# EXTENDING THE ROUGHNESS OF THE DATA VIA TRANSITIVE CLOSURES OF SIMILARITY INDEXES

F.X. Bertran, N. Clara, J.C. Ferrer

Girona University

One main assumption in the theory of rough sets applied to information tables is that the elements that exhibit the same information are indiscernible (similar) and form blocks that can be understood as elementary granules of knowledge about the universe. We propose a variant of this concept defining a measure of similarity between the elements of the universe in order to consider that two objects can be indiscernible even though they do not share all the attribute values because the knowledge is partial or uncertain. The set of similarities define a matrix of a fuzzy relation satisfying reflexivity and symmetry but transitivity thus a partition of the universe is not attained. This problem can be solved calculating its transitive closure what ensure a partition for each level belonging to the unit interval [0,1]. This procedure allows generalizing the theory of rough sets depending on the minimum level of similarity accepted. This new point of view increases the rough character of the data because increases the set of indiscernible objects. Finally, we apply our results to a not real application to be capable to remark the differences and the improvements between this methodology and the classical one.

***Keywords:*** *indiscernibility relation, rough sets, fuzzy similarity index, fuzzy relation, transitive closure*

## 1. INTRODUCTION

The theory of rough sets ([12], [13], [14], [16], [24]) has been very developed in the last two decades with a great success in the domains of engineering, applied sciences ([3]) and, specially, in artificial intelligence ([17], [18], [23]). This theory has a strong relation with fuzzy sets ([22], [25], [26]) and we seek to combine it with the theory of fuzzy relations and fuzzy similarity indexes ([10]). Obviously, we will focus our attention in those aspects related to social sciences ([7], [8], [21]).

Following the judgements and ideas of several authors in the area of social sciences it is widely accepted that the rough sets theory has some advantages in economical forecasting research with regard to traditional mathematical tools as mathematical functions or statistical models. This theory does not need any external information because it works with the original data, it is capable to analyze qualitative attributes, the information is given by the natural language of decision rules, and the results are easy to understand without requiring an interpretation of technical parameters as is the case of credit scoring, utility function or outranking relation.

The knowledge is usually described by qualitative valuations in an information table. This methodology has been extensively applied in economical fields. Interested results are usually reported in business failure prediction ([4], [5], [19], [20]), database marketing ([6], [11], [15]) and financial investment ([2]). Summarising, in models intending to predict bankruptcy ([9]), the elements (or objects) are the enterprises, the conditional attributes can be the Current Ratio and the Net income, and the decision attribute the fact that there is or not bankruptcy. In applications to Sightseeing Expenditures ([1]) objects are guests and attributes some characteristics about the hotel reservation (room type, number of nights stayed, payment method, computerized reservation system, and so on).

Knowledge and perception of the elements of the universe is a basis of the definition of a set in the rough set theory. From this point of view is possible that two different elements were seen as the same so they are indiscernible, which happens when all the attribute values are the same. In other terms means that the similarity between the objects is one. That is a consequence that in fact objects are only represented by their attribute values. What we propose is to softer this condition allowing to be indiscernible elements with similarity smaller than the unit. This new approach gives as a result different partitions depending on the level of similarity accepted by the experts.

### 1.1. INFORMATION SYSTEM

Let $U = \{x_1,\ldots,x_n\}$ be an universe of elements or objects, $A = \{a_1,\ldots,a_m\}$ a set of attributes with $A = C \cup D$ where $C$ and $D$ are the sets of conditional and decision attributes (usually $\text{card}(D) = 1$) respectively. Let $V = \bigcup_{i=1}^{m} V_{a_i}$ be the whole set of attribute values where $V_{a_i}$ is the set of values that takes $a_i$. Finally, let $\rho : U \times A \to V$ be a map with set domain all the pairs formed by elements and attributes and set image (or range) all the possible values. Under these suppositions $(U, A, V, \rho)$ is called an information system and the table grouping all the values an information table.

