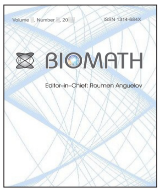
Biomath Forum

BIOMATH

http://www.biomathforum.org/biomath/index.php/biomath/



Rough Sets in Biomedical Informatics

Antony Popov and Simeon Stoykov
Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd., 1164 Sofia, Bulgaria
E-mail:simeon@microdicom.com

Received: 30 October 2012, accepted: 12 December 2012, published: 21 January 2013

***Abstract*—The intent of this paper is to face the essentials of granular computing and in its major component—the rough sets theory, introduced by Pawlak, since any rough set represents an information granule. As a part of modern soft computing paradigm, rough sets have been introduced as an interval-like extension of the usual sets with main applications in the intelligent systems. The proposed rough approach provides efficient algorithms for finding hidden patterns in data, finds minimal sets of data (data reduction), evaluates significance of data. Applications in medicine via DICOM standard are presented, as well as ideas for applications to microbiology.**

***Keywords*-Rough sets; mathematical morphology; molecular biology classifier; medical diagnosis; medical imaging**

I. Introduction

In classical set theory a set is uniquely determined by its elements. In other words, it means that every element must be uniquely classified as belonging to the set or not. That is to say the notion of a set is a precise, or crisp one. For instance, the set of integer numbers is crisp because every number can be uniquely represented by its decimal digits. In mathematics traditionally crisp notions are mainly use to ensure precise reasoning. However philosophers and natural scientists for many years were interested also in imprecise notions like feelings, moral categories, beauty, including also many biological features like the color of the skin or a flower. The rough set approach makes the vagueness of the data possible. It provides efficient algorithms for finding hidden patterns in data, finds minimal sets of data (data reduction), evaluates significance of data. Applications in medicine via DICOM standard are presented in this paper, as well as applications to microbiology and biometrics. Strictly speaking, any rough set represents an information granule. As an example, in gray scale images boundaries between object regions are often ill defined because of grayness or spatial ambiguities. This uncertainty can be effectively handled by describing the different objects as rough sets with upper (or outer) and lower (or inner) approximations as follows:

Let the universe U be an image consisting of a collection of pixels. Then if we partition U into a collection of non-overlapping windows of size $m \times n$, each window can be considered as a granule G. Given this granulation, object regions in the image can be approximated by rough sets. A rough image is a collection of pixels and the equivalence relation induced partition of an image into sets of pixels lying within each non-overlapping window over the image.

II. Information systems and rough sets

Let U be a non-empty, finite set called the universe and A is a non-empty, finite set of attributes, that is every $a \in A$ is a mapping of the form $a : U \to V_a$, where V_a is called a value set of a. The elements of U are called objects and interpreted as, e.g. cases, states, processes, patients, observations. Attributes are interpreted as features, variables, characteristic conditions, etc. Every information system $\mathcal{A} = (U, A)$ and non-empty set $B \subseteq A$ determine a B-information function defined by $\mathrm{Inf}_B(x) = \{(a, a(x)) : a \in B\}$. The set $\{\mathrm{Inf}_A(x) : x \in U\}$ is called A-information set and it is denoted by $\mathrm{INF}(A)$. With every subset of attributes $B \subseteq A$, an equivalence relation, denoted by $\mathrm{IND}_A(B)$

Citation: A Popov, S Stoykov, Rough Sets in Biomedical Informatics, Biomath 2 (2013), 1212127, http://dx.doi.org/10.11145/j.biomath.2012.12.127

(or IND(B)) called a B-indiscernibility relation, is associated and defined by

$$\text{IND}(B) = \{(s, s') \in U^2 : \text{for every } a \in B, a(s) = a(s')\}.$$

Any minimal subset $B \subseteq A$ such that IND(A) =IND(B) is called a reduct in the information system. In fact, microcalcification on X-ray mammogram is a significant mark for early detection of breast cancer. Texture analysis methods can be applied to detect clustered microcalcification in digitized mammograms [2]. In order to improve the predictive accuracy of the classifier, the original number of feature set is reduced into smaller set using feature reduction techniques. In [5] have been introduced rough set based reduction algorithms based on the extracted features. The rough reduction algorithms are tested on mammograms from Mammography Image Analysis Society (MIAS) database [8].

