Clustering: A Rough Set Approach to Constructing Information Granules

James F. Peters[1], Andrzej Skowron[2], Zbigniew Suraj[3], Wojciech Rząsa[4], Maciej Borkowski[1]

Abstract
This article introduces an approach to constructing clusters based on rough set theory. An algorithm for finding clusters of sample sensor signal values is introduced using a measure of closeness of information granules and a distance metric defined relative to a partition of the universe into equivalence classes containing elements that are considered indistinguishable from each other. The elements of each equivalence class are associated with what is known as a cell (or box) of a δ-mesh. The idea of an indistinguishability relation is briefly presented. It is this indistinguishability relation that underlies the construction each δ-mesh. Closeness of information granules is determined by measuring the separation between cells in a δ-mesh. The "center" of every cluster is a cell in a δ-mesh that is found using a measure of maximal rough inclusion. A cluster is constructed by starting with a particular cell in a δ-mesh and then gathering in all elements along the borders of the center and neighboring cells in the δ-mesh. The harvest of clusters continues until all non-empty cells in a δ-mesh have been considered. The problem of discovering clusters themselves as well as the size of a population clusters has been motivated by recent studies of rough neural networks and the classification of sensor signals. A sensor signal is a non-empty, finite, temporally ordered set of sample sensor signal values. Classification of sensor signals requires measurements of sample signal values over subintervals of time. The contribution of this article is the introduction of a rough set approach to constructing clusters.

Keywords: Closeness, cluster, inclusion, indistinguishability, information granule, measure, rough sets, sensor.

"Jeeves," I said, "you have heard?"
"Yes, sir."
"The position is serious."
"Yes, sir."
"We must cluster round."
"Yes, sir."
—P.G. Wodehouse, *Carry on, Jeeves!*, 1927

1. Introduction

A method of constructing clusters based on rough set theory is introduced in this article. This research is part of rapidly growing research on information granulation, granular computing and computing with words introduced by Zadeh [9] and the calculus of granules [5, 6], [8]. In this research, a cluster is a temporally ordered set of real-world objects (e.g., sample sensor signal values). Such clusters are constructed from vectors of real numbers using a variant of the traditional indiscernibility relation in rough set theory [1]. Clusters are associated with indiscernibility classes containing sample signal values that occur in precisely defined temporal intervals. The study of clusters of temporally ordered data has been considered by others (see, e.g., [13]). The partition of a universe of objects into information granules containing equivalent objects (i.e., equivalence classes) provides a basis for an algorithm to construct clusters. This partition of the universe is accomplished with a parameterized indistinguishability relation

[1] Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Manitoba R3T 5V6 Canada, jfpeters@ee.umanitoba.ca
[2] Institute of Mathematics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland, skowron@mimuw.edu.pl
[3] University of Information Technology and Management, H. Sucharskiego 2, 35-225 Rzeszów, Poland, zsuraj@wenus.wsiz.rzeszow.pl
[4] Institute of Mathematics, Rzeszów University, Rejtana 16A 35-310 Rzeszów, Poland, wrzasa@univ.rzeszow.pl

Ing introduced in [12], which results in what is known as a δ-mesh. What is new in the method of construction of clusters of temporally ordered data presented in this paper is the discovery of clusters resulting from the measurements of inclusion and closeness applied to δ-mesh cells. The presentation is limited to a specialized model of cluster construction. Consideration of a generalized cluster construction method is outside the scope of this article. The new clustering construction method has a number of practical applications, e.g., data mining [11], performance maps [14], sensor fusion [4], signal analysis [3], robotics [3], and neurocomputing [7]. The contribution of this article is the introduction of a rough set approach to constructing clusters of temporally ordered elements.
This paper is organized as follows. Section 2 presents introduces the parameterized indistinguishability relation and measures of inclusion based on a rough membership set function. The idea of a δ-mesh and maximal rough inclusion are also briefly introduced in this section. A specialized model for the construction of clusters of temporally ordered data in the context of a δ-mesh is presented in Section 3. A sample construction of clusters is also given in this section.

