Journal of Universal Computer Science, vol. 15, no. 13 (2009), 2622-2628
submitted: 31/10/08, accepted: 13/6/09, appeared: 1/7/09 © J.UCS

Rough Classification – New Approach and Applications

Ngoc Thanh Nguyen
(Institute of Informatics
Wroclaw University of Technology, Poland
thanh@pwr.wroc.pl)

Abstract: Rough classification has been known as the concept of Pawlak within the Rough Set Theory. In this paper the novel rough classification approach and its applications in e-learning systems and user interface management for recommendation processes will be presented.

Keywords: Rough Classification, User Management, E-learning Systems
Categories: E.1, H.2.1, I.2.4, I.2.11

1 Introduction

Using the relational structures real world objects are represented by a set of attributes and their values. Owing to this representation these objects can be classified into classes and the criterion for classification is the combinations of attribute values such that two objects belong to the same class if they have the same values for each attribute. Pawlak [Pawlak, 99] worked out a method which enables to determine the minimal set of attributes (called a *reduct*) which generates the same classification as the whole set of attributes. From the practical point of view, this method is very useful because owing to it the same classification may be created on the basis of a smaller set of attributes instead of a large one.

Rough classification methods, in general, serve to determining a set of attributes which generate an approximate classification referring to a given classification. In the Pawlak's concept of rough classification for a given classification *C* of set *U* of objects a rough classification is the approximation of *C*. Assume that classification *C* is generated by set *B* of attributes, then the approximation of *C* is based on determining a proper subset *B*' of *B* such that the classification generated by *B*' differs "a little" from *C*. The small difference between these classifications is illustrated by the difference of their accuracy measures which should not be larger than some threshold [Pawlak, 99].

In our approach we consider other problem of rough classification: *For a given classification of set U which is generated by set A of attributes, one should determine such minimal set B of attributes from A that the distance between the classification generated by attributes from B and the given classification is minimal.*

This problem differs from the one formulated by Pawlak. Besides, for its solving we will use distance functions between classifications (in fact they are partitions of set *U*) to generate the minimal set of attributes. This approach has been originally presented in work [Nguyen, 05] and developed in [Nguyen, 06] and [Nguyen, 08]. In this paper we give a brief description of its solution and applications for e-learning systems and user interface management in recommendation processes.

2 Brief Description of New Rough Classification Problems

In information system stores the following to kind of information about its users: User data and usage data. User data often consist of demographic data of the user. Usage data, in turn, refer to the interaction process between the user and the system and his behavior during using the system. User data are included in so called the user profile, while usage data are included in the usage path.

User classification is very often performed in information systems because owing to it the system can better assign to the user a proper way for serving him.

The new conception of rough classification is based on the following assumption: It is the usage data which should decide about the criterion of user classification. Figure 1 represents this idea. For a new user the system creates a profile including his user data, on its basis the user is classified into a class. Next this user will use the system and the usage path is created which contains the effective way for this user for using the system. After a period of the system life the set of all usage paths is clustered giving a clustering of the set of all users. This clustering is next compared with the user classification made on the basis of a criterion created from user profiles. If the difference is large, this means that the classification criterion is not proper and the new one should be set on the basis of the usage paths' clustering.

We present some basic notions needed for defining the problems.

By a partition of set U we call a finite set of nonempty and disjoint with each other classes which are subsets of U, such that their union is equal to U. By $\pi(U)$ we denote the set of all partitions of set U.

Let P, $Q \in \pi(U)$, the product of P and Q (written as $P \cap Q$) is defined as:

$$P \cap Q = \{p \cap q: p \in P, q \in Q \wedge p \cap q \neq \emptyset\}.$$

Thus product $P \cap Q$ is also a partition of U.

Distance between partitions P and Q of set U as the minimal number of elementary transformations (*removal of an element or augmentation of an element*) needed to transform partition P into partition Q.

For P , $Q \in \pi(U)$ let $M(P) = [p_{ij}]_{n \times n}$ be such a matrix that

$$p_{ij} = \begin{cases} 0 & \text{if} \quad x_i, x_j \text{ are in different classes of } P \\ 1 & \text{if} \quad x_i, x_j \text{ are in the same class of } P \end{cases}$$

Matrix $M(Q) = [q_{ij}]_{n \times n}$ be defined in the similar way, where $n = card(U)$.