III. Rough set formal definition and main properties

Rough set theory can be viewed as a specific implementation of fuzzyness and vagueness, i.e., imprecision in this approach is expressed by a boundary region of a set, and not by a partial membership, like in fuzzy set theory. However a rough set can be expressed by a fuzzy membership function, as demonstrated below, but it many cases the textual and table representation of a rough set makes it easier and more efficient to practical implementations rather than the original fuzzy approach (3). Moreover, the attributes may be numeric, or in the most cases non-numeric (categorical) quantities, such as big, small, good, malignant, benign etc. As we said previously, we represent the rough. objects x in the universe by their information vector $\text{Inf}_A(x)$. Thus we can define an equivalence relation R between two objects x and y if their information representation coincides, i.e. $\text{Inf}_A(x) = \text{Inf}_A(y)$, so x and y belong to a same information granule. Then the lower approximation of a rough set X with respect to R is the set of all objects, which can be for certain classified as X with respect to this relation. The upper approximation of a rough set X with respect to R is the set of all objects which can be possibly classified as X with respect to R. The boundary region of a set X with respect to R is the set of all objects, which can be classified neither as X nor as not-X with respect to R. For any crisp set the boundary is empty. Therefore we should mainly work with rough sets for which the boundary region of X is nonempty. The equivalence class of R determined by element x will be denoted by $R(x)$. Formal definitions of approximations and the boundary region follow below.

- R-lower approximation of X:

$$R_*(x) = \bigcup_{x \in U} \{R(x) : R(x) \subseteq X\}.$$

- R-upper approximation of X:

$$R^*(x) = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq\}.$$

- R-boundary region of X:

$$RN_R(X) = R^*(X) - R_*(X).$$

It is easy to see that

$$R_*(x) \subseteq X \subseteq R^*(x).$$

Thus we can define a fuzzy membership function, i.e. a fuzzy representation of the rough set X in the universe X with respect to the relation R, see [4]:

$$\mu_R(X) = \frac{\#(R(x) \cap X)}{\#(X)}.$$

Note, that in the definitions above, X is a normal precise subset of the universe U, but we have constructed its rough representation with respect to the relation R - the pair $(R_*(X), R^*(X))$ and its fuzzy analog $\mu_R(X)$. Here the sign # means the cardinality (the number of the elements) of a set. The lower approximation is sometimes referred to as positive region, while the space of the universe outside the upper approximation is called also negative region.

It has been mentioned by Bloch [7] that there is an analogy between rough sets and mathematical morphology. Namely, the R-upper approximation is an analog of morphological dilation, while the R-lower approximation is an analog of morphological erosion. This fact is not surprising, since the relations between fuzzy sets and operations on them and morphology are well studied [1], and a relation between classical interval operations and morphological ones have been established. Moreover, it is evident that the rough approximation of a set is similar to an interval approximation of a real number. On the other hand, interval operations and mathematical morphology have demonstrated their capabilities in solving problems in biomedicine. As an example, due to a complex nature of biomedical images, it is practically impossible to select or develop automatic segmentation methods of generic nature, that could be applied for any type of images, namely for either micro- and macroscopic images, cytological and histological ones, MRI and X-ray, and so on. Medical image segmentation is an indispensable process in the visualization of human tissues. However, medical images always contain a large amount of noise caused by operator performance, equipment and environment. This leads to inaccuracy with segmentation. So, a robust segmentation technique is required. The basic idea behind introducing rough sets is that while some cases may be clearly distinguished as being in a set X (positive region in rough sets theory), and some cases may be clearly labeled as not being in set X (negative region). Since we can obtain limited information we are not able to label all possible cases clearly. The remaining cases cannot be distinguished and lie in the boundary region.

IV. Rough set specification by decision rules

For rough separation of the universe U one can use efficiently fuzzy C-means clustering [4]. If we want to separate m data elements into n clusters, by this algorithm we obtain n cluster centers and m n numbers between 0 and 1 showing the degree of membership of ith data element to the jth cluster. Thus if this number is not less than 0.75 then the element belongs to the positive region of the cluster, if it is less than 0.25 it belongs to negative region, otherwise it belongs to the border. Then by IF_THEN_ELSE rules we may specify the regions [6]. The rules can be applied to a set of unseen cases in order to estimate their classification power. Several application schemes can be envisioned. Let us consider one of the simplest which has shown to be useful in practice:

1. When a rough set classifier is presented with a new case, the rule set is scanned to find applicable rules, i.e. rules whose predecessors match the case.
2. If no rule is found (i.e. no rule is fired), the most frequent outcome in the training data is chosen.
3. If more than one rule fires, these may in turn indicate more than one possible outcome A voting process is then performed among the rules that fire in order to resolve conflicts and to rank the predicted outcomes.

