Abstract. One of the central problems in data mining is to filter large sets of discovered patterns. Our experience shows that this task should be done not for a single rule but by taking into considerations other similar rules. To fulfill this requirement the author proposes a new syntax-based distance measure dedicated for multilevel multidimensional rules as well as a rules’ neighbourhood with variable radius and a rule’s interestingness within the neighbourhood. Included example presents one of the possible usage of the proposed definitions in analysis of data from fault simulations.

Introduction

Calculating a distance between two rules is a widespread problem in the data mining field. The rules’ distance as a measure of a dissimilarity can be used for instance in a process of removing redundant rules or to define rules’ neighbourhood. At the same time finding the good distance measure for multidimensional environments, where dimensions are represented by nominative attributes is not a trivial task. The level of difficulty even raised when we have hierarchical attributes and multilevel rules.

With such a problem the author had to face when trying to define the neighbourhood of a rule for multilevel multidimensional decision rules build from data from fault injection experiments. As it was hard to find in literature a solution which fulfills all needs the author propose a new syntax-based distance for multilevel multidimensional decision rules. The usefulness of this measure was proven in researches connected with results of fault simulation experiments but the measure itself is general and can be used for decision rules mined from data from different domains.

Problem description

Institute of Computer Science (Warsaw University of Technology) has been conducting research in fault injections and fault effects analysis from years. A lot of fault simulations methods and techniques were invented as well as many different fault simulators were designed and developed [1, 4, 7, 13, 16]. All those are followed by research in the field of the analysis of fault effects observed during simulations experiments. This part of the research is focused not only on statistical methods but also on introducing data warehouses and data mining in analysis of fault simulations results [6, 14].

The current study is focused on preparing universal method of analysis of fault simulations results using data exploration techniques. The part of this method is preparing data mining models describing collected results using multilevel multidimensional decision rules and the way of choosing the most valuable and interesting rules to be analysed in details by domain experts. Detailed information about multilevel and multidimensional rules, their properties and algorithms of building them can be found in [9]. This kind of models were chosen based on characteristics of fault simulations results. Majority of data records which we received consist of one decision attribute (the result of a single simulation test) and many nominative or numerical conditional attributes. Additionally some of these conditional attributes create hierarchies so building common multidimensional rules would lead to many redundant patterns.

During this works the author observed that evaluating interestingness of every single rule is inefficient and insufficient. Even using suitable interestingness measures chosen based on widely known and examined criteria like those mentioned in [8, 10] may lead to choosing many similar or obvious rules and omitting important ones. This happens due to the fact that often rules with extreme values of interestingness measures are those with the most frequent value of the decision attribute. Beyond that in the case of analysing data from fault simulators we often do not have possibility to prepare training set with balanced values of the decision attribute.

Another problem is that domain experts find it difficult to reason on system dependability features based on a single rule. In most of the cases in order to describe one interesting feature more than one rule and some basic knowledge about training data statistics is needed.

To overcome this two shortcomings the rule’s neighbourhood must be defined. At the beginning the author tried to define the neighbourhood based on experts’ intuition which rules should be treated as similar. This led to 3 propositions:

• rules are similar when they have the same antecedent and different values of the decision attribute in the consequence,
• rules are similar when they have the same consequence and the same names of attributes in the antecedent but with different values of one or more conditional attributes (in the antecedent),
• rules are similar when they have the same consequence and more general or more detail antecedent.

In the second case we can additionally restrict range of values for every conditional attribute. For instance for numerical attributes we can set a maximum and a minimum value and for nominal hierarchical attributes we can take only values which have a common parent in the hierarchy tree. In the third proposition the author recognises a rule \( r_1 \) as more detailed than a rule \( r_0 \) in two cases. The first is when the antecedent of the rule \( r_1 \) is superset of the antecedent of the rule \( r_0 \). The second is possible only for a hierarchical attribute: when conditional attributes in antecedents of both rules are the same and have the same values except those cases when all following conditions are fulfilled:

• in both rules we have different attributes from the same
paper discusses three different distances. The first is a

