

Study on Various Defuzzification Methods for Fuzzy Clustering Algorithms to Improve ROIs Detection in Lung CTs

Alberto Rey
Faculty of Computer Science
University of A Coruna
A Coruna, Spain
Email: alberto.rey@udc.es

Bernardino Arcay
Faculty of Computer Science
University of A Coruna
A Coruna, Spain
Email: bernardino.arcay@udc.es

Alfonso Castro
Faculty of Computer Science
University of A Coruna
A Coruna, Spain
Email: alfonso.castro@udc.es

Abstract—The detection of pulmonary nodules is one of the most studied areas and challenging task in the field of medical image analysis, due the current relevance of the lung carcinoma. The difficulty and complexity of this task has led to the development of CAD systems for the automated detection of lung nodules in CT scans, which provides valuable assistance for radiologists and could improve the detection rate. A common phase of these systems is the detection of regions of interest (ROIs) that could be marked as nodules, in order to reduce the searching space problem. In this paper, we evaluate and compare the combination of various approaches of supervised vector machines (SVMs) with different kinds of fuzzy clustering algorithms, so as to improve the detection and segmentation of ROIs that could represent lung nodules in high resolution CT scans. These images are provided by the LIDC database (Lung Internet Database Consortium).

Keywords—Fuzzy Clustering, Support Vector Machines, Medical Image Analysis, Computer Aided Detection.

I. INTRODUCTION

The high complexity of the lung structure, with 23 airway subdivisions between trachea and alveolar sacs, and numerous lung diseases represented in images by pathological structures such as nodules, tumors, etc., have encouraged the analysis of computer pulmonary images. In this area, one of the most studied diseases is lung carcinoma, one of the most mortal carcinomas for which early detection is essential. The survival rate after 5 years for patients who were diagnosed and had surgery in a early stage is over 60%, but drops to 10% if the patients do not have surgery during these five years.[1]

Both primary and metastatic lung carcinoma most commonly manifest as pulmonary nodules, so researchers are focusing on the detection of these structures as a previous step for an early diagnosis.

The development of new modalities of image acquisition, such as high resolution CTs, has improved the detection and diagnosis of early nodules that, at their initial stage, may have a diameter below 3mm [2]. This technique offers a non-invasive way to observe the lung structure with great definition.

This work was supported by project 09SIN006PR from the I+D Suma Call of the Galician Plan for Research, Development and Technological Innovation (INCITE) from Xunta de Galicia.

Nevertheless, it is rarely used due to the high cost of the technology.

High resolution CTs studies may contain up to 600 images [2] whose examination and interpretation by radiologist is a very tedious and time-consuming task. Moreover, lung nodules are difficult to detect in CT images because they may appear attached to other non-pathological elements such as blood vessels or pleura, or have a small size and different shapes. For these reasons, currently new CADs (Computer Aided System) systems are being developed [3] [4] for the automated detection of lung nodules, providing support to the diagnosis of the experts. Our research group is developing such a system.

The first stage for these systems usually consists in preprocessing the CT images in order to isolate the lungs and reduce the search area where nodules could appear, removing those external elements that could interfere with the detection task. Lungs are therefore the only elements on each image of the CT scan. Our system follows this approach.

The second step consists in the detection of ROIs that could be treated as possible nodules. This paper presents the results obtained with a combination of different fuzzy clustering algorithms with various approaches of the SVM technique. The analysed fuzzy algorithms represent the state of the art in this field, obtaining good results for the detection of ROIs in lung CTs [5] [6]. The non-defuzzified output of these algorithms provides membership factors for each image pixel that represents with which grade they are classified into each segmented cluster. The principal aim of the study is to improve these results by applying the SVM technique to the membership factors, so as to obtain a better defuzzification phase with fewer rates of false positives in the segmentation of lung CT images. As this step is included in a CAD system, the analysis focuses on achieving a good method that reduces the false nodule candidate area, minimizing the deficiencies in the classification of the pixels provided by the fuzzy algorithms.

