Conceptual Equivalence for Contrast Mining in Classification Learning

Ying Yang †, Xindong Wu ‡, Xingquan Zhu §
† Australian Taxation Office, Australia
ying.yang@ato.gov.au
‡ Department of Computer Science, University of Vermont, USA, xwu@cs.uvm.edu
§ Department of Computer Science & Engineering, Florida Atlantic University, USA, xqzhu@cse.fau.edu

Abstract

Learning often occurs through comparing. In classification learning, in order to compare data groups, most existing methods compare either raw instances or learned classification rules against each other. This paper takes a different approach, namely conceptual equivalence, that is, groups are equivalent if their underlying concepts are equivalent while their instance spaces do not necessarily overlap and their rule sets do not necessarily present the same appearance. A new methodology of comparing is proposed that learns a representation of each group’s underlying concept and respectively cross-exams one group’s instances by the other group’s concept representation. The innovation is five-fold. First, it is able to quantify the degree of conceptual equivalence between two groups. Second, it is able to retrace the source of discrepancy at two levels: an abstract level of underlying concepts and a specific level of instances. Third, it applies to numeric data as well as categorical data. Fourth, it circumvents direct comparisons between (possibly a
large number of) rules that demand substantial effort. Fifth, it reduces dependency on the accuracy of employed classification algorithms. Empirical evidence suggests that this new methodology is effective and yet simple to use in scenarios such as noise cleansing and concept-change learning.

**Keywords:** classification learning, contrasting data groups, knowledge discovery and representation, conceptual equivalence

1 Introduction

Understanding the discrepancy between contrasting data groups is fundamental to data analysis [1, 2]. Given two groups of interest, a user often needs to know the following. Do they represent different concepts? To what degree do they differ? What is the discrepancy and where does it originate from? In the context of this paper, a **concept** is an issue about which a user wants to find information. For example, the university administration is interested in discovering what affects student enrolment and the police are interested in what affects crime rate.

Solving these problems is important in reality. Examples are evident in many areas, including social science and computer science. For instance, health organizations often conduct comparative studies amongst different groups of women to discover which factors may affect the risk of getting breast cancer. The findings have contributed to encouraging many women to lead more beneficial life styles. Another example: a popular topic among computer scientists is to automate noise cleansing in data [3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. To verify the effectiveness of a cleansing method, one method is to compare a first-corrupted-then-cleansed data set with its clean version. The two sets of data form a pair of contrasting groups. It is useful to quantify how well the cleansed data set resembles the clean one and hence measure the quality of the cleansing method. A third example stems from the research and practice of data stream classification [13, 14, 15, 16, 17, 18], where the concept underlying the data may change over time and the classifier should be updated accordingly. A crucial task here is to judge whether the concept has changed across instances streaming through time space.
1.1 Related work

Various methods have been proposed to discover discrepancy between groups.

A popular method is contrast mining that mines contrast sets, conjunctions of attribute values that differ meaningfully across groups [1, 2]. It searches through data groups to find all contrast sets \((cset)\) that satisfy: \(\exists ij P(cset = True|G_i) \neq P(cset = True|G_j)\), and \(\max_{ij} |\text{support}(cset, G_i) - \text{support}(cset, G_j)| \geq \delta\), where \(\delta\) is a user-defined threshold called the minimum support difference. This paper will propose a new approach named conceptual equivalence mining. There are two differences between contrast mining and conceptual equivalence mining. First, contrast mining fits within association mining where the user wants to learn correlations between arbitrary attributes. Conceptual equivalence mining, however, focuses on classification learning. There are many real-world cases where the user has a particular interest (class) in mind. For instance, medical workers want to discover whether the causal factors of diabetes are significantly different across ethnic groups. In this case the class is ‘diabetes’ or ‘not diabetes’. Contrast mining will return many items regardless of the class and hence uninteresting to medical workers, such as the association between age and weight attributes. Conceptual equivalence mining will be more effective by targeting the discrepancy associated with class only. Second, conceptual equivalence mining can improve privacy preservation. Contrast mining requires two groups to share their raw data in which sensitive personal information often lies. Conceptual equivalence mining needs only two groups to exchange their learned classifiers, apply the counterpart’s classifiers to their own data, and discover their differences.

A method closely related to conceptual equivalence mining is correspondence tracing [19]. Correspondence tracing discovers changes of classification characteristics as data change. Given an old classifier that represents previous knowledge (in terms of classification characteristics) about the old data, and the new data, this method traces the corresponding new rules for each old rule through the instances that they both classify and use the new rules to describe the changes of the old rule. To present changes, it ranks each pair of corresponding old-new rules
according to the improvement to classification accuracy. Conceptual equivalence mining and correspondence tracing are different but complement each other. The differences are two-fold. First, conceptual equivalence mining discovers ‘global’ discrepancy. Its quantitative measures and rankings relate to whole data groups. Correspondence tracing discovers ‘local’ discrepancy. Its quantitative measures and rankings relate to each pair of corresponding old-new rules. Second, conceptual equivalence mining is interested in discrepancies instead of classification accuracy. It actually aims at reducing the dependency on the accuracy of employed classification algorithms. Correspondence tracing takes the improvement to classification accuracy as one primary goal. It ranks all changes according to the accuracy improvement. A change is important to the extent that recognizing it can improve the classification accuracy. Nonetheless, the global information delivered by conceptual equivalence mining and the local information delivered by correspondence tracing are complementary. For instance, conceptual equivalence mining can report in general that the rule $R$ of the data group $G_1$ represents the biggest discrepancy between $G_1$ and $G_2$. Correspondence tracing can report in detail which rules in $G_2$ correspond to $R$ (either agreeing or contradictory).

A recent approach uses contingency tables to calculate the similarity of the two rules [20]. For syntactic similarity, values in the table correspond to the number of attribute-value pairs that match or do not match between the two rules. For semantic similarity, values correspond to the number of instances that match or do not match within the instance sets supported by each rule. A statistical measurement, such as $\chi^2$, is applied to this table to calculate the similarity. This approach may be less applicable if the compared instance sets seldom overlap. Nor does it handle numeric values well. This paper will propose a new method to calculate the conceptual equivalence between two groups. The groups are equivalent if their underlying concepts are equivalent while their instance spaces do not necessarily overlap and their rule sets do not need to syntactically match each other. Moreover, the approach applies to numeric data as well as categorical data.

An interesting technique uses fuzzy set theory to decide the similarity between two rule sets $R_1$ and $R_2$ [21]. Each rule in $R_1$ is converted to a fuzzy rule which has
the same syntax as the original rule but has fuzzy linguistic variables as attribute values. A fuzzy matching system then matches each (non-fuzzy) rule in $R_2$ against each fuzzy rule in $R_1$ to obtain a degree of similarity. The higher the degree, the higher the similarity. Substantial domain knowledge is required to define fuzzy linguistic variables and the matching work still focuses on directly comparing rules. In contrast, this paper can offer an effective and yet simple scoring scheme to measure the similarity between two rule sets. What is more, the scheme applies not only to classification rules, but also to other families of classifiers such as Bayesian networks that are commonly used in real-world applications.

Another useful method identifies ‘fundamental rule changes’ that cannot be explained by changes in other rules [22]. By this means, superficial changes can be discarded and the major shift in data can be captured. However, capturing rule changes does not offer a quantitative measure of the degree of discrepancy between contrasting data groups. Thus this paper will propose new approaches that can quantify the degree of discrepancy as well as discover concept changes between data groups.

1.2 Open challenges

With due respect to existing achievements, this paper suggests that some problems remain.

- The relationship is one-to-many between the concept a user is interested in and a classification rule set learned from a data group. A rule set is a representation of a concept. A concept can have different representations. Hence syntactic dissimilarities between two rule sets do not necessarily indicate discrepancies between their represented concepts. It is sometimes advisable to circumvent direct comparisons between rule sets when contrasting groups. But how?

- There is a lack of measures quantifying the degree of discrepancy. For example, there are two rules (syntactically) different between $G_0$ and $G_1$ while there are three rules different between $G_0$ and $G_2$. Which ($G_1$ or $G_2$) better
resembles $G_0$ then? The difficulty is that the two sets of different rules may not be commensurable. Existing methods have not offered clear solutions.

