Symulacja w Badaniach i Rozwoju
Vol. 6, No. 4/2015

Dariusz MAŁYSZKO
Bialystok University of Technology, Faculty of Computer Science
Wiejska 45A, 15-351 Bialystok, Poland
E-mail: d.malyszko@pb.edu.pl

Clustering Models for Generalized Covering Approximation Spaces

1 Introduction

In the forthcoming era of intelligent information systems, existing mathematical foundations are growing into new directions, giving innovative data analysis tools and methodologies. Development of advanced intelligent systems and robust data analysis methods depends upon gradual process of extending mathematical theories.Rough set theory is grounded on equivalence classes generated from partition [1, 2]. Generalized approximation spaces present abstract model that originates in rough set theory and operates on the concept of neighborhoods and set overlapping. In generalized approximation spaces neighborhoods are defined as coverings of the universe as described in [3, 4, 5, 6]. In previous papers [7, 8, 9] rough coverings have been examined in data thresholding and clustering.

Generalized covering approximation spaces are systems with coverings of universe, neighborhood system and overlap function. In this way, covering approximation spaces has been extended by overlap function. The main contribution of the paper is in

- introducing the concept of generalized covering approximation space,
- presentation of rough feature model based upon generalized covering approximation spaces,
- presentation of rough feature model rough, fuzzy and probabilistic coverings.

The paper has been structured such as to present the above three objectives. In Section 2 the concept of generalized covering approximation spaces has been presented. In Section 3 rough feature covering model has been described. Application of rough feature covering model for image data has been presented in Section 4 followed by concluding remarks.

2 Generalized Covering Approximation Spaces

The definition of generalized covering approximation space is derived from the notion of information systems and rough sets introduced by Z. Pawlak [10, 11]. Information systems have been generalized into covering space in [12, 11, 13]. Important properties of tolerance spaces are given in [14]. Neighborhood systems as model for approximation spaces have been introduced in the early 90s in [15, 16].

An information system $IS = (U, A)$consists of objects described by attributes. Information system with objects divided into classes by reflexive, symmetric and transitive equivalence relation R on U is called an approximation space [15, 16]. Lower and upper approximations of any subset X of U are defined as two sets (sums of equivalence classes) completely contained or partially contained in the set X.

This approach have been generalized by extending equivalence relations to tolerance relations, similarity relations, binary relations leading into formulation of the concept of generalized approximation spaces. At the same time, covering based approaches to rough sets are introduced as described in [3]. In order to extend partitions into more general structures two new functions are introduced neighborhood function and overlap function.

Definition 1. A mapping $N: U \to P(U)$ is called a neighborhood function. Neighborhood function is serial when N is defined for every x of U. If $x \in N(x)\ \forall x \in U$, N is called a reflexive neighborhood function.

Neighborhood system allows for multiple neighborhoods and is defined as

Definition 2. A mapping $NS: U \to P(P(U))$ is called a neighborhood system.

Definition 3. Overlap function $\upsilon : P(U) \times P(U) \to [0,1]$ describes the degree of inclusion of sets.

Definition 4. A generalized approximation space is a tuple $GAS = (U.N, \upsilon)$ where N is a neighborhood function and υ overlap function. The lower and upper approximation operations (denoted respectively by GAS_* and GAS^*) can be defined in a GAS by

$$GAS_*(X) = \{x \in U: \upsilon(N(x), X) = 1\},$$

$$GAS^*(X) = \{x \in U: \upsilon(N(x), X) > 0\}.$$

The concept of covering spaces adds more regular basis for generalized approximation spaces and gives way to introducing the concept of generalized covering approximation space.

Definition 5. Let C be a family of non-empty subsets of U. In case of $S\ C_i = U$ then C is called a covering of U. In order to fulfill requirement of covering all universe for arbitrary family of sets, one set $C_i = U$ containing all universe is added and cover is called extended covering.

Definition 6. Let C be arbitrary family of sets such as exists $C_i = U$ then C is called an extended covering of U. When it is possible to choose subset of disjoint sets of covering that covers all universe this subset is called covering partition.

Definition 7. Subset of C with disjoint sets that covers all universe is called covering partition.

Definition 8. Let C be a family of nonempty subsets of U creating a covering of U. The ordered pair $CAS = (C, U)$ is called a covering approximation space.

Starting from covering approximation spaces [17], [13] and defining more specialized neighborhoods with overlap function we obtain the following definitions. Given covering approximation space CAS, neighborhood function $N(x)$ is defined for each object x as arbitrary set of covering that contains that object.

