidence of data misuse. In fact, these are interesting rules according to the given thresholds for relevance metrics and involve sensitive attributes, thus suggesting the existence of possible correlations between these attributes and the final decision. Specifically, from these rules, we can derive potentially misused itemsets* $PMI = \{\{SpeakLanguage = N, Country = China\}, \{PrevRole = Employee, Age = \text{“}< 25\text{”}\}, \{Country = China\}, \{Age = \text{“}< 25\text{”}\}, \{Country = SA, PrevAppl = N\}\}$*. For instance, the first potentially misused itemset indicates that attributes* SpeakLanguage *and* Country *have likely been used in the decision process, where* Country *is a sensitive attribute.*

4. Data Misuse Detection and Rules Redundancy

This section introduces the problem of data misuse detection and shows that finding reliable evidence of data misuse is related to the problem of redundancy reduction. Then, we provide an overview of previous work on redundancy reduction and discuss the limitation of existing approaches with respect to data misuse detection.

ID	rule	supp	conf	lift
r_1	$SpeakLanguage = N \rightarrow Decision = N$	0.18	1.00	1.21
r_2	$SpeakLanguage = N, Country = China \rightarrow Decision = N$	0.02	1.00	1.21
r_3	$PrevRole = Employee, Age = \text{“}{<}25\text{”} \rightarrow Decision = N$	0.38	0.69	0.83
r_4	$Country = China \rightarrow Decision = N$	0.15	1.00	1.20
r_5	$PrevAppl = Y \rightarrow Decision = N$	0.02	1.00	1.34
r_6	$Age = \text{“}{<}25\text{”} \rightarrow Decision = N$	0.41	0.71	0.85
r_7	$Country = SA, PrevAppl = N \rightarrow Decision = N$	0.17	0.66	0.49
r_8	$PrevRole = Self\text{-}Employee, PrevAppl = N \rightarrow Decision = N$	0.27	0.76	0.56
r_9	$PrevRole = Employee, Instruction = Master \rightarrow Decision = N$	0.44	1.00	1.39

Table 3: Association rules for the example in Table 2 along with their *support*, *confidence* and *lift*.

4.1. Data Misuse Detection

The problem of detecting possible data misuse in a decisional process corresponds to determine whether sensitive itemsets have been used for decision making. Accordingly, it can be formulated as a class association rule mining problem, with additional constraints concerning the presence of sensitive attributes. More precisely, the problem is to find the set of rules $R = \{r_i : \Sigma_i, \Delta_i \rightarrow \Omega_i \mid \Sigma_i \neq \emptyset \wedge \Sigma_i \subseteq \Sigma \wedge \Delta_i \subseteq \Lambda \setminus \Sigma \wedge isInteresting(r_i)\}$, where $isInteresting(r_i)$ indicates that rule r_i is relevant with respect to one or more relevance metrics, like support, confidence and lift (see Section 2).

This formalization, however, might not provide *reliable* evidence of data misuse. Relevance metrics only evaluate the quality of single rules with respect to the dataset and do not account for possible correlations among sensitive and non-sensitive itemsets of different rules. The presence of these correlations can lead to the generation of rules involving sensitive attributes but that describe (almost) the same set of samples of rules not involving any sensitive attribute (see, e.g., rules r_2 and r_5 in Table 3). When this happens, it is not possible to determine whether the final class value was affected by the sensitive or by the non-sensitive attributes; clearly, this seriously hampers the reliability of the data misuse evidence inferred by association rules mining.

In literature, some steps towards this direction have been carried out by seminal

work on discrimination discovery [14], which can be considered as a particular case of data misuse detection. Here, the authors introduce the notion of *extended lift* to determine which rules are discriminatory. Given a rule $r_i : \Sigma_i, \Delta_i \rightarrow \Omega_i$ such that $isInteresting(r_i)$ holds, the authors compare its confidence with the confidence of the rule without the sensitive itemset, i.e. $r_j : \Delta_i \rightarrow \Omega_i$. This metric measures how the rule confidence varies with/without the sensible itemset, thus providing an evaluation of the relevance of this itemset. A rule r_i is said α-discriminatory if $elif(r_i, r_j) \geq \alpha$, where α is a user-defined threshold. The approach is further refined in [15], where additional metrics have been introduced together with approaches to test the statistical significance of the results. This approach represents, to the best of our knowledge, the first example in which significance metrics are used in combination with metrics that evaluate the differences between rules involving and not involving sensitive itemsets. However, the extended lift is not suitable to unveil correlations among sensitive and non-sensitive itemsets as it does not account for non-sensitive itemsets not occurring in r_i. Moreover, this metric only provides an indirect and rough estimate on possible correlations in the dataset.

In this work, we propose to detect these correlations by analyzing the effect they have on the generation of the ruleset, namely the generation of so-called *redundant* rules. In the following sections, we discuss the issues of rule redundancy in data misuse detection along with an overview of the limitations in previous work on redundancy reduction in the context of data misuse.

4.2. Challenges of Redundancy for Data Misuse Detection

Generally speaking, redundant rules are rules that provide the same semantic information, i.e. they describe the same (or very similar) set of samples. We observe that when looking for data misuse evidence, the presence of redundant rules poses relevant challenges on the reliability of the discovered evidence.

Given two rules $r_i : \Lambda_i \rightarrow \Omega_i$, $r_j : \Lambda_j \rightarrow \Omega_i$ such that $\Lambda_i, \Lambda_j \subset \Lambda$, the fact that they are redundant means that for a significant portion of dataset records we have $cov(\Lambda_i \cup \Omega_i) \approx cov(\Lambda_j \cup \Omega_i)$. In this situation, we cannot establish which itemset is actually correlated with the class, which means that we cannot consider these rules as evidence of possible data misuse.

We identify three possible types of redundancy, which differ in terms of degree of overlap (*total* or *partial*) and in terms of the presence of *inclusion relationships* among attribute itemsets:

1. $r_i : \Lambda_i \rightarrow \Omega_i$, $r_j : \Lambda_j \rightarrow \Omega_i$ such that $\Lambda_i \not\subset \Lambda_j$, $\Lambda_j \not\subset \Lambda_i$ and $cov(\Lambda_i \cup \Omega_i) = cov(\Lambda_j \cup \Omega_i)$. In this case, the attribute itemsets of r_i and r_j differ among each other, while their coverage overlaps completely. This makes r_i (and, consequently, r_j) clearly redundant. As an example, let us consider rules r_2 and r_5 in Table 3. It is easy to observe that $r_2 : SpeakLanguage = N$, $Country = China \rightarrow Decision = N$ is redundant with respect to $r_5 : PrevAppl = Y \rightarrow Decision = N$, since they actually describe the same set of records. This implies that we cannot determine whether $SpeakLanguage = N$,$Country = China$ or $PrevAppl = N$ are both correlated with $Decision = N$, or, instead, if only one of them is actually correlated with the class, while the other one is indirectly correlated with the class, as a result of a partial correlation between features $SpeakLanguage$, $Country$ and $PrevAppl$. Accordingly, since in our example $Country$ is a sensitive attribute while $PrevAppl$ is not, we cannot determine whether sensitive data have been misused or not.
2. $r_i : \Lambda_i \rightarrow \Omega_i$, $r_j : \Lambda_j \rightarrow \Omega_i$ such that $\Lambda_j \subseteq \Lambda_i$. In this case, r_j is a *super-rule* of r_i or, equivalently, r_i is a *sub-rule* of r_j. Clearly, r_j completely covers r_i. However, we argue that this is not enough to consider r_i as redundant. As observed in previous work on redundancy reduction (e.g., [16]), sub-rules can

actually provide a more accurate description of the decisional process compared to their super-rules. Therefore, to determine whether a sub-rule is redundant, we argue that one should check if it brings some *improvement* (e.g., in terms of confidence) with respect to its super-rules.

3. $r_i : \Lambda_i \rightarrow \Omega_i$, $r_j : \Lambda_j \rightarrow \Omega_i$ such that $\Lambda_j \not\subseteq \Lambda_i$ and $cov(\Lambda_i \cup \Omega_i) \cap cov(\Lambda_j \cup \Omega_i) \neq \emptyset$. In this case, r_j is not a sub-rule of r_i (though it might be a super-rule), while their coverage overlaps at least partially (but not necessarily completely). To determine whether r_i is redundant with respect to r_j, we argue that one should analyze to what extent they overlap. More precisely, one should evaluate how the capability of r_i to describe the records in its coverage varies with/without considering the itemsets in r_j. As an example, let us consider r_3 and r_4 in Table 3. r_4 covers 180 records, among which 150 are also covered by r_3. This means that without considering records in which $PrevRole = Employee$, $Age =$ "<25" occur, only few records remain for which the relation between $Country = China$ and $Decision = N$ still holds. This weakens the strength of the relation between this itemset and the class. On the other hand, if we compare rule r_6 with r_5, we can observe that r_5 covers only 30 records over the 480 records covered by r_6; thus, even without considering the samples in which $PrevApp = N$ occurs, we still have a significant number of records for which the relation between $Age =$ "<25" and $Decision = N$ holds.

Summing up, to address data misuse detection we need to infer class association rules that not only are *interesting* but also *not redundant*. Based on this observation, we formulate our problem statement as follows:

Problem Statement. Given a dataset D and a ruleset $R = \{r_i : \Sigma_i, \Delta_i \rightarrow \Omega_i \mid \Sigma_i \neq \emptyset \wedge \Sigma_i \subseteq \Sigma \wedge \Delta_i \subseteq \Lambda \setminus \Sigma \wedge isInteresting(r_i)\}$, determine the set of ***Strong Rules*** $SR = \{r_s \mid r_s \in R \wedge \neg isRedundant(r_i)\}$ where $isInteresting(\cdot)$ and

$isRedundant(\cdot)$ are computed with respect to D.

We present our approach to address this problem statement in Section 5. In the next section, we discuss the limits of previous work on redundancy reduction when applied to data misuse detection.