- Feasibility is a problem. A thorough syntactic or semantic comparison of rules requires substantial efforts. The feasibility of comparison can be sub-optimal if the concept of the data becomes complicated, the attribute values become omnifarious, or the domain knowledge becomes less sufficient. This potential is particularly unwelcome in real-world applications where the data are always highly diversified and the domain knowledge is often in short supply.

- Popular learning algorithms like Bayesian probabilistic classifiers have no explicit rules. Rule comparison is thus inapplicable.

### 1.3 Main contributions of this paper

In light of these challenges, this paper proposes a new methodology in the context of classification learning that carries out contrast mining by measuring the degree of conceptual equivalence (CE) between groups. It first learns a representation of each group’s concept. It then cross-exams one group’s data against the other group’s concept representation, resulting in support, conflict or no-match for each instance. Consulting the evidence, it (1) quantifies the degree of conceptual equivalence; and (2) retraces the data source that has produced the discrepancy.

By no means does this paper devalue the importance of existing achievements as discussed in Section 1.1. Rather, it offers a different perspective and proposes a novel methodology, which contributes to completing the picture of handling discrepancy between contrasting data groups.

The authors’ preliminary thoughts on quantifying CE have successfully contributed to noise cleansing [12] and concept-change learning [17, 18]. The methodology proposed here is a new version that is systematically more complete, theoretically more correct and empirically more accurate. This paper is the first comprehensive introduction to CE’s lines of reasoning, approaches, merits and functions.
The remainder of the paper is organized as follows. Section 2 defines terms used throughout this paper and differentiates conceptual equivalence from literal equivalence. Section 3 proposes a scoring mechanism that is able to quantify the degree of conceptual equivalence. Section 4 proposes a ranking mechanism that, informed by the scoring mechanism, retraces the source of discrepancy in terms of specific instances and abstract concepts. Section 5 conducts experiments to verify the effectiveness of proposed mechanisms. Section 6 gives a conclusion and refers to some interesting directions for future research.

2 Background knowledge

2.1 Terminology

This paper is set in the context of classification learning. A data group is composed of instances. Each instance is a vector of attribute values and has a class label.

This paper uses a classification rule set as a means to represent a concept. A classification rule is composed of an antecedent to indicate attribute values, in which ‘&’ means the logical ‘and’, ‘||’ means the logical ‘or’, and ‘⇒’ means ‘implication’; and a consequence to indicate a class label. The classification accuracy of each rule on its training data is indicated in a pair of brackets at the end of the rule.

Nonetheless, the new quantitative measure only utilizes the verdict of a classifier of instances regardless of the classification algorithm. Hence it is applicable beyond rules to variations like Bayesian probabilistic classifiers. From a practical point of view, it is often advisable to choose the classifier that achieves a good prediction accuracy on the data of interest and use it to represent the underlying concept.

2.2 Literal vs. conceptual equivalence

A counterpart of conceptual equivalence is literal equivalence. The two types of equivalence have different degrees of granularity and require different methodolo-
gies to measure. Literal equivalence focuses on individual instances, such as a
doctor caring about every single patient (an instance). In contrast, conceptual
equivalence focuses on general concepts, such as a department store manager be-
ning more interested in customers’ shopping trends (a concept) than keeping track
of single customers. Calculating the similarity between instances has long been
a practice in research domains like case-based reasoning [23, 24, 25, 26, 27]. One
commonly used method is to calculate the weighted Euclidean distance between
two instances, where the weights may be equal, provided by the domain knowledge,
or even optimized by genetic algorithms. By contrast, much less effort has been
paid to concepts. This problem is further compounded by the fact that in many
real-world applications, concepts are particularly interesting to users because they
provide instructive concise knowledge. Accordingly, this paper targets conceptual
equivalence.

3 Quantifying conceptual equivalence

A scoring mechanism that is able to quantify the degree of conceptual equiva-
ience between two contrasting groups is proposed in this section. The sub-optimal
aspects of possible alternative mechanisms are also analyzed.

3.1 The proposed scoring mechanism

A new scoring mechanism is proposed that, given two contrasting groups $G_1$ and
$G_2$, is able to quantify: (1) the degree that $G_1$ conceptually resembles $G_2$; (2)
the degree that $G_2$ conceptually resembles $G_1$; and (3) the degree of the concep-
tual equivalence between $G_1$ and $G_2$. This mechanism is effective and yet simple
to apply. It circumvents direct comparisons between rule sets. It also reduces
dependency on the learner’s accuracy.

As illustrated in Figure 1, a learner is applied to $G_1$ and $G_2$, learning concept
representatives $RS_1$ and $RS_2$ respectively. Each instance $I$ in $G_1$ or $G_2$ is classified
by both $RS_1$ and $RS_2$. The classification has three possible outcomes: (1) approve,
Figure 1: A scoring mechanism quantifies conceptual equivalence between $G_1$ and $G_2$.

The rule set classifies the instance into its true class; (2) *contradict$_x$*, the rule set classifies the instance into a class ($x$) that differs from its true class; and (3) *no-match*, the rule set does not cover this instance and no classification is given. Observations of these outcomes indicate $I$’s different status of representing conceptual equivalence between $G_1$ and $G_2$. Typically there can be three scenarios.

1. $RS_1$ and $RS_2$ agree on classifying $I$. Hence $I$ is a contribution to the conceptual equivalence between $G_1$ and $G_2$, and is allocated a score of +1. The classification results can be:
   - *approve* vs. *approve*;
   - *contradict$_x$* vs. *contradict$_x$*: both $RS_1$ and $RS_2$ classify $I$ into the same class $x$ that differs from $I$’s true class.

2. $RS_1$ and $RS_2$ disagree on classifying $I$. Hence $I$ has a negative impact on the conceptual equivalence between $G_1$ and $G_2$, and is allocated a score of -1. The classification results can be:
   - *approve* vs. *contradict*;
   - *contradict$_x$* vs. *contradict$_y$*: both $RS_1$ and $RS_2$ contradict $I$, but one classifies $I$ into the class $x$ while the other gives a different class $y$;

---

1For probabilistic classifiers that output probability distributions over classes, no-match means that the classifier returns equal probabilities for multiple classes.
• no-match vs. [approve or contradict]: one rule set does not cover I. The
other covers I and predicts its class. This indicates that information
about I is lacking in one group while it is statistically sufficient in the
other group to make a prediction. Hence I represents a discrepancy.

3. RS$_1$ and RS$_2$ provide no evidence to make a decision, when their classifi-
cation results are no-match vs. no-match. Various reasons and solutions
exist.

• The current learner is not good at learning I’s concept. A solution is
to (repeatedly) apply a different learning heuristic until no no-match-
no-match instances remain.

• The local pattern in no-match-no-match instances is overwhelmed by
the global pattern. A solution is to recursively learn in no-match-no-
match instances.

• The concept is difficult to learn and no-match-no-match instances still
exist after a reasonable number of learning trials. A solution is to form
a rule for each instance. If there are identical instances within the
same class, delete redundant rules but retain corresponding statistics.
If there are identical instances within different classes, form a rule for
each class and decreasingly sort the rules by their coverage to resolve
the conflict$^2$. This solution has an effect that changes comparisons from
the conceptual level to the literal level.

Figure 1 includes the procedure of the third case. The other two cases sim-
ply execute loops or recursive calls of the whole procedure. RS$_1$-Nomatch
and RS$_2$-Nomatch are the rule sets formed by matching each no-match-no-match
instance and resolving possible conflicts. If their verdicts of I still involve no-match,
it indicates that first, no pattern of I can be inferred from either $G_1$ or $G_2$; or
second, I literally appears in only one group. Hence, there is no way of judging
I’s effect on the conceptual equivalence between $G_1$ and $G_2$, and the score is 0.

$^2$If their coverages are equal, both classes are allowed.
If the result does not involve a no-match, a score of either +1 or -1 is allocated following the above-detailed scoring mechanism.