II. MATERIALS

In this paper, we used a training and simulation set that consists in several high-resolution CT lung images, provided

by the LIDC (Lung Image Database Consortium) database. This project, developed by the NCI (National Cancer Institute), mainly aims at developing and publishing a reference repository for the evaluation and implementation of CAD systems in lung nodules detection.[7]

The database consists in a repository of images in which each has associated relevant data regarding the presence or absence and the characteristics of the observed nodule(s) if present. This information is obtained from a process whereby four expert thoracic radiologists analyse and mark, regardless, each image. Later, if discrepancies in observations exist, they can be revised and corrected. The data is provided in a XML file that indicates its type and contour, specified by the pixels coordinates that conform it.

The provided studies are stored using the DICOM standard and have a size of 512x512 pixels, with a pixel size varying from 0,5 to 0,8 mm. The grayscale is of 12 bits in Hounsfield Units (HU).

The implementations of the different multi-class SVM approaches is provided by the BSVM library [8], that can be used freely for research purposes. The fuzzy algorithms were implemented with the ITK library, focused mainly on medical image analysis.

III. ANALYSED FUZZY ALGORITHMS

We selected three fuzzy clustering algorithms to carry out the detection task of the initial ROIs. They all provide less sensitivity to noise and outliers, are able to easily handle multidimensional problems, and provided the best results during the previous analysis [5] and [6] carried out on LIDC data sets.

A. SFCM (Spatial Fuzzy C-Means)

This algorithm [9] is a modified version of the FCM (Fuzzy C-Means) that incorporates spatial information in such a way that it alters the membership weighting value of each cluster after considering the distribution in the neighborhood. This spatial information provides the possibility to obtain more homogeneous regions and reduce the effect of noise.

In the first step, the algorithm calculates the initial membership values for each pixel, applying the standard FCM algorithm. Then, the spatial relationship is introduced by the following function:

$$h_{ij} = \sum_{k \in NB(x_j)} u_{ik} \quad (1)$$

where $NB(x_j)$ represents a resizable square window, centered on each pixel x_j under consideration. Therefore, the greater the value of this function, the greater the probability that pixel x_j belongs to cluster i th.

The next step calculates the new membership function using the previous spatial information:

$$u'_{ij} = \frac{u_{ij}^p h_{ij}^q}{\sum_{k=1}^C u_{kj}^p h_{kj}^q} \quad (2)$$

where p and q control the relevance of functions u_{ij} and h_{ij} .

In the last step, the new centroids are obtained following the same FCM procedure:

$$c_j = \frac{\sum_{i=1}^N u'_{ij}^m x_i}{\sum_{i=1}^N u'_{ij}^m} \quad (3)$$

The iteration process will stop when the maximal difference between two centroids lies below a determined threshold.

B. SKFCM (Spatial Kernelized Fuzzy C-Means)

This algorithm [11] introduces a penalty factor that contains spatial neighborhood information to the KFCM (Kernelized Fuzzy C-Means) algorithm proposed in the same work. Kernelized methods introduce a kernel function that allows us to transform the original low dimension input space into a higher dimensional feature space, where complex nonlinear problems can be treated and solved in a better way as probed by Cover [12]. The present paper only considers the Gaussian radial basis function kernel:

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right) \quad (4)$$

Therefore, modifying the objective function of the FCM in order to introduce the kernel function and add the penalty factor, we obtain the final objective function:

$$J S_m^\phi = \sum_{i=1}^c \sum_{k=1}^N u_{ik}^m (1 - K(x_k, v_i)) + \frac{\alpha}{N_R} \sum_{i=1}^c \sum_{k=1}^N u_{ik}^m \sum_{r \in N_k} (1 - u_{ir})^m \quad (5)$$

where N_k represents the square window which includes the neighbors of pixel x_k (without considering it), N_r is the cardinality of N_k , and α ($0 < \alpha < 1$) is a parameter that controls the effect of the penalty term. Deriving the Eq. 5 respect the variables u_{ik} and v_i obtains the following two conditions that minimize the function. An iterative algorithm can be derived from the above conditions.

In the initialization of the algorithm, the parameters c , i.e. the number of clusters, the initial class centroids, the initial memberships, and the threshold epsilon, must be determined.