4.3. Redundancy Reduction: limitations of previous work for data misuse detection

The presence of redundant rules is a well-known problem in association rule mining [17]. Here, we present the main ideas underlying existing approaches for redundancy reduction and refer to Section 7 for a detailed overview. Roughly speaking, we can group work on this subject in two categories. The first category encompasses approaches that relate redundancy to the inclusion relationship between itemsets, i.e. they compare each rule with the set of its super-rules to check whether the rule brings some improvement and, hence, it is not redundant (e.g., [16, 18, 19]). The second category includes approaches that consider redundancy as the overlap among rules' coverage and prune from the ruleset all the rules whose records are covered by other (combination of) rules [20, 21].

Both types of approaches for redundancy reduction present important limitations when applied to data misuse detection. First, none of them can deal with all three types of redundancy discussed above. Approaches of the first category would miss cases of redundancy of types (1) and (3); while approaches in the second category tend to favor super-rules, with the result that redundancy of type (2) is not handled by those approaches. Moreover, they might not able to handle redundancy of type (3). For instance, given two rules r_i, r_j such that r_i is a super-rule of r_j, these approaches tend to choose r_i over r_j. However, when the coverages of the rule and its super-rule largely overlap, e.g. like the coverages of r_3 and r_4 discussed in case (3), r_i should be considered redundant as well.

In general, redundancy reduction methods aim purely to find the minimal ruleset that describes the given dataset without losing relevant information. As such, their outcome might still contain redundant rules with respect to the decisional process. We claim that, in the context of data misuse, this is undesirable as it affects the reliability of the gathered evidence of misuse.

Moreover, existing approaches do not evaluate the *level* of redundancy between rules. For instance, as discussed in case (3), rule r_5 presents a different level of redundancy with r_4 and with r_5; we argue that an analyst should be able to measure such a level in order to determine which case is acceptable and which is not in her domain of analysis.

Finally, existing redundancy reduction methods do not distinguish between sensitive and non-sensitive attributes and itemsets when evaluating redundancy. We argue that for detecting data misuse one has to take into account the presence/absence of sensitive itemsets when comparing two rules. Indeed, a rule is not relevant for data misuse detection only if it is redundant to a rule that does not involve sensitive itemsets.

An approach tailored to data misuse detection has been recently proposed by Pedreschi et al. [4]. This approach introduces a more flexible concept of redundancy that distinguishes between sensitive and non-sensitive itemsets and accounts for partially overlap by means of the notion of *p-instance*. Given two rules $r_i : \Sigma_i, \Delta_i \rightarrow \Omega_i$ and $r_j : \Delta_i, \Delta_j \rightarrow \Omega_i$, r_i is a p-instance of r_j if *(a)* $\frac{conf(r_j)}{conf(r_i)} \geq p$, and *(b)* $conf(\Sigma_i, \Delta_i \rightarrow \Delta_j) \geq p$. The underlying idea is that if it is possible to find at least one rule that has (a) the same or similar predictive capabilities (i.e., confidence) of r_i, and (b) involves a non-sensitive itemset Δ_j that presents a significant correlation with the sensitive itemset Σ_i in a context Δ_i (i.e., in records in which we have Σ_i, Δ_i, likely we also have Δ_j, Δ_i) then r_i should not be considered as evidence of data misuse. The parameter p allows considering partially overlapping between rules; when $p = 1$ only rules that perfectly overlap with (or are perfectly covered by) other rules are discarded;

when $p = 0.8\%$ this relation has to hold at least the 80% of cases; when $p = 0.7\%$ this relation has to hold at least the 70% of cases; and so on.

Yet, Pedreschi's approach [4] still suffers from several limitations:

- *Class Mismatch*: It analyzes the correlation between Σ_i and Δ_j without taking into account the class itemset. This implies that we might keep/remove a rule because of the values assumed by Σ_i and Δ_j for class values different from the one of the analyzed rule. We claim that this effect is undesirable, since the impact of the sensitive itemset is analyzed outside the context in which it appears in the rule, which involves the class value as well. In other words, by ignoring the class item, one would account for records that actually do not belong to the coverage of r_i and r_j. Although this effect is partially mitigated by condition (a) of p-instance, it might still lead to undesired results.

- *Context Mismatch*: It constraints the check to those rules sharing the same context Δ_i. Therefore, it cannot recognize redundancy between rules that share only a portion of the context Δ_i.

- *Super-Rules*: It does not account for redundancy between a rule and its sub-rules.

These issues have a relevant impact on the finding of reliable evidence of data misuse, as shown in our experiments (Section 6).

5. Extraction of reliable evidence of data misuse

In this section, we introduce a novel framework to evaluate rule redundancy and, thus, find reliable evidence of data misuse.

5.1. Confidence Gain

Given two rules $r_i : \Sigma_i, \Delta_i \rightarrow \Omega_i$, $r_j : \Delta_j \rightarrow \Omega_i$, such that r_i is not a sub-rule of r_j, the idea underlying our approach consists in determining whether the relation

described by r_i holds regardless of the satisfaction of the constraints in r_j. The intuition is that a rule that shows a low degree of redundancy with respect to other rules captures a unique (or almost unique) correlation between its attribute itemsets and a class value. Therefore, it is reasonable to assume that the classification of corresponding records depends on those attribute itemsets. We claim that only these rules should be considered to extract PMIs.

To formalize this idea, we introduce the concept of *confidence gain*. Recall from Section 2 that $conf(r_i) = \frac{supp(r_i)}{supp(\Sigma_i \cup \Delta_i)} = \frac{|cov(\Sigma_i \cup \Delta_i \cup \Omega_i)|}{|cov(\Sigma_i \cup \Delta_i)|}$. We can rewrite confidence to highlight the contribution of the samples shared between r_i and r_j as follows:

$$\begin{aligned} conf(r_i) &= \frac{|cov(\Sigma_i \cup \Delta_i \cup \Omega_i) \setminus cov(\Delta_j)| + |(cov(\Sigma_i \cup \Delta_i \cup \Omega_i) \cap cov(\Delta_j)|}{|cov(\Sigma_i \cup \Delta_i)|} \\ &= \frac{|cov(\Sigma_i \cup \Delta_i \cup \Omega_i) \setminus cov(\Delta_j)|}{|cov(\Sigma_i \cup \Delta_i)|} + \frac{(cov(\Sigma_i \cup \Delta_i \cup \Omega_i) \cap cov(\Delta_j)|}{|cov(\Sigma_i \cup \Delta_i)|} \\ &= conf_{res}(r_i, r_j) + conf_{ov}(r_i, r_j) \end{aligned} \tag{1}$$

We name $conf_{res}(r_i, r_j)$ the *residual confidence* of r_i with respect to r_j, and $conf_{ov}(r_i, r_j)$ the *overlapping confidence* of r_i with respect to r_j. Residual confidence provides the contribution given to the rule confidence by the samples covered by r_i but not by r_j. In particular, $conf_{res}$ represents the probability, given itemsets Σ_i, Δ_i, to have Ω_i when Δ_j does not occur. If this probability is low, it means that in most cases where r_i holds, also r_j holds. This suggests the existence of a correlation between $\Sigma_i \cup \Delta_i$ and Δ_j, which, in turn, likely implies that the relation shown by the attribute itemsets in r_i is actually unreliable. Similarly, overlapping confidence provides the contribution given by the samples covered by both rules.

We can now formalize the notion of *confidence gain*:

$$conf_{gain}(r_i, r_j) = \frac{conf_{res}(r_i, r_j)}{conf(r_i)} \tag{2}$$

Intuitively, $conf_{gain}$ provides an indication of the degree of correlation between two rules, which represents the impact of the samples covered by r_i and not covered by

r_j on the overall confidence. From Equation 2, it follows that the value of $conf_{gain}$ ranges between 0, which means that the coverage of r_i is completely covered by the coverage of r_j, and 1, which means that the coverage of r_i and r_j are disjoint (hence, $conf_{res}(r_i, r_j) = conf(r_i)$). By definition, lower is $conf_{gain}$, more likely r_i is redundant with respect to r_j.

Example 2. *Let us consider rules* $r_5 : PrevAppl = Y \rightarrow Decision = N$ *and* $r_6 : Age = \text{“}< 25\text{”} \rightarrow Decision = N$ *in Table 3. We now compute the confidence gain of* r_6 *with respect to* r_5 *based on the dataset in Table 2.* r_6 *covers 480 samples, among which 30 are also covered by* r_5*. Therefore,* $conf_{res}(r_6, r_5) = 0.66$ *and* $conf_{gain}(r_6, r_5) = 0.94$*. The high value obtained for the confidence gain suggests that, although some overlapping exists between the two rules,* r_6 *can be assumed to be independent (i.e., not redundant) from* r_5*.*

$conf_{gain}$ is in essence a “local” metric, i.e. it estimates the redundancy between rules pairwise. To evaluate the overall strength of a rule, we need to introduce some global metrics to compare a rule with a set of rules. In this respect, it is worth noting that, when determining the overall degree of redundancy of a rule r_i, we do not compare r_i with *all* the other rules in the rules set. First, note that super-rules of r_i should not be considered in this computation. This is because, as discussed in Section 4, comparing the coverage of a rule with those of its super-rules does not provide a fair evaluation of the relevance of the rule. Instead, we propose to use previous work in sub-rules redundancy to evaluate the redundancy of a rule with respect to its super-rules. Furthermore, we consider only rules that share at least one sample with r. Finally, we are interested in comparing r_i with rules that do not involve sensitive itemsets, since our aim is to determine whether itemsets in r_i are correlated with non-sensitive itemsets.

According to these observations, to determine the global redundancy of a rule r_i,

we first introduce the notion of *competitor rules* of r_i, i.e. the set of rules against which we intend to compare r_i. Let $R = \{r \mid isInteresting(r)\}$ be a set of association rules fulfilling a set of relevance metrics and $R_{sup_{r_i}} \subseteq R$ the set of rules that are super-rules of a rule $r_i : \Sigma_i, \Delta_i \rightarrow \Omega_i$. The set of competitor rules of r_i is $R_{comp_{r_i}} = \{r_k : \Delta_k \rightarrow \Omega_i \mid r_k \in R \setminus R_{sup_{r_i}} \wedge cov(r_i) \cap cov(r_j) \neq \emptyset\}$, i.e. the set of rules that: 1) do not involve sensitive itemsets; 2) are not super-rules of r_i; 3) share at least one record with r_i.