As a result, this scoring mechanism gives a score of +1, 0 or -1 to each instance in $G_1$ or $G_2$. Suppose the number of instances in $G_1$ and $G_2$ are $n_1$ and $n_2$ respectively. The following formulae are employed to calculate $score_1$ (the degree of $G_1$ conceptually resembling $G_2$), $score_2$ (the degree of $G_2$ conceptually resembling $G_1$), and $score_3$ (the degree of conceptual equivalence between $G_1$ and $G_2$).

\[
score_1 = \frac{\sum_{i=1}^{n_1} \text{score of } I_{n_1i}}{n_1}; \quad (1)
\]

\[
score_2 = \frac{\sum_{i=1}^{n_2} \text{score of } I_{n_2i}}{n_2}; \quad (2)
\]

\[
score_3 = score_1 \times \frac{n_1}{\sum_{i=1}^{n} n_i} + score_2 \times \frac{n_2}{\sum_{i=1}^{n} n_i}. \quad (3)
\]

This mechanism has several properties. First, the degree of the conceptual equivalence is quantified as a real number within [-1, +1]. The more two groups conceptually resemble each other, the greater the value. Second, $score_1$ is not necessarily equal to $score_2$. This indicates that the similarity from one group to the other does not have to be symmetric [26]. Third, it is applicable to both rule and non-rule learners. This is desirable because non-rule learners such as Bayesian classifiers are common in real-world applications. In the context of 0-1 loss classification learning, one can directly couple Bayesian classifiers with the proposed scoring mechanism.

Another result of this scoring mechanism is forming a set of discrepancy instances, each scored as -1. This set can help further retrace the source of discrepancy, which is to be addressed in Section 4.

3.2 Alternative scoring mechanisms

There exist alternative mechanisms to quantify conceptual equivalence. However, they can be less practical or less meticulous compared with the above proposed method.
3.2.1 Using rule comparisons

It is intuitive to measure conceptual equivalence by comparing the rule sets learned from each group because rules are summarized knowledge. As mentioned in Section 1, one sub-optimal scenario occurs when multiple valid representations of a concept exist. For example, one concept of the often-cited Monk’s Problem from the UCI machine learning repository [28] is $(A5=3 \& A4=1) \lor (A5\neq4 \& A2\neq3) \Rightarrow \text{Class}=1$; otherwise $\Rightarrow \text{Class}=0$. As in Table 1, although syntactically different, two rule sets (one from C4.5rules [29] and the other from PART [30]) can equally represent this concept. Another sub-optimal scenario occurs when numeric attributes are involved. For example, one rule is $A1<0.41 \& A3<0.39 \Rightarrow \text{Class}=+$ and the other is $A1<0.33 \& A3<0.47 \Rightarrow \text{Class}=+$. It is difficult to judge how similar they are without sufficient background knowledge to tell the meanings and scales of the numeric values. The complexity that numeric data bring into rule comparison will be further demonstrated in Section 5.3.

<table>
<thead>
<tr>
<th>A rule set learned by C4.5rules</th>
<th>A4=1 &amp; A5=3 ⇒ Class=1 [100.0%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A5=4 ⇒ Class=0 [100.0%]</td>
<td>Default: ⇒ Class=1</td>
</tr>
<tr>
<td>A2=3 &amp; A5=1 ⇒ Class=0 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A2=3 &amp; A5=2 ⇒ Class=0 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A2=3 &amp; A4=2 ⇒ Class=0 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A2=3 &amp; A4=3 ⇒ Class=0 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A2=1 &amp; A5=1 ⇒ Class=1 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A2=1 &amp; A5=2 ⇒ Class=1 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A2=1 &amp; A5=3 ⇒ Class=1 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A2=2 &amp; A5=1 ⇒ Class=1 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A2=2 &amp; A5=2 ⇒ Class=1 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A2=2 &amp; A5=3 ⇒ Class=1 [100.0%]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A rule set learned by PART</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A2=3 &amp; A5=1 ⇒ Class=0 [100.0%]</td>
<td>A4=1 &amp; A5=3 ⇒ Class=1 [100.0%]</td>
</tr>
<tr>
<td>A5=4 ⇒ Class=0 [100.0%]</td>
<td>Default: ⇒ Class=1</td>
</tr>
<tr>
<td>A2=1 ⇒ Class=1 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A2=2 ⇒ Class=1 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A4=2 ⇒ Class=0 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A4=3 ⇒ Class=0 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>A5=2 ⇒ Class=0 [100.0%]</td>
<td></td>
</tr>
<tr>
<td>Otherwise ⇒ Class=1 [100.0%]</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: A single concept may have different representations.
3.2.2 Using classification accuracies

A naive method that measures conceptual equivalence without directly comparing rules was proposed by the authors in previous work [12]. Figure 2 illustrates the process. Suppose there are two contrasting groups $G_1$ and $G_2$. First, a rule set $RS_1$ is learned from $G_1$ and is used to classify data in $G_2$, obtaining classification accuracy $acc_2$. Second, a rule set $RS_2$ is learned from $G_2$ and is used to classify data in $G_1$, obtaining classification accuracy $acc_1$. Assume the numbers of instances in $G_1$ and $G_2$ are $n_1$ and $n_2$ respectively. The weighed mean of $acc_1$ and $acc_2$ is used to indicate the degree of conceptual equivalence between $G_1$ and $G_2$:

\[
score = \frac{\sum_{i=1}^{2}(acc_i \times n_i)}{\sum_{i=1}^{2} n_i}.
\]

Although straightforward, this method is not meticulous since it depends heavily on the learner's classification accuracy. For example, $G_1$ and $G_2$ are identical. A learner's best accuracy on them is 85%. This method will result in 0.85 out of 1.00 as the degree of the conceptual equivalence, which is incorrect. Another example: a learned rule is ‘condition A $\Rightarrow$ class B’ with 90% accuracy. An instance ‘condition A, class C’ appears in both $G_1$ and $G_2$. This is not a discrepancy although it is contradicted by the learned rule, which is not perfect anyway. Nonetheless, this method will use this instance to decrease the equivalence, which is not correct.

4 Retracing discrepancy source

It is useful to retrace the discrepancy source. After being informed that there is a certain degree of discrepancy between two contrasting groups, a user often asks
questions like: what is the discrepancy and where does it originate from? This paper proposes to present the discrepancy source at two levels. A lower level lies within specific instances. A higher level relates to concepts.

4.1 Ranking discrepancy instances

Although every discrepancy instance is scored as -1 by the scoring mechanism in Section 3.1, it may reflect a different degree of discrepancy. Imagine that an instance $I_1$ is approved by $G_1$ but gets a ‘no-match’ verdict in $G_2$; and another instance $I_2$ is approved by $G_1$ but is contradicted by $G_2$. It is reasonable to assume that $I_2$ represents a bigger discrepancy. A ranking system is proposed to rank instances according to their represented degrees of conceptual discrepancy.

Recall that in the scoring procedure (Section 3.1) each discrepancy instance can be associated with a pair of rules, one learned from its own data group and the other from its contrasting data group. The former is represented as $R_{self}$ and the latter as $R_{contrast}$. Each rule itself is associated with some statistics such as coverage and accuracy. If the instance gets a ‘no-match’ verdict in a group, it is deemed to be associated with a ‘null’ rule whose statistics are all 0. In order to rank discrepancy instances, a measurement named $advantage^3$ is calculated. Given a rule $R$, its statistic $correct$ equals the number of instances that it approves, and its statistic $wrong$ equals the number of instances that it contradicts. The $advantage$ of $R$ with regard to an instance $I$ is defined as follows. If it approves $I$, $R$ has a positive$^4$ strength associated with $I$. If it contradicts $I$, $R$ has a negative strength. If it does not cover $I$, $R$ has no strength.

$$advantage(R, I) = \begin{cases} correct - wrong & \text{(if approve)} \\ -(correct - wrong) & \text{(if contradict)} \\ 0 & \text{(if no-match)} \end{cases}$$

The rank of each discrepancy instance is calculated as follows, where the function ‘abs’ obtains the absolute value. By this means, each discrepancy instance

---

$^3$It is different from C4.5rules’ statistic $advantage$.

$^4$The $correct$ is always larger than the $wrong$. Otherwise this rule will not be learned.
obtains a rank. The more powerful the two disagreeing rules in their own groups, the higher the instance is ranked.

\[
rank(I) = \text{abs}(\text{advantage}(R_{\text{self}}, I)) + \text{abs}(\text{advantage}(R_{\text{contrast}}, I)).
\]

However, the discrepancy can be difficult to understand if there is a large number of discovered instances. It will help if the discrepancy information can be abstracted into concepts, which is the topic of the following section.

### 4.2 Abstracting discrepancy concepts

Unlike discrepancy instances that contain individual information, discrepancy concepts represent a higher level of information. They are more abstract for an easier understanding and more instructive for further action. An approach is proposed in Table 2 that needs only a one-pass linear scan to abstract one data group’s rules (for instance \( G_1 \)) that differ from another group’s (for instance \( G_2 \)).