In the first step of the iterative process, the memberships are calculated as follows:

$$u_{ik} = \frac{\left((1 - K(x_k, v_i)) + \frac{\alpha}{N_R} \sum_{r \in N_k} (1 - u_{ir})^m \right)^{\frac{-1}{(m-1)}}}{\sum_{j=1}^c \left((1 - K(x_k, v_j)) + \frac{\alpha}{N_R} \sum_{r \in N_k} (1 - u_{jr})^m \right)^{\frac{-1}{(m-1)}}} \quad (6)$$

Finally, the centroids are updated as follows:

$$v_i = \frac{\sum_{k=1}^n u_{ik}^m K(x_k, v_i) x_k}{\sum_{k=1}^n u_{ik}^m K(x_k, v_i)} \quad (7)$$

As in the other algorithms, repeat these steps until condition $\max_{i,k} \|u_{ik}^t - u_{ik}^{t-1}\| \leq \epsilon$ is satisfied, where epsilon is a determined threshold.

C. MKSFCM (Modified Kernelized Spatial Fuzzy C-Means)

This algorithm was proposed by Castro et Al. in [6], and consists in the combination of the algorithms SFCM, introduced in section III-A, and the FKCM (Fuzzy Kernel C-Means) proposed in [13]. The principal purpose is to join the main advantages of both methods: homogeneity and robustness.

Initially, this algorithm calculates the kernelized membership function, using the same equation as the FKCM:

$$u_{jk} = \frac{(1/d^2(x_j, V_k))^{1/(q-1)}}{\sum_{j=1}^C (1/d^2(x_j, V_k))^{1/(q-1)}} \quad (8)$$

where $d^2(x_j, V_k) = K(x_j, x_j) - 2K(x_j, V_k) + K(V_k, V_k)$, q is a fuzzy index that determines the fuzziness of the clusters, C is the number of clusters, V are the centroids, and K is the kernelized function.

The second step introduces the spatial information as in the SFCM procedure through the spatial function h_{ij} in Eq. 1, after which the Eq. 2 is applied to obtain the membership factors.

Finally, the new kernel matrix $K(x_j, \hat{V}_k)$ and $K(\hat{V}_k, \hat{V}_k)$ is calculated as follows:

$$K(x_j, \hat{V}_k) = \phi(x_j) \cdot \phi(\hat{V}_k) = \frac{\sum_{i=1}^N (u_{ik})^q K(x_i, x_j)}{\sum_{i=1}^N (u_{ik})^q} \quad (9)$$

To halt the iteration process, the error must lie below a determined threshold, otherwise a new iteration will start and Eq. 8 is again applied.

IV. SUPPORT VECTOR MACHINE FOR CLASSIFICATION PROBLEMS

The standard Support Vector Machines (SVM) were designed specifically for binary classification and regression estimation. However, they have also been extended to multi-class problems [14]. We opted for the multi-class approach in order to realize a more refined classification of the lung structures segmented in the previous stages, rather than dividing the marked zones into nodule or non-nodule.

These problems are usually solved by two different approaches. One consists of decomposing and combining several binary classifiers that the binary SVM can solve, the other consists of solving it directly by considering all data in one formulation. This paper compares the two methods outlined above. In the first approach, we selected the one-against-one Kreβel implementation [15], in the other we selected the one proposed by Weston and Watkins [16]. Note that there are other implementations for multi-class SVM [14], but the proposal of this paper is not so much to compare different SVMs but rather to obtain an idea of the improvements that these SVM approaches can provide us.

A. One-against-one

This method constructs $k(k-1)/2$ classifiers, where k is the number of classes. The i th SVM is trained on data from two classes, resulting in l training data $(x_1, y_1), \dots, (x_l, y_l)$, where $x_i \in R^n$, $i=1, \dots, l$ and $y_i \in \{1, \dots, k\}$ is the class of

x_i . Training data from the i th and the j th classes are solved according to the following classification problem:

$$\begin{aligned} w^{ij}, b^{ij}, \xi^{ij} \quad & \frac{1}{2} (w^{ij})^T w^{ij} + C \sum_t \xi_t^{ij} (w^{ij})^T \\ & (w^{ij})^T \phi(x_t) + b^{ij} \geq 1 - \xi_t^{ij}, \quad \text{if } y_t = i \\ & (w^{ij})^T \phi(x_t) + b^{ij} \leq -1 + \xi_t^{ij}, \quad \text{if } y_t = j \\ & \xi_t^{ij} \geq 0 \end{aligned} \quad (10)$$

where the training data x_t are mapped to a higher dimensional space applying function ϕ , and parameter C represents the penalty.