Example 3. *Let us consider rule* $r_3 : PrevRole = Employee, Age = \text{“}{<}25\text{”} \rightarrow Decision = N$ *in Table 3. From the dataset in Table 2 and the ruleset in Table 3, it is easy to observe that* $R_{comp_{r_3}} = \{r_4, r_9\}$.

Having defined the set of competitor rules for r_i, we now define *Average Confidence Gain* (AVG_{cg}), *Minimum Confidence Gain* (MIN_{cg}) and *Weighted Average Confidence Gain* ($WAVG_{cg}$) as follows:

$$AVG_{cg}(r_i) = \frac{\sum_{k=1}^{N} conf_{gain}(r_i, r_k)}{|R_{comp_{r_i}}|} \tag{3}$$

$$MIN_{cg}(r_i) = \min_{r_k \in R_{comp_{r_i}}} conf_{gain}(r_i, r_k) \tag{4}$$

$$WAVG_{cg}(r_i) = \frac{\sum_{k=1}^{N} conf_{gain}(r_i, r_k) \times \omega(conf_{gain}(r_i, r_k))}{|R_{comp_{r_i}}|} \tag{5}$$

with $\omega(conf_{gain}(r_i, r_k)) = 2 \times conf_{gain}(r_i, r_k) - 1$.

$AVG_{cg}(r_i)$ and $MIN_{cg}(r_i)$ values range in the interval $[0, 1]$ and provide an estimation of the general performance of the rule. $MIN_{cg}(r_i)$ denotes the minimum probability that rule r_i has been actually exploited during the decision process; similarly, $AVG_{cg}(r_i)$ represents the average probability of the use of r_i.

$WAVG_{cg}$, instead, represents a weighted average of the confidence gain and its values range in the interval $[-1, 1]$. The rationale behind this metric is that the simple

average tends to overestimate the confidence gain. In fact, when a rule scores very poorly with respect to few rules, but quite well with respect to all the others, we can obtain relatively high values of the average confidence gain. However, for our purposes, low values of confidence gain are more important than high ones. Indeed, even if a rule scored poorly in very few cases, we want to be aware of this. $MIN_{cg}(r_i)$ provides us some insights in that sense, reporting the worst case. However, we argue that it would be desirable for an analyst that the impact of poor confidence gain is reported even in the average trend. To address this issue, we define the weighted average of confidence gain, using function ω as a weight. Intuitively, ω has the effect of amplifying low values of confidence gain, mapping them to negative values. For instance, when the confidence gain is 0, the weighted value is -1.

The analyst can select the metric that better fits her needs. It is also possible to use all of them and compare the results to obtain additional insights about the average trend of the rule.

Example 4. *By applying the global redundancy metrics to rule r_3 of Example 3, we have:*

- $AVG_{cg}(r_3) = AVG(conf_{gain}(r_3, r_4), conf_{gain}(r_3, r_9))$=$AVG(0.67, 0)$=0.335.
- $MIN_{cg}(r_3) = 0$.
- $WAVG_{cg}(r_3) = AVG(\omega(0.67), \omega(0))$= $AVG(0.34, -1)$=-0.33

$MIN_{cg}(r_3)$ is equal to 0, since r_9 covers all the samples of r_3. Since r_2 has only two competitor rules, and one of them shows a complete overlap with r_3, we obtain a low value of $AVG_{cg}(r_3)$ and $WAVG_{cg}(r_3)$.

5.2. Deriving μ-strong rules

The global metrics introduced above can be used to provide a practical approach to assess rule redundancy as defined in Section 4, in combination with the metrics for

Algorithm 1: Find μ-strong rules

Input : Set of sensitive attributes Σ, set of association rules R, minimum cg threshold μ

Output : μ-strong rules μ-SR

1 μ-$SR = \emptyset$;
2 **foreach** $r \in R$ **do**
3 $\quad redundant = checkSubRuleRedundancy(r)$;
4 $\quad$ **if** $redundant == true$ **then**
5 $\quad\quad MIN_{cg}(r) = 0$;
6 $\quad$ **else**
7 $\quad\quad$ generate R_{comp_r};
8 $\quad\quad CG_{list} = []$;
9 $\quad\quad$ **foreach** $r_c \in R_{comp_r}$ **do**
10 $\quad\quad MIN_{cg}(r) = computeMinCG(CG_{list})$;
11 $\quad$ **if** $MIN_{cg}(r) > \mu$ **then**
12 $\quad\quad \mu$-$SR = \mu$-$SR \cup \{r\}$;
13 **return** μ-SR

evaluating sub-rule redundancy proposed in the literature.

Specifically, let R be a set of interesting rules and μ a user-defined redundancy threshold. A rule $r_i : \Sigma_i, \Delta_i \rightarrow \Omega_i \in R$ is redundant if $\exists\ r_j : \Delta_j \rightarrow \Omega_i \in R$ such that $(\Delta_j \subseteq \Delta_j \wedge conf(r_i) \leq conf(r_j))$ or $MIN_{cg}(r_i) \leq \mu$. Note, however, that the $AVG_{cg}(r_i)$ and $WAVG_{cg}(r_i)$ metrics can be adopted as well.

Algorithm 1 describes our approach to extract the set of μ-strong rules, i.e. rules that are not redundant with respect to a given threshold μ. For each rule r in a given ruleset R, the algorithm first checks if the rule is redundant with respect to its super-rules that do not involve sensitive itemsets according to sub-rules redundancy analysis (line 3). If a rule is redundant, its confidence gain is set to 0. Otherwise, first the algorithm generates the set of competitor rules $R_{comp}(r)$ (line 7), by filtering from the set of interesting rules all those rules that do not fulfill the three constraints of $R_{comp}(r)$

set listed in the subsection above. Then, it computes the minimum confidence gain value using Eq. 4 (lines 9 and 10). If the value is above the user-defined threshold μ, r is added to output ruleset. At the end, the set of μ-strong rules is returned.

p-instance filtering vs μ-filtering. Our approach (hereafter, called μ-filtering) has similarities with p-instance proposed in [4] (an overview of p-instance is given in Section 4.3). They are both guided by the notion of *rule overlapping*, although the filtering logic of the two approaches can be considered one the "inverse" of the other. While p-filtering aims to remove all rules whose p (percent) or more of their samples is covered by other rules, μ-filtering keeps all those rules for which at most μ (percent) of their samples is covered by other rules. Nonetheless, there are also significant differences between these metrics:

- μ-filtering is based on a notion of overlap that is constrained by the class value. This reflects our claim that, without considering constraints on the class value, redundancy analysis might account for records that actually do not fulfill the rule and, thus, provides misleading results;
- μ-filtering employs sub-rules redundancy to also deal with redundancy of type (2), introduced in Section 4;
- μ-filtering allows evaluating redundancy among rules with different contexts, which are instead neglected by p-instance filtering.

Based on these observations, we expect some types of rule redundancy to be addressed by μ-filtering but not by p-instance filtering, as shown by the following examples. This intuition is confirmed by the results of our experiments as shown in Section 6.

The first type involves rules whose attribute itemsets occur in combination with different class itemsets in different records of the dataset.

Example 5. *Let us consider rules r_3 and r_9 in Table 3: $r_3 : PrevRole = Employee$, $Age =$ "< 25" $\rightarrow Decision = N$ and $r_9 : PrevRole = Employee, Instruction = Master \rightarrow Decision = N$. Based on the dataset in Table 2, we have $conf(r_9) > conf(r_3)$; thus, condition (a) of p-instance filtering is respected (see Section 4.3). However, if we consider rule $r : PreviousRole = Employee, Age =$ "< 25" $\rightarrow PrevRole = Employee, Instruction = Master$ as required by condition (b) of p-instance, we have $conf(r) = 0.69$. Therefore, rule r_3 is considered a p-instance of r_9 (i.e., it is filtered out) for values of p lower than 0.69. However, it is easy to observe from Table 3 that r_3 is completely covered by r_9. This is due to the fact that, when computing $conf(r)$, p-instance filtering takes also into account records in which $Age =$ "< 25", $PrevRole = Employee$ and $Decision = Y$. In particular, these itemsets have a negative impact on the confidence value of the rule, thus affecting the quality of the filtering. In particular, it prevents the filtering of a rule completely covered by another because records with a class value different from the one of the rule are taken into account when evaluating redundancy.*

The presence of itemsets occurring with different class values can also lead p-instance filtering to discard reliable evidence of data misuse, as shown in the next example.

Example 6. *Consider rules r_7: $Country = SA$, $PrevApp = N \rightarrow Decision = N$ and r_8: $PrevRole$ = $Self\text{-}Employee$, $PrevApp$=$N \rightarrow Decision = N$ (from Table 3). We assess the redundancy level of r_7 with respect to r_8 using p-instance filtering (refer to Section 4.3 for the notion of p-instance). First, since $conf(r_8) > conf(r_7)$, condition (a) is always true. With respect to condition (b), we have to assess the confidence of rule $r : Country = SA$, $PrevApp = N \rightarrow PrevRole = Self\text{-}Employee$. Based on the samples in Table 2, we have $conf(r) = 0.8$. Therefore, for values of p lower than 0.8, rule r_7 is considered a p-instance of r_8. On the other hand, by applying*

our approach with $\mu = 0.2$, r_7 would not be filtered out. In fact, we can observe from Table 2 that r_8 actually covers around 70% of the records covered by r_7. The difference in the results obtained by these filtering approaches is due to the presence of 100 records in the dataset (see Table 2) in which $Country = SA$, $PrevApp = N$, $PrevRole = Self\text{-}Employee$, $Decision = Y$. These records are not considered by μ-filtering to assess the redundancy of r_7 because of the different class value. In contrast, these records contribute to the value of $conf(r)$ in p-instance filtering and, thus, the impact of r_8 on r_7 is overestimated. We argue that this is undesirable, since it hides the real degree of overlap between the two rules.

Another type of rule redundancy that cannot be managed by p-instance filtering involves rules that do not share the same context Δ_i.