For each rule \( R \) learned from \( G_1 \), the algorithm counts the number of instances in \( G_1 \) that are classified by \( R \) but classified differently by \( G_2 \)’s rule set. As a result, each rule is associated with a value indicating the number of discrepancy instances contributed by this rule. Since the information of score and result has already been collected in the scoring procedure as in Section 3.1, the process of abstracting is straightforward and fast. The final stage of this process is to rank each rule by this value. The higher the rank, the greater discrepancy this rule stands for between contrasting groups.

### 5 Experiments

Experiments are conducted to verify three hypotheses. First, the scoring mechanism proposed in Section 3.1 can properly reflect the degree of conceptual equivalence between groups. Second, the ranking mechanism proposed in Section 4 can retrace the source of discrepancy between groups. Third, the whole system can tackle scenarios where previous methods are less capable. The employed learner is C4.5rules [29] since it is one of the most commonly used in practice [21].
input: $G_1$, $RS_1$ // $RS_1$ is the rule set learned from $G_1$.
output: $DRS_1$ // $DRS_1$ is a sorted discrepancy rule set of $G_1$.

foreach ruleIndex (1 .. $|RS_1|$)
    array_count[ruleIndex] = 0;

for each instance $I \in G_1$
    if (score($I$) $\neq$ -1) goes to the next instance without processing anything;
    result = classify $I$ by $RS_1$;
    ruleIndex = index of the rule that classifies $I$;
    if (result $\neq$ no-match) array_count[ruleIndex]++;

$DRS_1$ = decreasingly sort $RS_1$ according to array_count;

Table 2: Abstracting discrepancy concepts

5.1 Testing the scoring mechanism

In the research area of noise cleansing, the clean data and the corrupted data form a pair of contrasting groups. Empirically, the higher the amount of noise, the greater the discrepancy between the two. If it can produce scores consistent with this trend, the proposed scoring mechanism is useful.

5.1.1 Design

The test is designed as follows. A data set $G$ is divided into two exclusive groups $G_1$ and $G_2$ of the same size and the same class proportion so that they support the same concept. Keep $G_1$ untouched. Randomly corrupt $G_2$ into various noisy data sets. The degree of corruption is increasing in two dimensions $<X,Y>$, where $X \in [0\%,100\%]$ and $Y \in [0, \min(\text{threshold}, \text{total attribute number})]$. The $X$ dimension controls the percentage of instances to be corrupted. The $Y$ dimension controls the number of attributes, each corrupted into a value that is different to that of its original. The parameter $\text{threshold}$ is used for the user’s flexibility. The scoring mechanism calculates the conceptual equivalence between $G_1$ and each of $G_2$’s noisy versions.
5.1.2 Data

The experimental data are benchmark data sets from the UCI machine learning repository [28], whose details are as follows.

The ‘Car’ data set is derived from a hierarchical decision model that evaluates cars according to six input attributes: price of buying, price of maintenance, number of doors, capacity in terms of persons to carry, size of boot, and estimated safety. There are 1728 instances, each with one of four classes: unacceptable, acceptable, good and very good.

The ‘KR-vs-KP’ data set describes the chess King+Rook versus King+Pawn endgame. There are 3196 instances. Each instance is a board description for a chess endgame using 36 attributes. There are two classes: white-can-win and white-cannot-win.

The ‘Monks’ problem is the basis of a first international comparison of learning algorithms. It makes several target concepts and accordingly produces data sets of two classes, six attributes and 473 instances to test a wide range of induction algorithms. For instance, one target concept is: \((A_1 = A_2)\lor (A_5 = 1)\).

The ‘Mushroom’ data set has 8124 mushroom instances drawn from “The Audubon Society Field Guide to North American Mushrooms”. Each instance is described in terms of 22 (physical) attributes, and has a class as either ‘edible’ or ‘poisonous’. Logical rules have been provided that seem to be the simplest possible for the data and therefore can be treated as benchmark results [31, 32].

The ‘Nursery’ data set is derived from a hierarchical decision model originally developed to rank applications for nursery schools. The nursery rank relates to eight attributes: parents’ occupation, child’s nursery, form of the family, number of children, housing conditions, financial standing of the family, social conditions and health conditions. There are 12,960 instances.

The ‘Splice’ data set is composed of 3190 instances from Genbank 64.1. Given a position in the middle of a window of 60 DNA sequence elements (attributes), the task is to decide if this is an intron→exon boundary (IE) or an exon→intron boundary (EI) or neither. Introns are the parts of the DNA sequence that are
spliced out and exons are the parts of the DNA sequence retained after splicing. In the biological community, IE borders are referred to as “acceptors” while EI borders are referred to as “donors”.

The ‘LED’ data set predicts the decimal digit according to the light-emitting diodes in LED displays. It has seven attributes (diodes), 10 classes (digits 0-9) and 500 instances.

The ‘Vote’ data set includes votes for each of the US House of Representatives Congressmen on 16 key votes. The task is to predict whether a congressman is a democrat or republican according to his/her votes. Hence there are 16 attributes, 2 classes and 435 instances.

The ‘Tic-tac-toe data set encodes the complete set of possible board configurations at the end of tic-tac-toe games. It has 9 attributes (each corresponding to one tic-tac-toe square), 2 classes (win for x or otherwise) and 958 instances.

5.1.3 Result analysis

Figure 3 illustrates some representative results. The ‘CE’ dimension is the score of conceptual equivalence. The ‘Instances’ and ‘Atts’ dimensions are respectively the percentages of instances and attributes to be randomly chosen and corrupted.

The resulting score starts with 1 when no noise exists and almost monotonously decreases while the corruption becomes more severe. On rare occasions, the score does not decrease when the corruption degree becomes higher. A probe into the corrupted data suggests that because the corruption is random, an instance is occasionally changed from one valid instance to another valid instance. Hence, it does not contribute to decreasing the conceptual equivalence. These results favor the proposed scoring mechanism and support the hypothesis that it can properly reflect the degree of conceptual equivalence between data groups.
Figure 3: The score of conceptual equivalence reflects the degree of corruption in data.
5.2 Testing the ranking system

This section of experiments comprises two parts. In Section 5.2.1, the concept changes are manually generated. Thus one can know exactly what concept has changed to what, and can verify how the proposed ranking system performs. In Section 5.2.2, the concept changes are inherent to the data. In this way, one can systematically produce a large suite of contrasting data groups and conduct statistical evaluations on the proposed system.

5.2.1 Case study

This section of experiments changes the underlying concept of a data set, reproduces instances accordingly and tests whether the proposed ranking system can retrace this change. Two UCI benchmark data sets with documented underlying concepts are used, the synthetic ‘Monk-1’ and the natural ‘Mushroom’. Complete details (such as logical rules) describing Monk-1 and Mushroom can be found on the UCI machine learning repository website [28].

The test is designed as follows. Each data group $G$ is divided into two exclusive groups $G_1$ and $G_2$ of the same size. To challenge the new methodology, $G_1$ and $G_2$’s class distributions are not necessarily identical. Keep $G_1$ untouched. Randomly change the concept and apply it to $G_2$, revising its instances’ classes accordingly. The unchanged $G_1$ and the changed $G_2$ are put into the ranking system. Some example results are given below.

Cases 1 to 3 test the ranking scheme that ranks discrepancy concepts. In the report, each line describes a classification rule with its classification accuracy on the data group. The rank of each rule is marked in front of the rule. The greater discrepancy a rule represents between contrasting groups, the higher its rank.

Case 4 tests the ranking scheme that ranks discrepancy instances. In the report, each line describes an instance from the data group. The rank of each instance is marked in front of the instance. The more powerful an instance’s $R_{self}$ and $R_{contrast}$ in their own groups, the higher the instance is ranked.
Test Case 1  For Monk-1’s $G_2$, an underlying concept $A_1 = A_2 || A_5 = 1 \Rightarrow Class = 1$ is changed to $A_1 = A_2 || (A_5 = 1 || 2) \Rightarrow Class = 1$. The ranking system returns the following.

