Agnieszka KOMOROWSKA

Warsaw University of Technology

doi:doi:10.15199/48.2015.03.07

Syntax-based distance for multilevel multidimensional decision rules

Abstract. *One of the central problem 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 fulfil 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.*

Streszczenie. *Badania pokazują, że wybór istotnych wzorców z dużych zbiorów reguł powinien być dokonywany nie na podstawie pojedynczej reguły, ale w powiązaniu z regułami podobnymi. Aby zrealizować to zadanie, autorka proponuje nową syntaktyczną miarę odległości miedzy wielopoziomowymi wielowymiarowymi regułami decyzyjnymi i definicje: sąsiedztwa reguły ze zmiennym promieniem oraz atrakcyjności reguły w sąsiedztwie. Artykuł zawiera także przykład wykorzystania zaproponowanych definicji w analizie danych z symulatorów błędów.* ***(Syntaktyczna odległóść między wielowymiarowymi wielopoziomowymi regułami decyzyjnymi)***

Keywords: multilevel multidimensional decision rules, rules neighbourhood, rules interestingness, syntax-based rules distance, fault injection experiments
Słowa kluczowe: wielopoziomowe wielowymiarowe reguły dezyzyjne, sąsiedztwo reguł, atrakcyjność reguł, syntaktyczna odległość między regułami, symulacja błędów

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 fulfils 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 mention 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 treat 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

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 neighbourhood 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 section 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 distance 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 symmetric difference of all items in the two rules, the symmetric difference of the antecedents of the two rules and the symmetric 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 unexpected confidence or spare neighbourhood. The confidence of the rule is unexpected when it deviates from the average confidence in rule's neighbourhood much more than standard deviation. 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 fulfil support and confidence threshold) is relatively small. 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 symmetric difference of two attributes from one hierarchy.

In [15] authors deal with the problem of reducing a number of created association rules not by commonly used pruning techniques but by grouping. To enable association rules grouping a normalised distance metric is proposed. The paper discusses three different distances. The first is a semantic-based distance called *Absolute Market-basket Difference Distance*. Its most important disadvantage is that it is strongly correlated with support. The improvement is *Conditional Market-basket Probability Distance* based on the conditional probability that one rule is false when the another one is true. The range of this metric is between 0 and 1 which is unnatural for a distance. The improvement of this metric is *Conditional Market-basket Log-likelihood Distance*. The proposed distance metric is used to group rules by using *Dimensionless Agglomerative Chain Clustering* and *Self Organizing Feature Map Clustering*. In the result authors received clusters of groups of correlated rules which can be analysed.

The authors of [11] notice similar problem of domain experts while analysing a single rule that was discover during analysis of rules from results of fault injections experiments. While working on finding cellular phone call failures using association rules manning the authors observed that the most of existing interestingness methods treats rules individually while users rarely find a single rule interesting. The rule is only interesting in the context of other rules or in the group of other rules. Though the similar problem the paper proposes quite different solutions. Rules are organised in a multidimensional structure and are explored by using OLAP (On-Line Analysis Processing) operations and general impression mining as well as visualisation techniques.

[2] is another example of organising of association rules using clustering. Authors designed an algorithm named k-means-AR which is an adaptation of k-means clustering algorithm. The idea is to visualise association rules in clusters based on their methodological similarity. To achieve this goal authors propose a similarity measure for association rules. Similarity measures can be tread as an opposite to distance measures which represents dissimilarity. The proposed measure is a difference between the number of all items in both rules and the number of common items in both rules. As it grows with the number of items that are different between both rules it is rather a distance measure.

[12] shows an example of interestingness measures for multilevel association rules based on diversity and peculiarity. A peculiarity measure is an enhancement of the distance measure introduced in [3] and it allows one to use it with multilevel data. To enable this a cardinality of the symmetric difference is replaced by a diversity of the symmetric difference which takes into account the hierarchical difference of the items in the symmetric difference. The diversity measure is a measure of the difference or the distance between the items within a rule. Both proposed measures consist of two distances: *hierarchical relationship distance* and *concept level distance*. The new measures make an assumption that all items have a common ancestors. That makes those measures useless for multidimensional data where every dimension 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 information 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 environments the number of records supporting a rule will differ depends on hierarchy levels from which attributes come. Because of this the rules with attributes from one path in a hierarchy 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, +∞) is metric on DR, such that for any a, b, c $\in$ DR the following holds:*