Definition 9. Given covering approximation space CAS, neighborhood system NS is defined for each object x as all elements of covering C of U that contain that object $NS(x) = \{C_i \in C: x \in C_i\}$.

Neighborhood system NS with two sets L, U of covering C for each universe object is called approximation neighborhood system.

Definition 10. Given covering approximation space CAS, approximation neighborhood system NS assigns for each object x two sets L, U of covering C containing that object $NS = \{L, U \in C : x \in L, U\}$.

Approximation neighborhood system RS with two sets L, U of covering C for each universe object satisfying condition $L \subset U$ is called rough approximation neighborhood system RS.

Definition 11. In covering approximation space CAS, rough approximation neighborhood system $RS(x) = \{L, U \in C : x \in L, U\}$ for each object x defines two sets L, U of covering C containing x, satisfying $L \subset U$.

Definition 12. A generalized covering approximation space is a system $GCS = (U, C, RS, \upsilon)$ where RS is a approximation neighborhood system defined on U with covering C, and υ is the overlap function measuring inclusion of each two sets of covering C.

Introduced concept of generalized covering approximation spaces gives theoretical foundations into creation of rough (feature) covering model described in the next section. Metric spaces define the distance function that makes it possible to compare objects, their similarity, relations, data structure and extract fuzzy and probabilistic properties of the objects of the universe giving at the same time interoperability of different data models.

3 Rough feature covering model

In presented feature coverings data object properties and structures are analyzed by means of their relation to the selected set of data objects from the data space. This reference set of data objects performs as the set of thresholds or the set of cluster centers. Feature coverings basically consist of two interrelated approaches, namely clusters and thresholds. Cluster centers are regarded as representatives of the clusters.

Definition 13. Rough feature covering model defines coverings based upon measures assigned for universe objects according to metric, fuzzy and probabilistic measures from selected set of representative objects called cluster centers or threshold centers.

In the process of the inspection of the data assignment patterns in different parametric settings it is possible to reveal or describe properly data properties. The following properties are included during calculations

- data objects, selected number of cluster centers,
- threshold type: threshold, difference,
- threshold metric: standard, fuzzy, probabilistic, fuzzified probabilistic,
- approximation measure: standard, fuzzy, probabilistic, fuzzified probabilistic.

The distances for two dimensional features in the range (0, 255) have been presented in Fig. 1 (for 8 cluster centers, rescaled with lighter color meaning greater similarity).

One additional cover is added to the C set, with all feature space. In this way, the above feature model is covering approximation space.

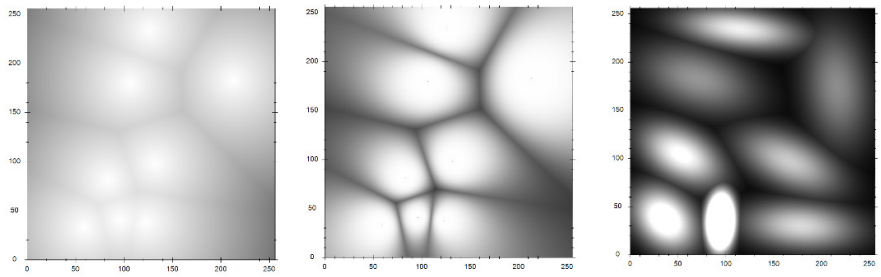

Fig. 1. Feature coverings distances: (a) standard Minkowski, (b) fuzzy similarity, (c) probabilistic similarity

Rys. 1. Feature coverings distances: (a) standard Minkowski, (b) fuzzy similarity, (c) probabilistic similarity

Standard coverings - standard coverings are created on the base of the Minkowski metric. Objects closest to the given cluster center or sufficiently close (relative to the selected threshold type) are assigned to this cluster lower and upper approximations based upon Minkowski distance as follows

$$d_{cr}(x_i, C_m) = \left(\sum_{j=1}^{d}\left(x_{ij} - c_{mj}\right)^p\right)^{\frac{1}{p}}, \tag{1}$$

where d_{cr} represents Minkowski distance of object x_i to cluster center C_m, object $x_i = (x_1, \ldots, x_d) \in R^d$, the cluster center $C_m = (c_{m1}, \ldots, c_{md}) \in R^d$ for $m \in (1..k)$.

Fuzzy coverings-fuzzy coverings are created on the base of fuzzy membership value of objects in clusters, calculated as function of fuzzy membership in clusters. Fuzzy membership value $\mu_{Cm}(x_i) \in [0,1]$ for object $x_i \in U$ in cluster C_m is given as

$$d_{fz}(x_i, C_m) = \mu_{C_m}(x_i) = \frac{d(x_i, C_m)^{-2/(\mu-1)}}{\sum_{j=1}^{k} d\left(x_i, C_j\right)^{-2/(\mu-1)}} \tag{2}$$

with real fuzzifier value μ > 1 representing degree of fuzziness and $d\left(x_i, C_j\right)$ denoting distance between data object x_i and cluster (center) C_j and k number of clusters.