(1) $G_2$’s concept that contrasts with $G_1$:
Rank=45: $A_5=2 \Rightarrow Class=1$ [100.0%]

(2) $G_1$’s concept that contrasts with $G_2$:
Rank=7: $A_1=1 \& A_2=3 \& A_5=2 \Rightarrow Class=0$ [100.0%]
Rank=6: $A_1=1 \& A_2=2 \& A_5=2 \Rightarrow Class=0$ [100.0%]
Rank=5: $A_1=2 \& A_2=3 \& A_5=2 \Rightarrow Class=0$ [100.0%]
Rank=5: $A_1=3 \& A_2=1 \& A_5=2 \Rightarrow Class=0$ [100.0%]
Rank=4: $A_1=2 \& A_2=1 \& A_5=2 \Rightarrow Class=0$ [100.0%]
Rank=2: $A_1=3 \& A_4=1 \& A_5=2 \& A_6=2 \Rightarrow Class=0$ [100.0%]
Rank=1: $A_2=2 \& A_4=1 \& A_5=2 \& A_6=1 \Rightarrow Class=0$ [100.0%]

It is observed that the system accurately captures the concept discrepancy between $G_1$ and $G_2$: $A_5=2$ implies $Class=1$ in $G_2$ while $A_5=2$ always relates to $Class=0$ in $G_1$.

Test Case 2  For Mushroom’s $G_2$, an underlying concept $Spore-print-color = green \Rightarrow Class = poisonous$ is changed to $Spore-print-color = green \Rightarrow Class = edible$. The ranking system returns the following.

(1) $G_1$’s concept that contrasts with $G_2$:
Rank=38: Spore-print-color=green $\Rightarrow$ Class=poisonous [100.0%]

(2) $G_2$’s concept that contrasts with $G_1$:
Rank=34: Odor=none & Gill-size=broad $\Rightarrow$ Class=edible [100.0%]

The system correctly reports that $Spore-print-color = green$ relates to a conceptual discrepancy between $G_1$ and $G_2$. $Spore-print-color = green$ implies $Class = poisonous$ in $G_1$ but that is not the case in $G_2$. Meanwhile, the reason that $G_2$’s report is not directly represented as ‘Odor=none & Gill-size=broad $\Rightarrow$ Class=edible’ is because it becomes a subset of ‘Odor=none & Gill-size=broad $\Rightarrow$
Class=edible' and hence is absorbed. Nonetheless, the discrepancy is explicitly reflected in $G_1$’s report.

**Test Case 3** For Mushroom’s $G_2$, an underlying concept $\text{Odor} = \text{NOT}(\text{almond} \mid \text{anise} \mid \text{none}) \Rightarrow \text{Class} = \text{poisonous}$ is changed to $\text{Odor} = \text{NOT}(\text{almond} \mid \text{none}) \Rightarrow \text{Class} = \text{poisonous}$. The ranking system returns the following.

(1) $G_2$’s concept that contrasts with $G_1$:
   Rank=181: Odor=anise ⇒ Class=poisonous [100.0%]

(2) $G_1$’s concept that contrasts with $G_2$:
   Rank=195: Odor=anise ⇒ Class=edible [85.7%]

The system precisely discovers the discrepancy between $G_1$ and $G_2$: Odor=anise implies Class=poisonous in $G_2$ while it implies Class=edible in $G_1$.

**Test Case 4** Meanwhile, the ranks of instances of discrepancy can be produced. When the two conceptual changes in test cases 2 and 3 both take place in Mushroom’s $G_2$, example results are as follows. Mushroom has 22 attributes and a class. The UCI machine learning repository website uses abbreviations for each attribute value. Only the names of the two involved attributes are recovered. For more information, please refer to the UCI machine learning repository website. For instance, the ranking system returns the following.

**Instances in $G_1$ that contrast with $G_2$**

(1) Rank=1679: b,s,w,t,Odor=none,f,c,b,g,e,b,s,s,w,p,p,t,p,
    Spore-print-color=green,v,g,Class=poisonous.

(2) Rank=400: x,y,y,t,Odor=anise,f,c,b,w,e,r,s,y,w,p,w,o,p,
    Spore-print-color=black,y,g,Class=edible.

The first instance of $G_1$ represents a greater discrepancy and has a higher rank. The explanation is as follows. Mushroom’s rules for the poisonous class are *disjunctive*, that is, the mushroom is poisonous as long as one rule is satisfied. Note that this is in contrast to each single rule whose left hand side (the antecedent) comprises *conjunctive* conditions to determine the class label (the consequence).
Hence although Instance(1)’s odor is ‘none’, the mushroom is still poisonous because it satisfies the rule ‘Spore-print-color=green ⇒ poisonous’. However, in $G_2$, after the rule ‘Spore-print-color=green ⇒ poisonous’ is changed to ‘Spore-print-color=green ⇒ edible’, all instances with ‘odor=none’ become edible, and the population is very large. Hence this instance represents a major discrepancy and is correctly ranked higher than the instance listed below.

5.2.2 General results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Annealing</td>
<td>898</td>
<td>38</td>
<td>5</td>
<td>Hypothyroid</td>
<td>3772</td>
<td>29</td>
<td>2</td>
</tr>
<tr>
<td>Audiology</td>
<td>226</td>
<td>69</td>
<td>24</td>
<td>KR-vs-KP</td>
<td>3196</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>Automobile</td>
<td>205</td>
<td>25</td>
<td>6</td>
<td>LaborNegotiation</td>
<td>57</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>Balloons</td>
<td>16</td>
<td>5</td>
<td>2</td>
<td>LED</td>
<td>1000</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Contraceptive</td>
<td>1473</td>
<td>9</td>
<td>3</td>
<td>Monk’sProblem</td>
<td>432</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Echocardiogram</td>
<td>131</td>
<td>6</td>
<td>2</td>
<td>Mushroom</td>
<td>8124</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>GermanCredit</td>
<td>1000</td>
<td>20</td>
<td>2</td>
<td>SolarFlare</td>
<td>1389</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>HeartDisease</td>
<td>270</td>
<td>13</td>
<td>2</td>
<td>Soybean</td>
<td>683</td>
<td>35</td>
<td>19</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>155</td>
<td>19</td>
<td>2</td>
<td>Yeast</td>
<td>1484</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>HorseColic</td>
<td>368</td>
<td>21</td>
<td>2</td>
<td>Zoo</td>
<td>101</td>
<td>17</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 3: Each experimental data set’s name, number of instances (Ins.), number of attributes (Att.) and number of classes (Cls.)

In this section of experiments, concept changes are not manipulated manually but are inherent to the data. Given a data set and one of its binary-valued nominal attributes, instances are partitioned into two groups $G_1$ and $G_2$ according to this attribute. $G_1$ comprises instances that bear the second value of this attribute and $G_2$ comprises instances that bear the first value of this attribute. A large suite of 20 commonly-used UCI benchmark data sets [28] are employed, whose statistical details are listed in Table 3. Evaluations are given with statistical evidence on the proposed system’s performance of discrepancy discovery.
Experimental results are observed and analyzed as in Table 4. The first column is the name of the data set from which $G_1$ and $G_2$ are produced. The second column is one of the top-ranked discrepancy rules returned by the proposed ranking mechanism. The third column is an evaluation whether this discovery is indeed a major discrepancy between $G_1$ and $G_2$. Because concept changes here are inherent instead of manipulated, one needs to manually verify what is going on in $G_1$ and $G_2$ in order to evaluate the results. Finally explanations and analysis are given to each result.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Reported Discrepancy</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annealing</td>
<td>$G_1$: Hardness&gt;60 ⇒ Class=U [92.2%]</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Explanation: This data set contains 898 instances, each of which is described by attributes like family, steel, shape and hardness. Each instance belongs to one of six alternative classes. The partition attribute is Shape. $G_1$ has instances of shape as sheet while $G_2$ as coil. Among sheets, if hardness is bigger than 60, the class is always U (17 out of 17 instances). However, among coils, the class is U when hardness is bigger than 80 whereas the class is 3 when hardness is less than 80. Hence the hardness range of [60,80] produces a discrepancy between $G_1$ and $G_2$. This discrepancy has been correctly reported by the system.

| Audiology | $G_1$: History-nausea=false & History-noise=false & Tymp=a ⇒ Class=cochlear-age [93.2%] | Yes |

