Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 38(4), 2011, Pages 50–62
ISSN: 1223-6934, Online 2246-9958

Rough Sets and Gaussian Mixture Model in Medical Image Diagnosis

Anca Loredana Ion

Abstract. This paper proposes methods for modeling the content of medical images to help physicians in diagnosing. The Gaussian mixture model is used for medical image segmentation, while the rough set theory is a powerful approach that permits the searching for patterns in medical images using the minimal length principles. Searching for models with small size is performed by means of many different kinds of reducts that generate the decision rules capable for identifying the medical diagnosis.

2010 Mathematics Subject Classification. Primary 68U10; Secondary 68T10.
Key words and phrases. medical image diagnosis, rough sets, Gaussian mixture model, image colour, image texture, image shape.

1. Introduction

In the medical domain, the physicians have to make feasible judgments on the diagnoses of patients by analyzing images offered by different technologies like endoscopy, radiology, magnetic resonance, etc. Therefore, strong methods are needed to support the physicians in making diagnosis decisions by dividing patients into different categories of risk.

Medical image content representation and retrieval are playing an increasing role in a large sphere of applications within the clinical process [1]. For the clinical decision-making process it can be beneficial and important to find other images of the same modality, the same anatomic region or the same disease [2]. Thus, a lot of researches were developed to investigate automated techniques for extracting the low-level features that could generate semantic descriptions of the medical image content. Among these techniques are the methods based on machine learning that manually annotate the test image datasets. Algorithms that recognize specific organs with different structures of the medical images are studied in [5]. FIRE application [4] and IRMA [7] use the sub-symbolic processing of images with good results. Also, in the medical domain, taxonomies, thesaurus and ontology were developed, varying from the general target, like UMLS [3], SNOMED CT [10], to the specific ones, like FMA [8] for anatomy, RadLex [6] for radiology, and AIM [9] a project developed at Stanford University.

In this paper, we exploit the opportunities of the medical domain, rich in formal representation of knowledge, the rough sets and the statistical framework for developing methods with automatic diagnosis capacity. So, the objective of this work is to improve the performance of existing approaches in the diagnosing of medical images.

Received July 11, 2011. Revision received October 12, 2011.
This work was supported by the strategic grant POSDRU/89/1.5/S/61968, Project ID 61968 (2009), co-financed by the European Social Fund within the Sectorial Operational Program Human Resources Development 2007 - 2013.

The first step of this complex study is the representation of medical images as a collection of homogenous non-overlapping regions. To resolve this task, the Gaussian mixture model is used in the unsupervised pixels classification of an image, due to its capability of resolving the uncertainty of the mixed pixels [2]. The success of methods for medical image analysis depends on the quality of segmentation process.

The second step is the discovering of patterns from medical images for establishing their diagnosis, adapting the rough set approach to our need. Among methods proposed for modeling the imperfect knowledge [17], the rough set theory is an interesting attempt to solve this problem. This theory is based on an assumption that objects are recognized by partial information about them and some objects can be indiscernible. From this fact, it follows that some sets cannot be exactly described by the available information about objects [16], [18]. The methods based on rough set theory have an important utilization in many real life applications. Among the rough set based software systems are ROSETTA [19], RSES [20], and LERS [21], which have been applied to knowledge discover problems.

2. Image Segmentation and Feature Computation

The diagnosis of medical images is directly related to the visual features (colour, texture, shape, position, dimension, etc.), because these attributes capture the information about the semantic meaning. The computation of the low-level features of sick regions involve two steps:

- image segmentation, which takes into account the color feature and produces homogeneous color regions,
- the computation of mathematical descriptor as texture, shape, dimension, position of each detected region.

2.1. Image segmentation. To resolve the problem of image segmentation, each region is represented as a parameterized Gauss distribution. A *Gaussian mixture model* is a powerful framework that estimates the probability density function of a variable and was widely used in statistic, image and signal processing, physic, biology, finance, web information extraction, etc.

The image is modeled as a mixture of Gaussian distribution, where an individual distribution is used to specify a region of pixels. Thus, the image is modeled as a "random field" [12], being composed from two collections of two random variables Y and X. The values of the first variable correspond to the classes/regions, while the values of the second variables correspond to "measurements" or "observations" of the pixels. The problem of segmentation consists in determining Y, knowing X.

In our case, the image is represented in the *RGB* color space. Then an image pixel, x_n, is represented as a colour vector of dimension m, where $m = 3$. If we want to attach the pixel, x_n, to one of the clusters/components $z_1, .., z_k$, we have to determine the conditional probability $p(z_k|x_n)$.

