Discovering Non-Standard Semantics of Semi-Stable Attributes

Angelina A. Tzacheva and Zbigniew W. Ras

University of North Carolina – Charlotte
College of Information Technology, Computer Science Department
Charlotte, NC 28223, USA
aatzache@uncc.edu and ras@uncc.edu

Abstract

A new class of rules, called action rules, show what actions should be taken to improve the profitability of customers. Action rules introduced by (Ras and Wieczorkowska, 2000a) and investigated further by (Ras and Gupta, 2002a) assume that attributes in a database are divided into two groups: stable and flexible. These reflect the ability of a business user to influence and control their change for a given consumer. In this paper, we introduce a new classification of attributes partitioning them into stable, semi-stable, and flexible. Values of stable attributes can not be changed for a given consumer (for instance "maiden name" is an example of such an attribute). So, stable attributes have only one interpretation. If values of an attribute change in a deterministic way as a function of time (for instance values of attribute "age" or "height"), we call them semi-stable. All remaining attributes are called flexible. Clearly, in the process of action rule extraction, stable attributes are highly undesirable. What about semi-stable attributes? Although, they seem to be quite similar to stable attributes, the difference between them is quite essential. Semi-stable attribute may have many different interpretations but among them only one interpretation is natural and it is called standard. All its other interpretations are called non-standard. In a non-standard interpretation, a semi-stable attribute can be classified as flexible. In a single database we may easily fail to identify attributes which have non-standard interpretation. In this paper, we show how distributed information system introduced by (Ras, 2000b, 2001a) can be used as a tool to identify which semi-stable attributes have non-standard interpretation so they can be classified as flexible. This way, by decreasing the number of stable attributes in a database we may discover action rules which would not be discovered otherwise.

Introduction

In (Ras and Wieczorkowska, 2000a) the notion of special type of rules, called action rules, was introduced. These rules can be constructed from classification rules to suggest a way to re-classify objects (for instance customers) to a desired state. In e-commerce applications, this re-classification may mean that a consumer not interested in a certain product, now may buy it, and therefore may fall into a group of more profitable customers.

These groups are described by values of classification attributes in a decision table schema. In paper by (Ras and Wieczorkowska, 2000a), all attributes are divided into stable and flexible. This time, a new subclass of attributes called semi-stable attributes is introduced. Semi-stable attributes are typically a function of time, and undergo deterministic changes (for instance attribute "age" or "height"). Different interpretations, called non-standard, of such attributes may exist, and in such cases all these attributes can be treated the same way as flexible attributes. In the algorithm of action rule extraction, proposed by (Ras and Wieczorkowska, 2000a) attributes which are not flexible are highly undesirable. By identifying which semi-stable attributes have non-standard interpretation, we increase the number of flexible attributes and the same may increase the confidence of generated action rules.

Assuming that attribute is flexible, we may find a way to change its value for a given object. However, quite often, such a change cannot be done directly to a chosen attribute (for instance to the attribute "profit"). In that situation, definitions of such an attribute in terms of other attributes have to be learned. These definitions are used to construct action rules showing what changes in values of attributes, for a given consumer, are needed in order to re-classify this consumer the way business user wants. In a distributed system, we may search for definitions of these flexible attributes looking at either local or remote sites for help.

The application of semi-stable attributes to the process of action rules mining involves detection of nonstandard interpretations of semi-stable attributes. At local system level detection is possible, but limited to dependencies existing between local attributes. At distributed information systems level the detection of nonstandard interpretations involves discovering semantic inconsistencies, addressed by (Ras and Dardzinska, 2002b).

Information Systems and Decision Tables

An information system is used for representing business knowledge. (Pawlak, 1985a) gives the following definition:

By an information system we mean a pair $S = (U, A)$, where:

1. U is a nonempty, finite set of objects (called customer identifiers),
2. A is a nonempty, finite set of attributes i.e. $a: U \rightarrow V_a$ for $a \in A$, where V_a is called the domain of a.