The decision of the ensemble is made using the max-wins voting strategy, in which each classifier is considered to be a voting. In this approach, Hsu and Lin use $f_{ij}(x) = ((w^{ij})^T \phi(x) + b^{ij})$ as a decision function. Let us denote the voting function by:

$$V_i(x) = \sum_{j=1, \dots, q, j \neq i} s(f_{ij}(x)) \quad (11)$$

where the function $s(t)$ maps the decision value to $\{0, 1\}$

$$s(t) = \begin{cases} 1, & \text{if } t \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

Therefore, when element x_t is in the i th class ($f_{ij}(x_t) > 0$) rather than class j , the vote for this element will increase its value by one in accordance with Eq. 11; otherwise it contributes 0. At the end of the voting process, the value of $V_i(x)$ will be the sum of all votes made by all classifiers that cast one vote for the i th class.

Consequently, the final decision is made as follows:

$$\text{Class Label of } i^*(x) = \underset{i=1, \dots, q}{\operatorname{argmax}} V_i(x) \quad (13)$$

i.e element x is classified into class i , which is the class with most votes from the classifiers. Remark that if two or more classes have the same number of votes, the class with the smallest index is selected.

B. Multi-class SVM

This kind of approach was introduced simultaneously by Vapnik [17] and Weston and Watkins [16]. In essence, these formulations are essentially identical and propose an approach to solve multi-class problems in a single optimization.

Like the one-against-one approach, this approach constructs k two-class decision rules where the m th function $w_m^T \phi(x) + b$ separates training vectors of the m class from the other are constructed. However, unlike the previous method, even though there are k decisions functions, all are obtained by solving a single optimization problem. The formulation of this is as follows:

$$\begin{aligned} w^{ij}, b^{ij}, \xi^{ij} \quad & \frac{1}{2} \sum_{m=1}^k w_m^T w_m + C \sum_{i=1}^l \sum_{m \neq y_i} \xi_i^m w_{y_i}^T \phi(x_i) \\ & + b_{y_i} \geq w_m^T \phi(x_i) + b_m + 2 - \xi_i^m, \\ & \xi_i^m \geq 0, \quad i = 1, \dots, l, \quad m \in \{1, \dots, k\} \setminus y_i \end{aligned} \quad (14)$$

The decision function is:

$$\operatorname{argmax}_{m=1,\dots,k}(w_m^T\phi(x) + b_m) \quad (15)$$

V. METHODOLOGY

For this paper we selected 23 preprocessed images of CT scans with isolated lungs, containing a representation of different kinds of lung nodules (initial phase, juxtaleural or wall-attached, isolated, etc.), as well as examples from different thoracic zones. These images, provided by LIDC database, each have an XML file with the nodule information, as specified in the material section. These files are converted into mask images so as to facilitate the measurement task of the quality results.

The process to combine the different techniques is initiated by the execution of the clustering algorithms, who perform a classification of the preprocessed images, obtaining the initial ROIs. Attending to the results in [5] and [6], it can be observed that the better results are obtained from partitioning the image into 3 clusters that represent lung tissue, regions marked as nodules, and background.

The construction of the training set, formed by the non-defuzzified output of the fuzzy algorithms as indicated in section I, follows a process that consists of several steps:

- 1) Select a set of pixels of each cluster, proportional to the percentage of each cluster in the segmented image. For example, if one cluster represents 20% of a segmented image, select only 20% of pixels of this cluster. Thereby, not only the training set is reduced in size, but the original characteristics of the input space are maintained as well. We observed during this analysis that better results were obtained by separating the ROI marked as nodule into two sets: one set with the true nodule pixels, another set with the remaining marked pixels, representing the false positives. Note that the results of this analysis follow this approach.
- 2) Obtain the membership factors for the previously selected pixels, using the non-defuzzified output of the fuzzy algorithms. These factors for each pixel are disposed as a row in the train set and represent one input for the SVM.
- 3) Repeat steps 1 and 2 for each image that serves as an input for the training set.