Here are some rough rules which in fact form a decision table:

- *IF Gene A is* ***up-regulated*** *AND Gene D is* ***down-regulated*** *THEN Tissue is* ***healthy****;*
- *IF Transcription factor F* ***binds*** *AND Transcription factor V* ***binds*** *THEN Gene is* ***co-regulated*** *with Gene H;*
- *IF Protein* ***contains*** *motif J THEN Function is* ***magnesium ion binding*** *OR* ***copper ion binding****;*
- *IF Protein* ***contains*** *motif D AND Ligand water-octanol coeff.* $> c_1$ *THEN Binding affinity is* ***high****;*
- *IF* ***change in*** *frequency of alpha-helix at position* $X > c_2$ *THEN* ***Resistant to*** *drug W.*

V. DICOM format and its realization

ACR (the American College of Radiology) and NEMA (the National Electrical Manufacturers Association) formed a joint committee to develop a Standard for Digital Imaging and Communications in Medicine [9] . This Standard is developed in liaison with other Standardization Organizations including CEN TC251 in Europe and JIRA in Japan, with review also by other organizations including IEEE, HL7 and ANSI in the USA. This Standard is now designated for almost CT, PET, MRI, Ultrasound devices used in practice. It is applicable to a networked environment. The previous versions were applicable in a point-to-point environment only; for operation in a networked environment a Network Interface Unit (NIU) was required.

DICOM Version 3.0 supports operation in a networked environment using industry standard networking protocols such as OSI and TCP/IP. It specifies how devices claiming conformance to the Standard react to commands and data being exchanged. Previous versions were confined to the transfer of data, but DICOM Version 3.0 specifies, through the concept of Service Classes, the semantics of commands and associated data. DICOM Version 3.0 explicitly describes how an implementor must structure a Conformance Statement to select specific options. It is structured as a multi-part document. This facilitates evolution of the Standard in a rapidly evolving environment by simplifying the addition of new features. ISO directives which define how to structure multi-part documents have been followed in the construction of the DICOM Standard. A single DICOM file contains both a header (which stores information about the patient's name, the type of scan, image dimensions, etc), as well as all of the image data (which can contain information in three dimensions).

The DICOM header

The size of this header varies depending on how much header information is stored. Header represents an instance of a real world, referred to as Information Object. Header is constructed of Data Elements. Data Elements contain the encoded Values of Attributes of that object. The specific content and semantics of these

Attributes are specified in Information Object Definitions (see PS 3.3 of the DICOM Standard [9]). The construction, characteristics, and encoding of a Data Set and its Data Elements are discussed in PS 3.5 of the DICOM Standard. Pixel Data, Overlays, and Curves are Data Elements whose interpretation depends on other related elements. As seen below, the data elements can be interpreted as rough set attributes. The main part of a DICOM file is the image collection reffered by the textual part described above.

Data Elements

A Data Element is uniquely identified by a Data Element Tag. The Data Elements in header shall be ordered by increasing Data Element Tag Number and shall occur at most once in a Data Set. A DICOM attribute or data element is composed of:

- A *tag*, in the format of *group*, *element* (XXXX,XXXX) that identifies the element.
- A *Value Representation* (VR) that describes the data type and format of the attribute's value.
- A *value length* that defines the length of the attribute's value.
- A *value field* containing the attribute's data.

The basic attribute structure is shown below.

Tag	VR	Value Length	Value Field

A simple example for a single tag for a CT image is: (0028, 0004), Photometric Interpretation: MONOCHROME2

Here you can see also a genetic data representation:

Tag	Term	Frequency Gene(s)	Gene(s)
(6950,0001)	response to stress	16 of 106 15.1%	PRX1, HSP26, PHO5, HSP30, ...
(6810,0015)	transport	15 of 106 14.2%	GLK1, HXT7, HXT6, PIC2, STF2, ...

Each data element is described by a pair of numbers (group number, data element number). Even numbered groups are elements defined by the DICOM standard and are referred to as public tags. Odd numbered groups can be defined by users of the file format, but must conform to the same structure as standard elements. These are referred to as private tags. The ACR-NEMA Version 1 and 2 standards did not use object-oriented analysis or design. Instead, attributes (or elements, as they were called) were grouped according to use. For example, there were groups of elements that carried identifying information about the patient and others consisting of elements that described the methods of image acquisition. Because they were developed without an entity relationship (E-R) model, these groups do not conform to conventional object-oriented definitions. Note, that the E-R data model views the real world as a set of basic objects (entities) and relationships among these objects. For example, a collection of elements used in the ACR-NEMA Version 2 standard to identify and describe a computer tomographic (CT) image would also contain the patient name. In an entity relationship (E-R) model, however, the patient name is an attribute of the patient object, not of the image object. In other words, the patient name is not needed to describe the CT image, even though it would be needed to identify the image. One might also view these complex objects as consisting of parts of more than one entity in an E-R model. A novel free DICOM viewer called MICRODICOM has been created by the second author (10). It gives good opportunities for finding pathological objects. After clustering by 5 features (pixel intensity, mean and standard deviation in a 7×7 window, the two x and y Sobel operations) four tissue clusters are specified. Then by adding rules for finding the connected components associated with the pathology cluster based on contour tracing techniques, the tumor is located, see Figure 1.