### 1.2. SIMILARITY INDEXES

In the methodology that we propose it is essential to determine how similar the objects are. The theory of rough sets includes this point of view in the sense that if all the attribute values that define two objects are identical then the similarity between them is one. Many ways for calculating the similarity have been proposed, discussed, analyzed and used. Selecting an appropriate index of similarity will be fundamental to achieve suitable results.

Literature about this topic suggests different kind of measures of similarities according to types of scales in measuring data. Fuzzy data belong to the unit interval then a model based on the metric space can be taken in account. Associating to each object a m-dimensional vector and calculating their similarity by means of a decreasing function $s = f(d)$ (usually $s = 1 - d$) of their normalized distance $d$ seems a very logical procedure that provides a mathematical tool with appropriate geometrical properties.

On the other hand, for binary data the vectors associated to each object are successions of 0 and 1. In these conditions is preferable a set-theoretical model based on the cardinality of the common and distinctive features. Some well-known similarity measures are the simple matching coefficient, the Jackard coefficient and the Rao's coefficient. In our context we deal with fuzzy data which can be interpreted as a generalization of binary data. Several fuzzy similarity measures have been introduced from those concerning binary data simply generalizing the cardinal as the addition of the different values of the membership function and making some crisp simplifications. To deal with the fuzzy character of the data a weaker definition (in comparison of the metric space model) of fuzzy similarity measure is usually taken in account in some applications: $s : \tilde{\wp}(U) \times \tilde{\wp}(U) \to [0,1]$ is a fuzzy similarity measure if and only if $0 \le s(\tilde{x}_i, \tilde{x}_j) \le 1$ and $s(\tilde{x}_i, \tilde{x}_j) = s(\tilde{x}_i, \tilde{x}_j)$ for any pair of elements. The reflexive property ($\forall \tilde{x} \in \tilde{\wp}(U) \quad s(\tilde{x}, \tilde{x}) = 1$) is strongly recommended and used in almost all theoretical and applied papers. These kind of fuzzy relations are very important in fuzzy clustering for fuzzy relational data what is not surprising because of the closeness between the clustering and similarity concepts. We will refer to all them as fuzzy similarity indexes ([10]). These indexes define a symmetric similarity square matrix with the unit element in the main diagonal therefore is the matrix of a fuzzy proximity relation.

For applications we have chosen the fuzzy simple matching coefficient because is the generalization of the simple matching coefficient for binary data and is associated to the normalized distance of Hamming (the normalized $L^1$-distance in functional analysis) so verifies both approaches defined above. Moreover, is very often used in fuzzy economical applications. This index is defined as

follows:

$$s(\tilde{x}_i, \tilde{x}_j) = 1 - d_H(\tilde{x}_i, \tilde{x}_j) = 1 - \frac{1}{m} \sum_{k=1}^{m} \left| \mu_{\tilde{x}_i}(a_k) - \mu_{\tilde{x}_j}(a_k) \right| \qquad (1)$$

## 2. ROUGH SETS DEPENDING ON A LEVEL OF SIMILARITY

One main assumption in the theory of rough sets applied to information tables is that the elements that exhibit the same information are indiscernible (similar) and form blocks that can be understood as elementary granules of knowledge about the universe. In fact, from a mathematical point of view, two elements of the universe are related if their values for all attributes are the same. This relation is an equivalence relation called indiscernibility relation. Equivalence classes of the indiscernibility relation are referred to as elementary sets. Following this idea we can interpret that in certain conditions of imprecision about the language or the seizure of data, two elements can be thought as similar if their similarity is greater that a certain value belonging to the unit interval $[0,1]$ and not only if their similarity is equal to one.

### 2.1. VALUATION OF THE INFORMATION TABLE

In order to find the similarities between the elements of the universe we need to establish a valuation for all the elements of the information table. That means to define $v : U \times A \to [0,1]$ where $v(x_i, a_j)$ is the numerical value that takes the categorical value $\rho(x_i, a_j)$. Obviously, $v$ has to be a "monotone" function of the attributes defined by an expert namely, if the categorical values are ordered in an increasing sequence, for instance: no, very low, low, regular, high, very high and yes, then, the numerical values have to be an increasing list of numbers between 0 and 1. This valuation depends on the experts but as is applied to the entire table the differences of criteria are quite irrelevant. The substitution of attribute values for numbers between 0 and 1 defines a fuzzy subset $\tilde{x}_i = (v(x_i, a_l), \ldots, v(x_i, a_p))$ for each element of the universe. All the values define the matrix of a fuzzy relation between the universe and the set of attributes. From now on we identify elements with their fuzzy subsets defined in the valuation.