The authors propose also some variations of their rules' distance for

In [3] considers association rules interestingness in the

Related works

Although the problem of the analysis of large sets of

The proposed distance metric is used to group rules by using

The authors of [11] notice similar problem of domain ex-

Two main types of rules distance: a semantics-based dis-

In [15] authors deal with the problem of reducing a num-

In [3] considers association rules interestingness in the
term of rules' unexpectancy in their neighbourhoods. Authors
mention two types of rules distance: a semantics-based dis-

tance and a syntax-based distance. The semantics-based
distance is calculated based on the differences between
record sets matching both rules. The syntax-based distance
measures difference between itemsets from which rules were
built and should take into account three factors: the sym-
metric difference of all items in the two rules, the symmetric
difference of the antecedents of the two rules and the sym-
metric difference of the consequence. The paper proposes
the new distance measure of the second type for association
rules to determine rule's neighbourhoods. The rule is called
interesting within its neighbourhoods when it has unex-
pected confidence or spare neighbourhood. The confi-
deence of the rule is unexpected when it deviates from the average confi-
dence in rule's neighbourhood much more than standard de-
viation. This interestingness is of the unexpected type.
The second type is interestingness of the isolated type, that is
when the number of potential rules in rule's neighbourhood is
large but the number of mined rules (that is those which ful-
fil support and confidence threshold) is relatively small. Au-

tors propose also some variations of their rules' distance for
other kind of rules like Horn clauses or interval-based rules
but even with these modifications the measure is unsuitable
for multilevel data. The problem is how to calculate symmet-
ric difference of two attributes from one hierarchy.

In [15] authors deal with the problem of reducing a num-
ber of created association rules not by commonly used prin-
ting techniques but by grouping. To enable association rules

grouping a normalised distance metric is proposed. The

hierarchy,

• the attribute in the rule \( r_1 \) describes the lower level of

the hierarchy,

• a value of the attribute in the rule \( r_0 \) is an ancestor of a
value of the attribute in the rule \( r_1 \).

If the rule \( r_1 \) is more detailed than \( r_0 \), then the rule \( r_0 \) is
more general than \( r_1 \). Additionally we can combine the first
proposition with the second or the third.

Neighbourhoods being built based on above three
propositions may be sufficient for the analysis of some rules, but
for measuring of rule's interestingness within their neigh-
bourhood the more general and the more formal definition of
the rule's neighbourhood must be found. Such a definition
should be built based on a rules' distance. Unfortunately
no distance measures suitable for multilevel multidimensional
rules could be found in literature.

Related works

Although the problem of the analysis of large sets of
rules is well represented in literature it is hard to find solutions
designed for multilevel multidimensional rules. In this sec-
tion some works on association rules neighbourhoods and
association rules clustering are presented. These works also
include the problem of finding a good distance measure for
association rules and give a good background for our current
work.

[3] considers association rules interestingness in the
term of rules' unexpectancy in their neighbourhoods. Authors
mention two types of rules distance: a semantics-based dis-
tance and a syntax-based distance. The semantics-based
distance is calculated based on the differences between
record sets matching both rules. The syntax-based distance
measures difference between itemsets from which rules were
built and should take into account three factors: the sym-
metric difference of all items in the two rules, the symmetric
difference of the antecedents of the two rules and the sym-
metric difference of the consequence. The paper proposes
the new distance measure of the second type for association
rules to determine rule's neighbourhoods. The rule is called
interesting within its neighbourhoods when it has unex-
pected confidence or spare neighbourhood. The confi-
deence of the rule is unexpected when it deviates from the average confi-
deence in rule's neighbourhood much more than standard de-
viation. This interestingness is of the unexpected type.
The second type is interestingness of the isolated type, that is
when the number of potential rules in rule's neighbourhood is
large but the number of mined rules (that is those which ful-
fil support and confidence threshold) is relatively small. Au-

Authors propose also some variations of their rules' distance for
other kind of rules like Horn clauses or interval-based rules
but even with these modifications the measure is unsuitable
for multilevel data. The problem is how to calculate symmet-
ric difference of two attributes from one hierarchy.