The training set used in the SVMs analysis is constructed, following this strategy, with three images of the dataset that represent different kinds of lung nodules. Since initially the dataset is not bigger, this number is sufficient to represent the input space, but it should be increased to boost the generalization capacity in the case of a bigger dataset.

The feature vector is defined as numerical values ranging from 1 to 4, and is associated to the appropriate train row, depending on the cluster that it represents.

The quality of the obtained results is quantified by means of true positives (TP) and false positives (FP), because the mask images only have information about the nodule region, omitting the information referring to the other elements of the CT image.

The comparative is divided into two approaches: first, analyse each fuzzy algorithm along with its possible combinations with the two SVM implementations, in order to select the optimal parameters; then, compare all combinations methods, to form an idea of the general improvements.

The results are represented in the form of charts and tables that summarize the best results obtained with different selections of parameters of the analysed algorithms. The principal aim is to observe the enhancement capacity of the fuzzy algorithms and its stability when they are combined with the SVM technique.

VI. RESULTS

Table I presents the optimal parameters values obtained for the different combinations of the fuzzy algorithms with SVMs and their respective kernels. For each method, we search the optimal parameters for the SVM executions, focusing on the regularization cost C and the kernel parameter gamma. These were estimated by varying their values into the ranges of [0.05, 2] and [10, 600], respectively. Moreover, we adopt 3 different kernels for each kind of SVM: RBF, Sigmoid, and Polynomial with grade 3. Therefore, for each method we try 264 combinations.

We started by analysing the SFCM algorithm with its SVM combinations. The true and false positive rates for this algorithm can be observed in the first row of each image represented on the x-axis of Fig. 1 and Fig. 2. The success rates are 100% in most cases and 76% for the worst result. Even though the false positives rate is not too high in general, in seven images this rate is above 15%-20%, reaching 36% in image number 20.

In order to select the best combination for each SVM approach with the SFCM algorithm, we compare the results obtained for a selected group of images that are more representative of the dataset and present a significant variability in our problem. The results obtained for the remaining images are represented in the charts because of limited space and their smaller relevance in the decisions. These results are given in Table II.

This table shows that the rates obtained for the different approaches and kernels are quite similar, except for the sigmoid kernels, which obtain a better true positive rate, the improvement being approximately 12% for images 5 and 7, as well as a slightly higher false positive rate than the other kernels

TABLE I
OPTIMAL VALUES FOR (C,GAMMA) FOR EACH ALGORITHM COMBINATION

Svm Approaches	Fuzzy Algorithms		
	SFCM	SKFCM	MKSFCM
OA0 ¹ -RBF	(0.05,10)	(0.3,500)	(0.05,10)
OA0-SIGMOID	(0.8,10)	(0.5,500)	(0.05,10)
OA0-POLYNOMIAL	(0.1,10)	(0.08,150)	(0.1,400)
MSVM ² -RBF	(0.05,10)	(0.5,300)	(0.1,10)
MSVM-SIGMOID	(0.5,10)	(1.5,10)	(0.05,100)
MSVM-POLYNOMIAL	(0.05,150)	(0.08,200)	(0.1,300)

¹ one-against-one

² multi-class svm (Weston's approach)

TABLE II
RESULTS OBTAINED FOR SOME REPRESENTATIVE IMAGES FOR THE SFCM
COMBINATIONS

Image		SVM Approaches					
		OA0 -RBF	OA0 -SIG. ¹	OA0 -POLY. ¹	MSVM -RBF	MSVM -SIG.	MSVM -POLY.
5	TP	77.08	95.83	77.08	85.42	95.83	77.08
	FP	8.2	9.08	7.95	8.62	8.9	7.93
7	TP	50	64.29	52.86	50	62.86	51.43
	FP	9.02	11.78	8.96	9.67	11.03	8.72
13	TP	59.26	70.37	59.26	59.26	62.96	59.26
	FP	6.79	8.49	6.54	7.33	8.1	6.37
18	TP	88.89	95.68	91.98	89.51	93.21	90.74
	FP	14.19	18.5	14.11	14.73	16.94	13.67
20	TP	82.86	82.86	80	82.86	82.86	80
	FP	18.88	25.19	17.95	19.57	22.81	17.47
21	TP	81.69	94.37	90.14	83.1	92.96	84.51
	FP	12.14	16.7	12.14	12.72	14.94	11.76