Note that although this rule’s classification accuracy is 100% on its training data ($G_1$), C4.5 reports instead its estimated true accuracy, which is 92.2% here.
**Explanation:** This data set contains persons each of whom is described by attributes like age, air bone gap and history of buzzing. The task is to diagnose each person’s hearing condition. The partition attribute is whether one’s age is greater than 60. $G_1$ contains persons more than 60 years old and $G_2$ contains the others. Among people who are older than 60, when their tymp is of type ‘a’ and when they do not have a history of noise, the decisive factor is whether a person has a history of nausea. If not, they are very likely to be cochlear-age (54 out of 56 instances). In contrast, among people who are younger than 60, when their tymp is of type ‘a’ and when they do not have a history of noise, the diagnose is totally independent of the history of nausea. As a matter of fact, every classification in $G_2$ made by the above $G_1$ rule is a misclassification (altogether 72 instances). This discrepancy has been correctly reported by the system.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Reported Discrepancy</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automobile</td>
<td>$G_1$: City-mpg$&gt;22$ &amp; Bore$&gt;3.39$ ⇒ Class=2</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Explanation:** This data set contains cars each of which is described by various characteristics such as engine type and fuel type. The task is to measure the degree to which a car is more risky than its price indicates. The partition attribute is the number of doors. $G_1$ contains cars of two doors while $G_2$ contains cars of four doors. Among two-door cars, the attribute City-mpg is very important for judging the risk level of a car. For instance, it is the root attribute of the decision tree. In contrast, among four-door cars, the attribute City-mpg does not matter at all for the risk evaluation. Likewise, the attribute Bore is involved in decision making in $G_1$ but not in $G_2$. This discrepancy has been correctly reported by the system.

<p>| Balloons | $G_2$: Size=small ⇒ Class=true [70.7%] | Yes |</p>
<table>
<thead>
<tr>
<th>Data Set</th>
<th>Reported Discrepancy</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Explanation:</em> The underlying concept of this data set is (Color=yellow and Size=small) or (Age=adult and Act=stretch) ⇒ Class=true; otherwise Class=false. The partition attribute is Color. $G_1$ contains instances whose colors are all purple and $G_2$ contains instances whose colors are all yellow. Because of the conjunctive relation between the color ‘yellow’ and the size ‘small’, the attribute Size has a decisive influence in $G_2$ but does not matter at all in $G_1$. This discrepancy has been correctly reported by the system.</td>
<td></td>
</tr>
<tr>
<td>Contraceptive</td>
<td>$G_1$: Wifes-age&gt;36 ⇒ Class=no-use [87.8%]</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td><em>Explanation:</em> This data set is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The instances are married women who either were not pregnant or did not know if they were at the time of interview. The task is to predict the current contraceptive method choice (no use, long-term methods or short-term methods) of a woman based on her demographic and socioeconomic characteristics. The partition attribute is Media-exposure. $G_1$ contains contraceptive methods whose media exposure is not good, and $G_2$ contains the others. Among women who are more than 36 years old, if a contraceptive method is not reported good by media ($G_1$), it is very likely that those women won’t use this method (46 out of 50 instances). However, if a contraceptive method is reported good by media, women over 36 years old may choose it, depending on factors like the number of children ever born. This discrepancy has been correctly reported by the system.</td>
<td></td>
</tr>
<tr>
<td>Echocardiogram</td>
<td>$G_2$: Fractional-shortening&gt;0.24 &amp; wall-motion-index≤2 ⇒ Class=dead [90.1%]</td>
<td>No</td>
</tr>
</tbody>
</table>
Data Set | Reported Discrepancy | Correct?
--- | --- | ---
**HeartDisease**  
$G_1$: Chest-pain-type=asymptomatic & Number-of-major-vessels-colored-by-fluoroscopy>0 ⇒ Class=disease-presence [93.9%]  
Yes

*Explanation:* This data set contains patients who suffered heart attacks at some point in the past. Some are still alive and some are not. The task is to predict whether a patient could survive for at least one year following the heart attack. The partition attribute is Pericardial-effusion. $G_1$ contains patients whose pericardial-effusion is fluid and $G_2$ no fluid. In the original data, the concept is influenced by Pericardial-effusion. For instance, when wall-motion-index≤1.3, if Pericardial-effusion=no-fluid, the patient is most probably dead; however, if Pericardial-effusion=fluid, it is uncertain. The reported rule does not precisely catch the wall-motion-index threshold value of 1.3 and hence is evaluated as incorrect.

GermanCredit  
$G_1$: Job=unskilled ⇒ Class=good-credit [90.6%]  
Yes

*Explanation:* This data set classifies customers into ‘good’ or ‘bad’ in terms of credit risks. Each instance is a customer described by attributes like status of existing checking account and credit history. The partition attribute is Foreign-worker. $G_1$ contains local workers and $G_2$ contains foreign workers. If a customer is a local worker, the assessment is almost always good credit (34 out of 37 instances) even when his or her job is unskilled (14 out of 14 instances). In contrast, if a customer is a foreign worker, the assessment is often more negative. For instance, if Existing-checking-account=0DM & Duration-in-month>11, many (30) unskilled workers are assessed as bad credit. This discrepancy has been correctly reported by the system.
**Data Set**

**Reported Discrepancy**

**Correct?**

*Explanation:* This data set contains patients each described by attributes such as age and gender. The class refers to the presence of heart disease in each patient. The partition attribute is Gender. \( G_1 \) contains all male patients while \( G_2 \) all females. Among males, once Chest-pain-type=asymptomatic and Number-of-major-vessels-colored-by-flouroscopy>0, a patient is very likely to have heart disease (43 out of 45 instances). In contrast, among females under the same conditions, the outcome is not certain at all. Whether a female patient has heart disease depends on other factors like the slope of the peak exercise ST segment. This discrepancy has been correctly reported by the system.

<table>
<thead>
<tr>
<th>Hepatitis</th>
<th>( G_1: \text{Spiders=\text{yes} &amp; Ascites=\text{yes}} \Rightarrow\text{Class=Live} )</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[96.3%]</td>
<td></td>
</tr>
</tbody>
</table>

*Explanation:* This data set contains instances each of which is a hepatitis patient described by attributes such as liver conditions. The class is whether the patient survives or dies. The partition attribute is Malaise. \( G_1 \) contains patients who have malaise and \( G_2 \) contains the others. Among patients who have spiders and ascites, if he or she also has malaise (\( G_1 \)), the patient almost always lives (71 out of 72 instances). In contrast, if a patient does not have malaise (\( G_2 \)), the outcome is uncertain. It further depends on factors like protime. This discrepancy has been correctly reported by the system.

<table>
<thead>
<tr>
<th>HorseColic</th>
<th>( G_1: \text{Abdominal-distension=none} \Rightarrow \text{Class=surgical-lesion-no} )</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[81.5%]</td>
<td></td>
</tr>
</tbody>
</table>
### Data Set

**Reported Discrepancy**

**Correct?**

*Explanation:* The task related to this data set is to predict whether or not a horse lesion is surgical. Each instance is a horse described by attributes like age, pulse and temperature. The partition attribute is Surgery. \( G_1 \) contains horses that were treated without surgery and \( G_2 \) contains the others. When a horse had no surgery (\( G_1 \)), Abdominal-distension is a very important parameter. An animal with abdominal distension is likely to be in pain and have reduced gut motility. A horse with severe abdominal distension is likely to require surgery just to relieve the pressure. However, when a horse had surgery, Abdominal-distension has trivial impact. Instead, factors like Abdomen-pain matters more. This discrepancy has been correctly reported by the system.

<table>
<thead>
<tr>
<th>Hypothyroid</th>
<th>( G_1 ): On-thyroxine=false &amp; On-antithyroid-medication=false &amp; TSH&gt;6 &amp; FTI≤64 ⇒ Class=hypothyroid [94.4%]</th>
<th>Yes</th>
</tr>
</thead>
</table>

*Explanation:* This data set contains information of diagnosing whether a patient has hypothyroid according to this patient’s age, gender and other measurements. The partition attribute is TSH-measured. \( G_1 \) has instances whose TSH-measured is ‘yes’ while \( G_2 \) is ‘no’. When TSH>6 and FTI≤64, patients whose TSH-measured equals ‘no’ (\( G_2 \)) is almost always diagnosed as negative (22 out of 24 instances). In contrast if those patients’ TSH-measured equals ‘yes’ (\( G_1 \)), the diagnose can be hypothyroid as well as negative depending on other factors, among which the most number of patients are hypothyroid if their On-thyroxine=false and On-antithyroid-medication=false (126 out of 131 instances), which rule has been correctly reported here.