In [15] authors deal with the problem of reducing a num-
ber of created association rules not by commonly used prin-
ting techniques but by grouping. To enable association rules

grouping a normalised distance metric is proposed. The

The new measures make an assumption that all items have a common ancestors. That makes those mea-
sures useless for multidimensional data where every dimen-
sion have its own attributes' hierarchy.

Multilevel multidimensional decision rules distance

From two main types of rules' distance the syntax-based
definition will be used in this paper. This is mostly due to the
two main disadvantages of semantic-based definitions. The
first is that in order to calculate this distance we need infor-
mation about identifiers of all records in a training set that
supports rules, what would be time and space consuming. In
contrary, syntax-based distance can be calculate in runtime
based only on rules. The second is that in multilevel envi-
rnments the number of records supporting a rule will differ
depends on hierarchy levels from which attributes come. Be-
cause of this the rules with attributes from one path in a hie-
archy tree may have large distance because of a small set of
common supporting records.

A good distance metric should fulfil definitions 1.

Definition 1. For a set of decision rules DR function \( d: DR \times DR \rightarrow [0, +\infty) \) is metric on DR, such that for any a, b, c \( \in DR \) the following holds:
The author proposes the following conditions that a good distance measure for multilevel multidimensional rules should fulfill. These conditions were developed based on experts' intuition about rules similarity. For simplification they are presented for rules with one attribute in antecedent but they can be generalised for longer rules. For decision rules \( r_i : a_1 = w_1 \Rightarrow d_1, \ldots, a_i = w_i \Rightarrow d_i \) a distance measure \( d \) should meet the following conditions:

1. **non-negative**: \( d(a, b) \geq 0 \);
2. **identity of indiscernibles**: \( d(a, b) = 0 \iff a = b \);
3. **symmetry**: \( d(a, b) = d(b, a) \);
4. **triangle inequality**: \( d(a, b) \leq d(a, c) + d(c, b) \).

The distance measure for multilevel multidimensional rules should fulfills the following conditions:

- \( d((a_1, w_1), (a_2, w_2)) = 0 \iff a_1 = a_2 \land w_1 = w_2 \)
- \( d((a_1, w_1), (a_2, w_2)) = 1 \iff a_1 \neq a_2 \land w_1 = w_2 \)
- \( d((a_1, w_1), (a_2, w_2)) = 2 \iff a_1 \neq a_2 \land w_1 \neq w_2 \)

The above definition is most general if rules consist of attributes from all hierarchies but the majority of the rules' mining algorithms like Apriori or FP-Grown build rules of different lengths. To enable usage of this distance we have to extend all rules with missing attributes. Every attribute in a training set must belong to some hierarchy and every hierarchy must be organised in a tree with one root attribute with one value **ALL**. The **ALL** value can be also interpreted as **ANY** because any value of lower-level attributes is a child of **ALL**.

- If there are attributes that do not belong to any hierarchy or there are hierarchies that have many roots, a root attribute must be created. Root attributes from all hierarchies do not take part in rules mining and are not presented to a user as they do not add any new knowledge. They are only used to fill missing attributes hierarchies during the extension of rules before the calculation of the distance.

As it was mentioned above, the distance proposed in 2 fulfills the definition 1. Three first features from this definition are obvious. The last, the triangle inequality is more complicated to prove. The proposed distance is the sum of distances between all pairs of conditional and decision attributes from the same hierarchy. If the measure is the sum of distances between attributes from the same hierarchy in order to fulfill the triangle inequality it will be enough if inequality will be fulfilled for factors from every hierarchy separately.

The following shows the distance measure for two decision rules which contain one attribute from every hierarchy of attributes:

\[
d((a_1, w_1), (a_2, w_2)) \leq (d((a_1, w_1), (a_3, w_3)) + d((a_2, w_2), (a_3, w_3)))
\]

where \( (a_1, w_1), (a_2, w_2), (a_3, w_3) \) are attributes' names and values which belong to the same hierarchy. According to the definition 2 the left side of the inequality can take a value from a set \( \{0, 1, 2, 3\} \) and the right from a set \( \{0, 1, 2, 3, 4, 5, 6\} \). So we have to prove that the triangle inequality is fulfilled when the right side of the equation 1 takes values less than maximum value of the left side \( \{0, 1, 2\} \). Now we will analysed those three cases in details.