### 2.2. GENERATING AN INDISCERNIBILITY RELATION AT DIFFERENT LEVELS OF SIMILARITY

Let $x, y$ be elements of the universe and $B \subset A$ with $\text{card}(B) = r \leq m$. Each

fuzzy subset $\tilde{x}_i$ defines a r-dimensional vector and the similarity between elements can be calculated by means any fuzzy similarity index identifying elements with fuzzy subsets namely, $s(x_i, x_j) = s(\tilde{x}_i, \tilde{x}_j)$. Selecting α belonging to the unit interval [0,1] and defining the relation $x_i T x_j$ if and only if $s(x_i, x_j) \geq \alpha$ seems a logical procedure to deal with our objective. Even though this crisp relation is reflexive and symmetric unfortunately is not transitive, so is not an equivalence relation. Therefore we can not calculate directly with the primary similarities between elements because the elementary sets could not form a partition.

Let $R = (r_{ij})$ be the similarity matrix defined by the selected similarity index so $r_{ij} = s(\tilde{x}_i, \tilde{x}_j)$. Notice that the classical indiscernibility relation $x_i Ind(B) x_j$ if and only if $a(x_i) = a(x_j)$ for all $a \in B$ implies that $\tilde{x}_i = \tilde{x}_j$ so $s(x_i, x_j) = s(\tilde{x}_i, \tilde{x}_j) = 1$. This relation is a proximity fuzzy relation in the universe of objects but an equivalence fuzzy relation because does not verify transitivity. In order to achieve an equivalence fuzzy relation we calculate its max-min transitive closure. This strategy depends on the selection of the fuzzy similarity index and allows finding a partition of the universe depending on the level of similarity considered. A very important theorem proves that the partition obtained from the transitive closure is the same that with the hierarchical method of single linkage and the fuzzy connected components of the fuzzy graph defined by the matrix ([10]). The transitive closure of R is defined by:

$$R^* = \min\{S : R \subset S \text{ and } S \text{ is an equivalence fuzzy relation}\} \qquad (2)$$

We note $R^* = (r_{ij}^*)$. Many methods are available to find the transitive closure; the best known is the power's method which consists in finding when stabilizes the powers of the matrix that defines the fuzzy relation. Defining the powers of the matrix as $R^2 = (r_{ij}^2)$ where $r_{ij}^2 = \max_k(\min(r_{ik}, r_{kj}))$ and in a similar way for $R^n, n > 2$, it is easy to prove that $R \leq R^2 \leq R^3 \leq \ldots \leq R^n$. Therefore, the transitive closure is $R^* = R^m$ where m is the minimum value that verifies $R^m = R^{m+1}$. From this point of view:

$$R^* = \bigcup_{\alpha \in [0,1]} Ind(B(\alpha)) \qquad (3)$$

Relation $Ind(B(\alpha))$ can be written as $Ind(B(\alpha)) = \alpha Ind(B_\alpha)$. Relation $B_\alpha$ is a crisp equivalence relation defined in the universe. In analogy with the original ideas of Z. Pawlak we designate $B_\alpha$ as indiscernibility relation at level α.

### 2.3. MAIN DEFINITIONS

1. The equivalence classes are referred as $B(\alpha)-$elementary sets (or B elementary sets at level $\alpha$).
2. The equivalence class $B(\alpha)[x]$ is the block of the partition containing $x$.
3. Two elements $x$ and $y$ belonging to the same equivalence class are $B(\alpha)$-indiscernible (or B-indiscernible at level $\alpha$).
4. B-Lower approximation at threshold $\alpha$:

$$B_*(\alpha)(X) = \{x \in U : B(\alpha)[x] \subset X\} \tag{4}$$

5. B-Upper approximation at threshold $\alpha$:

$$B^*(\alpha)(X) = \{x \in U : B(\alpha)[x] \cap X \neq \varphi\} \tag{5}$$

6. Boundary region of a set $X$ at level $\alpha$:

$$BN_B(\alpha)(X) = B^*(\alpha)(X) - B_*(\alpha)(X) \tag{6}$$

7. Rough Set at threshold $\alpha$:

$$(B_*(\alpha)(X), B^*(\alpha)(X)) \tag{7}$$

8. Accuracy of the approximation at level $\alpha$:

$$\lambda_{B(\alpha)}(X) = \frac{|B_*(\alpha)(X)|}{|B^*(\alpha)(X)|} \tag{8}$$

Notice that, in applications, it is necessary that an expert defines the minimum accepted level for considering two elements as similar in order to select the appropriated partition. We will call this parameter lower threshold of similarity.

Summarising, the first step is to assign to each element of the information table a number in the unit interval. The second step is selecting an index of similarity that allows calculating similarities between the elements of the universe. These similarities define a fuzzy proximity relation. After that, we calculate its transitive closure which provides a partition of the universe depending on the alpha level belonging to the unit interval. Once fixed the lower threshold of similarity we deduce a unique partition (its elements are the equivalence classes). It is relevant to remark that when the lower threshold of similarity is equal to one we obtain the same partition (or another less fine) that we would obtain with the usual theory of rough sets what means that this new approach includes the classical one.

We would define in a similar way other concepts concerning the reduction of attributes: reduct at level α, reduct D-α-indispensable, attribute D-α-dispensable and attribute D-dispensable for an element of the universe. Future researches will study in depth on these concepts.

## 3. EXAMPLE

Finally, we apply our results to a not real application to be capable to remark the differences and the improvements between this new methodology and the usual one, namely, without and with the introduction of the similarities between the elements of the universe, the calculus of the transitive closure of the matrix of similarities, the selection of the lower threshold of similarity parameter and the generalization of the main concepts of rough sets.

Suppose we are given some data about six economical subjects (companies, guests…) so $U=\{x_1,x_2,x_3,x_4,x_5,x_6\}$ and $A=\{a_1,a_2,a_3,a_4\}$ where $C=\{a_1,a_2,a_3\}$ are conditional attributes and $D=\{a_4\}$ a decision attribute. The different values of the data are shown in the information table represented in Table 1 and the corresponding valuation in Table 2.

**Table 1**

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|
| $x_1$ | No | Yes | High | Yes |
| $x_2$ | Yes | No | High | Yes |
| $x_3$ | Yes | Yes | Very high | Yes |
| $x_4$ | No | Yes | Normal | No |
| $x_5$ | Yes | No | High | No |
| $x_6$ | No | Yes | Very high | Yes |

**Table 2**

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|
| $x_1$ | 0 | 1 | 0.7 | Yes |
| $x_2$ | 1 | 0 | 0.7 | Yes |
| $x_3$ | 1 | 1 | 0.9 | Yes |
| $x_4$ | 0 | 1 | 0.5 | No |
| $x_5$ | 1 | 0 | 0.7 | No |
| $x_6$ | 0 | 1 | 0.9 | Yes |

As similarity index we have chosen the normalized distance of Hamming (1).

Taken $C = B = \{a_1, a_2, a_3\}$ the similarity matrix is the following:

$$R = \begin{pmatrix} 1 & 0.33 & 0.6 & 0.93 & 0.33 & 0.93 \\ 0.33 & 1 & 0.6 & 0.27 & 1 & 0.27 \\ 0.6 & 0.6 & 1 & 0.53 & 0.6 & 0.66 \\ 0.93 & 0.27 & 0.53 & 1 & 0.26 & 0.86 \\ 0.33 & 1 & 0.6 & 0.26 & 1 & 0.26 \\ 0.93 & 0.27 & 0.66 & 0.86 & 0.26 & 1 \end{pmatrix}$$