In conformity with Bayes theorem [14], the conditional probability is defined as in equation 1:

$$p(z_k|x_n) = \frac{p(x_n|z_k)p(z_k)}{p(x_n)} \tag{1}$$

The pixel x_n could be in one of clusters, having the initial probabilities: $w_1 = p(z_1), w_2 = p(z_2), .., w_k = p(z_k)$. The conditional probability of x_n, for a given z_k is

modeled by a Gaussian distribution G parameterized by two parameters V_k and μ_k as in equation 2:

$$p(x_n|z_k) = G(x_n|\mu_k, V_k) \tag{2}$$

In conformity with the product rule [14], the joint probability is defined as in equation 3:

$$p(x_n, z_k) = p(z_k|x_n)p(x_n) \tag{3}$$

In conformity with the sum rule [14], the marginal probability is defined as in equation 4:

$$p(x_n) = \sum_{z_k} p(x_n, z_k) = \sum_{z_k} p(z_k|x_n)p(x_n) = \sum_{z_k} p(z_k)p(x_n|z_k) \tag{4}$$

Thus, from Equations 2 and 4, the mixture density function is defined as in equation 5:

$$p(x_n) = \sum_{z_k} p(z_k)p(x_n|z_k) = \sum_{k=1}^{K} w_k G(x_n|\mu_k, V_k), \tag{5}$$

where, μ_k is the mean of dimension m and V_k is the covariance matrix of dimension $m \times m$ of the Gaussian component.

Suppose the image is a data set of pixels $X = \{x_1.., x_N\}$, then the likelihood function is defined as in equation 6:

$$p(X|w, \mu, V) = \Pi_{n=1}^{N} \sum_{k=1}^{K} w_k G(x_n|\mu_k, V_k) \tag{6}$$

where

$$G(x_n|\mu_k, V_k) = \frac{1}{(2\pi)^{m/2}} \frac{1}{|V_k|^{1/2}} \cdot exp(-\frac{1}{2}(x_n - \mu_k)^T V_k^{-1}(x_n - \mu_k)), x_n \in X \tag{7}$$

The function $p(X|w, \mu, V)$ has to be maximized using the *expectation-maximization (EM) algorithm.* An advantage of expectation-maximization method is its capability for handling uncertainties due to mixed pixels [11], [14].

In the next steps, the EM algorithm for the Gaussian mixture model is described:

Step 1: Initialize the means μ_k, co-variances V_k and evaluate initial value of likelihood function.

Step 2: E-step: Evaluate the conditional probability of z_k, $p(z_k|x_n)$, for a given $x_n \in X$, as in equation 8:

$$\gamma(z_{nk}) = p(z_{nk}|x_n) = \frac{w_k G(x_n|\mu_k, V_k)}{\sum_{k=1}^{K} w_k G(x_n|\mu_k, V_k)} \tag{8}$$

Step 3: M-step: Re-estimate parameters as in equations 9, 10, 11:

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \tag{9}$$

$$V_k^{new} = \frac{1}{N_k} \sum_{n-1}^{N} \gamma(z_{nk})(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T \tag{10}$$

$$w_k^{new} = \frac{N_k}{N}; N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{11}$$

Step 4: Evaluate the log-likelihood from eqaution 12 and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied, return to step 2:

$$ln(p(X|w,\mu,V)) = \sum_{n=1}^{N} ln \sum_{k=1}^{K} w_k G(x_n|\mu_k, V_k) \tag{12}$$

2.2. Kullback-Leibler (KL) divergence. In the GMM-KL framework, the distance measure between two images is defined as a distance measure between the two Gaussian mixture distributions GMM_i and GMM_j obtained from the images, as in Equation(13) [11], [12]:

$$D(GMM_i, GMM_j) = \int GMM_i(x) log \frac{GMM_i(x)}{GMM_j(x)} dx \tag{13}$$

If GMM_i and GMM_j are two multivariate Gaussian distributions parameterized by their means, μ_i and μ_j and by their covariance matrices, V_i and V_j, the equation leads to a closed form expression of the KL distance as in (14) [2]:

$$D(GMM_i, GMM_j) = \frac{1}{2} log \frac{|V_j|}{|V_i|} + \frac{1}{2} tr(V_j^{-1} V_i) + \frac{1}{2} (\mu_j - \mu_j)^T V_j^{-1} (\mu_j - \mu_j) - \frac{n}{2} \tag{14}$$

2.3. Hierarchical clustering. Although Gaussian mixture models are used in many research domains from image processing to machine learning, this statistical mixture modeling is usually complex and need to be simplified [13], [14].