Probabilistic coverings-probabilistic coverings are created from probability distributions in clusters. Coverings should have appropriate probability distribution values for.Gauss distribution has been selected as probabilistic distance metric for data point $x_i \in U$ to cluster center C_m calculated as follows

$$d_{pr}(x_i, C_m) = (2\pi)^{-d/2}|\sum_m|^{-1/2} exp\left(-\frac{1}{2}(x_i - \mu_m)^T \sum_m^{-1}(x_i - \mu_m)\right), \tag{3}$$

where $|\sum_m|$ is the determinant of the semi-positively defined covariance matrix $\sum_m$ and the inverse of covariance matrix for the C_m cluster is denoted as$\sum_m^{-1}$, data dimensionality is denoted as d.

Neighborhood of the object x consists of coverings sufficiently similar to x according to Equations 1, 2, 3. Given above distances (similarity), the feature coverings are calculated as presented in the Table 1 for threshold based coverings and in Table 2 for similarity based coverings, for data point x C_m is the closest covering (represented by cluster center) and C_l are other similar coverings.

Table 1. Threshold coverings

Tabela 1. Pokrycia progowe

Threshold coverings		
Covering	Threshold	Condition
T	std	$d_{cr}(x_i, C_m) \leq \varepsilon_{cr}$
FT	fuzzy	$d_{fz}(x_i, C_m) \geq \varepsilon_{fz}$
PT	pr	$d_{pr}(x_i, C_m) \geq \varepsilon_{pr}$

Definition 14. In standard coverings based upon Minkowski metric, threshold neighborhoods are of two types - hypersphere and hypercube type.

Table 2. Similarity coverings

Tabela 2. Pokrycia podobieństwa

Threshold coverings		
Covering	Threshold	Condition
CD	std	$\lvert d_{cr}(x_i, C_m) - \lvert d_{cr}(x_i, C_l)\rvert \leq \varepsilon_{cr}$
FD	fuzzy	$\lvert d_{fz}(x_i, C_m) - \lvert d_{fz}(x_i, C_l)\rvert \geq \varepsilon_{fz}$
PD	pr	$\lvert d_{pr}(x_i, C_m) - \lvert d_{pr}(x_i, C_l)\rvert \geq \varepsilon_{pr}$

Upper and lower neighborhoods - upper and lower neighborhood approximations are defined on the base neighborhood type and feature type. In this way, lower approximation and upper approximation are defined independently. In feature covering model, the data object are assigned to the cluster with two strategies. Threshold strategy consists in assignment data object when its distance (similarity) to the cluster center C_m is within threshold value. In similarity strategy, the data points are assigned to the closest cluster C_l and to clusters that are sufficiently close to the nearest cluster. The threshold can define two types of covers - hypersphere covers and hypercube covers. Hypercube covers are obtained by standard data thresholding.

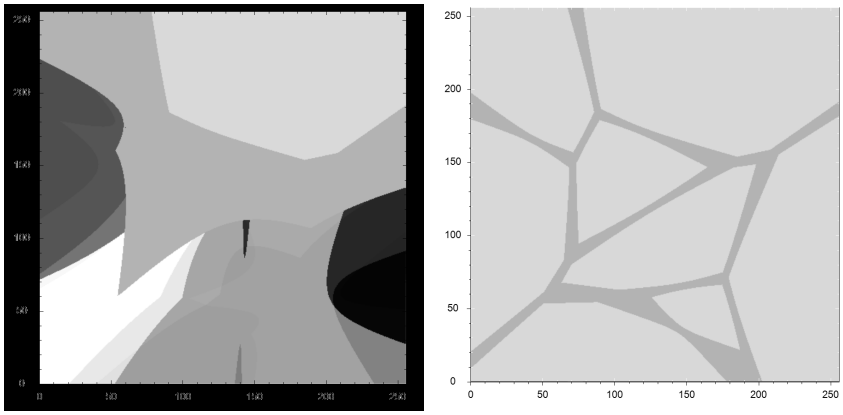

Fig. 2. Standard feature coverings for random 8 cluster centers: (a) lower and upper approximations in gray scale; (b) all lower approximations in light color and upper approximations darker color

Rys. 2. Standardowe pokrycia z losowymi 8 środkami grup: (a) dolne i górne aproksymacje w kolorze szarym; (b) wszystkie aproksymacje dolne w kolorze jasnym oraz górne w kolorze ciemniejszym

4 Rough feature coverings

Standard feature coverings

For standard feature coverings presentation, eight cluster centers has been generated, two coverings - lower and upper approximation has been created on the base of Equation 1 and selected thresholds. Results are presented in Fig. 2.