The distance between partitions P and Q is defined as follows:

$$d(P,Q) = ½ \sum_{i.j=1}^{n} \left| p_{ij} - q_{ij} \right| .$$

As stated earlier, if U denotes the set of learners and A is the set of attributes from user data a learner may be represented by a tuple of type A. For $a \in A$ we define the following binary relation P_a on set U: For $x_1, x_2 \in U$ pair $<x_1, x_2>$ belongs to P_a if and only if learners x_1 and x_2 are assigned with same value referring to attribute a. It is easy to prove that P_a is an equivalence relation, and therefore it is a partition of set U.

More general, a set B of attributes determines an equivalent relation P_B of set U as follows: a pair of 2 elements from U belongs to $P_{\mathbf{B}}$ if and only if referring to each attribute $b \in B$ these elements are assigned with the same value. Thus P_B is also a partition of U and the following property is true:

$$P_{\mathbf{B}} = \bigcap\nolimits_{b \in B} P_b \ .$$

It is obvious that $P_B \subseteq P_a$ for each $b \in B$.

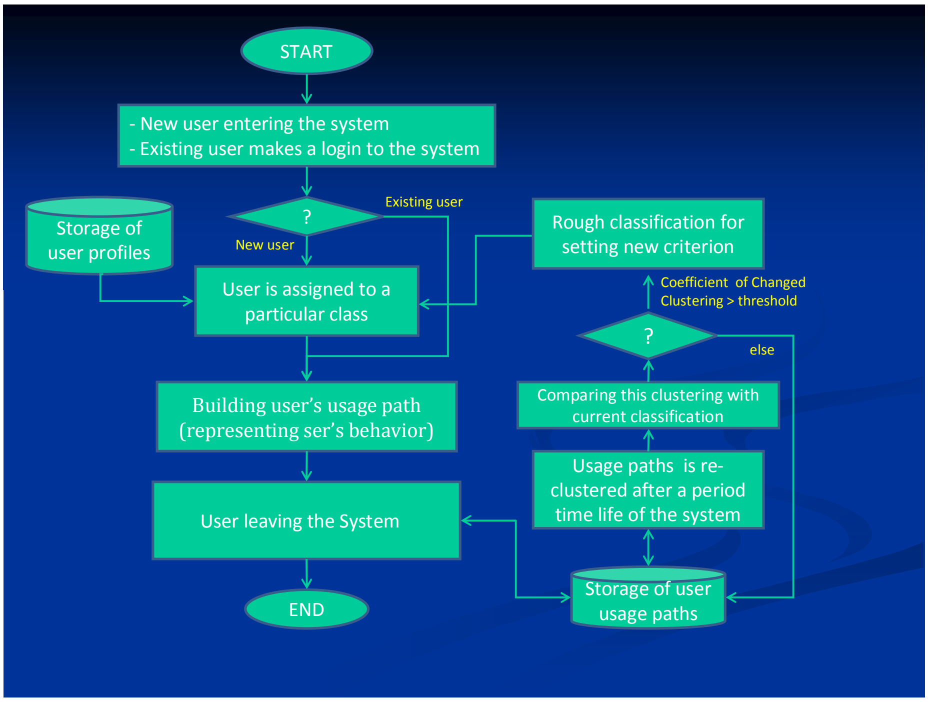
START
- New user entering the system
- Existing user makes a login to the system
?
Existing user
New user
Storage of user profiles
User is assigned to a particular class
Rough classification for setting new criterion
Coefficient of Changed Clustering > threshold
?
else
Building user's usage path (representing ser's behavior)
Comparing this clustering with current classification
Usage paths is re-clustered after a period time life of the system
User leaving the System
END
Storage of user usage paths

Figure 1: General idea of rough classification

Very often user profiles are described by a set of attributes. Thus the user classification criterion is a subset of this attributes set. Owing the clustering of the set of usage paths one can determine a subset of attributes which best reflects this clustering. This idea is presented in Figure 2.