Information systems can be seen as generalizations of decision tables (Pawlak, 1985a). Partition of the set of attributes into conditions and decisions is given in any decision table. We assume that the set of conditions is partitioned into stable, semi-stable, and flexible conditions. Attribute $a \in A$ is called stable for the set U if its values assigned to objects from U can not be changed by a business user. An attribute is called semi-stable, if it is a function of time and it is changing, in its standard interpretation, in a deterministic way. Otherwise, it is called flexible. Date of birth is an example of a stable attribute. Age is an example of semi-stable attribute (its value "young" in a non-standard interpretation may mean "behaving as a young person"). Interest rate on any customer account is an example of a flexible attribute. For simplicity reason, we will consider decision tables with only one decision. We adopt the following definition of a decision table:

By a decision table we mean any information system of the form $S = (U, A_1 \cup A_2 \cup A_3 \cup \{d\})$, where $d \notin A_1 \cup A_2 \cup A_3$ is a distinguished attribute called the decision. Elements of A_1 are called stable conditions, the elements of A_3 are called semi-stable, and $A_2 \cup \{d\}$ are called flexible conditions.

The assumption that attribute d is flexible is quite essential. Otherwise we would be unable to re-classify objects in U from the point of view of attribute d. So, if d is flexible and we want to change its value for a given object, values of some attributes from $A_2 \cup A_3$ have to be changed as well.

Before we proceed, certain relationships between values of attributes from A_2 and A_3 and values of the attribute d have to be presented first.

Action Rules

(Ras and Wieczorkowska, 2000a) proposed a method to construct action rules from a decision table containing both stable and flexible attributes. In this section, let us assume that for each attribute in A_3 we know its semantics and also we know if it is stable or flexible. So $A_3 = \emptyset$, which means some semi-stable attributes are moved to A_1, some to A_2.

Assume now that for any two collections of sets X, Y, we write, $X \leq Y$ if $(\forall x \in X)(\exists y \in Y)[x \subseteq y]$. Let $S = (U, A_1 \cup A_2 \cup \{d\})$ be a decision table and $B \subseteq A_1 \cup A_2$. We say that attribute d depends on B if $CLASS_S(B) \leq CLASS_S(d)$, where $CLASS_S(B)$ is a partition of U generated by B (Pawlak, 1985a).

Assume now that attribute d depends on B where $B \subseteq A_1 \cup A_2$. The set B is called d-reduct in S if there is no proper subset C of B such that d depends on C. The concept of d-reduct in S was introduced, in rough sets theory (Pawlak, 1985a), to identify minimal subsets of $A_1 \cup A_2$ such that rules describing the attribute d in terms of these subsets are the same as rules describing d in terms of $A_1 \cup A_2$. It was shown that in order to induce rules in which THEN part consists of the decision attribute d and IF part consists of attributes belonging to $A_1 \cup A_2$, only subtables $(U, B \cup \{d\})$ of S where B is a d-reduct in S can be used for rules extraction.

By L(r) we mean all attributes listed in IF part of rule r. For example, if $r = [(a1,3)*(a2,4) \rightarrow (d,3)]$ is a rule then $L(r) = \{a1,a2\}$. By d(r) we denote the decision value of a rule r. In our example d(r) = 3. Similarly, a1(r)=3.

If r1, r2 are rules and $B \subseteq A_1 \cup A_2$ is a set of attributes, then r1/B = r2/B means that the conditional parts of rules r1, r2 restricted to attributes B are the same. For example if $r1 = [(a1,3) \rightarrow (d,3)]$, then $r1/\{a1\} = r/\{a1\}$.

Algorithm for constructing action rules, implemented as system DAR was given by (Ras and Wieczorkowska, 2000a).