2. Indistinguishability and Information Granule Inclusion

In laying the groundwork constructing clusters in the context of rough set theory [1], we introduce a parameterized indistinguishability relation *Ing* to partition universes of reals. That is, consider a universe that is a subset of the reals, where set approximation and measurement can be carried out relative to a partition of such a universe into equivalence classes (subintervals that are sources of granules of information about the sensorial world). This partition is accomplished using *Ing*. To see this, let $S = (U, A)$ be an information system where U is a non-empty set and A is a non-empty, finite set of attributes, where $a: U \to V_a$ *and* $V_a \subseteq \Re$ for every $a \in A$. Let $a(x) \geq 0$, $\delta > 0$ and let $\lfloor a(x)/\delta \rfloor$ denote the greatest integer less than or equal to $a(x)/\delta$ ("floor" of $a(x)/\delta$). In case $a(x) < 0$ $\lfloor a(x)/\delta \rfloor$ denotes the greatest integer bigger then $a(x)/\delta$. For each $B \subseteq A$, there is associated an equivalence relation $Ing_{A,\delta}(B)$ defined as follows:

$$Ing_{A,\delta}(B) = \{(x, x') \in U \times U \mid \forall a \in B.\ \lfloor a(x)/\delta \rfloor = \lfloor a(x')/\delta \rfloor\}$$

Remark. In this paper, clustering is a relation defined on vectors of real numbers. If $B = \{a_1,\ldots,a_m\}$ and $U = \Re^{m+1}$ then an indistinguishability relation is defined on objects $(x, x_1,\ldots, x_m) = (x(t), a_1(x(t)), ..., a_m(x(t))$ and the indistinguishability classes are cells. Two vectors $(x, x_1,\ldots, x_m)$ and $(y, y_1,\ldots, y_m)$ are indistinguishable if, and only if,

$\lfloor x/\delta \rfloor = \lfloor y/\delta \rfloor$ and $\lfloor a_i(x_i)/\delta \rfloor = \lfloor a_i(y_i)/\delta \rfloor$ for $i = 1, \ldots, m$.

The parameter δ serves as a means of computing a "neighborhood" size on real-valued intervals. Elements within the same subinterval bounded by $k\delta$ and $(k+1)\delta$ for integer k are considered δ–indistinguishable. Each partition $U/\ Ing_{A,\delta}(B)$ is called a δ-mesh. Every equivalence class (also called a δ-mesh cell) is numbered by a pair of indices $E_{n,m}$ where $n = \lfloor x/\delta \rfloor$ and $m = \lfloor a(x)/\delta \rfloor$. If $(x, x') \in Ing_{A,\delta}(B)$, we say that objects x and x' are indistinguishable from each other relative to attributes from B. The notation $[x]_B^\delta$ denotes equivalence classes of $Ing_{A,\delta}(B)$ in $U/\ Ing_{A,\delta}(B)$. Let ρ be a measure on $\wp(U)$, where $\wp(U)$ is the powerset of U. For any $X \in \wp(U)$, define $\mu_x^{B,\delta} : \wp(U) \to [0,1]$ as in (1).

$$\mu_x^{B,\delta}(X) = \frac{\rho(X \cap [x]_B^\delta)}{\rho([x]_B^\delta)} \quad (1) \qquad \mu_x^{B,\delta}(X) = \frac{\rho(X \cap [x]_B^\delta)}{\rho([x]_B^\delta)} = \frac{\int_{X \cap [x]_B^\delta} 1 d\rho}{\int_{[x]_B^\delta} 1 d\rho} \quad (2) \qquad \mu_{\max}^{B,\delta}(X) = \max_{x' \in U}\{\frac{\rho(X \cap [x']_B^\delta)}{\rho([x']_B^\delta)}\}$$

(3)

Also, $\mu_x^{B,\delta}$ is a measure of rough inclusion of X in $[x]_B^{\delta}$. If $\rho([x]_B^{\delta})=0$, then $\rho(X \cap [x]_B^{\delta}) = 0$ and we define 0/0 (division by 0) to be equal to 0. For example, $\mu_x^{B,\delta}$ in (1) can be computed using (2). A model for computing maximal rough inclusion of set X in a δ-mesh is given in (3). It should also be noted that each choice of δ results in a particular δ-mesh.