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 author proposes the following conditions that a good distance measure for multilevel multidimensional rules should fulfil. These conditions ware 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_1*:* a_1 = w_1 -> d_1, r_2*:* a_2 = w_2 -> d_2 i r_3*:* a_3 = w_3 -> d_3 a distance measure *d* should meet the following conditions:

- d(r_1, r_2) = 0 $\iff$ $a_1 = a_2 \wedge w_1 = w_2 \wedge d_1 = d_2$;
- $d_1 = d_2 = d_3 \wedge a_1 = a_2 \wedge a_1 \neq a_3 \Rightarrow$ d(r_1, r_2) < d(r_1, r_3);
- $d_1 = d_2 \neq d_3 \wedge a_1 = a_2 = a_3 \wedge w_1 = w_2 = w_3 \Rightarrow$ d(r_1, r_2) < d(r_1, r_3);
- $d_1 = d_2 = d_3 \wedge a_1 \neq a_2 \neq a_3 \wedge a_1$, a_2 belong to the same hierarchy $\wedge$ a_1, a_3 do not belong to the same hierarchy $\Rightarrow$ d(r_1, r_2) < d(r_1, r_3);
- $d_1 = d_2 = d_3 \wedge a_1 \neq a_2 \neq a_3 \wedge a_1$, a_2, a_3 belong to the same hierarchy $\wedge$ w_1, w_2 are on the same path from the root to a leaf in the hierarchy tree $\wedge$ w_1, w_3 are not on the same path from the root to a leaf in the hierarchy tree $\Rightarrow$ d(r_1, r_2) < d(r_1, r_3).

Definition 2 contains the proposition of a new distance, which fulfils the definition 1 and above conditions.

Definition 2. *The distance measure between two decision rules which contain one attribute from every hierarchy of attributes is a sum of distances between all pairs of conditional and decision attributes from the same hierarchy, where each attribute in a pair comes from a different rule. The distance between two attributes, each represented by an attribute's name and attribute's value,* (a_1, w_1) *and* (a_2, w_2)*, where* a_1, a_2 *belongs to the same hierarchy, is calculated as follow:*

- d((a_1, w_1), (a_2, w_2)) = 0 $\iff$ $a_1 = a_2 \wedge w_1 = w_2$*;*
- d((a_1, w_1), (a_2, w_2)) = 1 $\iff$ $a_1 = a_2 \wedge w_1 \neq w_2$*;*
- d((a_1, w_1), (a_2, w_2)) = 2 $\iff$ $a_1 \neq a_2 \wedge w_1$, w_2 *are on the same path from the root to a leaf in the hierarchy tree;*
- d((a_1, w_1), (a_2, w_2)) = 3 $\iff$ $a_1 \neq a_2 \wedge w_1$, w_2 *are not on the same path from the root to a leaf in the hierarchy tree.*

The above definition assumes that rules consist of attributes from all hierarchies but the majority of the rules' mining algorithms like *Apriori* or *FPGrown* 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 fulfils 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 fulfil the triangle inequality it will be enough if inequality will be fulfilled for factors from every hierarchy separately. This condition can be expressed as follow:

$$d((a_1,w_1),(a_2,w_2)) \leq \\ d((a_1,w_1),(a_3,w_3)) + d((a_2,w_2),(a_3,w_3)), \tag{1}$$

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

$$\begin{gathered} d((a_1,w_1),(a_3,w_3)) + d((a_2,w_2),(a_3,w_3)) = 0 \iff \\ d((a_1,w_1),(a_3,w_3)) = 0 \wedge d((a_2,w_2),(a_3,w_3)) = 0 \\ d((a_1,w_1),(a_3,w_3)) = 0 \wedge d((a_2,w_2),(a_3,w_3)) = 0 \iff \\ a_1 = a_3 \wedge w_1 = w_3 \wedge a_2 = a_3 \wedge w_2 = w_3 \\ a_1 = a_3 \wedge w_1 = w_3 \wedge a_2 = a_3 \wedge \\ w_2 = w_3 \Rightarrow a_1 = a_2 \wedge w_1 = w_2 \\ a_1 = a_2 \wedge w_1 = w_2 \iff \\ d((a_1,w_1),(a_2,w_2)) = 0 - \textit{the triangle inequality is fulfilled} \end{gathered}$$