For each pair of rules (r1, r2) satisfying the conditions $r1/A_1 = r2/A_1$, d(r1)=k1, d(r2)=k2 where k1< k2, if (b1, b2,…, bp) was a list of all attributes in $L(r1) \cup L(r2) \cup A_2$ on which r1, r2 differ and r1(b1)= v1, r1(b2)= v2,…, r1(bp)= vp, r2(b1)= w1, r2(b2)= w2,…, r2(bp)= wp then the algorithm DAR generates the following (r1,r2)-action rule:

If $[(b1, v1 \rightarrow w1) \wedge (b2, v2 \rightarrow w2) \wedge ... \wedge (bp, vp \rightarrow wp)](x)$ then the ranking profit of customer x is expected to change from *k1* to *k2*.

Object x supports this action rule if it satisfies the properties: b1(x)=v1, b2(x)=v2, ….bp(x)=vp, d(x)= k1.

By the support of an action rule r we mean all the objects supporting that rule. It is denoted by Sup(r).

If the change of values of attributes of object $x \in Sup(r)$ will match the term

$[(b1, v1 \rightarrow w1) \wedge (b2, v2 \rightarrow w2) \wedge ... \wedge (bp, vp \rightarrow wp)](x)$

and the resulting object is either classified in S as k2 or not classified at all, then the action rule r successfully suports x. The set of objects successfully supported by r is denoted by SSup(r). By the confidence of rule r we mean Conf(r) = SSup(r)/ Sup(r).

Example 1. Assume that $S = (\{x1,x2,x3,x4,x5,x6,x7,x8\}, \{a,c\} \cup \{b\} \cup \{d\})$ is a decision table represented as Table 1. The set {a,c} contains stable attributes, b is a flexible attribute and d is a decision attribute.

It can be easily checked that {b,c}, {a,b} are the only two d-reducts in S.

Applying, for instance, LERS discovery system (Chmielewski and Grzymala-Busse, 1993a) the following definitions are extracted from S:

$(a,0) \rightarrow (d,L)$, $(c,0) \rightarrow (d,L)$,
$(b,R) \rightarrow (d,L)$, $(c,1) \rightarrow (d,L)$,
$(b,P) \rightarrow (d,L)$, $(a,2) \wedge (b,S) \rightarrow (d,H)$,
$(b,S) \wedge (c,2) \rightarrow (d,H)$.

Now, let us assume that (a, v → w) denotes the fact that the value of attribute a has been changed from v to w. Similarly, the term (a, v → w)(x) means that a(x)=v has been changed to a(x)=w. Saying another words, the property (a,v) of object x has been changed to property (a,w).

	a	b	c	d
x1	0	S	0	L
x2	0	R	1	L
x3	0	S	0	L
x4	0	R	1	L
x5	2	P	2	L
x6	2	P	2	L
x7	2	S	2	H
x8	2	S	2	H

Table 1

If we take r = [(b, R → S) ∧ (d, L → H)] as the action rule, then Sup(r)={x2,x4}, Conf(r) = 1.

Semi-Stable Attributes

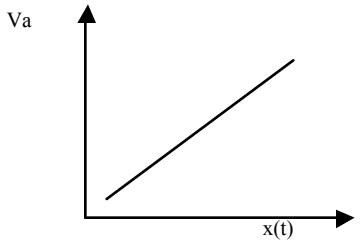
The notion of action rules introduced by (Ras and Wieczorkowska, 2000a) divides attributes into two groups: stable and flexible. These reflect the ability of a business user to influence and control their change for a given consumer. In the process of action rule extraction stable attributes are highly undesirable. We introduce a new classification of attributes into stable, semi-stable, and flexible taking into consideration semantics of atributes which clearly may differ from database to database.

Value of a stable attribute *a* for a given object cannot be changed by a business user in any interpretation of *a*. All such interpretations are called standard.

An example of such an attribute is "date of birth". Standard interpretations of this attribute may differ in a granularity level. It is possible that one stable attribute implies another one.

There is a special subset of attributes called semi-stable, which at first impression may look stable, but they are a function of time and undergo changes in a deterministic way. Therefore, they cannot be called stable. The change is not necessarily in a linear fashion (see Graph 1). An attribute may be stable for a period of time, and then begin changing in certain direction as shown on Graph 2.