In this paper, we present a simplification method based on a hierarchical clustering algorithm. This algorithm provides a hierarchical representation of the initial Gaussian mixture model and experiments on medical image processing are reported. Given a set of k Gaussian distribution $GMM_1, .., GMM_k$ to be reduced and a $k \times k$ similarity matrix, the basic process of hierarchical cluster is this:

1. Start by assigning each Gaussian distribution to a cluster, so we now have N clusters, each containing just one item. The distances (similarities) between the clusters are the same as the distances (similarities) between the items they contain.

2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.

3. Compute distances (similarities) between the new cluster and each of the old clusters $D(GMM_i, GMM_j)$.

4. Repeat steps 2 and 3 until all items are clustered into k cluster.

Step 3 is done using single-linkage clustering, in which the distance between one cluster and another cluster is equal to the shortest distance from any member of one cluster to any member of the other cluster.

2.4. Segmentation results. The image collections used in our experiments were taken from free repositories [27], [28]. The experiments were carried out on images diagnosed with: duodenal ulcer, gastric ulcer, gastric cancer, esophagitis, etc. Through our experiments, we considered 4 components for the mixture model to observe the results.

In this section, we present experiments realized on images diagnosed with duodenal ulcer and colon cancer. The duodenal ulcers can come in different shapes, sizes, and textures [28], increasing the complexity to diagnose them. For example, the image from Figure 1(a) shows a single, white-based ulcer. The segmentation results on this image can be observed in Figure 1(b).

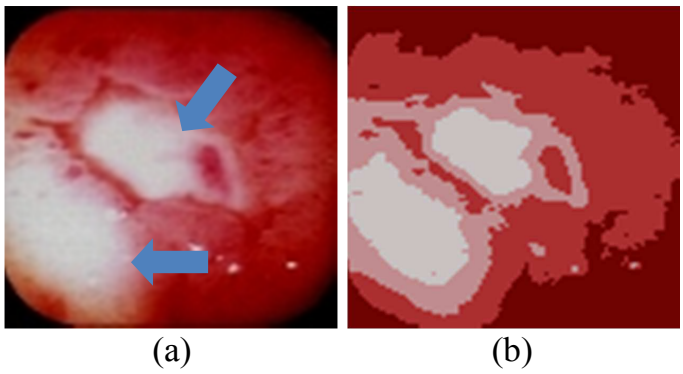

FIGURE 1. Results of segmentation on an image diagnosed with duodenal ulcer: (a) Original image; (b) Image segmented with 4 regions.

FIGURE 2. Regions of interest extracted from the image segmented with 4 components.

The interested regions of the duodenal ulcer image segmented into 4 components are observed in Figure 2. There are only 2 extracted regions of interest in different hues, because the mixture model was reduced to 4 components.

The image from Figure 3(a) shows an advanced cancer in the right colon [28] and the sick region come in different yellow hues. The segmentation results on this image can be observed in Figure 3(b).

The interested region of the colon cancer image segmented into 4 components are observed in Figure 4.

By analyzing the results of medical images segmentation using the Gaussian mixture model, a good delimitation of regions of interest can be observed.

2.5. Low-level features computation. After the detection of image regions, the following visual features of each region are computed [23]:

- The colour, which is represented in the RGB colour space and it is the mean of an image region/component(see Section 2).
- The spatial coherency, which measures the spatial compactness of the pixels of the same colour.
- A seven-dimension vector (maximum probability, energy, entropy, contrast, cluster shade, cluster prominence, correlation), which represents the texture characteristics.

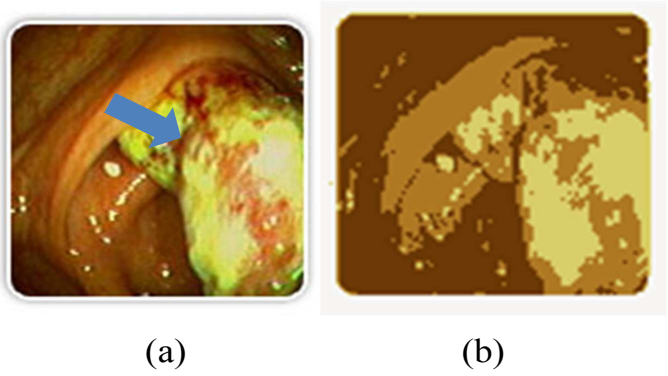

FIGURE 3. Results of segmentation on an image diagnosed with colon cancer: (a) Original image; (b) Image segmented with 4 regions.