**Case 1**

\[
d((a_1, w_1), (a_3, w_3)) + d((a_2, w_2), (a_3, w_3)) = 0 \iff d((a_1, w_1), (a_3, w_3)) = 0 \land d((a_2, w_2), (a_3, w_3)) = 0
\]

\[
d((a_1, w_1), (a_2, w_2)) = 0 \iff a_1 = a_2 \land w_1 = w_2 \land a_3 \land w_3
\]

**Case 2**

\[
d((a_1, w_1), (a_3, w_3)) + d((a_2, w_2), (a_3, w_3)) = 1 \iff d((a_1, w_1), (a_3, w_3)) = 1 \land d((a_2, w_2), (a_3, w_3)) = 1
\]

\[
d((a_1, w_1), (a_2, w_2)) = 1 \iff a_1 \neq a_2 \land w_1 \neq w_2 \land a_3 \land w_3
\]

In the following we will consider only the first part of the alternative because considerations for the second are analogous.

**Case 3**

\[
d((a_1, w_1), (a_3, w_3)) = 1 \iff a_1 = a_3 \land w_1 = w_3 \land a_2 = a_3 \land w_2 \neq w_3
\]

\[
d((a_1, w_1), (a_2, w_2)) = 1 \iff a_1 = a_2 \land w_1 \neq w_2
\]
contain only rules that have something in common with the rule in the centre of the neighbourhood. For instance a common attribute or at least attributes from the same hierarchy in the antecedent or a common consequence. In this case the radius can not be set up as constant. In this paper the new definition 4 for the neighbourhood with variable radius is presented.

**Definition 4.** An X-neighbourhood of a rule \( r_0 \), denoted as \( S_x(r_0) \) is the following set:

\[
\{ r_n : d(r_0, r_n) \leq (3)a* (\max(|r_0.ant| + |r_n.ant|, N_H)) \land r_n \neq r_0, \}
\]

where \( r_n \) is a potential rule, \( |r.n| \) is a cardinality of the rule antecedent before extension, \( N_H \) is a number of the conditional attributes’ hierarchies, \( a \) is a real number from the interval (0, 2]. This neighbourhood can be called the neighbourhood with variable radius.

The definition 4 is based on the number of attributes’ hierarchies in a training set. That is the number of hierarchies of the conditional attributes plus one for the decision attribute’s hierarchy. However, we do not add one in the formula 3 because we want rules in the neighbourhood to have at least one common attribute with the neighbourhood’s centre. Because many rules are rather short and have, before extension, much less attributes in the antecedent than the total number of conditional attributes’ hierarchies the definition 4 assumes that the number of hierarchies in the condition cannot be higher than rules antecedents cardinality sum. This provides that we do not take into considerations hierarchies which are not present in both rules before extension.

The maximum value of the multiplier is \( a = 2 \). It describes a situation when all pairs of attributes from the same hierarchy, except one which is common, have values on the same path from the root to a leaf in the hierarchy tree. Of course, this minimal condition to add a rule to other rule’s neighbourhood is a kind of averaging. In reality some pairs of attributes may have values from different paths while others will have common values.

Now we can move to the definition 5 which defines the rule’s interestingness within the neighbourhood.

**Definition 5.** A rule \( r \) is interesting within its neighbourhood \( S(r) \), if:

\[
M(r) \notin AVG_M(S(r)) - STEDEV_M(S(r)),
\]

where \( M \) is a rule interestingness measure, \( AVG_M(S(r)) \) is an average value of the measure \( M \) for rules in \( S(r) \), \( STEDEV_M(S(r)) \) is a standard deviation of the measure \( M \) for rules in \( S(r) \).

In the above definition we can use as well the neighbourhood with constant \( S_2(r) \) or with variable radius \( S_2(r) \). This definition is similar to the definition of the rule’s interestingness of the unexpected confidence type presented in [3] but it is more flexible thanks to the possibility of using different interestingness measures.