3. Specialized Model for Construction of Clusters

A specialized model for constructing clusters of sample sensor signal values is introduced in this section. Intuitively, the method of construction of clusters in a δ-mesh starts with the identification of the δ_0-mesh cell $[x_1]_{\{a\}}^{\delta_0}$ with the maximal overlap with a set X measured using (3). Then $[x_1]_{\{a\}}^{\delta_0}$ is the "center" of cluster C_1. Nearby elements of mesh cells along the border of C_1 are then incorporated into C_1. After removing the elements of C_1 from the δ_0-mesh, the remaining clusters are constructed in the same manner until no non-empty δ_0-mesh cells are left. In what follows, we consider only the special case where $i \in \{0, 1, 2, 3\}$ and set B of attributes contains one element a. We assume δ_i is of the form δ_0/k for arbitrary positive integer k and $i = 1, 2, 3$. To begin, let $\lfloor x/\delta \rfloor = n_1$, $\lfloor \lceil a(x) \rceil / \delta \rfloor = m_1$, $\lfloor x'/\delta \rfloor = n_2$, $\lfloor \lceil a(x') \rceil / \delta \rfloor = m_2$ and let $d_0([x]_{\{a\}}^{\delta_0}, [x']_{\{a\}}^{\delta_0}) = \max\{|m_1 - m_2|, |n_1 - n_2|\}$. For simplicity, a small selection of d-values δ_0, δ_1, δ_2, δ_3 are considered where $\delta_{i+1} \le \delta_i$. Then consider a set of δ_i-mesh cells U_i such that $U_i = \{[x']_{\{a\}}^{\delta_i} \mid d_i\left([x']_{\{a\}}^{\delta_i}, [x]_{\{a\}}^{\delta_0}\right) \le \sum_{j=1,..,i} \delta_j\}$ where $[x_1]_{\{a\}}^{\delta_0}$ is a center of cluster and d_i is a measure of distance separating a single cell $[x']_{\{a\}}^{\delta_i}$ and set of cells $\{[x'']_{\{a\}}^{\delta_i} \mid x'' \in [x]_{\{a\}}^{\delta_0}\}$ in a δ_i-mesh, given by formula $d_i\left([x']_{\{a\}}^{\delta_i}, [x]_{\{a\}}^{\delta_0}\right) = \min_{x'' \in [x]_{\{a\}}^{\delta_0}} \left\{d_i([x']_{\{a\}}^{\delta_i}, [x'']_{\{a\}}^{\delta_i})\right\}$.

Algorithm [Specialized Cluster Construction]

Input: Information system $S = (U, A)$, set $X \subseteq U$ and $\{a\} = B \subseteq A$, numbers δ_0, δ_1, δ_2, δ_3 (decide about 'sizes' of equivalence classes), cardinality measure ρ on the set U, maximum metric d_i on the set $U / Ing_{A,\delta_i}(B)$ for $i = 0, 1, 2, 3$.

Output: Clusters of X.

Step 1. $j := 1$.

Step2. If $X \neq \emptyset$, find $\mu_{\max}^{\{a\},\delta_0}(X)$ given by (3) and one of $x \in U$ realizes the maximum and set $i := 1$, else go to Step 6.

Step3. If $i \le 3$, find $\mu_x^{\{a\},\delta_i}(X)$ given by (1) else go to Step 5.

Step4. If $\mu_x^{\{a\},\delta_i}(X) \ge 0.5 \cdot \mu_{\max}^{\{a\},\delta_0}(X)$ then $i := i + 1$ and go to Step 3.

Step5. $C_j := U_{i-1}$, $X := X - C_j$, $j := j + 1$ and go back to Step 2.