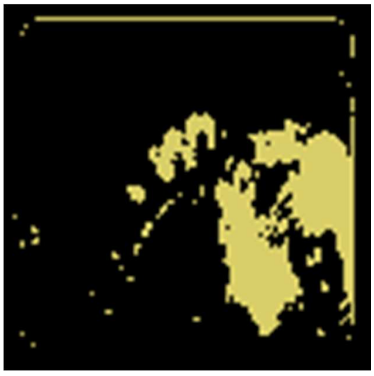

FIGURE 4. Regions of interest extracted from the image segmented with 4 components.

- The region dimension descriptor, which represents the number of pixels from region.
- The spatial information which is represented by the centroid coordinates of the region and by minimum bounded rectangle.
- A two-dimensional vector (eccentricity and compactness), which represents the shape feature.

2.6. Low-Level Feature Discreteness. The visual features of sick regions were computed over intervals, using the concept of *semantic indicators*, which are visual elements: the colour (colour-light-red, etc.), spatial coherency (spatial coherency-weak, spatial coherency-medium, spatial coherency-strong), texture (energy-small, energy-medium, energy-big), dimension (dimension-small, dimension-medium, dimension-big), position (vertical-upper, vertical-center, vertical-bottom, horizontal-upper, etc.), shape (eccentricity- small, compactness-small, etc.). The values of each semantic descriptor are mapped to a value domain, which corresponds to the mathematical descriptor [22], [23].

A medical image is represented by means of the terms *figure(ListofRegions)*, where *ListofRegions* is a list of images' sick regions. The term *region(ListofDescriptors)* is used for region representation, where the argument is a list of terms used to specify the *semantic indicators*. The term used to specify the semantic indicators is of form *descriptor(DescriptorName, DescriptorValue)*.

The mapping between the values of low-level (mathematical) descriptors and the values of *semantic indicators* is based on experiments effectuated on images from different categories and the following facts are used: *mappingDescriptor(Name, SemanticValue, ListValues)*. The argument *Name* is the semantic indicator name, *SemanticValue* is the value of the semantic indicator, *ListValues* represents a list of mathematical values and closed intervals, described by the following terms: *interval(InferiorLimit, SuperiorLimit)* [24].

3. Modeling Image Diagnosis using Rough Sets

3.1. Rough sets foundations. Rough sets theory is an intelligent mathematical tool and it is based on the concept of approximation space [16], [17], [25]. In this section, we recall some basic definitions from literature [16], [17], [25], [18].

Let U denote a finite non-empty set of objects (sick image regions) called the universe. Further, let A denote a finite non-empty set of attributes. For every attribute $a \in A$, there is a function $a : U \to V_a$, where V_a is the set of all possible values of a, called the domain of a. A pair $IS = (U, A)$ is an information system. Usually, the specification of an information system can be presented in tabular form. Each subset of attributes $B \subseteq A$ determines a binary B-indiscernibility relation $IND(B)$ consisting of pairs of objects indiscernible with respect to attributes from B like in equation 15:

$$IND(B) = \{(x, y) \in U \times U : \forall a \in B, a(x) = a(y)\} \tag{15}$$

$IND(B)$ is an equivalence relation and determines a partition of U, which is denoted by $U/IND(B)$. The set of objects indiscernible with an object $x \in U$ with respect to the attribute set, B, is denoted by $I_B(x)$ and is called B-indiscernibility class:

$$I_B = \{y \in U : (x, y) \in IND(B)\} \tag{16}$$

$$U/IND(B) = \{I_B(x) : x \in U\} \tag{17}$$

It is said that a pair $AS_B = (U, IND(B))$ is an approximation space for the information system $IS = (U, A)$, where $B \subseteq A$.

The information system from Table 1 represents the sick regions of images from different diagnoses described in terms of semantic indicators values. For simplicity we consider only two semantic indicators as attributes, namely the colour and texture-entropy. So, our information systems is $IS = (U, B)$, where $U = \{R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8, R_9, R_{10}, R_{11}\}$ and $B = \{colour, texture - entropy\}$. Some examples of partitions defined by indiscernibility relations for the information system in Table 1 are given in Table 2.

In rough sets theory, the approximations of sets are introduced to deal with inconsistency. A rough set approximates traditional sets using a pair of sets named the lower and upper approximations of the set. Let $W = \{w_1, .., w_n\}$ be the elements of the approximation space $AS_B = (U, IND(B))$. We want to represent X, a subset of U, using attribute subset B. In general, X cannot be expressed exactly, because

TABLE 1. Medical Information System.

U	Colour	Texture-entropy	Diagnosis
R1	light-red	Small	gastric-ulcer
R2	light-red	Small	gastric-ulcer
R3	light-red	Small	gastric-ulcer
R4	light-red	Big	gastric-ulcer
R5	light-yellow	Big	gastric-ulcer
R6	light-yellow	Medium	duodenal-ulcer
R7	light-yellow	Medium	duodenal-ulcer
R8	medium-yellow	Small	duodenal-ulcer
R9	medium-yellow	Small	duodenal-ulcer
R10	dark-yellow	Small	duodenal-ulcer
R11	dark-yellow	Small	duodenal-ulcer

TABLE 2. Partitions Defined by Indiscernibility Relations.