Calculating we deduce $R \neq R^2$, $R^2 \neq R^3$ but $R^3 = R^4$ so $R^* = R^3$ and therefore:

$$R^* = \begin{pmatrix} 1 & 0.6 & 0.66 & 0.93 & 0.6 & 0.93 \\ 0.6 & 1 & 0.6 & 0.6 & 1 & 0.6 \\ 0.66 & 0.6 & 1 & 0.66 & 0.6 & 0.66 \\ 0.93 & 0.6 & 0.66 & 1 & 0.6 & 0.93 \\ 0.6 & 1 & 0.6 & 0.6 & 1 & 0.6 \\ 0.93 & 0.6 & 0.66 & 0.93 & 0.6 & 1 \end{pmatrix}$$

From the previous matrix we obtain the elementary sets at different levels as defined in (3).

For $0.93 < \alpha \leq 1$, the elements of the partition are: $\{x_1\}$, $\{x_2, x_5\}$, $\{x_3\}$, $\{x_4\}$, $\{x_6\}$.
For $0.66 < \alpha \leq 0.93$ the elements of the partition are: $\{x_1, x_4, x_6\}$, $\{x_2, x_5\}$, $\{x_3\}$.
For $0.6 < \alpha \leq 0.66$ the elements of the partition are: $\{x_1, x_3, x_4, x_6\}$, $\{x_2, x_5\}$.
For $\alpha \leq 0.66$ only one element define the partition $\{x_1, x_2, x_3, x_4, x_5, x_6\}$

Choosing $0.66 < \alpha \leq 0.93$ and $X = \{x_1, x_2, x_3, x_6\}$ what means flu, we obtain the lower-approximation, upper-approximation, boundary region and accuracy following definitions (4), (5), (6) and (8).

B-Lower approximation at threshold $\alpha$: $aB_*(\alpha)(X) = \{x_3\}$.
B-Upper approximation at threshold $\alpha$: $B^*(\alpha)(X) = \{x_1, x_2, x_3, x_4, x_5, x_6\}$.
Boundary region: $BN_B(\alpha)(X) = \{x_1, x_2, x_4, x_5, x_6\}$.
Rough Set at threshold $\alpha$: $(\{x_3\}, \{x_1, x_2, x_3, x_4, x_5, x_6\})$.

Accuracy at level $\alpha$: $\lambda_{B(\alpha)}(X) = 1/6$.

This methodology could be done for any other subset $B \subset A$.

## 4. CONCLUSIONS

Considering as similar the objects that have not only the same valuations for all the attributes but those that their attribute values have a certain degree of similarity, we are capable to build a partition depending on the selected level of similarity. From this new point of view we achieve a generalization of the main concepts related with the theory of rough sets. Considering the maximum level of similarity equal to one we conclude that the classical methodology of Z. Pawlak is contained in this new procedure. Our hope is that this point of view will contribute in an increase of flexibility and adaptability in applications.