Step6. Return all C_j determined up to now, then STOP.

Example. Consider information system $S = (U, A)$ where $U = [0, 1] \times [0, 0.8]$. For a distinguished set X represented in Fig. 1, the cluster algorithm is applied. In this example, the following parameters are used: $\delta_0 = 0.2$, $\delta_1 = 0.5\delta_0$ $\delta_2 = \delta_3 = 0.25\delta_0$, and measure ρ denotes the number of points of a cluster. For simplicity, only elements of X are shown in the plot in Figs. 1 and 2. In addition, it is also assumed that every equivalence class determined by relation $\mathrm{Ing}_{A,\delta}(B)$ in Fig. 1 contains 8 points distributed in a uniform way. Table 1 provides a trace of the execution of the cluster algorithm on the partition shown in Fig. 2, where the "center" of C_1 is $E_{1,0}$.

As a result of applying the cluster algorithm to the partition in Fig. 1, we obtain clusters C_1, C_2, and C_3 shown in Fig. 1.

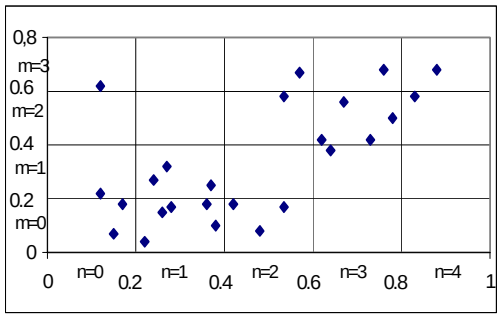

Fig. 1 Sample δ-mesh

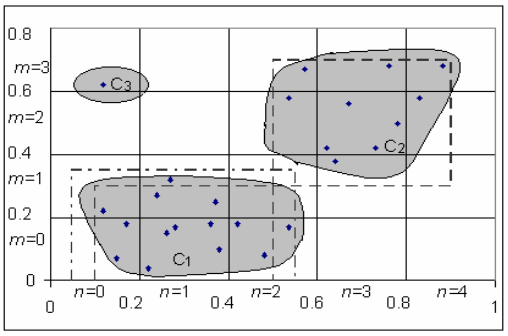

Fig. 2 Sample Clusters of Sensor Values

Table 1 Trace of Computation Cycles in Cluster Algorithm

variables	Computation Cycle in Cluster Algorithm					
j	1			2		3
$\mu_{\max}^{\{a\},\delta}(X)$	5/8=0.625			4/8=0.5		¼=0.25
i	1	2	3	1	2	1
$\mu_{x}^{\{a\},\delta_i}(X)$	12/24=0.5	14/34=0.41	14/48=0.29	10/32=0.31	10/50=0.2	1/18=0.06
	0.5>0.3125	0.41>0.3125	0.29<0.3125	0.31>0.25	0.2<0.25	0.06<0.125

4. Conclusion

An approach to constructing clusters of temporally ordered data been presented in the context of rough set theory. The parameterized indistinguishability relation *Ing* that is a slight variation of the traditional indiscernibility relation has been introduced to make it possible to identify elements that are considered "indistinguishable" from each other. The partition of a universe using *Ing* results in a mesh of cells (called a δ-mesh), where each cell of the δ-mesh represents an equivalence class. The configuration of cells in a δ-mesh yields a number of useful measures, namely, inclusion and closeness. A measure inclusion of a subset of the universe is defined relative to cells of a δ-mesh. A measure of closeness of a pair of information granules contained in cells of the δ-mesh results from determining the number of cells separating members of the pair using a distance metric. For the sake of illustration and understanding, the method of construction of clusters has been highly specialized in this paper. In later work, the intention is to consider the implications of the cluster construction method, its generalization, and its application. It should also be observed that other cluster construction methods and their application in rough neurocomputing are possible (see, e.g., [3]).