IND(B)	Partitions U/IND(B)
IND(Colour)	R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11
IND(Colour,Texture-entropy)	R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11

the set may include and exclude objects which are indistinguishable on the basis of attributes B, so we could define X using the lower and upper approximation.

The B-lower approximation X, $\underline{B}X$, is the union of all equivalence classes in $IND(B)$ which are contained by the target set X. The lower approximation of X is called the positive region of X and is denoted $POS(X)$.

$$\underline{B}X = \bigcup \{w_i | w_i \subseteq X\} \tag{18}$$

The B-upper approximation $\overline{B}X$is the union of all equivalence classes in $IND(B)$ which have non-empty intersection with the target set X.

$$\overline{B}X = \bigcup \{w_i | w_i \bigcap X \neq \emptyset\} \tag{19}$$

Example: Let $X = \{R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8\}$ be the subset of U that we wish to be represented by the attributes set $B = \{colour, texture - entropy\}$. We can approximate X, by computing its B-lower approximation and B-upper approximation. So, $\underline{B}X = \{\{R_1, R_2, R_3\}, \{R_4\}, \{R_5\}, \{R_6, R_7\}\}$ and $\overline{B}X = \{\{R_1, R_2, R_3\}, \{R_4\}, \{R_5\}, \{R_6, R_7\}, \{R_8, R_9\}\}$. The tuple $(\underline{B}X, \overline{B}X)$ composed of the lower and upper approximation is called a rough set; thus, a rough set is composed of two crisp sets, one representing a lower boundary of the target set X, and the other representing an upper boundary of the target set X. The accuracy of a rough set is defined as: $cardinality(\underline{B}X)/cardinality(\overline{B}X)$. If the accuracy is equal to 1, then the approximation is perfect.

3.2. Dispensable features, reducts and core. An important notion used in rough set theory is the *decision table*. Pawlak [16], [17] gives also a formal definition of a decision table: an information system with distinguished conditional attributes and decision attribute is called a decision table. So, a tuple $DT = (U, C, D)$ is a decision table. The attributes $C = \{colour, texture - entropy\}$ are called conditional attributes, instead $D = \{diagnosis\}$ is called decision attribute. The classes

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}
R_1	-	-	-	-	-	$C^{light\text{-}red}$ T^{small}	$C^{light\text{-}red}$ T^{small}	$C^{light\text{-}red}$	$C^{light\text{-}red}$	$C^{light\text{-}red}$	$C^{light\text{-}red}$
R_2	-	-	-	-	-	$C^{light\text{-}red}$ T^{small}	$C^{light\text{-}red}$ T^{small}	$C^{light\text{-}red}$	$C^{light\text{-}red}$	$C^{light\text{-}red}$	$C^{light\text{-}red}$
R_3	-	-	-	-	-	$C^{light\text{-}red}$ T^{small}	$C^{light\text{-}red}$ T^{small}	$C^{light\text{-}red}$	$C^{light\text{-}red}$	$C^{light\text{-}red}$	$C^{light\text{-}red}$
R_4	-	-	-	-	-	$C^{light\text{-}red}$ T^{big}	$C^{light\text{-}red}$ T^{big}	$C^{light\text{-}red}$ T^{big}	$C^{light\text{-}red}$ T^{big}	$C^{light\text{-}red}$ T^{big}	$C^{light\text{-}red}$ T^{big}
R_5	-	-	-	-	-	T^{big}	T^{big}	$C^{light\text{-}yellow}$ T^{big}	$C^{light\text{-}yellow}$ T^{big}	$C^{light\text{-}yellow}$ T^{big}	$C^{light\text{-}yellow}$ T^{big}
R_6	$C^{light\text{-}yellow}$ T^{medium}	$C^{light\text{-}yellow}$ T^{medium}	$C^{light\text{-}yellow}$ T^{medium}	$C^{light\text{-}yellow}$ T^{medium}	T^{medium}	-	-	-	-	-	-
R_7	$C^{light\text{-}yellow}$ T^{medium}	$C^{light\text{-}yellow}$ T^{medium}	$C^{light\text{-}yellow}$ T^{medium}	$C^{light\text{-}yellow}$ T^{medium}	T^{medium}	-	-	-	-	-	-
R_8	$C^{medium\text{-}yellow}$	$C^{medium\text{-}yellow}$	$C^{medium\text{-}yellow}$	$C^{medium\text{-}yellow}$ T^{small}	$C^{medium\text{-}yellow}$ T^{small}	-	-	-	-	-	-
R_9	$C^{medium\text{-}yellow}$	$C^{medium\text{-}yellow}$	$C^{medium\text{-}yellow}$	$C^{medium\text{-}yellow}$ T^{small}	$C^{medium\text{-}yellow}$ T^{small}	-	-	-	-	-	-
R_{10}	$C^{dark\text{-}yellow}$	$C^{dark\text{-}yellow}$	$C^{dark\text{-}yellow}$	$C^{dark\text{-}yellow}$ T^{small}	$C^{dark\text{-}yellow}$ T^{small}	-	-	-	-	-	-
R_{11}	$C^{dark\text{-}yellow}$	$C^{dark\text{-}yellow}$	$C^{dark\text{-}yellow}$	$C^{dark\text{-}yellow}$ T^{small}	$C^{dark\text{-}yellow}$ T^{small}	-	-	-	-	-	-