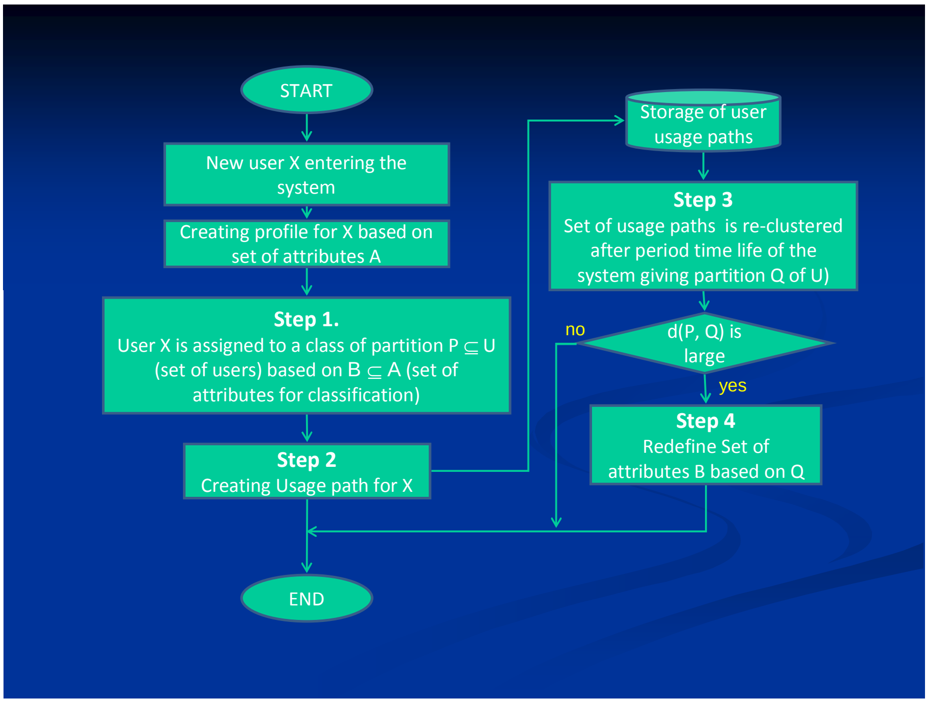
START
New user X entering the system
Creating profile for X based on set of attributes A
Step 1.
User X is assigned to a class of partition P ⊆ U (set of users) based on B ⊆ A (set of attributes for classification)
Step 2
Creating Usage path for X
Storage of user usage paths
Step 3
Set of usage paths is re-clustered after period time life of the system giving partition Q of U)
no
d(P, Q) is large
yes
Step 4
Redefine Set of attributes B based on Q
END

Figure 2: Rough classification process using attributes form user profiles

Let *A* be the set of attributes used for representing real world objects from set *U*. Let P_B be the classification of set *U* generated by attributes from set $B \subseteq A$ and let *Q* be some given classification of set *U*.

If P_B differs from *Q* (or their distance is greater than a given threshold) then we say that the system is weakly adaptive. The smaller is the difference between P_B and *Q* the more adaptive is the system. A system is then fully adaptive if $P_B = Q$ [Nguyen, 08].

To make a system more adaptive, one should solve the problem of determining set *B* on the basis of classification *Q*. Notice that P_B is dependent on *B*, this means that a set *B* determines exactly one classification P_B. Besides, we know that different sets of attributes may determine the same classification, thus the problem relies on selection of a minimal set *B* (i.e. a set with minimal number of elements) such that the distance between P_B and *Q* is minimal.

Let *d* be a distance function between classifications (partitions) of set *U*. The problems of Rough Classification (RC) can formally be defined as follows [Nguyen, 06], [Nguyen, 08]:

Problem RC-1:

For a given partition Q of set U one should determine set $B \subseteq A$ *such that* $P_B \subseteq Q$ *and* $d(P_B,Q)$ *is minimal.*

In this case set B generates such classification which is equal or minimally more exact than Q.

Problem RC-2:

For a given partition Q of set U one should determine set $B \subseteq A$ such that $Q \subseteq P_B$ and $d(P_B,Q)$ is minimal.

In this problem set B generates such classification which is equal to or minimally differs from Q.

Problem RC-3:

For a given partition Q of set U one should determine set $B \subseteq A$ such that $d(P_B,Q)$ is minimal.

Solutions of this problem may be needed if solutions of problems RC-1 and RC-2 do not exist. Thus set B generates a rough classification referring to Q. This problem should be solved in the majority of practical cases.

It should be emphasized that the difference between our approach and Pawlak's approach to rough classification is that we do not use the upper and lower approximations of partitions defined by Pawlak, but we use the distance functions between partitions to determining the nearest partition to the given. The problems RC-1, RC-2 and RC-3 defined in this section are novel and their solutions should help in determining the most effective (and also the most economic) set of attributes, which describe the user interfaces.

For these problems several algorithms have been worked out and presented in [2, 3] and their effectiveness has been analyzed [Kozierkiewicz, 07]. It is worth to note that problem RC-3 has been proved to be a NP-complete problem [Musial, 89]. Therefore, there have been worked out several heuristic algorithms, which have also been statistically verified referring to their effectiveness.