<table>
<thead>
<tr>
<th>KR-vs-KP</th>
<th>( G_1 ): Bxqsq=taken ⇒ Class=White-cannot-win [99.8%]</th>
<th>Yes</th>
</tr>
</thead>
</table>
Data Set | Reported Discrepancy | Correct?
--- | --- | ---
**Explanation:** This data set describes the board of the chess endgame King+Rook versus King+Pawn on A7. Each instance is a board description using 36 board positions. The class is either White wins or White cannot win. The partition attribute is the board position Rimmx. $G_1$ contains instances where White has not taken Rimmx and $G_2$ contains the others. If White has taken Rimmx in the endgame, White always wins (584 out of 584 instances). If White has not taken Rimmx in the endgame, the outcome depends on other positions, such as Bxqsq. For instance, if Bxqsq is taken, White cannot win, which covers a large number of 743 instances. This discrepancy has been correctly reported by the system.

<table>
<thead>
<tr>
<th>LaborNegotiation</th>
<th>$G_1$: Wage-increase-first-year≤2.5 ⇒ Class=unacceptable [88.2%]</th>
<th>Yes</th>
</tr>
</thead>
</table>
**Explanation:** This data set includes all collective agreements reached in the business and personal services sector for locals with at least 500 members (teachers, nurses, university staff, police and so on) in Canada in Year 1987 and in the first quarter of Year 1988. Each instance is a final settlement in labor negotiations described by 16 attributes such as wage increase in the first year of the contract and the number of working hours during the week. The class is either acceptable contract or unacceptable contract. The partition attribute is Education-allowance. $G_1$ contains workers who have no education allowance while $G_2$ contains workers who have. When the wage increase of the first year is no more than 2.5, if workers have no education allowance, the bargaining always falls into the class of ‘unacceptable’ (11 out of 11); in contrast, if workers have education allowance, the bargaining result can be acceptable as well as unacceptable. This discrepancy has been correctly reported by the system.

<p>| LED | $G_1$: Upper-right-vertical=on &amp; Lower-right-vertical=on ⇒ Class=7 [97.8%] | Yes |</p>
<table>
<thead>
<tr>
<th>Data Set</th>
<th>Reported Discrepancy</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Explanation:</strong> This data set predicts the decimal digit according to the light-emitting diodes in LED displays. Each instance describes a display of seven light-emitting diodes (attributes). The class is a digit between 0 to 9 (inclusive). The partition attribute is the top-horizontal diode. ( G_1 ) comprises instances with this diode on and ( G_2 ) comprises instances with this diode off. Hence ( G_1 ) contains digits like 7 which ( G_2 ) does not, whereas ( G_2 ) contains digits like 1 which ( G_1 ) does not. Because the most number of instances belong to 7, the reported discrepancy is correct.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monk’sProblem</td>
<td>None</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Explanation:</strong> This data set’s underlying concept is: ((A_1 = A_2) \text{ or } (A_5 = 1) \Rightarrow \text{Class} = 1; \text{otherwise } \text{Class} = 0.) It has two binary attributes ( A_3 ) and ( A_6.) Hence none of the binary attributes affects the concept. In particular, instances here are partitioned according to ( A_3.) As a result, ( G_1 ) and ( G_2 ) are conceptually identical. It is correct to report that no discrepancy exists.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mushroom</td>
<td>( G_2: \text{Odor=} \text{pungent} \Rightarrow \text{Class=} \text{poisonous} \text{ [99.5%]} )</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Explanation:</strong> This data set presents whether a mushroom is edible or poisonous. Each instance describes a mushroom in terms of its physical attributes such as odor and cap color. The partition attribute is Stalk-shape. ( G_1 ) contains mushrooms whose stalk-shape is tapering while ( G_2 ) contains those whose stalk-shape is enlarging. In ( G_2, ) the attribute Odor is the most decisive factor. For example, if its order is pungent, a mushroom is always poisonous (256 out of 256 instances). However, in ( G_1 ) this concept does not apply at all because it turns out that no instance in ( G_1 ) has Odor=pungent. As a result, this concept covers a large number of instances in ( G_2 ) whereas always finds ‘nomatch’ in ( G_1 ) (96 instances). This discrepancy has been correctly reported by the system.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### SolarFlare

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Reported Discrepancy</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>SolarFlare</td>
<td>$G_1$: Spot-distribution=O $\Rightarrow$ Class=common-flares [84.7%]</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*Explanation:* This data set contains three classes of solar flares occurred in a 24-hour period. Each instance represents the captured features (attributes) for one active region on the sun. The partition attribute is Activity. $G_1$ contains active regions on the sun whose activities are unchanged while $G_2$ contains regions whose activities are reduced. Among $G_2$, when the previous 24-hour flare activity code equals 3 (more activity than one M1), the class is almost fixed to moderate-flares (9 out of 10 instances). In contrast, among $G_1$, when the previous 24-hour flare activity code equals 3, the class is diverse according to different values of the attribute Spot-distribution, and 27 out of 46 instances do not belong to moderate-flares. This discrepancy has been correctly reported by the system.

### Soybean

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Reported Discrepancy</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soybean</td>
<td>$G_1$: Leafspot-size=greater-than-1/8 &amp; Stem=norm $\Rightarrow$ Class=downy-mildew [93.3%]</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*Explanation:* This data set describes soybean diseases. Each instance is a soybean described by its attributes such as leaves, stem and fruit spots. The class is the disease diagnose. The partition attribute is Mold-growth. $G_1$ contains soybeans for which mold growth is present and $G_2$ contains soybeans for which mold growth is absent. When the leafspot size is greater than $\frac{1}{8}$ inch and the stem is normal, if there is mold, soybeans are always downy-mildew (20 out of 20); in contrast, if there is no mold, soybeans are diverse depending on factors like date. For example, if it is September, soybeans tend to be alternarialeaf-spot (40 out of 44). This discrepancy has been correctly reported by the system.

### Yeast

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Reported Discrepancy</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>$G_1$: Mce&gt;0.69 $\Rightarrow$ Class=erl [75.8%]</td>
<td>Yes</td>
</tr>
</tbody>
</table>

32
**Explanation:** This data set describes proteins from yeasts. The class contains 10 protein localization sites. This data set has only one binary attribute ERL. \(G_1\) contains yeasts which present the HDEL substring while \(G_2\) contains the others. The HDEL substring is thought to act as a signal for retention in the endoplasmic reticulum lumen. The partition results in very imbalanced groups, where \(G_1\) has 14 instances but \(G_2\) has 1470 instances. In \(G_1\), if \(\text{Mcg} > 0.69\), a yeast always belongs to ERL (5 out of 5). However, in \(G_2\), no yeast presents the HDEL substring and hence no instance belongs to the class ERL. As a result, this rule of \(G_1\) misclassifies 131 out 131 instances in \(G_2\). This discrepancy has been correctly reported by the system.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Reported Discrepancy</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zoo</strong></td>
<td>(G_1): Milk=true (\Rightarrow) Class=mammal [96.7%]</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Explanation:** This data set describes seven types of animals including mammal, bird, reptile, fish, amphibian, insect and invertebrate. Each instance is an animal described by its name and 16 attributes including hair, backbone, feather, egg, milk and so on. One may predict an animal's type according to its attributes. The partition attribute is Backbone. \(G_1\) contains animals that have backbones while \(G_2\) contains the others. Among \(G_1\) the attribute Milk is decisive. If it has milk, an animal always belongs to mammal (41 out of 41 instances). In contrast, among \(G_2\), the attribute Milk is not predictive. The data suggest that milk and backbone are correlated. Instances in \(G_2\) have neither backbone nor milk. They belong to either insect or invertebrate. This discrepancy has been correctly reported by the system.

Table 4: Experimental results and analysis for discrepancy discovery.

Accordingly, 20 trials have been conducted to test conceptual equivalence mining, out of which 19 trials witness correct discrepancy discovery (win) and 1 trial
incorrect (loss). When a two-tailed binomial sign test is applied to the 19 wins versus 1 loss out of 20 trials, the result is less than 0.01. Hence, the wins against losses are statistically significant at the critical level of 0.01, supporting the claim that conceptual equivalence mining has a systematic (instead of by chance) advantage in finding conceptual discrepancy between contrasting data groups.