FIGURE 5. The discernibility matrix.

$U/IND(C)$ and $U/IND(D)$ are called condition and decision classes, respectively. The C-Positive region of D is given by (20):

$$POS_C(D) = \bigcup_{X \in IND(D)} \underline{C}X \tag{20}$$

Let $c \in C$ a feature. It is said that c is dispensable in the decision table DT, if $POS_{C-\{c\}}(D) = POS_C(D)$, otherwise the feature c is called indispensable in DT. If c is an indispensable feature, deleting it from DT makes it to be inconsistent. A set of features R in C is called a reduct, if DT'= (U, R, D) is independent and $POS_R(D) = POS_C(D)$. In other words, a *reduct* is the minimal feature subset preserving the above condition. The set of all features indispensable in C or the set of all reducts of C is denoted by $CORE(C)$.

3.3. Producing rules by discernibility matrix. Decision rules are generated from reducts. The rule generation algorithm has the flowing steps:

-construct the decision table and discernibility matrix,

-obtain the discernibility function and the prime implicants,

-apply the Boolean algebra rules,

-compute the reducts,

-produce the rules using the reducts.

Firstly, we transform the decision table into discernibility matrix to compute the reducts. Let $DT = (U, C, D)$ be the decision table, with $U = \{R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8, R_9, R_{10}, R_{11}\}$. By a discernibility matrix of DT, denoted DM(T), we will mean $n \times n$ matrix defined as in (21):

$$m_{ij}^{a(R_i)} = \{(a \in C : a(R_i) \neq a(R_j)) and (d(R_i) \neq d(R_j))\} \tag{21}$$

where $i, j = 1, 2..11$ We construct the discernibility matrix, DM(DT) as in Figure 5, where the colour and texture-entropy are denoted by C, respectively T. The items within each cell are aggregated disjunctively, and the individual cells are then aggregated conjunctively.

To compute the *reducts* of the discernibility matrix we use the following theorems that demonstrate equivalence between *reducts* and prime implicants of suitable Boolean functions [18], [25]. For every object $R_i \in U$, the following Boolean function is defined:

$$g_{R_i}(Colour, Texture) = \bigwedge_{R_j \in U} \bigvee_{a \in m_{ij}} a \tag{22}$$

The following conditions are equivalent [18]:

1. $\{a_{i1}, .., a_{in}\}$ is a reduct for the object R_i, i = 1..n.
2. $a_{i1} \bigwedge a_{i2} \bigwedge .. \bigwedge a_{in}$ is a prime implicant of the Boolean function g_{R_i}.

Next, from each decision matrix we form a set of Boolean expressions, one expression for each row of the matrix.

For the gastric ulcer we obtain the following rules based on the table reducts:

1. $(C^{light-red} \bigvee T^{small}) \bigwedge (C^{light-red})$
2. $(C^{light-red} \bigvee T^{big}) \bigwedge (C^{light-red})$
3. $(T^{big}) \bigwedge (C^{light-yellow} \bigvee T^{big})$

For the duodenal ulcer we obtain the following rules based on the table reducts:

1. $(C^{light-yellow} \bigvee T^{medium}) \bigwedge T^{medium}$
2. $(C^{medium-yellow}) \bigwedge (C^{medium-yellow} \bigvee T^{small})$
3. $(C^{dark-yellow}) \bigwedge (C^{dark-yellow} \bigvee T^{small})$

On Boolean expression the absorption Boolean algebra rule is applied. The absorption law is an identity linking a pair of binary operations. For example: $a \bigvee (a \bigwedge b) = a \bigwedge (a \bigvee b) = a$.