Example 7. *Let us consider rules r_2 and r_5 in Table 3, described within case (1) of redundancy in the previous section. Although r_5 describes exactly the same set of records of r_2, p-instance filtering would not filter r_2 out, since the non-sensitive itemset $SpeakLanguage = N$ does not occur in r_5.*

We conclude by observing that, intuitively, whether a rule is to be removed or kept depends on the threshold values selected for support and confidence thresholds, in combination with high values of parameter p. However, by increasing values for these thresholds, we significantly limit the level of overlap that can be checked by p-instance filtering. Moreover, neither increasing the support-based thresholds nor using high values for p can make p-instance filtering able to evaluate the redundancy between two rules that do not share the same non-sensitive context Δ_i.

6. Experiments

We implemented our proposed metrics and our algorithm to find μ-strong rules as a Java application. We performed a set of experiments using both synthetic and

real-life datasets. The aim of the evaluation with the synthetic data is to perform controlled experiments with a known ground truth. We then used two real-life datasets to show that the approach is able to handle a dataset with real-life complexity. In the experiments, we compared our approach with p-instance filtering (see Section 4.3) as a comparative method. To the best of our knowledge, this is the only approach that provides a method that accounts for the presence of sensible attributes and possible redundancy.

Moreover, we show an application of our approach to enhance existing approaches aiming at *evaluating the degree of discrimination* in a given trace of data. In particular, we evaluate the ability of our approach to improve the seminal method proposed by Ruggeri et al. [14] on the discovery of discriminatory groups (an overview of this method is presented in Section 4.1). To this end, we show the differences in the set of α-discriminatory rules obtained when applying the approach of Ruggeri and colleagues on the set of rules (a) only fitting support-based user-defined thresholds, (b) returned by p-instance filtering and (c) returned by μ-filtering. We show how this difference varies for various values of α, p and μ.

Settings. In both experiments, given a relational dataset and a set of sensitive items, we first extracted the set of class association rules using the R tool (`https://www.r-project.org/`). We adopted the standard support-based metrics introduced in Section 2 to obtain the set of interesting rules. In particular, we set the thresholds for support and confidence to $\rho_{supp} = 1\%$ and $\rho_{conf} = 80\%$, as done in similar experiments performed in previous work on discrimination discovery (e.g., [14]). Moreover, we set a threshold for lift, namely $\rho_{lift} = 1.5$, in order to obtain a set of rules showing correlation between the attribute itemsets and the class itemset.

To test our approach, we varied parameter μ between 0 and 0.5 by increasing the value of 0.1 at each step. Moreover, we adopted the approach proposed by Bayardo and

colleagues [16] to evaluate sub-rule redundancy. Bayardo's approach filters out each rule r_i that has a super-rule r_j such that both r_i and r_j fit the user-defined thresholds for support-based metrics and $conf(r_i) \leq conf(r_j)$.

To compare p-instance filtering with μ-filtering, we had to choose appropriate values for parameter p. As observed in Section 5.2, the two approaches are both based on the notion of rule overlapping but implement opposite filtering logic. Therefore, to obtain from μ-filtering a filtering behavior similar to the one exhibited by p-instance, we have to set $\mu = 1 - p$. Indeed, when we set $\mu = 0$, we remove only those rules whose coverage is completely covered by others; hence, we can compare the rules obtained using μ-filtering with those obtained using p-instance filtering with $p = 1$. Similarly, if we set $\mu = 0.2$, we filter those rules whose coverage is covered by others for 80% or more, thus allowing us to compare the results with those obtained by setting $p = 0.8$; and so on. Following these observations, we varied p between 1 and 0.5 by decreasing the value of 0.1 at each step.

With respect to the experiments concerning the discovery of α-discriminatory rules, we varied α between 1 and 1.5, increasing the value of 0.25 at each step. Specifically, we considered three values of α: $\alpha = 1$ can be considered as the minimum value from which the discrimination level becomes significant; $\alpha = 1.25$ is derived from the so called *fourth-fifths* rule[2] that is a rule to determine discrimination according to U.S. legislation [22]; finally, $\alpha = 1.5$ can be considered as representative of strong discriminatory rules. Note that these values of α were also used in previous work on discrimination discovery (e.g., [4, 14]).

[2]The *fourth-fifths* rule states that "a selection rate for any race, sex, or ethnic group which is less than four-fifths (or eighty percent) of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact". We follow Pedreschi and colleagues approach [4], and translate this rule in a numerical value for α equal to 1.25.

Evaluation Framework. The goal of our evaluation is to demonstrate the capability of μ-filtering to identify and remove redundant rules. To this end, we define a metric, named *Coverage Redundancy* (CR), to evaluate the degree of redundancy of a ruleset obtained using μ-filtering and to compare our approach with similar approaches. In particular, this metric measures how many rules describe each sample covered by the ruleset. Let D be a dataset and R a ruleset describing the decisional process underlying D. Let $\overline{D} = \bigcup\{cov(r) \mid r \in R\}$ be the set of records in the dataset covered by at least one rule in R. Give a record $d \in \overline{D}$, the coverage redundancy of r is defined as:

$$CR(d) = |\{r \mid d \in cov(r) \wedge r \in R\}| \tag{6}$$

The values returned by $CR(d)$ range from 1, meaning that a sample is described by exactly one rule, to $|R|$, meaning that all the rules in the ruleset describe the sample. To evaluate the overall coverage redundancy of a ruleset, we plot values of $CR(d)$ using boxplot, offering a standardized way of displaying the distribution of data. Note that we excluded records that are not covered by any rule. This is twofold. First, we argue that records that are not covered by any rule are not interesting for the analysis. In fact, these are records covered by rules that do not fulfill minimum support-based metric requirements and/or do not involve sensitive itemsets and/or are redundant with respect to other rules. Morever, considering these records may lead to misleading values for the CR metric. To get an intuition of this effect, let us assume to have two rulesets R_a, R_b obtained using two rules mining approaches a and b on a dataset D. Let us assume that each rule in R_a and in R_b covers exactly one record in the dataset; however, while rules in R_a cover all the records in D, rules in R_b cover only a percentage of the original dataset. If we consider also records not covered by any rule, in this situation, R_b would score better than R_a with respect to the CR metric; however, this is clearly undesired, since R_a actually represents an optimal situation in which all records are covered without any redundancy.

Besides coverage redundancy, we also use a number of metrics to obtain detailed statistics on the rulesets obtained by the filtering approaches (e.g., number of obtained rules) and on their comparison (e.g., number of common rules).

6.1. Synthetic experiments

This section presents our experiments on a synthetic dataset. We first describe how the dataset was generated; then, we discuss the results of the experiments.

6.1.1. Dataset

We built a synthetic dataset of 1000 records by simulating a decisional process concerning the selection of candidates for a job position. To this end, we defined a set of features along with the corresponding feature domain (i.e., the values that a feature can take). Table 4 shows the features we considered for each applicant along with the feature domains and the probability distribution of feature values. The last column of the table shows whether a feature was considered sensitive in our experiment or not. In particular, we considered sensitive all itemsets involving at least one of the following attributes: *Country*, *Age*, *Gender* and *PersonalStatus*.

We modeled the decisional process as a *decision tree*, yielding a decision – "yes" or "no" – on the basis of the feature values. We set different features to have a different impact on the final decision. Specifically, some features directly influence the decision, while other features become relevant only in certain "context", defined as a combination of feature values (i.e., an itemset containing more than one item). For instance, applications of candidates that do not speak the required language are immediately rejected. If the applicant is able to speak that language, other features are taken into account; for instance, in this context applicants with a high degree of instruction (Doctorate level) are always hired. Similarly, applicants whose age is over 50 are rejected if they have a low level of instruction (Bachelor); while if they

Feature	Values	Probability Distribution (%)	Sensitive
Instruction	Bachelor	40	No
	Master	30	
	Doctorate	20	
	HighSchool	10	
SpeakLanguage	No	20	No
	Yes	80	
PreviousRole	Unemployed	15	No
	Employee	25	
	Manager	30	
	Self-Employee	30	
Age	$[0, 25)$	20	Yes
	$[25, 50]$	50	
	$(50, +\infty)$	30	
Country	China	15	Yes
	Europe	30	
	India	20	
	SA	15	
	USA	20	
Gender	F	60	Yes
	M	40	
PersonalStatus	married	30	Yes
	relationship	30	
	single	40	
Hobbies	No	60	No
	Yes	40	
PreviousApplications	No	50	No
	Yes	50	
FullTime	No	50	No
	Yes	50	

Table 4: Features for each applicant in the synthetic dataset, along with the probability distribution of their values and the sensitivity of the feature.

have at least a Master degree, they might be hired or not depending on their previous position. In the creation of the dataset, we simulated the decisional process as realistic as possible to not test our approach in a too favorable situation.

6.1.2. Results

We first analyze the performance of μ-filtering and compare it to p-instance filtering on the synthetic dataset. Then, we evaluate the impact of those filtering approaches when combined with techniques for discrimination level evaluation.

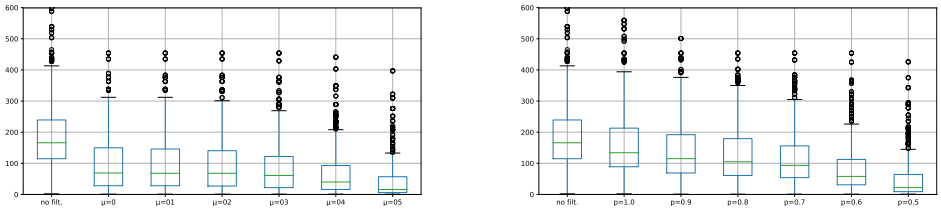
μ-filtering		p-instance filtering		Comparison		
Par Value	$\#(R_\mu)$	**Par Value**	$\#(R_p)$	$\#(R_p \cap R_\mu)$	$\#(R_p \setminus R_\mu)$	$\#(R_\mu \setminus R_p)$
$\mu = 0.0$	4498	$p = 1.0$	8590	4498	4092	0
$\mu = 0.1$	4409	$p = 0.9$	7577	4409	3168	0
$\mu = 0.2$	4230	$p = 0.8$	6819	4230	2589	0
$\mu = 0.3$	3741	$p = 0.7$	5914	3741	2173	0
$\mu = 0.4$	2853	$p = 0.6$	4260	2853	1407	0
$\mu = 0.5$	1750	$p = 0.5$	2520	1748	772	0

Table 5: Results obtained for the synthetic dataset.