Semi-stable attributes may have many interpretations, some of which might be nonstandard. We denote by $M_s(a)$ the set of standard interpretations of attribute a and by $M_n(a)$ the set of non-standard interpretations of a. If the attribute a has a nonstandard interpretation, $I(a) \in M_n(a)$, then it can be changed, and thus it may be seen as a flexible attribute in action rule extraction.

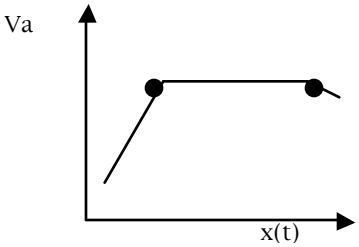


Graph 1. Semi-stable attribute "Age"

For instance, if a = "age" and Dom(a) = {*young, middle-aged, old*}, the author of the database may indeed input *young* for a person who behaves as young when their actual age is *middle-aged*. Then the interpretation is nonstandard. The business user can therefore influence this attribute. For example, if the following action rule was mined for object x

r1 = [[(a, young → middle-aged)](x) ∧ [(d, L → H)](x)]

with respect to decision attribute d (ex. *loyalty*) the business user would like to change the attribute value "young" to "middle-aged" for object x. Since the database author interpretation is nonstandard related to the behavior associated with certain age, if the object is put into special conditions that can affect its behavior, such as top university, the attribute value can be changed, and the same object x might be re-classified from *low loyalty* to *high loyalty*.



Graph 2. Semi-stable attribute "Height"

Many cases of nonstandard interpretations could be found in databases. It is particularly important for those to be detected when mining for global rules in distributed knowledge systems. An example is the attribute "height". Consider the following situation: Chinese people living in the mountains are generally taller than majority of Chinese population. If Dom(a) = {*short, medium, tall*} for attribute "height", and a system S1, contains data about Chinese population in the mountains. The author of the database may consider a certain Chinese person living in the mountains *medium* height in relation to the rest. Now assume another system S2 containing data about Chinese people living in popular urban area. In global action rule extraction, if S2 is mined for rules, the interpretation would

regard the height value *medium* from S1 as *tall*. Therefore, the interpretation in S1 is nonstandard.

Numeric attributes may possess nonstandard interpretations as well. Consider for instance the attribute "number of children". When one is asked about the number of children one has, that person may count step-children, or children who have died. In such a case, the interpretation is nonstandard.

A flexible attribute is an attribute which value for a given object varies with time, and can be changed by a business user. Also, flexible attributes may have many interpretations. "Interest rate" is an example of a flexible attribute.

Assume that S = (U,A) is an information system which represents one of the sites of a distributed information system (DIS). Also, let us assume that each attribute in A is either stable or flexible but we may not have sufficient temporal information about semi-stable attributes in S to decide which one is stable and which one is flexible. In such cases we will search for additional information, usually at remote sites for S, to classify uniquely these semi-stable attributes either as stable or flexible.

Discovering Semantic Inconsistencies

Different interpretations of flexible and semi-stable attributes may exist. Semi-stable attributes, in a non-standard interpretation, can be classified as flexible attributes and therefore can be used in action rule extraction. We discuss a detection method of nonstandard interpretations of a semi-stable attribute at local information system level, and next at distributed information systems level.

Detection of a nonstandard interpretation at local level is limited to the dependency of one semi-stable attribute to another semi-stable attribute for which it is known that its interpretation is standard. Attribute related to time must be available in the information system, such as the attribute "age". Furthermore, information about certain breakpoints in attribute behavior is required, such as the break points shown in Graph 2. This information can be placed in the information system ontology.