Fuzzy feature coverings

For fuzzy feature coverings presentation, eight cluster centers has been generated, two coverings - lower and upper approximation has been created on the base of Equation 2 and selected thresholds. Results are presented in Fig. 3.

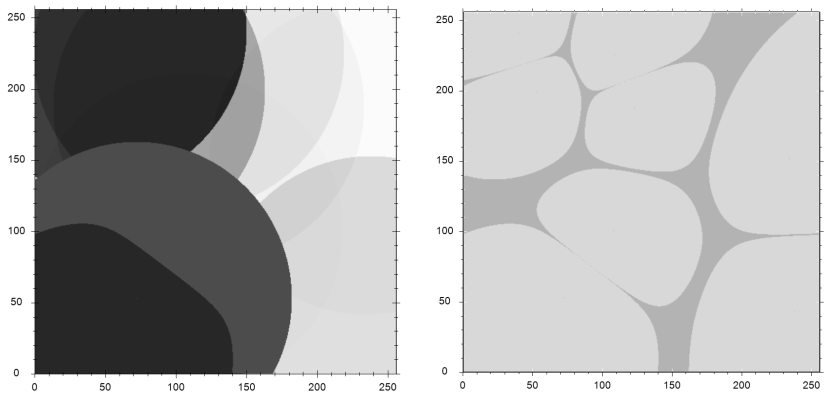

Fig. 3. Fuzzy feature coverings for random 8 cluster centers: (a) lower and upper approximations in gray scale; (b) all lower approximations in light color and upper approximations darker color

Rys. 3. Rozmyte pokrycia z losowymi 8 środkami grup: (a) dolne i górne aproksymacje w kolorze szarym; (b) wszystkie aproksymacje dolne w kolorze jasnym oraz górne w kolorze ciemniejszym

Probabilistic feature coverings

For standard feature coverings presentation, eight cluster centers has been generated, two coverings - lower and upper approximation has been created on the base of Equation 3 and selected thresholds. Results are presented in Fig. 4.

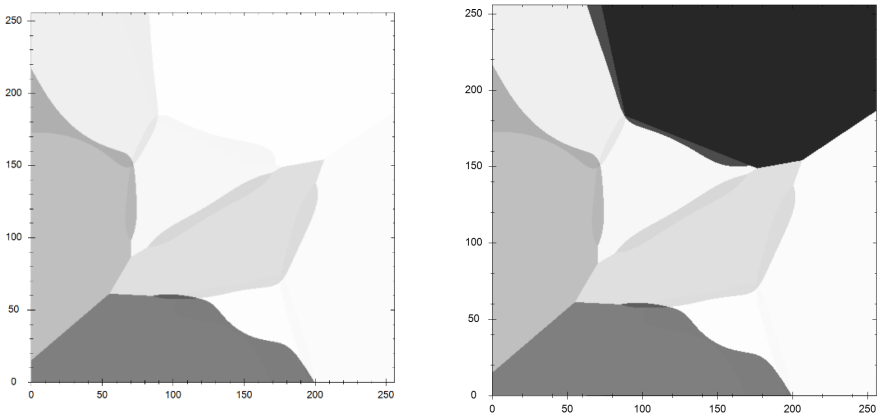

Fig. 4. Probabilistic feature coverings for random 8 cluster centers: (a) lower and (b) upper approximations in gray scale

Rys. 4. Probabilistyczne pokrycia z losowymi 8 środkami grup: (a) dolne i górne aproksymacje w kolorze szarym; (b) wszystkie aproksymacje dolne w kolorze jasnym oraz górne w kolorze ciemniejszym

Conclusions and Future Research

In the study, the concept of generalized covering approximation space has been introduced. Generalized covering approximation spaces extend covering approximation spaces by added overlap function defining degree of set inclusion such as in generalizedapproximation spaces. The introduced concept makes possible creation of distinct rough, fuzzy and probabilistic coverings. The definition, detailed analysis and presentation of rough feature covering model has been presented. Rough covering models has been embedded into generalized approximation spaces. Introduced approximation spaces create generalized covering approximation spaces. Introduced concept of general ized covering approximation spaces gives theoretical foundations into creation of rough feature covering model. Metric spaces define the distance function that make it possible to compare objects, their similarity, relations, data structure and extract fuzzy and probabilistic properties of the objects of the universe giving at the same time interoperability of different data models. The introduced concept has theoretical impact on rough set theory and practical aspect as tool for data clustering and thresholding.