Analysis of μ-filtering. We obtained 10677 association rules fitting the support-based thresholds, among which 9990 rules involve at least one sensitive itemset. Table 5 shows the results obtained using μ-filtering (R_μ) and p-instance filtering (R_p) on these rules. Column **Par Value** shows the configuration of parameters μ and p, and columns $\#(R_\mu)$ and $\#(R_p)$ report the size of the rulesets obtained by applying these filtering approaches. In addition, the table reports comparison metrics, i.e. the intersection ($R_p \cap R_\mu$) and the differences between the rulesets ($R_p \setminus R_\mu$ and $R_\mu \setminus R_p$).

From the table, we can observe that, in most cases, there is a significant gap in the number of potentially misuse (PM) rules returned by the two approaches. In every configuration tested, the ruleset obtained using μ-filtering is a subset of the ruleset extracted with p-instance filtering. This means that the p-instance filtering approach failed to detect a significant number of redundant rules. The table shows that around 40-50% of the rules obtained by p-instance filtering in all configurations actually turned out to be redundant. This result is also confirmed by the coverage redundancy of the two approaches reported, shown in Figure 1.

Table 6 provides details on the redundant rules filtered by μ-filtering. For each value of the threshold μ, column $R \setminus R_\mu$ reports the total number of filtered rules, while column **Redundancy Type** reports the number of redundant rules for each type of redundancy discussed in Section 3. Specifically, column **Type 1** reports the number of rules that completely overlap with at least one of their competitor rules in R. Column **Type 2** reports the number of rules for which a super rule with a better



(a) μ-filtering (b) p-instance filtering

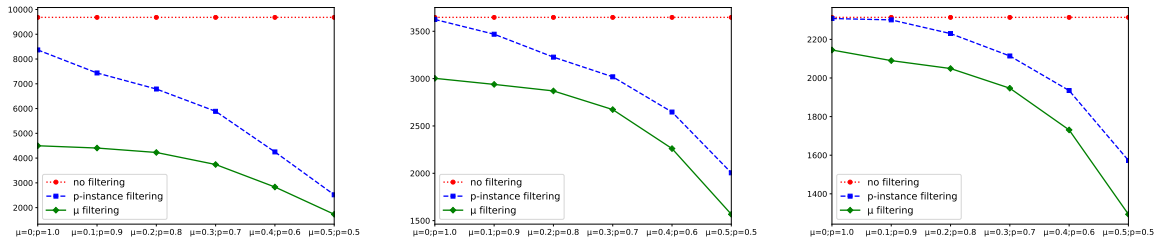
Figure 1: Coverage Redundancy for the synthetic dataset

		Redundancy Type			**Rules missed by p-instance filtering**			
μ	$\#(R \setminus R_\mu)$	**Type 1**	**Type 2**	**Type 3**	**Super Rules**	**Context Mismatch**	**Class Mismatch**	**Multiple reasons**
0	5492	1657	3835	0	2677	766	77	572
0.1	5584	1657	3835	92	2368	392	8	400
0.2	5725	1657	3835	233	2029	243	1	316
0.3	6249	1657	3835	757	1584	215	0	374
0.4	7158	1657	3835	1666	805	159	0	443
0.5	8240	1657	3835	2748	169	90	2	511

Table 6: Analysis of the redundancy types detected by μ-filtering, together with redundant rules missed by p-instance filtering for the synthetic dataset

confidence exists. Clearly, these values are the same for all values of μ. Columns **Type 3** reports the number of rules whose confidence gain was lower than the given threshold for at least one competitor rule. It is straightforward to observe that stronger is the filtering, larger is the number of rules filtered out because of partial overlap.

The last three columns of Table 6 focus on rules that are filtered by μ-filtering but not by p-instance filtering (i.e., $R_p \setminus R_\mu$) for the given value of μ (and, hence, of p); the number of these rule was reported in column $\#(R_p \setminus R_\mu)$ in Table 5. These results can be explained based on the observations made in Sections 4.3 and 5.2. Column **Super Rules** reports the number of rules in $R_p \setminus R_\mu$ that are redundant due to the presence of (at least) one super-rule amongst its competitor rules. Column **Context Mismatch** reports the number of rules in $R_p \setminus R_\mu$ that are redundant due to the presence of (at least) one competitor rule with whom the rule shares only a portion of its non-sensitive itemsets. Column **Class Mismatch** reports the number of rules in $R_p \setminus R_\mu$ that are



(a) $\alpha = 1$ (b) $\alpha = 1.25$ (c) $\alpha = 1.5$

Figure 2: Number of α-discriminatory rules mined from the synthetic dataset.

not marked as redundant by p-instance filtering because they neglect the class itemset value. Finally, note that a rule might be redundant for more than one of the reasons above, since it might have more than one competitor rule and each competitor rule can make the rule redundant for a different reason. The number of such rules is reported in column **Multiple reasons**.

Determining α-discriminatory rules. We now analyze the impact of μ-filtering and p-instance filtering on the mining of α-discriminatory rules on the synthetic dataset. Here, for sake of space, we only report the number of mined rules; the other metrics exhibit a trend similar to the one observed in the previous experiment and therefore are omitted.

Fig. 2 shows the number of α-discriminatory rules obtained with and without filtering at the varying of α. In particular, for each tested α value, the figure shows the number of α-discriminatory rules: a) on the set of rules fitting support, confidence and lift thresholds (red dotted line); on the set of rules returned using p-instance filtering (blue dashed line); on the set of rules obtained using μ-filtering (green full line). Parameters p and μ were varied as in the previous experiment.

It is easy to observe how the use of p-instance filtering and μ-filtering has a relevant impact on the number of α-discriminatory rules that are obtained. For all tested values of α, without using any filtering we obtained in output thousands of rules marked as

discriminatory while they are actually redundant with respect to non-discriminatory rules. Not surprisingly, this difference becomes even larger when stronger filtering approach is applied. For instance, the number of the discriminatory rules for $\alpha = 1$ drops from 9682 to 4227 for $\mu = 0.2$. Note that, also in these experiments, the p-instance filtering returned a larger ruleset compared to μ-filtering, in which many rules are actually redundant. The difference between the number of rules returned by the two filtering approaches is particularly evident when $p = 1, \mu = 0$. Increasing the filtering power, the differences between the two filtering approaches become less prominent.

6.2. Experiments on Speed Dating dataset

We also evaluated our approach on real-life datasets. For the first set of experiments, we used a dataset about individual preferences in dating [23]. This dataset was collected in experimental speed dating events at the Columbia University in 2002-2004. During the events, attendees had a series of four minute "first date" with every other participant of the opposite sex. At the end, participants were asked if they would like to see their date again. Personal data regarding participants, like their age and race, as well as their beliefs, aspirations and the personal characteristics that they look in a partner, were collected by means of questionnaires.

6.2.1. Dataset

We pre-processed the dataset in [23] by removing missing values and by discretizing numerical attributes. In particular, we dropped the features with large number of missing values (more than 40% of the total samples), since the presence of a large number of missing values suggests that some issues might have occurred during data collection; then, we removed users' records with one or more missing values on the remaining features. Finally, we discretized numerical attributes using quarterly's; namely, we divided the feature range in four equal groups, each comprising one quarter

of the data. The resulting dataset involves 6468 samples with 39 features grouped according to the following categories: personal information on the participants to the experiment (age, field of study, race, gender); participants' rating on some of their own as well as on the date partner's characteristics (ambition, attractiveness, funniness, intelligence, sincerity), over a scale from 0 to 10; preferences expressed by the participants over the same characteristics, i.e. an evaluation from 0 to 100 of how important each characteristic is when looking for a partner; similarities between interests and beliefs of the participants; the number of people that every participant expects to be interested in him/her; attributes representing whether the date participants have the same race, the same field of study, and whether the partner expects more people to be interested in him/her than in the subject;[3] and the final decision of the subject of the experiment. A complete description of the experiment set-up and data collection can be found in [24].

The choice of a partner is clearly related to an individual's preferences and orientation. Our aim is to investigate whether a choice for a partner is affected by intrinsic characteristics related to individual's origins and/or personal preferences rather than the impression that the subject had about the date's personality during the meeting. Accordingly, we marked as sensitive the following attributes: $partner\_race$, $date\_same\_race$, $partner\_age$, $partner\_gender$, $partner\_field\_category$, $same\_field$.

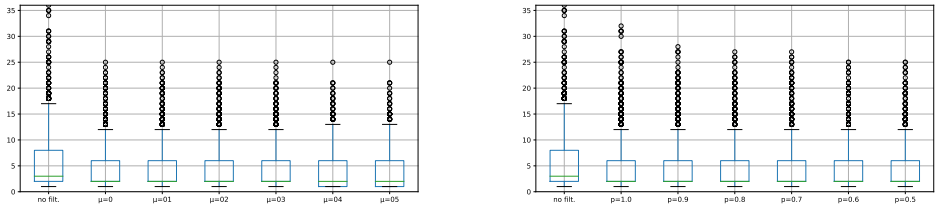
6.2.2. Results

We first analyze the performance of μ-filtering and compare it to p-instance filtering on the Speed Dating dataset. Then, we evaluate the impact of those filtering

[3] At the beginning of each dates session, participants were asked to estimate the number of people who they believed would have shown interest in them; this attribute is derived by comparing the numbers provided by the participants.

μ-filtering		p-instance filtering		Comparison		
Par Value	$\#(R_\mu)$	**Par Value**	$\#(R_p)$	$\#(R_p \cap R_\mu)$	$\#(R_p \setminus R_\mu)$	$\#(R_\mu \setminus R_p)$
$\mu = 0.0$	91	$p = 1.0$	111	91	20	0
$\mu = 0.1$	89	$p = 0.9$	105	89	16	0
$\mu = 0.2$	89	$p = 0.8$	103	89	14	0
$\mu = 0.3$	83	$p = 0.7$	103	83	20	0
$\mu = 0.4$	73	$p = 0.6$	98	73	25	0
$\mu = 0.5$	67	$p = 0.5$	95	67	28	0

Table 7: Results obtained for the Speed Dating dataset



(a) μ-filtering (b) p-instance filtering

Figure 3: Coverage Redundancy for the Speed Dating dataset

approaches when combined with techniques for discrimination level evaluation.