3 Applications

In a system where user management process is needed, the rough classification methods can be used. The user management process can be presented as follows: After registering to the system, a new user is classified by the system into an appropriate class of users according to his personal data, which are the basis for determining the best way for serving the user. We can distinguish the following sets:

- U – the set of users of the system;
- A – the set of all potential attributes, which may be used in representing the user profiles.

The best ways for serving users can then generated by means of a classification of users (SC) independently from the user classification (UC) generated by the values of attributes from set A. We may notice that it is the classification SC which should decide about the proper classification of the system users. The question is: *Which attributes from set A should be crucial in the classification UC, so that it is most similar to classification SC*? It follows that UC should be a rough classification referring to SC. The answer to this question can be very useful since owing to it the system will be more adaptive and the user classification process will be more economical.

The application of the rough classification method for intelligent e-learning systems is based on determining the minimal set of user data attributes which generate a similar classification to that generated by a learners clustering process. After these attributes of set *B* have been determined, they should enable to classify new learners more properly and as the consequence, to assign good scenarios to them.

An approach for using rough classification methods to perform the recommendation processes in intelligent e-learning systems has been proposed. Rough classification in this case is related to inconsistency aspect of knowledge of the system. The inconsistency here appears in two aspects: In the first aspect inconsistency refers to difference of the passed scenarios of similar learners (belonging to the same class of the classification). In this case to determine an opening scenario for a new learner it is needed to calculate the consensus of the passed scenarios of the members of the class. The second aspect of inconsistency refers to the fact that assumed to be similar learners (belonging to the same class of the classification) may have very miscellaneous passed scenarios. This, in turn, may cause a lack of efficiency of the procedure proposed for the first aspect. Here we propose to use a rough classification based method to redefine the criterion for classification [Nguyen, 08]. It turned out that the usage paths of the system consist of learning scenarios which have successfully used by the learners. On the basis of the clustering of learning scenarios the classification criterion has been generated.

Rough classification methods have also been used in designing and managing adaptive user interfaces [Nguyen, 06], [Sobecki, 07]. In this application the user data contain the geographic data, and the usage data contain user interface settings generated by the user choice during using the system. The clustering of interface settings helped in setting the proper set of attributes being the criterion for user classification.

4 Conclusions

The future works should concern more complex structures of user data. As to now their atomicity has been assumed. However, for representing the complexity and uncertainty other structures such as interval, set etc. are often used. For these structures problems RC-1, RC-2 and RC-3 require working our new algorithms. Besides, new application, for example in medical knowledge systems [Tabakov, 07] could be investigated.

Acknowledgement

This paper was partially supported by Grant no. N N519 407437 funded by Polish Ministry of Science and Higher Education (2009-2012).

References

[Kozierkiewicz, 07] Kozierkiewicz, A., Nguyen, N.T.: *The Statistical Verification of Rough Classification Algorithms*. In: Proc. of KES (1) 2007, Lecture Notes in Artificial Intelligence 4692, 2007, 238-245.

[Musial, 89] Musial, K., Nguyen, N.T.: *On the Nearest Product of Partitions*. Bull. of Polish Academy of Sci., 36(5-6), 1989, 333-338.

[Nguyen, 08] Nguyen, N.T.: Advanced Methods for Inconsistent Knowledge Management. Springer-Verlag London 2008.

[Nguyen, 05] Nguyen, N.T., Sobecki, J.: *Rough Classification Used for Learning Scenario Determination in Intelligent Learning Systems*. In: Kłopotek M et al. (eds) Intelligent Information Processing and Web Mining. Series "Advances in Soft Computing". Physica-Verlag, 2005, 107-116.

[Nguyen, 06] Nguyen, Sobecki, N.T., J.: *Determination of User Interfaces in Adaptive Systems Using a Rough Classification Based Method*. Journal of New Generation Computing 24, 2006, 377-402.

[Pawlak, 99] Pawlak, Z.: *Rough Classification*. Int. J Human-Computer Studies 51, 1999, 369-383.

[Sobecki, 07] Sobecki, J: *Web-Based System User Interface Hybrid Recommendation Using Ant Colony Metaphor*. In: Proceedings of KES 2007. Lecture Notes in Artificial Intelligence 4694, 2007, 1033-1040.

[Tabakov, 07] Tabakov, M.: *A fuzzy segmentation method for Computed Tomography images*. International Journal of Intelligent Information and Database Systems 1(1), 2007, 79-89.