**Example of usage**

Usage of the proposed distance measure and the neighbourhood definition will be presented based on the analysis of fault effects in zlib compression and decompression library [17]. Five versions of the library (1.1.4, 1.2.1, 1.2.5, 1.2.6, 1.2.7) compiled with three Microsoft Visual C++ compilers (2008, 2010, 2012) were tested. Faults were injected.
Detailed investigations show that different measures choose different rules and values of $R$ or $a$. Also, certain interestingness measures are not suitable in filtering large sets of multilevel multidimensional decision rules. As similar we mean with close distance between rules or fulfilling expert's propositions mentions in problem description.

Tables 3 and 4 show statistics for neighbourhood's size for different values of $R$ or $a$. As it was expected the number of rules in the neighbourhood rises with values of $R$ or $a$. Additionally we can see that standard deviations for the neighbourhood with constant radius are less than for the neighbourhood with variable radius. For the smallest values of the parameter $a$ we have even empty neighbourhoods for all rules with one attribute in the antecedent.

Presented experimental results show that the proposed distance measure and neighbourhoods' definitions are effective in filtering large sets of multilevel multidimensional decision rules. Depending on used interestingness measure we can reduce original rules set even by more than 95%. The problem of choosing the right interestingness measure should be studied in more details in cooperation with domain experts. It is possible that for different problems different interestingness measures will be suitable.

### Conclusion

This article contains the proposal of the new distance metric for multilevel multidimensional decision rules. As far as author knows there is no similar measure mentioned in literature. In addition the new definition of the rule's neighbourhood which depends on rule's length, called the neighbourhood based on the same measure. Another interesting property is that in a set of rules that are interesting within their neighbourhood we can find group of similar rules. As similar we mean with close distance between rules or fulfilling expert's propositions mentions in problem description.

Tables 3 and 4 show statistics for neighbourhood's size for different values of $R$ or $a$. As it was expected the number of rules in the neighbourhood rises with values of $R$ or $a$. Additionally we can see that standard deviations for the neighbourhood with constant radius are less than for the neighbourhood with variable radius. For the smallest values of the parameter $a$ we have even empty neighbourhoods for all rules with one attribute in the antecedent.

Presented experimental results show that the proposed distance measure and neighbourhoods' definitions are effective in filtering large sets of multilevel multidimensional decision rules. Depending on used interestingness measure we can reduce original rules set even by more than 95%. The problem of choosing the right interestingness measure should be studied in more details in cooperation with domain experts. It is possible that for different problems different interestingness measures will be suitable.

### REFERENCES

ing association rules with meta-rules using knowledge cluster-
ing, 11th IEEE International Symposium on Programming and 

rules in terms of neighbourhood-based unexpectedness, Re-
search and Development in Knowledge Discovery and Data 

vironment for Complex Experiments, 14th IEEE International 

[5] Gawkowski, P., Ławryńczuk, M., Marusak, P. M., Tatjewski, 
P., Sosnowski, J.: On improving dependability of the numeri-
cal GPC algorithm, European Control Conference, pp. 1377–
1382, 2009.

fects Analysis and Reporting System for Dependability Evalu-

M.: LRFI – Fault Injection Tool for Testing Mobile Software, 
Emerging Intelligent Technologies in Industry, Studies in Com-
putational Intelligence, 369, pp. 269–282, 2011.

mining: A survey, ACM Computing Surveys (CSUR), 38(3), 9 

[9] Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and 
Techniques, Morgan Kaufmann, USA, 2012.

Rule Interestingness Measures: Experimental and Theoretical 
Studies, Quality Measures in Data Mining, Studies in Computa-

Analysis Using OLAP Operations, 12th ACM SIGKDD Interna-
tional Conference on Knowledge Discovery and Data Mining, 

for Multi-Level Association Rules, Innovations in Intelli-
gent Machines-4, Studies in Computational Intelligence, 514, 


warehouse for simulation experiments, RSEISP 2007, LNAI, 


Distributed ABS System with Fault Injection, Innovations in 
Computing Sciences and Software Engineering, pp. 201-206, 
2010.

2014.]

Author: M. Sc. Agnieszka Komorowska, Institute of 
Computer Science, Faculty of Electronics and In-
formation Technology, Warsaw University of Technology, 
Nowowiejska 15/19, 00-662 Warszawa, Poland, email:
A.Komorowska@ii.pw.edu.pl