Case 2

$$\begin{gathered} d((a_1,w_1),(a_3,w_3)) + d((a_2,w_2),(a_3,w_3)) = 1 \iff \\ (d((a_1,w_1),(a_3,w_3)) = 0 \wedge d((a_2,w_2),(a_3,w_3)) = 1) \vee \\ (d((a_1,w_1),(a_3,w_3)) = 1 \wedge d((a_2,w_2),(a_3,w_3)) = 0) \end{gathered}$$

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

$$\begin{gathered} d((a_1,w_1),(a_3,w_3)) = 0 \wedge d((a_2,w_2),(a_3,w_3)) = 1 \iff \\ a_1 = a_3 \wedge w_1 = w_3 \wedge a_2 = a_3 \wedge w_2 \neq w_3 \\ a_1 = a_3 \wedge w_1 = w_3 \wedge a_2 = a_3 \wedge w_2 \neq w_3 \Rightarrow \\ a_1 = a_2 \wedge w_1 \neq w_2 \\ a_1 = a_2 \wedge w_1 \neq w_2 \iff \\ d((a_1,w_1),(a_2,w_2)) = 1 - \textit{the triangle inequality is fulfilled} \end{gathered}$$

Case 3

$$\begin{gathered} d((a_1,w_1),(a_3,w_3)) + d((a_2,w_2),(a_3,w_3)) = 2 \iff \\ (d((a_1,w_1),(a_3,w_3)) = 1 \wedge d((a_2,w_2),(a_3,w_3)) = 1) \vee \\ (d((a_1,w_1),(a_3,w_3)) = 0 \wedge d((a_2,w_2),(a_3,w_3)) = 2) \vee \\ (d((a_1,w_1),(a_3,w_3)) = 2 \wedge d((a_2,w_2),(a_3,w_3)) = 0) \end{gathered}$$

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

Case 3a

$$d((a_1, w_1), (a_3, w_3)) = 1 \wedge d((a_2, w_2), (a_3, w_3)) = 1 \iff$$
$$a_1 = a_3 \wedge w_1 \neq w_3 \wedge a_2 = a_3 \wedge w_2 \neq w_3$$
$$a_1 = a_3 \wedge w_1 \neq w_3 \wedge a_2 = a_3 \wedge w_2 \neq w_3 \Rightarrow$$
$$a_1 = a_2 \wedge (w_1 \neq w_2 \vee w_1 = w_2)$$
$$a_1 = a_2 \wedge (w_1 \neq w_2 \vee w_1 = w_2) \iff$$
$$d((a_1, w_1), (a_2, w_2)) = 1 \vee$$
$$d((a_1, w_1), (a_2, w_2) = 0) - \textit{the triangle inequality is fulfilled}$$

Case 3b

$$d((a_1, w_1), (a_3, w_3)) = 0 \wedge d((a_2, w_2), (a_3, w_3)) = 2 \iff$$
$$a_1 = a_3 \wedge w_1 = w_3 \wedge a_2 \neq a_3 \wedge$$
$$w_2, w_3 \textit{are on the same path}$$
$$\textit{from the root to a leaf in the hierarchy tree}$$
$$a_1 = a_3 \wedge w_1 = w_3 \wedge a_2 \neq a_3 \wedge$$
$$w_2, w_3 \textit{are on the same path}$$
$$\textit{from the root to a leaf in the hierarchy tree} \Rightarrow$$
$$a_1 \neq a_2 \wedge w_1, w_2 \textit{are on the same path}$$
$$\textit{from the root to a leaf in the hierarchy tree}$$
$$a_1 \neq a_2 \wedge w_1, w_2 \textit{are on the same path}$$
$$\textit{from the root to a leaf in the hierarchy tree} \iff$$
$$d((a_1, w_1), (a_2, w_2)) = 2 -$$
$$- \textit{the triangle inequality is fulfilled}$$

As it was shown above the triangle inequality is always fulfilled for one attribute so it will be also fulfilled for the sum.

Rule's interestingness within neighbourhood

Rules' distance is not an independent tool while analysing interestingness. Before the presentation of the usage of the proposed distance measure the rule's neighbourhood and rule's interestingness within neighbourhood must be defined.

In [3] the following definition was presented.