5.3 Comparison with alternative methods

Compared with correspondence tracing, both conceptual equivalence mining and correspondence tracing have a ranking system. The two systems complement each other. Take the Horse Colic data in Section 5.2.2 as example. Conceptual equivalence mining returns the following. Thus conceptual equivalence mining reports in general that Rule 7 of \( G_1 \) contributes 51 discrepancy instances and represents the biggest discrepancy between \( G_1 \) and \( G_2 \). The second most different concept is reflected by Rule 12 of \( G_2 \) whose rank is 41, and so on.

\( (1) \ G_1 \)'s concept that contrasts with \( G_2 \):

- o7: Rank=51: Abdominal-distension=none \( \Rightarrow \) Class=surgical-lesion-no [81.5%]
- o15: Rank=35: Abdominal-distension=slight \( \Rightarrow \) Class=surgical-lesion-no [76.2%]
  …

\( (2) \ G_2 \)'s concept that contrasts with \( G_1 \):

- n12: Rank=41: Total-protein \( \leq \) 58 \( \Rightarrow \) Class=surgical-lesion-yes [90.7%]
- n4: Rank=25: Abdomen=distended-small-intestine \( \Rightarrow \) Class=surgical-lesion-yes [94.3%]
  …

In comparison, correspondence tracing reports which rules in \( G_2 \) (new data) are correspondent to which rules in \( G_1 \) (old data). It returns the following results where the old-new rule pairs are sorted according to their improvements to classification accuracy (their ranks). Thus correspondence tracing reports that instances
in $G_2$ used to be classified by the old rule 7 ($o7$) from $G_1$ are now classified by the new rule 12 ($n12$) from $G_2$. The change $<o7, n12>$ is important to the extent of increasing the estimated classification accuracy by 5.6% and is ranked highest. The second most important change on the ranked list is $<o7, n4>$, and so on.

\[
\Delta(o7, n12) = 0.056 \\
\Delta(o7, n4) = 0.052 \\
\Delta(o15, n12) = 0.031 \\
\ldots
\]

Hence conceptual equivalence mining and correspondence tracing consistently point out discrepancy rules like $o7$, $n12$, $o15$, and $n4$. Combining these two methods, a user can capture conceptual discrepancy between two data groups both in a general and in a detailed format.

Compared with literal equivalence, conceptual equivalence is more capable of dealing with groups whose instance spaces do not necessarily overlap. For data sets such as Mushroom and Nursery, there are few duplicate instances between $G_1$ and $G_2$. As a result, for tests illustrated in Figure 3, the literal equivalence is always close to 0 regardless of the noise status, which is not indicative. Likewise in many real-world applications, contrasting groups involve different individual customers, where conceptual equivalence is of greater use.

Compared with the measure using classification accuracy in Section 3.2.2, conceptual equivalence is less biased by the employed concept learner. Take the ‘Car’ data set as an example. When two identical copies of the data set are compared, because C4.5rules can only reach a learning accuracy of 96%, the accuracy measure only results in an equivalence of 0.96 according to Formula (4). By comparison, conceptual equivalence can still properly measure the equivalence as 1. The improved independency of concept learners is highly desirable in the sense that in the real world, seldom can learning achieve 100% accuracy.

Compared with rule comparison, conceptual equivalence can be more applicable. For example, the Hyperplane data [17, 18, 33] is a benchmark data set in time-series research: $\sum_{i=1}^{d} w_i x_i = w_0$ where each instance $<x_1, ..., x_d>$ is
<table>
<thead>
<tr>
<th>29 Rules learned from the first sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5: A1 &gt; 0.9 &amp; A2 &gt; 0.61 ⇒ Class=+ [100.0%]</td>
</tr>
<tr>
<td>8: A1 &gt; 0.18 &amp; A2 &gt; 0.85 &amp; A3 &gt; 0.26 ⇒ Class=+ [100.0%]</td>
</tr>
<tr>
<td>9: A1 &gt; 0.53 &amp; A2 &gt; 0.61 &amp; A3 &gt; 0.26 ⇒ Class=+ [100.0%]</td>
</tr>
<tr>
<td>15: A1 &gt; 0.68 &amp; A2 &gt; 0.31 &amp; A3 &gt; 0.45 ⇒ Class=+ [100.0%]</td>
</tr>
<tr>
<td>16: A1 &gt; 0.53 &amp; A2 &gt; 0.31 &amp; A3 &gt; 0.52 ⇒ Class=+ [100.0%]</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>32 Rules learned from the second sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>4: A1 &gt; 0.85 &amp; A2 &gt; 0.4 &amp; A3 &gt; 0.41 ⇒ Class=+ [100.0%]</td>
</tr>
<tr>
<td>7: A1 &gt; 0.22 &amp; A2 &gt; 0.98 ⇒ Class=+ [100.0%]</td>
</tr>
<tr>
<td>14: A2 &gt; 0.81 &amp; A3 &gt; 0.32 ⇒ Class=+ [100.0%]</td>
</tr>
<tr>
<td>17: A1 &gt; 0.37 &amp; A2 &gt; 0.68 &amp; A3 &gt; 0.26 ⇒ Class=+ [100.0%]</td>
</tr>
<tr>
<td>21: A1 &gt; 0.21 &amp; A2 &gt; 0.36 &amp; A3 &gt; 0.56 ⇒ Class=+ [100.0%]</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

Table 5: Direct rule comparison can be very difficult.

randomly generated and uniformly distributed in multi-dimensional space \([0,1]^d\). Instances satisfying \(\sum_{i=1}^{d} w_i x_i \geq w_0\) are labeled as positive, and otherwise negative. Two groups of 1000 instances\(^6\) are produced for the same hyperplane and are of the same class distribution. Since the random function takes the time as seed, the two groups involve many different instances. As in Table 5, the rule sets, each learned by C4.5rules from a group, contain many syntactically very different rules. A thorough semantic analysis may uncover more similarities. It can, however, be difficult since the semantics of numerical values requires a great deal of background knowledge. By comparison, conceptual equivalence can accurately judge the two groups equivalent to the degree of 0.95.

\(^6\)The sample size is sufficient to avoid the difference caused by classification variance.
6 Conclusion and Future Research

An important approach to acquiring knowledge is to learn through comparison. However, how to compare is not a trivial task. Given two contrasting groups, evaluating literal equivalence or conceptual equivalence takes place depending on whether one is interested in specific instances or abstract concepts. This paper seeks solutions to the latter in the context of classification learning, where the discrepancy between concepts is of interest. Existing work in this topic focuses on directly comparing the classification rule sets learned from each group. With due respect to previous contributions, this paper handles this issue from a different perspective, namely conceptual equivalence. First, it explains that a learned rule set is only a representation of the underlying concept and a concept may have various representations. Second, it argues that rule comparison might not be able to quantify the degree of discrepancy because different rules are not always commensurable. Third, it suggests that judging the equivalence of groups through the accuracy of classification can often be inaccurate and misleading.

To overcome these problems, a novel quantitative approach is proposed to carry out contrast mining in the context of classification learning. The new methodology of conceptual equivalence is composed of a scoring mechanism and a ranking system. It offers several attractive features. First, contrasting data groups is achieved without comparing two (sometimes large and complicated) rule sets. This results in a simple and elegant solution to an otherwise complex problem. Second, the scoring measure is adept at handling numeric data as well as categorical data. Third, the system collects evidence by checking the verdict of classifiers on each specific instance instead of utilizing the sometimes misleading classification accuracy. This improves mining’s independency of a learner’s accuracy and makes the contrast more objective. Last but not least, the scoring mechanism can be coupled with various classification algorithms besides rule learners, such as Bayesian probabilistic classifiers that are commonly used in real-world applications but produce no explicit rules.

Using conceptual equivalence to contrast data groups is a new topic and there
are various interesting issues to further investigate. For instance, current mechanisms handle two groups. It could be useful to adapt them to contrast multiple groups. Another instance: the current approach calculates the degree of conceptual equivalence between two groups by averaging \( \text{score}_1 \) and \( \text{score}_2 \). It may be interesting to explore other formulae to combine \( \text{score}_1 \) and \( \text{score}_2 \) in order to calculate the equivalence degree more accurately in Formula (3).

7 Acknowledgments

This research was supported by the Australian Research Council (ARC) Discovery Grant DP0770741. Xingquan Zhu was supported by the National Science Foundation of China (NSFC) Grant 60674109.

References


[22] Liu, B., Hsu, W., Ma, Y.: Discovering the set of fundamental rule changes. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2001) 335–340