VI. Conclusion

We tried to explain the power of rough modeling in biomedicine. The MICRODICOM project is under development and further intelligent capabilities based on soft computing, and especially on rough sets theory will be included.

Acknowledgments

This work was supported by Ministry of Education, Youth and Technologies under contract DO 02-359/2008 “Computer Simulation and Innovative Model-Based Study of Bioprocesses”.

References

[1] Popov A. T. (2007) General Definition of Fuzzy Mathematical Morphology Operations. Image and Non-image Applications, In: Nachtegael, M., Van der Weken, D., Kerre, E.E., Philips, W. (Eds.), Soft Computing in Image Processing - Recent Advances, Series: Studies in Fuzziness and Soft Computing, Vol. 210, Springer, 355–384.

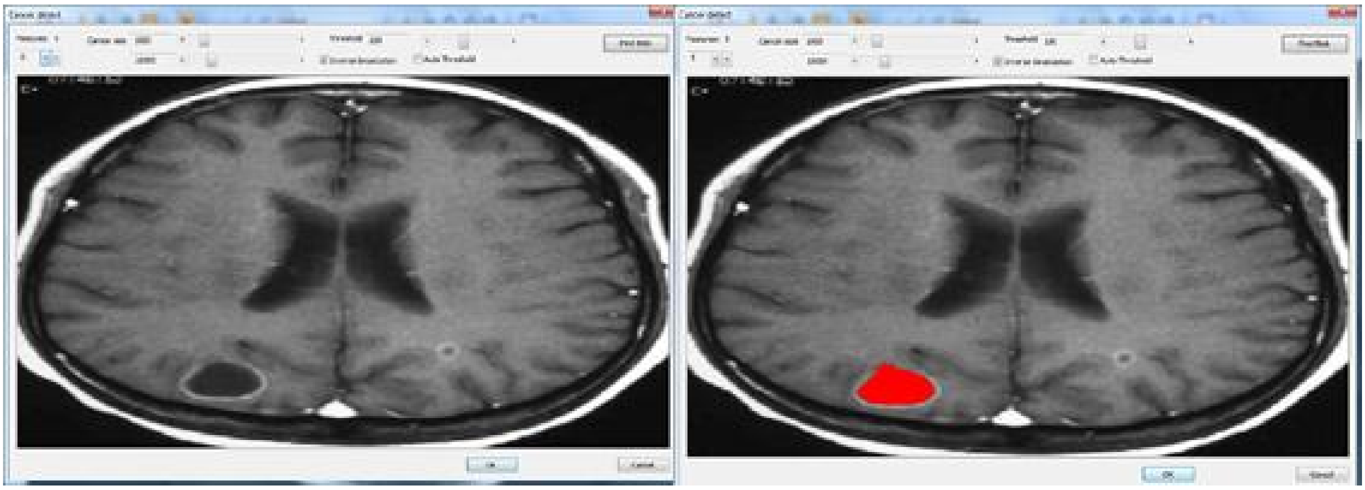


Figure 1. MRI slice of a human brain with a tumor detected by MICRODICOM software

[2] Baeg S., S. Batman, E. R. Dougherty, V. G. Kamat, N. Kehtarnavaz, Seunghan Kim, A. Popov, K.Sivakumar, R. Shah, (1999) Unsupervised morphological granulometric texture segmentation of digital mammograms, Journal of Electronic Imaging 8(1), 65–75.
http://dx.doi.org/10.1117/1.482685

[3] Pawlak, Z. (1991) Rough Sets: Theoretical Aspects of Reasoning about Data. Volume 9 of Series D: System Theory, Knowledge Engineering and Problem Solving, Kluwer.

[4] Nguyen H. T., E. A.Walker (2000) A first course in fuzzy logic (2nd edition), CRC Press.

[5] Thangavel K., Karnan M., and Pethalakshmi A. (2005) Performance Analysis of Rough Reduct Algorithms in image Mammogram, ICGST International Journal on Graphics, Vision and Image Processing 8, 13–21

[6] Hvidsten, T.R., Lgreid, A., Komorowski, J. (2003) Learning rule-based models of biological process from gene expression time profiles using gene ontology. Bioinformatics 19, 1116–1123
http://dx.doi.org/10.1093/bioinformatics/btg047

[7] Bloch, I (2000) On links between mathematical morphology and rough sets, Pattern Recognition 33, 1487–1496
http://dx.doi.org/10.1016/S0031-3203(99)00129-6

[8] http://http://www.mammoimage.org/databases/;
http://peipa.essex.ac.uk/info/mias.html

[9] http://www.dclunie.com/dicom-status/status.html

[10] http://www.microdicom.com