**REFERENCES**

[1] AU, N.; LAW, R. (2000). "The Application of Rough Sets to Sightseeing Expenditures". *Journal of Travel Research,* Vol. 39, No. 70, p. 70-77.
[2] BAZAN, J.G.; SKOWRON, A.; SYNAK, P. (1994). "Market Data Analysis: a Rough Sets Approach". *ICS Research Reports,* 6/94, Warsaw University of Technology.
[3] COLETTE, T.W.; SZLADOW, A.J. (1994). "Use of Rough Sets and Spectral Data for Building Predictive Models of Reaction Rate Constants". *Applied Spectroscopy,* Vol. 48, No. 11, p. 1379-1386.
[4] DIMITRAS, A.I.; SLOWINSKI, R.; SUSMAGA, R.; ZOPOUNIDIS, C. (1999). "Business Failure Prediction Using Rough Sets". *European Journal of Operational Research,* Vol. 114, No. 2, p. 263-280.
[5] GRECO, B.; MATARAZZO, B.; SLOWINSKI, R. (1998). "A New Rough Set Approach to Evaluation of Bankruptcy Risk". *Operational Tools in the Management of Financial Risk*, p. 121-136. Boston, Kluwer Academic Publishing.
[6] KOWALCZYK, W. (1998). "Rough Data Modelling: a New Technique for Analyzing Data". *Rough Sets in Knowledge Discovery* 1, p. 400-421. Heildelberg, Physica-Velag.
[7] KOWALCZYK W.; PIASTA Z. (1998). "Rough Set-Inspired Approach to Knowledge Discovery in Business Databases". *Proceedings of the 2nd Pacific-Asia Conference (PAKDD-98)*. Springer, p. 186-197.
[8] LIXIANG, S.; TONG, H. (2004). "Applying Rough Sets to Market Timing Decisions". *Decision Support Systems, Data mining for financial decision making,* Vol. 37, No. 4, p. 583-597.
[9] MCKEE, T.E. (2000). "Developing a Bankruptcy Prediction Model via Rough Sets Theory". *International Journal of Intelligent Systems in Accounting, Finance & Management,* Vol. 9, p. 159-173.
[10] MIYAMOTO, S. (1990). *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Dordrecht, Kluwer Academic Publishers.
[11] MROZEK, A.; SKABEK, K. (1998). "Rough Sets in Economic Applications". *Rough Sets in Knowledge Discovery* 2, p. 238-271. Heildelberg, Physica-Verlag.
[12] PAWLAK, Z. (1982). "Rough Sets". *International Journal of Computer and Information Sciences*, Vol. 11, No. 5, p. 341-356.
[13] PAWLAK, Z. (1991). *Rough Sets-Theoretical Aspects of Reasoning about Data*. Dordrecht, Kluwer Academic Publishers.
[14] PAWLAK, Z.; SKOWRON, A. (1994). *Rough Membership Functions, Advances in the Dempster*

*Shafer Theory of Evidence*, p. 251-271. New York, John Wiley & Sons, Inc.
[15] POEL, D. (1998). "Rough Sets for Data Base Marketing". *Rough Sets in Knowledge Discovery 2,* p. 324-335. Heildelberg, Physica-Verlag.
[16] POLKOWSKI, L. (2002). *Rough Sets-Mathematical Foundations, Advances in Soft Computing*. Physica Verlag, Springer-Verlag Company.
[17] SKOWRON, A.; RAUSCER, C. (1992). "The Discernibility Matrices and Functions in Information Systems, Intelligent Decision Support". *Handbook of Applications and Advances of the Rough Set Theory*, p. 311-362. Dordrecht, Kluwer Academic Publishers.
[18] SKOWRON, A. (2002). "Rough Set Perspective on Data and Knowledge". *Handbook of Data Mining and Knowledge Discovery*, p. 134-149. Oxford University Press.
[19] SLOWINSKI, R.; ZOPOUNIDIS, C., (1995). "Application of the Rough Set Approach to Evaluation of Bankruptcy Risk". *Intelligent Systems in Accounting, Finance and Management,* Vol. 4, No. 1, p. 27-41.
[20] SLOWINSKI, R.; ZOPOUNIDIS, C.; DIMITRAS, A.I.; SUSMAGA, R. (1999). "Rough Set Predictor of Business Failure". *Soft Computing in Financial Engineering*, p. 402-424. New York, Physica-Verlag.
[21] SZLADOW, A.; MILLS, D. (1993). "Tapping Financial Databases". *Business Credit,* Vol. 95, No. 7, p. 8.
[22] YAO, Y.Y. (1995). "On Combining Rough and Fuzzy Sets". *Proceedings of the CSC'95 Workshop on Rough Sets and Database Mining*, San Jose State University.
[23] YAO, Y.Y.; LIN, T.Y. (1996). "Generalization of Rough Sets Using Modal Logic". *Intelligent. Automation. And Soft Computing International Journal*, Vol. 2, p. 103-120.
[24] YAO, Y.Y.; WONG, S.K.M.; WANG, L.S. (1995). "A Non-Numeric Approach to Uncertain Reasoning". *International Journal of General Systems*, Vol. 23, No. 4, p. 343-359.
[25] ZADEH, L.A. (1965). "Fuzzy Sets". *Inform. & Control*, Vol. 8, p. 338-353.
[26] ZIARKO, W.P. (1994). *Rough Sets, Fuzzy Sets and Knowledge Discovery*. London, Springer-Verlag.