By applying the absorption rule on the prime implicants, the following rules are generated:

Rule 1: $(Colour = light - red) \rightarrow$ gastric ulcer;
Rule 2: $(Texture - entropy = big) \rightarrow$ gastric ulcer;
Rule 3: $(Texture - entropy = medium) \rightarrow$ duodenal ulcer;
Rule 4: $(Colour = dark - yellow) \rightarrow$ duodenal ulcer.

3.4. Evaluation of decision rules and classification of new images. Decision rules can be evaluated along at least two dimensions: performance (prediction) and explanatory features (description). The performance estimates how well the rules classify new images. The explanatory feature estimates how interpretable the rules are [18].

Let be our decision table $DT = (U, C, D)$. We use the set-theoretical interpretation of rules that relates a rule to data sets from which the rule is discovered. Also, using the cardinalities of sets, we may define the support (s) and accuracy (a) of a decision rule as in equations 23 and and 24:

$$s(rule) = cardinality(featureSet \bigcap diagnosisSet) \tag{23}$$

$$a(rule) = \frac{cardinality(featureSet \bigcap diagnosisSet)}{cardinality(featureSet)} \tag{24}$$

where the set $featureSet \bigcap diagnosisSet$ is composed from image regions which have a certain featureSet and a certain diagnosis. In term of set theory, the accuracy is the degree in which the set of features rule is included in the set of diagnosis rule.

The coverage(c) of a rule is defined by:

$$c(rule) = \frac{cardinality(featureSet \bigcap diagnosisSet)}{cardinality(diagnosisSet)} \tag{25}$$

TABLE 3. Results recorded for different diagnoses

Diagnosis	Accuracy(%)	Specificity(%)
Duodenal Ulcer	96.4	95.2
Gastric Ulcer	96.7	95.4
Gastric Cancer	95.9	93
Rectocolitis	96.4	95.3

The coverage of a rule is the degree in which the set of diagnosis rule is included in the features set of rule.

For the generated Rule 1, the support is 4, accuracy is 4/4 and coverage is 4/5. Ryszard et al [26] suggests that high accuracy and coverage are requirements of decision rules.

After rule generation the process of image classification includes the following steps:
- collect all the decision rules in a classifier,
- compute for each rule the support, accuracy and coverage,
- eliminate the rules with the support less than the minimum defined support,
- order the rules by accuracy, than by coverage,
- if an image matches more rules select the first one: an image matches a rule, if all the semantic indicators, which appear in the body of the rule, are included in the characteristics of the image regions.

4. Experiments

The image collections used in our experiments were taken from free repositories on the Internet [27],[28]. Two image databases are used for learning and diagnosing process. The database used to learn the correlations between images and digestive diagnoses, contains 200 images and is categorized into the following diagnoses: duodenal ulcer, gastric ulcer, gastric cancer, esophagitis, and rectocolitis. The system learns each concept by submitting about 20 images per diagnosis. After classification, we counted: the number of true positives (images correctly diagnosed with a given diagnosis); the number of false positives (images incorrectly diagnosed with a given diagnosis); the number of true negatives (images correctly diagnosed with a different diagnosis); the number of false negatives (images incorrectly diagnosed with a different diagnosis). Based on this information we compute:

- the accuracy of classification, which measures the proportion of true results,
- the specificity of classification, which measures the capability of diagnosis rules not to miss the correct images, and not to diagnose images with a different diagnosis.

For the database diagnoses, the counted results are presented in Table 3, where we can observe that the sets of rules are very specific and with good accuracy.

5. Conclusion

Methods proposed and developed in this study could assist physicians by doing automatic diagnosing based on visual content of medical images. An important improvement of this paper is in the generation of rules with very high specificity using the Gaussian mixture model and rough set theory. We performed experiments on a medical image database, which includes endoscopies of the digestive apparatus.

References

[1] H. Muller, N. Michoux, D. Bandon and A. Geissbuhler, A review of content-based image retrieval systems in medical aplication-clinical benefits and future directions, *International Journal of Medical Informatics* **73** (2004), no. 1, 1–23.

[2] H. Greenspan and A.T. Pinhas, Medical Image Categorization and Retrieval for PACS Using the GMM-KL Framework, *IEEE Transaction on Information Technology in Biomedicine* (2007).

[3] O. Bodenreider, The Unified Medical Language System (UMLS): Integrating biomedical terminology, *Nucleic Acids Research* **32** (2004), 267–270.

[4] T. Deselaers, D. Keysers and H. Ney, FIRE - flexible image retrieval engine: ImageCLEF 2004 evaluation, *Multilingual Information Access for Text, Speech and Images* **3491** (2004), 688–698.