Definition 3. *An* R-neighbourhood *of a rule* r_0 *(*R > 0*), denoted as* $S_R(r_0)$ *is the following set:*

$$\{r_n : d(r_0, r_n) \leq R \wedge r_n \neq r_0\}, \quad (2)$$

where r_n *is a potential rule and* R *is a radius of the neighbourhood. This neighbourhood can be called the neighbourhood with a constant radius.*

The neighbourhood from the definition 3 is intuitive and easy to calculate as it represents a sphere with the rule r_0 in the centre in the multidimensional space of rules. Unfortunately in multilevel environments rules even for quite small radius like *R = 5* may be found by users as irrelevant. For example in *5-neighbourhood* of a rule with one attribute in the antecedent we will find all other rules of the same length. At the same time a rule with 4 attributes in the antecedent, including one with the same name and value as in the first rule will not be included in considered neighbourhood. However, users find the rule with 4 attributes in the antecedent as more detailed rule in comparison to the rule in the centre of the neighbourhood. Besides, rules with one attribute from a different hierarchy in antecedents are considered to be irrelevant. To meet this intuition the neighbourhood must 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:*

$$\{r_n : d(r_0, r_n) \leq a * (max(|r_0.ant| + |r_n.ant|, N_H)) \wedge r_n \neq r_0\}, \quad (3)$$

where r_n *is a potential rule,* |r.ant| *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:*

$$M(r) \notin < AVG_M(S(r)) - STEDEV_M(S(r)), \quad AVG_M(S(r)) + STEDEV_M(S(r)) >, \quad (4)$$

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 $S_R(r)$ or with variable radius $S_x(r)$. 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

Table 1. A number of rules interesting within constant radius neighbourhood for different values of *R*.

Measure	R = 3	R = 4	R = 5	R = 6	R = 7	R =8	R = 9
Piatetsky-Shapiro	440	295	466	431	551	543	644
Information gain	870	903	1061	1108	1242	1246	1325
Confidence	905	944	1110	1144	1285	1294	1372
Support	221	157	225	217	258	289	304
Cosine	729	722	746	738	737	733	770
Odds ratio	483	261	197	183	178	178	178
ϕ coefficient	790	716	642	631	603	579	578
Two-way support	439	293	455	417	542	533	641
Certain factor	885	928	1095	1145	1295	1291	1373
from all measures	1892	1776	1952	1944	2045	2029	2050

Table 2. A number of rules interesting within variable radius neighbourhood for different values of *a*.

Measure	a=0.5	a=1	a=1.5	a=2
Piatetsky-Shapiro	701	670	769	724
Information gain	917	1093	1312	1370
Confidence	976	1244	1458	1473
Support	565	387	417	387
Cosine	918	825	898	855
Odds ratio	713	185	178	178
ϕ coefficient	1001	610	614	605
Two-way support	703	685	765	751
Certain factor	936	1176	1398	1416
from all measures	2015	2033	2121	2074

using FITS simulator [4, 13] into a sample text compression and decompression application and (from version 1.2.5) into code coverage tests. Results from experiments were loaded in into FEARS data warehouse [6] and multilevel multidimensional decision rules were built using author's modification of Apriori algorithm. The minimal support level was set to 1% and the confidence to 20%. The decision attribute was a test result: correct application execution, application returned incorrect result, application stopped after a system exception or application terminated after a fixed time of inactivity. The training set contains 13 conditional attributes (root attributes from hierarchies trees are not included) grouped in 7 hierarchies. Obtained model contains 4469 rules.

Tables: 1, 2, 3 and 4 present statistical results of using the definition 5 with different variations of the neighbourhood. For the neighbourhood with a constant radius ($S_R(r)$) seven different values of the radius (*R*) were evaluated and for the neighbourhood with a variable radius ($S_x(r)$) four values of multiplier *a* were tested. Tables 1 and 2 contain a number of the interesting rules obtained by using different interestingness measures in the definition 5. Definitions of used measures and their properties can be found in [8, 10]. The last rows in both tables contain sum of interesting rules sets for all measures. Based on this data we can observe that the number of interesting rules differs for different measures and for different values of *R* or *a*. At the same time it is impossible to find monotonic relationship between the number of interesting rules and values of *R* or *a* for the most of the measures. Detailed investigations show that different measures choose different rules as interesting. Besides not all rules with extreme values of a specific measure are interesting within their 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.

Table 3. Constant radius neighbourhood's size statistics for different values of *R*.