Bibliography

1. Pawlak Z., Skowron A.: Rudiments of rough sets. *Information Sciences*, 177(1): 3-27, 2007
2. Kreinovich V., Pedrycz W., Skowron A.: *Handbook of Granular Computing.* John Wiley & Sons, 2008
3. Yao B Yao Y.: Covering based rough set approximations. *Information Sciences*, 200: 91-107, 2012
4. Zhu. K. Liu G. The relationship among three types of rough approximation pairs. *Knowl.-Based Syst.*, 60: 28-34, 2014
5. Liu G. The relationship among different covering approximations. *Information Sciences*, 250:178-183, 2013
6. Gomez J., Restrepo M., Cornelis C. Partial order relation for approximation operators in covering based rough sets. *Information Sciences*, 284: 44-59, 2014
7. Stepaniuk J., Malyszko D.: Adaptive multilevel rough entropy evolutionary thresholding. *Information Sciences*, 180(7): 1138-1158, 2010
8. Stepaniuk J., Malyszko D.: Standard and fuzzy rough entropy clustering algorithms in image segmentation. Lecture Notes in Computer Science 5306, 5306:409–418, 2008
9. Stepaniuk J., Malyszko D.: Adaptive rough entropy clustering algorithms in image segmentation. *Fundamenta Informaticae*, 98(2-3): 199-231, 2010
10. Pawlak Z.: Rough sets. *International Journal of Computer and Information Sciences*, 11: 341-356, 1982
11. Pawlak Z.: *Systemy informacyjne. Podstawy teoretyczne.* Wydawnictwa NaukowoTechniczne, Warszawa 1983
12. Zakowski W.: Relational interpretations of neighborhood operators and rough set approximation operators. *Demonstratio Mathematica*, 16 (3): 761769, 1983
13. Pomykala J.A.: On definability in the nondeterministic information system. *Bulletin of the Polish Academy of Science Mathematics*, 36:193-210, 1988

14. Stepaniuk J., Skowron A.: Tolerance approximation space. *Fundamenta Informaticae*, 27: 245-253, 1996
15. Yao Y.Y., Lin T.Y.: Neighborhood systems and approximation in database and knowledge base systems. *Fourth International Symposium on Methodologies of Intelligent Systems*, page 75-86, 1989
16. Yao Y.Y.: Relational interpretations of neighborhood operators and rough set approximation operators. *Information Sciences*, 111:239–259, 1998
17. Pomykala. J.A.: Approximation operations in approximation space. *Bulletin of the Polish Academy of Science Mathematics*, 35 (9-10): 653–662, 1987

Summary

Mathematical foundations are steadily extended and pushing rough set theory into incorporating new data analysis methods and data models. Generalized approximation spaces present abstract model useful in understanding unknown and undefined data structure leading into creation many new robust and intelligent approaches. Covering approximation spaces present data by means of coverings of the universe. In the paper, these two approaches have been put together introducing the concept of generalized covering approximation space. Further rough coverings model for generalized covering approximation spaces has been presented. Proposed rough covering models are based upon clustering and thresholding of feature space, are embedded in generalized approximation spaces, simultaneously spanning standard, fuzzy and probabilistic data models.

Keywords: generalized approximation spaces, covering approximation spaces, rough sets, fuzzy sets, probabilistic sets

Grupowanie w uogólnionych aproksymacyjnych przestrzeniach pokryć

Streszczenie

Tematem pracy jest przedstawienie modelu grupowania w rozszerzenym pojęciu uogólnionych przestrzeni aproksymacyjnych, polegającym na zdefiniowaniu pokryć 9 w tych przestrzeniach. W ten sposób uogólniona przestrzeń aproksymacyjna, posiadająca z definicji sąsiedztwa oraz funkcję zawierania się zbiorów, posiada dodatkowo zdefiniowany system pokryć - czyli jest także przestrzenią pokryć. Praca wprowadza model grupowania w uogólnionych aproksymacyjnych przestrzeniach pokryć obejmujący pokrycia standardowe, rozmyte oraz probabilistyczne. W części prezentacyjnej przedstawione zostały przykłady wybranych uogólnionych aproksymacyjnych przestrzeni pokryć.

Słowa kluczowe: uogólnione przestrzenie aproksymacyjne, przestrzenie pokryć, zbiory przybliżone, zbiory rozmyte, zbiory probabilistyczne