Analysis of μ-filtering. From this dataset, we obtained 213 rules fitting the support-based thresholds, 135 of which involve sensitive itemsets. Table 7 shows the results obtained on this ruleset using μ-filtering and p-instance filtering. The trend of coverage redundancy for each configuration is reported in Fig. 3.

We note that the results do not change significantly while varying μ and p parameters. Having a closer look at the results, we can observe that most redundant rules show an almost complete overlap with some other rule in the ruleset; hence, they were filtered out already in the first configuration (i.e., $\mu = 0.0, p = 1.0$). In particular, we can observe in Fig. 3 that the coverage redundancy of μ-filtering and p-instance filtering maintains a constant trend for all configurations. However, it is worth noting that the number of outliers (i.e., the dots above the upper limit) decreases when the value of μ increases (and the one of p decreases) as well as when comparing μ-filtering (Fig. 3a) with p-instance filtering (Fig. 3b). Besides, results are consistent

		Redundancy Type			Rules missed by p-instance filtering			
μ	$\#(R \setminus R_\mu)$	**Type 1**	**Type 2**	**Type 3**	**Super Rules**	**Context Mismatch**	**Class Mismatch**	**Multiple reasons**
0	44	8	36	0	1	19	0	0
0.1	46	8	36	2	1	15	0	0
0.2	46	8	36	2	1	13	0	0
0.3	52	8	36	8	1	19	0	0
0.4	62	8	36	18	1	24	0	0
0.5	68	8	36	24	0	27	0	1

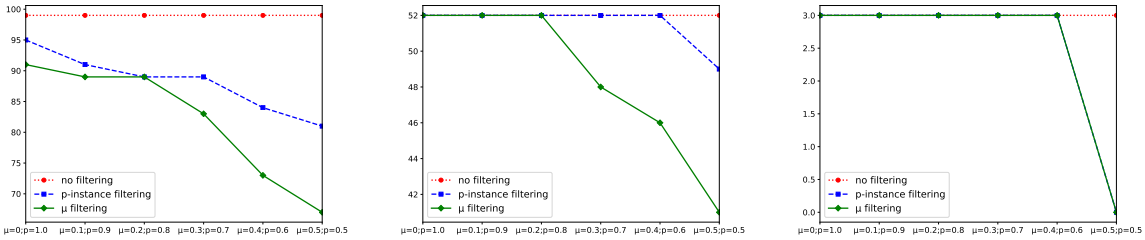
Table 8: Analysis of the redundancy types detected by μ-filtering, together with redundant rules missed by p-instance filtering for the Speed Dating dataset

with those obtained for the synthetic dataset. Yet, the ruleset obtained using μ-filtering is a subset of the ruleset obtained using p-instance filtering.

Table 8 details the different types of redundancy identified and removed by μ-filtering. A deeper analysis of the rules filtered by μ-filtering but missed by p-instance filtering (column $\#(R_p \setminus R_\mu)$ in Table 7) shows that for this dataset we did not find any rule missed because of mismatches of the class itemset. Instead, for all configuration of μ and p, we found just one rule that was not captured by p-instance filtering due to the existence of a super rule, while all remaining rules missed by p-instance filtering are due to mismatches in the context. It is interesting to note that, when increasing the level of overlap (i.e., $\mu = 0.5$), we obtain one rule that is not captured by p-instance filtering because of both the existence of a super-rule and context mismatch.

Determining α-discriminatory rules. We now analyze the impact of μ-filtering and p-instance filtering on the mining of α-discriminatory rules on the Speed Dating dataset. As for the experiments on the synthetic dataset, we only report the number of mined rules.

From Fig. 4, we can observe that, for all three values of α (i.e., 1, 1.25 and 1.5), we obtained the same or similar ruleset in the first configuration(s) regardless of filtering approach that is applied. On the other hand, this trend changes when a stronger filtering is applied. When $\alpha = 1$, we note that, without applying any filter, we



(a) $\alpha = 1$ (b) $\alpha = 1.25$ (c) $\alpha = 1.5$

Figure 4: Number of α-discriminatory rules mined from the Speed Dating dataset.

obtain a ruleset involving at least 10% redundant rules. This percentage increases when the filtering is stricter. Moreover, it is worth noting that, while for the initial configurations p-instance filtering provides a ruleset similar to those returned using μ-filtering (although they only converge on one configuration, i.e. $\mu = 0.2, p = 0.8$), for lower degree of overlap, the performance of p-instance filtering is worse, since it misses several redundant rules, which on the other hand are caught using μ-filtering. For $\alpha = 1.25$, we obtained a ruleset in which no rule overlaps with other rules for more than 30%. Therefore, both μ-filtering and p-instance filtering returned the same ruleset of the non-filtering approach for the first three configurations. We also note μ-filtering was able to capture some redundant rules, diverging from the case where only support-based user-defined thresholds are applied; whereas p-instance filtering is not able to detect any redundancy until the last configuration, where in any case it still missed most redundant rules. Finally, we obtain only three rules for $\alpha = 1.5$, all with an overlapping degree ranging between 50% and 60%. Therefore, these rules have been filtered out using both μ-filtering and p-instance filtering only for $\mu = 0.5, p = 0.5$.

6.3. German Credit dataset

For the second set of experiments on real-life datasets, we used the German Credit dataset [25], which has been largely studied in previous work [4, 14].

6.3.1. Dataset

This dataset consists of 1000 records representing the good/bad credit class of bank account holders. There are 21 features in total, grouped according to the following categories: personal properties, i.e. checking account status, duration, savings status, property magnitude, type of housing; properties related to past/current credits and requested credit, i.e. credit history, credit request purpose, credit request amount, installment commitment, existing credits, other parties, other payment plan; properties related to the employment status, i.e. job type, employment since, number of dependents, own telephone; finally, personal attributes, i.e. personal status and gender, age, resident since, foreign worker. We discretized the numeric attributes as suggested in [14].

While in previous experiments we analyzed data misuse with respect to entire features, i.e. by considering all itemsets as sensitive, here we show the capability of the approach to perform also a more fine-grained analysis, relating possible misuses only to a subset of the itemsets. More precisely, consistent with [4], we marked as sensitive any itemset that contains one of the following items: $personal\_status = female\_div\_or\_dep\_or\_mar$, $age = [52, +\infty)$, $foreign\_worker = yes$, which means that the fact that the holder is a female, is older than 52 years and/or is a foreign worker should not influence the final decision.

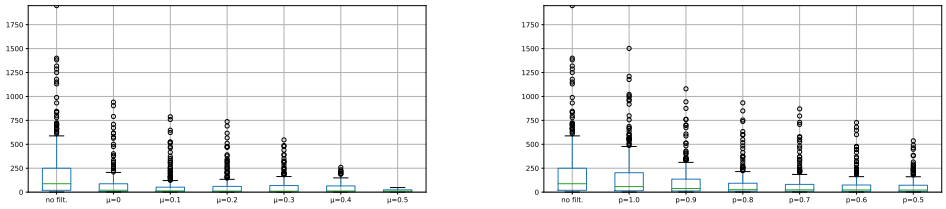
6.3.2. Results

We first analyze the performance of μ-filtering and compare it to p-instance filtering on the German credit dataset. Then, we evaluate the impact of those filtering approaches when combined with techniques for discrimination level evaluation.

Analysis of μ-filtering. From the German credit dataset, we obtained 8016 rules fulfilling the support-based thresholds, among which 4788 involving sensitive itemsets. Table 9 shows the results obtained using μ-filtering and p-instance filtering on this dataset, and Fig. 5 shows their coverage redundancy.

μ-filtering		p-instance filtering		Comparison		
Par. Value	$\#(R_\mu)$	**Par. Value**	$\#(R_p)$	$\#(R_p \cap R_\mu)$	$\#(R_p \setminus R_\mu)$	$\#(R_\mu \setminus R_p)$
$\mu = 0.0$	1536	$p = 1.0$	3660	1536	2124	0
$\mu = 0.1$	1247	$p = 0.9$	2586	1247	1339	0
$\mu = 0.2$	1110	$p = 0.8$	1994	1110	884	0
$\mu = 0.3$	843	$p = 0.7$	1858	843	1015	0
$\mu = 0.4$	448	$p = 0.6$	1687	448	1239	0
$\mu = 0.5$	100	$p = 0.5$	1464	100	1364	0

Table 9: Results obtained on the German credit dataset.



(a) μ-filtering (b) p-instance filtering

Figure 5: Coverage Redundancy for the German credit dataset

We can observe that the results are in line with the results obtained in the previous experiments. Also in this case, p-instance filtering missed a significantly large number of redundant rules and obtained much higher values of coverage redundancy in all configurations. In particular, around half of the rules returned by p-instance filtering in the first four configurations turned out to be redundant when checked with μ-filtering. This difference increases significantly if we consider the last two configurations, where the number of the rules returned using p-instance filtering are around four times (for $\mu = 0.4, p = 0.6$) and fourteen times (for $\mu = 0.5, p = 0.5$) the number of rules returned by μ-filtering. Also in this case, we did not find rules filtered out by p-instance filtering but not by μ-filtering. Details on redundant rules are reported in Table 10.