Statistic	R = 3	R = 4	R = 5	R = 6	R = 7	R =8	R = 9
min size	19	60	167	372	629	1085	1653
max size	187	618	1124	1966	2958	3675	4242
average size	74	194	445	857	1453	2169	2908
median	69	186	428	826	1419	2146	2906
std. deviation	23	60	128	232	342	421	445

Table 4. Variable radius neighbourhood's size statistics for different values of *a*.

Statistic	a=0.5	a=1	a=1.5	a=2
min size	0	42	498	875
max size	114	1563	3545	4427
average size	45,14	760,83	2438,55	4124,18
median	47	785	2558	4223
std. deviation	21,41	327,55	605,59	345,14

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 with variable radius, as well as definitions of the rule's interestingness within the neighbourhood were also presented. Included example shows that those elements can be successfully used to filter large sets of the multilevel multidimensional rules.

The measure was so far used to define the neighbourhood. It can be also used to choose group of the similar rules which can be presented to domain experts for detailed analysis. In the future, it should be evaluated if it is possible to use the proposed distance measure in decision rules grouping task.

REFERENCES

[1] Chyłek, S.: Emulation based software reliability evaluation and optimization, PRZEGLĄD ELEKTROTECHNICZNY, 90(2), pp. 121–124, 2014.

[2] Djenouri, Y., Drias, H., Habbas, Z., Chemchem, A.: Organizing association rules with meta-rules using knowledge clustering, 11th IEEE International Symposium on Programming and Systems (ISPS), pp. 109-115, 2013.
[3] Dong, G.,Li, J.: Interestingness of discovered association rules in terms of neighbourhood-based unexpectedness, Research and Development in Knowledge Discovery and Data Mining, LNCS, 1394, pp. 72–86, 1998.
[4] Gawkowski, P., Sosnowski, J.: Developing Fault Injection Environment for Complex Experiments, 14th IEEE International On-Line Testing Symposium, pp. 179–181, 2008.
[5] Gawkowski, P., Ławryńczuk, M., Marusak, P. M., Tatjewski, P., Sosnowski, J.: On improving dependability of the numerical GPC algorithm, European Control Conference, pp. 1377–1382, 2009.
[6] Gawkowski, P., Kuczyńska, M. A., Komorowska A.: Fault Effects Analysis and Reporting System for Dependability Evaluation, RSCTC 2010, LNCS, 6086, pp. 524–533, 2010.
[7] Gawkowski, P., Pawełczyk, P., Sosnowski, J., Cabaj, K., Gajda M.: LRFI – Fault Injection Tool for Testing Mobile Software, Emerging Intelligent Technologies in Industry, Studies in Computational Intelligence, 369, pp. 269–282, 2011.
[8] Geng, L., Hamilton, H. J.: Interestingness measures for data mining: A survey, ACM Computing Surveys (CSUR), 38(3), 9 (2006)
[9] Han, J., Kamber, M., Pei, J,: Data Mining: Concepts and Techniques, Morgan Kaufmann, USA, 2012.
[10] Lenca, P., Vaillant, B., Meyer, P., Lallich, S.: Association Rule Interestingness Measures: Experimental and Theoretical Studies, Quality Measures in Data Mining, Studies in Computational Intelligence, 43, pp. 51–76, 2007.
[11] Liu, B., Zhao, K., Benkler, J., Xiao, W.: Rule Interestingness Analysis Using OLAP Operations, 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 297–306, 2006.
[12] Shaw, G., Xu, Y., Geva, S.: Interestingness Measures for Multi-Level Association Rules, Innovations in Intelligent Machines-4, Studies in Computational Intelligence, 514, pp. 47–74, 2014.
[13] Sosnowski, J., Lesiak, A., Gawkowski, P., Włodawiec, P.: Software Implemented Fault Inserters, IFAC Work. On Progr. Dev. and Sys., pp. 293–298, 2003.
[14] Sosnowski, J., Zygulski, P., Gawkowski, P.: Developing data warehouse for simulation experiments, RSEISP 2007, LNAI, 4585, pp. 543–552, 2007
[15] Strehl, A., Gupta, G. K., Ghosh, J.: Distance based clustering of association rules. ANNIE 1999, 9, pp. 759–764, 1999.
[16] Trawczyński, D., Sosnowski, J., Gawkowski, P.: Testing Distributed ABS System with Fault Injection, Innovations in Computing Sciences and Software Engineering, pp. 201-206, 2010.
[17] zlib [web page] `http://zlib.net/`. [Accessed on 16 Jun. 2014.].

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