Acknowledgements

The research of James F. Peters has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) research grant 194376 and grants from Manitoba Hydro and the University of Information Technology and Management (UITM), Rzeszów, Poland. The research of Andrzej Skowron and Zbigniew Suraj has been supported by grant 8 T11C 025 19 from the State Committee for Scientific Research (KBN) in Poland. Moreover, the research of Andrzej Skowron has been supported by a grant from the Wallenberg Foundation in Sweden.

References

[1] Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning About Data*. Boston, MA, Kluwer Academic Publishers, Dordrecht 1991.

[2] Pawlak, Z., Skowron, A.: Rough membership functions. In: R. Yager, M. Fedrizzi, J. Kacprzyk (Eds.), *Advances in the Dempster-Shafer Theory of Evidence*, NY, John Wiley & Sons, 1994, pp. 251-271.

[3] Peters, J.F., Degtyaryov, V., Borkowski, M., Ramanna, S.: Line-crawling robot navigation: Rough neurocomputing approach. In: C. Zhou, D. Maravall, D. Ruan (Eds.), *Fusion of Soft Computing and Hard Computing for Autonomous Robotic Systems*. Berlin: Physica-Verlag, 2002 [to appear].

[4] Peters, J.F., Ramanna, S., Skowron, A., Stepaniuk, J., Suraj, Z., Borkowski, M.: Sensor fusion: A rough granular approach. In: *Proc. Joint 9th International Fuzzy Systems Association (IFSA) World Congress and 20th North American Fuzzy Information Processing Society (NAFIPS) Int. Conf.*, Vancouver, British Columbia, Canada, 25-28 June 2001, pp. 1367-1372.

[5] Polkowski, L., Skowron, A.: Calculi of granules based on rough set theory: Approximate distributed synthesis and granular semantics for computing with words. In: N. Zhong, A. Skowron, S. Ohsuga (Eds.), New Directions in Rough Sets, Data Mining, and Granular Soft Computing (RSFDGrC'99), *Lecture Notes in Artificial Intelligence* 1711, 1999, pp. 20-28.

[6] Polkowski, L., Skowron, A.: Towards adaptive calculus of granules. In: *Proc. of the Sixth Int. Conf. on Fuzzy Systems (FUZZ-IEEE'98)*, Anchorage, Alaska, 4-9 May 1998, pp. 111-116.

[7] Pal, S.K., Polkowski, L., Skowron, A.: *Rough-neuro computing: Techniques for Computing with Words*. Berlin: Springer-Verlag 2002 [to appear].

[8] Skowron, A.: Toward intelligent systems: Calculi of information granules. In: S. Hirano, M. Inuiguchi, S. Tsumoto (Eds.), *Bulletin of the International Rough Set Society*, vol. 5, no. 1 / 2, 2001, pp. 9-30.

[9] Zadeh, L.A.: Fuzzy logic = computing with words, *IEEE Trans. on Fuzzy Systems*, vol. 4, 1996, pp. 103-111.

[10] Alpigini, J.J., Peters, J.F.: Measures of closeness of performance map information granules: A rough set approach. In: RSCTC'02 [submitted].

[11] Ramanna, S., Peters, J.F., Ahn, T.C: Software quality knowledge discovery: A rough set approach. In: *COMPSAC'02*, Oxford, UK, August 2002 [to appear].

[12] Peters, J.F., Ramanna, S., Suraj, Z., Borkowski, M.: Rough neurons: Petri net models and Applications. In: L. Polkowski, A. Skowron (Eds.), *Rough-Neuro Computing*. Berlin: Springer, 2002, pp. 472-491.

[13] Roddick, J.F., Hornsby, K., Spilopoulou, M.: An updated bibliography of temporal, spatial and spatio-temporal data mining research. In: J.F. Roddick and K. Hornsby (Eds.), *Post-Workshop Proceedings of the International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining (TSDM2000)*, LNAI 2007, Springer-Verlag, Berlin 2000, pp. 147-163.

[14] Alpigini, J.J.: Measures of closeness of performance map information granules: A rough set approach. In: RSCTC'02 [to appear].