Assume that S=(U,A) is an information system and I is the interpretation used for attributes in S. Also, assume that both a ,b ⊆ A are semi-stable attributes, I(a) ∈ M_s and the relation $\Pi_{\{a, b\}}(S)$ = {(v_a , v_b): v_a ∈ Dom(a), v_b ∈ Dom(b)} is obtained by taking projection of the system S on {a,b}. The ontology information about break points for attributes a and b in S, represented in the next section as relation $R_{I(a),I(b)}$, is assuming that the interpretation I is standard for attribute b. It is possible that some tuples in $\Pi_{\{a, b\}}(S)$ do not satisfy the break points requirement given. In such a case the interpretation of B is nonstandard, I(B) ∈ M_n.

Consider the following situation:

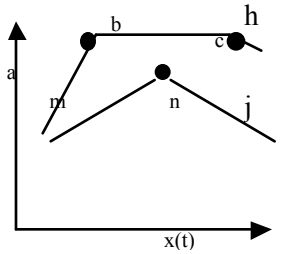
Example 3. Assume that S_1 = (U_1, {a}, {h, j}, {b}) is an information system represented by Table 2, where {a} is a set of stable attributes, {h, j} is a set of semi-stable attributes, and {b} is a set of flexible attributes, where h is "height" and j is "number of cousins". The interpretation of j is known to be standard, I(j) ∈ M_s. The system represents a local site.

	a	h	b	j
x1	0	a	S	m
x2	0	b	R	m
x3	0	b	S	n
x4	0	c	R	m
x5	2	b	P	n
x6	2	b	P	n
x7	2	a	S	m
x8	2	c	S	m

Table 2

Graph 3 shows the break points defined by the system's ontology for attributes h and j as a function of time t. The number of cousins grows as the height grows, since the person is young, and the parents' brothers and sisters have newborn children. The number of cousins decreases, as the height becomes constant or shrinks, since for a person who is middle-aged or old, the number of his/her cousins naturally decreases as they die. Therefore,

if I(h) ∈ M_s and I(j) ∈ M_s,
then $R_{I(h),I(j)}$ = {(a, m), (b, m), (b, n), (c, n)}
is placed in the ontology layer for system S_1.



Graph 3. Dependency relation between attributes h and j assuming standard interpretations for both of them.

From Figure 3, we see that relation instance (c,m) ∈ $\Pi_{\{h,b\}}(S_1)$ representing objects x4, x8 does not belong to $R_{I(h),I(j)}$. Therefore, I(h) ∈ M_n

In other words, objects x4, x8 do not satisfy the break point requirement given on Graph 3, thus the interpretation of attribute "height" is nonstandard.

Single information systems provide limited capability of detecting nonstandard semantics. Distributed information systems supply greater ability to detect nonstandard semantics. They also give the opportunity to business users to seek alternative solution at remote sites. This is particularly important in a situation when they are either not willing or not able to undertake the suggested actions from their local site. In a distributed information systems

(DIS) scenario semantic inconsistencies can be detected even if temporal information is not available. With large number of sites containing similar attributes in DIS, certain trends can be observed, such as association rules with high confidence and support common for all the sights. In such a situation, it is also possible that a small number of sites do not support those common rules, or even contradict them. This case presents a hint for nonstandard attribute interpretation, and semantic inconsistencies.

Assume that S, S_1, S_2, … , S_n where $S_i = (U_i, B_i)$, $(i=1,2,\ldots,n)$ are information systems which are parts of DIS. Association rule mining is performed on all systems in DIS. Rules satisfying the minimum support *ms* and minimum confidence *mc* thresholds are mined. If a rule

$$[w_a \wedge w_b \wedge \ldots \wedge w_c] \Rightarrow w_d \quad [ms, mc] \quad (1)$$

where $d \in A$ is a semi-stable attribute in $S = (U,A)$ is extracted from S and supported by many sites in DIS, it is called a *trade*, or a common rule for DIS. If rule (1) is supported either only by S or by a very small number of sites in DIS and at the same time rule (2) is supported by many sites in DIS

$$\neg([w_a \wedge w_b \wedge \ldots \wedge w_c] \Rightarrow w_d) \quad [ms, mc] \quad (2)$$

then the attribute d has nonstandard interpretation in S.