[5] W. Hong, B. Georgescu, X.S. Zhou, S. Krishnan, Y. Ma and D. Comaniciu, Database-guided simultaneous multi-slice 3D segmentation for volumetric data, *Proceedings of 9th European Conference on Computer Vision*, Graz, Austria, 2006, 397–409.

[6] C. Langlotz, Radlex: A new method for indexing online educational materials, *RadioGraphics* **26** (2006), 1595–1597.

[7] T. Lehmann, M. Gld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohne, H. Schubert and B. Wein, Content-based image retrieval in medical applications, *Methods Inf Med* **43** (2004), no.4, 354–361.

[8] C. Rosse and J.L.V. Mejino, A reference ontology for bioinformatics: The Foundational Model of Anatomy, *Journal of Biomedical Informatics* **36** (2003), 478–500.

[9] D. L. Rubin, P. Mongkolwat, V. Kleper, K. Supekar, D.S. Channin, Annotation and Image Markup: Accessing and Interoperating with the Semantic Content in Medical Imaging, *IEEE Intelligent Systems* **24** (2009), no. 1, 57–65.

[10] M. Stearns, C. Price, K. Spackman and A. Wang, SNOMED clinical terms: overview of the development process and project status, *Proceedings of American Medical Informatics Association Symposium*, 2001, 662–666.

[11] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society* **39** (1977), 1–38.

[12] H. Greenspan,Probabilistic models for generating, modelling and matching image categories, *Proceedings of the International Conference on Pattern Recognition*, 2002.

[13] J. Goldberger, H. Greenspan and J. Dreyfuss, Simplifying Mixture Models Using the Unscented Transform, *IEEE Transactions on Pattern Analysis and Machine Intelligence archive* **30** (2008), no.8.

[14] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, 2006.

[15] V. Garcia, F. Nielsen, Simplification and hierarchical representations of mixtures of exponential families, *Signal Processing* **90** (2010), no.12, 3197–3212.

[16] Z. Pawlak, Rough Relations, *Bulletin of the Polish Academy of Sciences, Technical Sciences* **34** (1986), no. 10, 587–590.

[17] Z. Pawlak and A. Skowron, Rough Membership Functions. In *Ronald R. Yager , Janusz Kacprzyk, Mario Fedrizzi (ed) Advances in the Dempster-Shafer Theory of Evidence*, Wiley, New York (1994), 251–271.

[18] J. Stepaniuk, *Rough Granular Computing in Knowledge Discovery and Data Mining*, Springer-Verlag, 2008.

[19] A.Ohrn, J. Komorowski, A. Skowron, et al., The Design and Implementation of a Knowledge Discovery Toolkit Based on Rough Sets - The Rosetta System. In *L.Polkowski, A. Skowron(ed.) Rough Sets in Knowledge Discovery 1, Methodology and Applications*, Physica-Verlag, Heidelberg (1998), 376-399.

[20] J. Bazan, M. Szczuk, The Rough Set Exploration System, Transactions on Rough Sets III, *LNCS* **3400** (2005), 37–56.

[21] J.W. Grzymala-Busse, *LERS-A Data Mining System*, Springer US, 2005.

[22] A.L. Ion, S. Udristoiu, Image Mining for discovering medical diagnosis, *Information Technology and Control* **39** (2010), no. 1, 123–129.

[23] A.L. Ion, A Framework for semantic modeling of images, *Annals of the University of Craiova, Mathematics and Computer Science Series* **37** (2010), no. 4, 37–49.

[24] A.L. Ion, S. Udristoiu, Automation of the Medical Diagnosis Process using Semantic Image Interpretation, *Advances in Databases and Information Systems, Lecture Notes in Computer Science* **6295** (2010), 234–246.

[25] A.E.Hassanien, A. Abraham, J.F. Peters et al., Rough Sets in Medical Imaging: Foundations and Trends. In *Computational Intelligence in Medical Imaging: Techniques and Applications*, CRC Press, USA, (2008), 47–87.
[26] A. Ryszard, S. Michalski, A Theory and Methodology of Inductive Learning. In *Readings in knowledge acquisition and learning*, Morgan Kaufmann, San Francisco (1993), 323–348.
[27] The Gatrolab Image Library, `http://www.gastrolab.net/`.
[28] Gastroenterology, `http://gicare.com/Endoscopy-Center/Endoscopy-images.aspx`.

(Anca Loredana Ion) Department of Software Engineering, Faculty of Automation, Computers and Electronics, Bvd. Decebal, Nr. 107, 200440, Craiova, Dolj, ROMANIA
E-mail address: `anca.ion@software.ucv.ro`