These results show how redundancy has a huge impact on the reliability of the evidence of data misuse, highlighting the limitations of p-instance filtering. Also for the German Credit dataset, p-instance filtering returned a significant amount of rules suggesting potential misuse of (sensitive) data, while they are actually redundant with

		Redundancy Type			Rules missed by p-instance filtering			
μ	$\#(R \setminus R_\mu)$	**Type 1**	**Type 2**	**Type 3**	**Super Rules**	**Context Mismatch**	**Class Mismatch**	**Multiple reasons**
0	3252	329	2922	0	1630	60	7	427
0.1	3569	329	2922	217	792	79	8	460
0.2	3687	329	2922	389	352	118	4	410
0.3	3945	329	2922	693	147	197	0	671
0.4	4341	329	2922	982	115	262	1	861
0.5	4688	329	2922	1432	41	245	0	1078

Table 10: Analysis of the redundancy types detected by μ-filtering, together with redundant rules missed by p-instance filtering for the German credit dataset

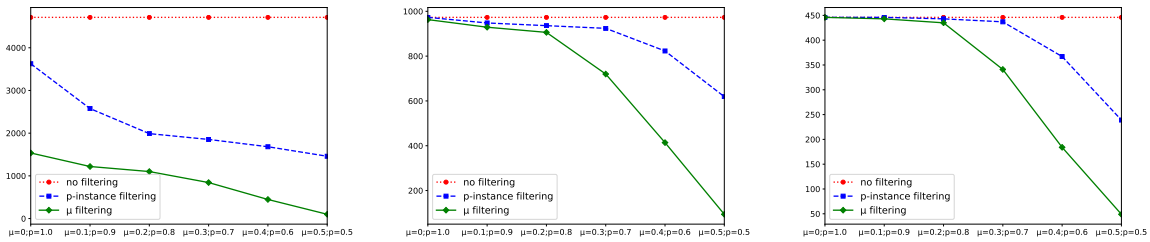
respect to rules without a sensitive itemset.

Determining α-discriminatory rules. We now analyze the impact of μ-filtering and p-instance filtering on the mining of α-discriminatory rules on the German credit dataset. As for the experiments on the other datasets, we only report the number of mined rules.

Fig. 6 shows the impact of the filtering approaches on the discovery of α-discriminatory rules for various value of α. The benefits of μ-filtering are particularly evident from these experiments. For $\alpha = 1$ (Fig. 6a), μ-filtering returned a set of rules significantly smaller than p-instance filtering for all configurations, especially for $\mu = 0.5$. For $\alpha = 1.25$ (Fig. 6b) and $\alpha = 1.5$ (Fig. 6c), the results of both the approaches tend to be similar for the first configurations. However, we can observe a significant improvement for higher values of μ, while we cannot see a similar trend for lower values of p.

6.4. Discussion

Our results show that μ-filtering outperforms previous methods in most of the tested datasets and filtering configurations both in terms of removal of redundant rules and reduction of coverage redundancy. In particular, for all experiments, the ruleset obtained using μ-filtering is a subset of the ruleset obtained using p-instance filtering,



(a) $\alpha = 1$ (b) $\alpha = 1.25$ (c) $\alpha = 1.5$

Figure 6: Number of α-discriminatory rules mined from the German Credit dataset.

proving that the limitations discussed in Section 4.3 have a relevant impact in terms of redundancy reduction.

Nevertheless, we also observed that, even when applying a strong redundancy filtering, we might still obtain a large ruleset. Although this is mainly due to the technique used to mine association rules (e.g., for the synthetic dataset we obtained 10677 rules), a large set of rules is likely hard to analyze for a human analyst. In the remanding of the section, we sketch possible strategies to deal with this issue.

A first possible solution consists of increasing the thresholds for the support-based metrics. By doing so, it is possible to reduce the size of the interesting ruleset and, consequently, of the final filtered set. It is worth noting, however, that this solution might lead to miss potential discrimination, especially when potentially discriminated groups are a minority in the dataset, which is a quite common situation. To address this issue, researchers have started studying the rules mining problem with unbalanced datasets [26]. A direction for future work is to investigate their application in the context of data misuse.

Another strategy consists of applying an additional post-processing step to *(i)* further simplify the set of rules and *(ii)* carry a more fine grained analysis of the sensitive information that was used in the decision process.

With respect to (i), let us consider as an example two of the μ-strong rules

returned with $\mu = 1$ from the synthetic dataset, namely r_{4757} : *Instruction* $=$ $Master$, $PreviousRole = Self\text{-}Employee$, $Country = Europe$, $FullTime = N \rightarrow Decision = Y$ and r_{856} : $Instruction = Master$, $PreviousRole = Self\text{-}Employee$, $Country = Europe$, $PreviousApplication = Y \rightarrow Decision = Y$. An analysis of the dataset shows that these rules actually cover the same samples. However, since both rules involve the *same* sensitive itemset, i.e. $Country = Europe$, we do not compare them with each other to determine whether they are redundant. This is in line with the goal of this work. Indeed, the presence of multiple rules involving the same sensitive itemsets does not contradict the fact that such an itemset has been exploited in the decision making process. Different rules show different possible contexts in which the itemset has been exploited. Nevertheless, if one is interested in a further reduction of size of the obtained ruleset and in a more fine grained detection of the misuse contexts, a post-processing step can be employed to "merge" redundant rules involving the same sensitive itemset.

In reference to (ii), it is worth noting that in the current setting, we might not be able to point out *which* sensitive item has been used in the decision process. Let us consider two rules obtained using μ-filtering: r_{4757} : $Instruction = Master, PreviousRole = Self\text{-}Employee$, $Country = Europe$, $FullTime = N \rightarrow Decision = Y$ and r_{4767} : $Instruction = Master$, $PreviousRole = Self\text{-}Employee$, $Age = [25, 50]$, $FullTime = N \rightarrow Decision = Y$. These rules describe the same samples, but they are not filtered out because there does not exist any other rule describing the same samples without considering a sensitive itemset. This suggests that either $Age = [25, 50]$ or $Country = Europe$ has an impact on the decision, providing evidence that some sensitive data were misused; however, we cannot determine which one was actually used. To determine which sensitive itemsets are used in the decision making process, one can relax the constraints on the rules to be compared with a given rule to determine the redundancy of that rule. In particular, one may ana-

lyze each rule within the ruleset but those involving exactly the same sensitive itemset.

Finally, we point out that applying the discussed post-processing strategies would also reduce the coverage redundancy of the obtained ruleset. Indeed, the high values we obtained in the experiments are due to the presence of multiple rules that describe the same samples but are not filtered out because they all involve sensitive itemsets.

7. Related work

In this work, we have presented a framework to determine reliable evidence of data misuse in the context of data analytics. Data misuse detection presents relevant similarities with the problem of *anti-discrimination* in data mining. Discrimination refers to the unfair treatment of people due to their belonging to same category (e.g., ethnic group, gender) [14]. Although the problem of discrimination has been largely studied (see, e.g., [27] for a survey), its relation with data mining has started being investigated only recently. In particular, data mining was recognized as a possible source of discrimination [28]. Since then, research on anti-discrimination in data mining has been developing across two main trends: *discrimination prevention* and *discrimination discovery*. Discrimination prevention techniques [14, 29, 30, 31, 32, 33, 34] aim to build discrimination-free classifiers, starting from dataset biased by discriminant choices.

Our work can be positioned in the area of discrimination discovery. The goal of discrimination discovery is to detect the presence of discriminatory decisions, possibly hidden within a dataset of historical decisions. A seminal work in this area is the one of Ruggieri et al. [5], who model the problem in terms of association rules mining and formalize the notion of discrimination by means of a set of metrics able to quantify the level of discrimination in the rules discovered by association rules mining algorithms. This idea has been further refined in subsequent studies [4, 14, 15, 35, 36], where additional metrics for discrimination evaluation and a framework to assess the

statistical significance of results are proposed. However, these approaches assess the discrimination levels of association rules without accounting for possible redundancy between these rules. As discussed in Section 4, this might result in misleading findings. Qureshi and colleagues recently introduced an approach for discrimination discovery in [3]. This work proposes the use of propensity score analysis to filter out the effects of so-called "confounding variables", i.e. variables (attributes) linked to both the class and other variables, which might introduce bias in the discovered association. Qureshi's approach requires an analyst to know which are the confounding variables to filter out. Feature selection techniques can be applied when this knowledge is not available; however, this introduces a bias in the obtained results, since the output depends on the adopted feature selection technique and the thresholds used for feature selection. Moreover, this approach aims to determine whether a group has been discriminated, rather than showing the contexts within which the discrimination has occurred.

Similar to [4, 5, 14, 15], we modeled an analytical process as a classification problem; however, in contrast to those works, we reduce the problem of data misuse to the discovery of redundancy in a set of class association rules.

Redundancy reduction has largely studied in association rule mining discipline. Table 11 provides an overview of existing approaches in this area that are related to our work, i.e. approaches that: *i)* consider redundancy in terms of rules with the same semantic meaning and *ii)* neither require users' intervention nor background knowledge to filter redundant rules out. For every approach, we report: the *redundancy types* they are able to address fully/partially (see Section 4); whether they account for the presence of *sensitive itemsets*, explicitly involving them for redundancy filtering; and whether they allow evaluating the *level* of redundancy of a rule with respect to other rules in the set.

A popular trend in redundancy reduction consider redundancy as an inclusion relation between rules. Aggarwal and Yu [37] define two types of redundant rules, i.e. *simple redundant* rules, which are rules having the same frequent itemsets of another

	Redundancy Types			Sensitive Itemsets	Redundancy Level
	Type 1	Type 2	Type 3		
Agrawal et al. [37]		●			
Zaki [19]		●			
Bastide et al. [38]		●			
Ashrafi et al. [39]		●			
Bayardo et al. [16]		●			
Jaroszewicz et al. [18]		●			
Liu et al. [40]		●			
Cristofor et al. [41]		●			
Pasquier et al. [42]		●			
Toivonen et al. [21]	●		◑		
Brijs et al. [20]	●		◑		
Pedreschi et al. [4]	◑		◑	●	●

● : fully supported ◑ : partially supported

Table 11: Previous work on redundancy reduction.

rule but a lower confidence value, and *strictly redundant* rules, which are sub-rules of another rule in the ruleset. Zaki [19] proposes to mine association rules from *closed* frequent itemsets, i.e. itemsets for which there does not exist any other frequent itemset that includes them and fits the desired support thresholds. Frequent closed itemsets are also used in [38], where rules with minimal antecedents and maximal consequent are selected. Ashrafi et al. [39] and Bayardo et al. [16] remove rules for which a super-rule with the same or a better confidence exists. Jaroszewicz et al. [18] exploit the maximum entropy principle, considering a rule as redundant if it does not introduce an improvement in terms of entropy with respect to its super-rules. Similarly, Liu et al. [40] determine this improvement using the χ^2 test. Moreover, Liu et al. introduce the notion of *directional rule* to capture rules that show a change of direction of the correlation between the rule consequent and their super-rules. Other works [41, 42] propose to extract minimal *covers*, i.e. a subset of association rules from which all other rules can be derived by some *inference rules*. It is worth noting that approaches that evaluate the redundancy

of a rule by exploiting only its super-rules, can only address redundancy of type (2).