Assume that we do not know if the interpretation of a semi-stable attribute $d \in B_1$ at site $S_1 = (U_1, B_1)$ is standard. We have discussed the case when attribute d is defined in S_1 in terms of a semi-stable attribute b, which has standard interpretation in S_1. Namely, if we identify an object in S_1 which description contradicts information about attributes d, b stored in S_1 ontology, then attribute d has non-standard interpretation in S_1. However, it can happen that we do not have any information about the interpretation of b in S_1. We can either look for a definition of d in terms of another semi-stable attribute in S_1 or look for a definition of d in terms of attribute b at another site of DIS. If we cannot find any attribute other than d, which is semi-stable and has non-standard interpretation in S1, we contact another site.

Let B_1^{ss} be the set of all semi-stable attributes in S_1. We search for sites S_i such that $d \in B_i$ and $B_i^{ss} \cap B_1^{ss} \neq \emptyset$. Let I_d be the collection of such sites and $b \in B_1^{ss} \cap B_i^{ss}$, where $i \in I_d$.

In the case where, the interpretation of both attributes d, b is standard, if $I \subseteq I_d$ satisfies the property that any $b \in B_1^{ss} \cap B_i^{ss}$ has standard interpretation in S_i, then i is not considered. Thus, we need to observe another site from I_d.

This algorithm was tested on a DIS consisting of thirteen sites representing thirteen Insurance Company Datasets, with a total of 5000 tuples in all DIS sites. Semi-stable attributes with non-standard interpretation have been detected and used jointly with flexible attributes in action rules mining. The confidence of these action rules is usually higher than the confidence of the corresponding action rules based only on flexible attributes.

Now let us assume that $I_{\{d, b\}}$ is the set of sites in I_d such that $b \in B_i \cap B_1^{ss}$. We extract rules at sites $I_{\{d, b\}}$ describing d and having b on their left side. Either association rules discovered at site S_1 will support association rules discovered at majority of sites $I_{\{d, b\}}$ or conflict many of them. We claim that the interpretation of attribute d is standard in the first case. In the second case, it is non-standard.

References

Chmielewski M. R., Grzymala-Busse J. W., Peterson N. W., Than S., 1993a. The Rule Induction System LERS - a Version for Personal Computers, *Foundations of Computing and Decision Sciences*, 18, No. 3-4

Pawlak Z., 1985a. Rough Sets and Decision Tables, *Lecture Notes in Computer Science* 208**,** Springer-Verlag, 186-196

Ras, Z., Wieczorkowska, A., 2000a. Action Rules: how to increase profit of a company, in Principles of Data Mining and Knowledge Discovery, (Eds. D.A. Zighed, J. Komorowski, J. Zytkow), Proceedings of PKDD'00, LNCS/LNAI, No. 1910, Springer-Verlag, 587-592, Lyon, France.

Ras, Z., Zytkow, J., 2000b. Mining for attribute definitions in a distributed two-layered DB system, *Journal of Intelligent Information Systems*, Kluwer, 14, No. 2/3, 2000, 115-130

Ras, Z., 2001a. Query Answering based on Distributed Knowledge Mining, in Intelligent Agent Technology, Research and Development, IAT'2001 Proceedings, 17-27, Maebashi City, Japan

Ras, Z., Gupta, S., 2002a. Global Action Rules in Distributed Knowledge Systems, in CS&P 2001 Special Issue (Eds. L. Czaja, H.-D. Burkhard, P. Starke), *Fundamenta Informaticae Journal*, IOS Press, 51, No. 1-2, 175-184

Ras, Z., Dardziska A., 2002b. Handling semantic incosistencies in distributed knowledge systems using ontologies, in Foundations of Intelligent Systems, ISMIS'02 Proceedings, LNCS/LNAI, Vol. 2366, Springer-Verlag, 66-74, Lyon, France