Few studies have addressed redundancy between rules that are not involved in an inclusion relationship. For instance, Toivonen et al. [21] interpret rule redundancy as the overlap between the set of records covered by the rules and propose an algorithm to remove from the ruleset all the rules whose coverage is covered by some other rule. Brijs et al. [20] adopt a similar approach; however, they constrain the rule removal process to minimize the mixing of rules describing objects related to different class values. These approaches are able to deal with complete and partial overlap of rules not involved in an inclusion relationship (redundancy of types (1) and (3)). However, they favor rules with the highest support and, thus, sub-rules tend to be filtered out even when they actually might better characterize their records. Therefore, these approach cannot handle redundancy of type (2). Moreover, their goal is to determine a small number of rules able to describe the dataset; thus, the obtained rules might be redundant with respect to rules that have been filtered out, failing to meet our analysis goal.

It is worth noting that approaches for redundancy reduction are typically not devised for data misuse detection and, thus, they do not account for the presence of sensitive itemsets nor allow evaluating the level of redundancy. To the best of our knowledge, the work in [4] is the only one that addresses the problem of redundancy in association rules in the context of discrimination discovery. However, as discussed in Section 5, it presents some severe limitations in filtering redundant rules and is able to partially deal with both redundancy types (1) and (3) and is not able to deal with redundancy type (2). Compared with previous work, our approach is able to deal with all three types of redundancy.

8. Conclusion

In this paper, we focused on data misuse in the context of classification tasks, wherein evidence of data misuse is extracted by means of association rule mining. We discussed issues related to reliability of the evidence of misuse in presence of redundant rules; we also studied various types of redundancy and investigated their effect on the reliability of data misuse evidence. Compared to previous work, our approach *i)* adopts a notion of redundancy tailored to the detection of data misuse, which takes into account the presence/absence of sensitive itemsets; *ii)* is able to address redundancy both between overlapping and partially overlapping rules and between rules involved in inclusion relationships, thus being able to address all three cases of redundancy discussed in Section 4; *iii)* allows the evaluation of the level of redundancy of a rule with respect to a dataset; *iv)* accounts for the class value when assessing rule redundancy. We validated our approach using both synthetic and real-life datasets, and compared it with state-of-the-art solutions. Our approach outperformed existing methods in all datasets, as it performs a significantly higher reduction of redundant rules.

As future work, we plan to develop post-processing techniques to further filter the rules returned by our approach to fit specific analysts' needs (e.g., to group all rules that share the same sensitive itemsets). Moreover, we plan to enhance our approach to further quantify the statistical significance of the results. Finally, we plan to study a possible generalization of the proposed metrics to compare each rule with respect to *combinations* of other rules, rather than performing one-to-one comparisons.

Acknowledgments. This work is partially supported by the ITEA2 project M2MGrids (13011).

References

[1] P. Guarda, N. Zannone, Towards the development of privacy-aware systems, Information & Software Technology 51 (2) (2009) 337–350.

[2] H. Fienberg, FTC Warns Against Use and Misuse of Big Data Analytics, http://www.insightsassociation.org/article/ftc-warns-against-use-and-misuse-big-data-analytics, 2016.

[3] B. Qureshi, F. Kamiran, A. Karim, S. Ruggieri, Causal Discrimination Discovery Through Propensity Score Analysis, arXiv preprint arXiv:1608.03735, 2016.

[4] D. Pedreschi, S. Ruggieri, F. Turini, Integrating induction and deduction for finding evidence of discrimination, in: Proceedings of International Conference on Artificial Intelligence and Law, ACM, 157–166, 2009.

[5] D. Pedreshi, S. Ruggieri, F. Turini, Discrimination-aware data mining, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 560–568, 2008.

[6] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial intelligence 97 (1-2) (1997) 273–324.

[7] L. Allodi, F. Massacci, Comparing vulnerability severity and exploits using case-control studies, ACM Transaction on Information and System Security (TISSEC) 17 (1) (2014) 1:1–1:20.

[8] E. E. Eljadi, Z. A. Othman, Anomaly detection for PTM's network traffic using association rule, in: Proceeding of Conference on Data Mining and Optimization, IEEE, 63–69, 2011.

[9] M. V. Mahoney, P. K. Chan, Learning Rules for Anomaly Detection of Hostile Network Traffic, in: Proceedings of IEEE International Conference on Data Mining, IEEE, 601–, 2003.

[10] C. Feng, V. R. Palleti, A. Mathur, D. Chana, A Systematic Framework to Generate Invariants for Anomaly Detection in Industrial Control Systems, in: Proceedings of Network and Distributed Systems Security Symposium, 2019.

[11] R. Agrawal, T. Imieliński, A. Swami, Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Rec. 22 (2) (1993) 207–216.

[12] P.-N. Tan, V. Kumar, J. Srivastava, Selecting the right objective measure for association analysis, Information Systems 29 (4) (2004) 293–313.

[13] B. L. W. H. Y. Ma, B. Liu, Integrating classification and association rule mining, in: Proceedings of International Conference on Knowledge Discovery and Data Mining, AAAI Press, 80–86, 1998.

[14] S. Ruggieri, D. Pedreschi, F. Turini, Data mining for discrimination discovery, ACM Transactions on Knowledge Discovery from Data 4 (2) (2010) 9:1–9:40.

[15] D. Pedreschi, S. Ruggieri, F. Turini, Measuring discrimination in socially-sensitive decision records, in: Proceedings of International Conference on Data Mining, SIAM, 581–592, 2009.

[16] R. J. Bayardo, R. Agrawal, D. Gunopulos, Constraint-based rule mining in large, dense databases, in: Proceedings of International Conference on Data Engineering, IEEE, 188–197, 1999.

[17] S. Kotsiantis, D. Kanellopoulos, Association rules mining: A recent overview, GESTS International Transactions on Computer Science and Engineering 32 (1) (2006) 71–82.

[18] S. Jaroszewicz, D. A. Simovici, Pruning redundant association rules using maximum entropy principle, in: Advances in Knowledge Discovery and Data Mining, vol. 2336 of *LNCS*, Springer, 135–147, 2002.

[19] M. J. Zaki, Generating non-redundant association rules, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 34–43, 2000.

[20] T. Brijs, K. Vanhoof, G. Wets, Reducing redundancy in characteristic rule discovery by using integer programming techniques, Intelligent Data Analysis 4 (3, 4) (2000) 229–240.

[21] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hätönen, H. Mannila, Pruning and grouping discovered association rules, in: Proceedings of Workshop on Statistics, Machine Learning, and Discovery in Databases, 47–52, 1995.

[22] U.S. Federal Legislation, (a) Equal Credit Opportunity Act, (b) Fair Housing Act, (c) Intentional Employment Discrimination, (d) Equal Pay Act, (e) Pregnancy Discrimination Act., `http://www.usdoj.gov`, 2009.

[23] Speed Dating Dataset, `https://www.kaggle.com/annavictoria/speed-dating-experiment`, accessed: 2018-01-15, 2009.

[24] R. Fisman, S. S. Iyengar, E. Kamenica, I. Simonson, Gender differences in mate selection: Evidence from a speed dating experiment, The Quarterly Journal of Economics 121 (2) (2006) 673–697.

[25] UCI, Machine Learning Repository, `http://archive.ics.uci.edu/ml`, accessed: 2019-04-11, 2007.

[26] L. Gu, J. Li, H. He, G. Williams, S. Hawkins, C. Kelman, Association rule

discovery with unbalanced class distributions, in: Australasian Joint Conference on Artificial Intelligence, Springer, 221–232, 2003.

[27] A. Romei, S. Ruggieri, A multidisciplinary survey on discrimination analysis, The Knowledge Engineering Review 29 (5) (2014) 582–638.

[28] C. Clifton, Privacy preserving data mining: How do we mine data when we aren't allowed to see it, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[29] S. Hajian, J. Domingo-Ferrer, A methodology for direct and indirect discrimination prevention in data mining, IEEE Transactions on Knowledge and Data Engineering 25 (7) (2013) 1445–1459.

[30] A. A. Hintoglu, A. Inan, Y. Saygin, M. Keskinoz, Suppressing data sets to prevent discovery of association rules, in: Proceedings of IEEE International Conference on Data Mining, IEEE, 645–648, 2005.

[31] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (1) (2012) 1–33.

[32] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (5) (2002) 571–588.

[33] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, E. Dasseni, Association rule hiding, IEEE Transactions on Knowledge and Data Engineering 16 (4) (2004) 434–447.

[34] K. Wang, B. C. Fung, P. S. Yu, Template-based privacy preservation in classification problems, in: Proceedings of IEEE International Conference on Data Mining, IEEE, 466–473, 2005.

[35] B. T. Luong, S. Ruggieri, F. Turini, k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention, in: Proceedings of SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 502–510, 2011.

[36] L. Genga, L. Allodi, N. Zannone, Unveiling Systematic Biases in Decisional Processes. An Application to Discrimination Discovery, in: Proceedings of ACM ASIA Conference on Computer and Communications Security, ACM, 2019.

[37] C. C. Aggarwal, P. S. Yu, A new approach to online generation of association rules, IEEE Transactions on Knowledge and Data Engineering 13 (4) (2001) 527–540.

[38] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, L. Lakhal, Mining Minimal Non-redundant Association Rules Using Frequent Closed Itemsets, in: Computational Logic, vol. 1861 of *LNCS*, Springer, 972–986, 2000.

[39] M. Z. Ashrafi, D. Taniar, K. Smith, Redundant Association Rules Reduction Techniques, Int. J. Bus. Intell. Data Min. 2 (1) (2007) 29–63, ISSN 1743-8195.

[40] B. Liu, W. Hsu, Y. Ma, Pruning and summarizing the discovered associations, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 125–134, 1999.

[41] L. Cristofor, D. Simovici, Generating an informative cover for association rules, in: Proceedings of IEEE International Conference on Data Mining, IEEE, 597–600, 2002.

[42] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, Closed sets based discovery of small covers for association rules, in: Proceedings of International Conference on Advanced Databases, 361–381, 1999.