¹ Sigmoid

² Polynomial

(less than 4% for most cases). These results are extended to the one-against-one approach and the approach presented by Weston and Watkins. Therefore, Fig. 1 and Fig. 2 compare the initial rates for the fuzzy algorithm with the best results obtained for each approach of the SVM, in this case, using the sigmoid kernel.

Fig. 1 shows that, in general, the success rate is good and maintains the fuzzy results with values above 80%. The worst true positive rate is for images 7 and 13, reducing the success rate with respect to the initial segmentation above 20%.

In Fig. 2 it can be observed that the false positive rates is reduced for all the datasets, except for image 5, where the rate is slightly higher but still below 0.2%, so it is not relevant. For five images, the reduction reaches 10%.

Therefore, this combination provides good results, maintaining the true positives rate of the initial fuzzy segmentation, and realizes a significant reduction of the false positive ratio for most images.

The second analysed algorithm was the SKFCM and its combinations with the approaches of the SVM. The results, indicated in the first row of each image in Fig. 3, show a high

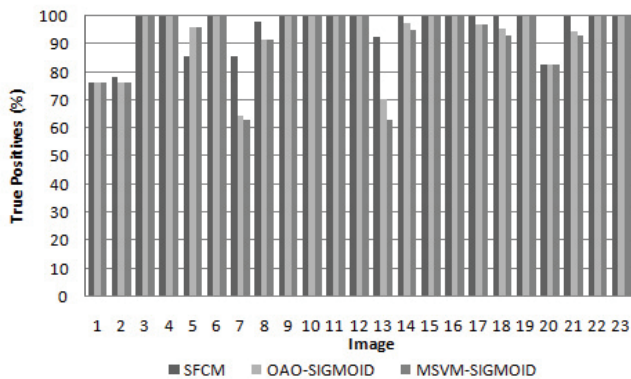


Fig. 1. True positive rates for the SFCM algorithm and the best combinations.

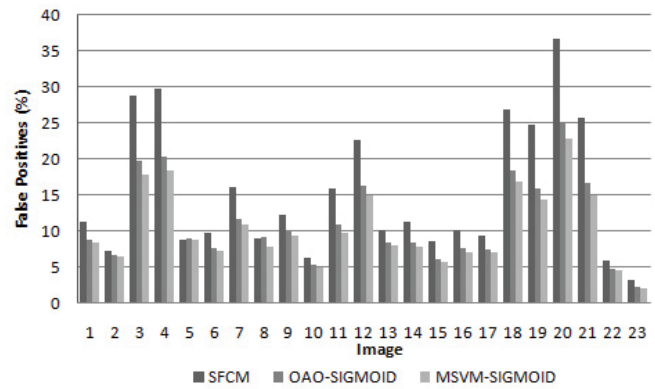


Fig. 2. False positive rates for the SFCM algorithm and the best combinations.

true positive rate for most images, the results being almost identical to those results of the SFCM method, except for small variations. Nonetheless, the false positive rate, illustrated in Fig. 4, has increased for all the images that do not fall in the 10%-20% range and reach 37% and 42% for images 22 and 2, respectively.

Table III shows the results obtained for the group of variability images. These reveal a classification problem with images number 7 and 21 in most SVM combinations with their different kernels, because the success rate decreases below 10%. This indicates that the nodules are misclassified in other cluster, probably because these images have smooth nodules with small intensity variations that are difficult to detect and distinguish from the other structures.

The best combinations for the SKFCM algorithm are the one-against-one approach with polynomial kernel and the Weston and Watkins approach with sigmoid kernel. For the one-against-one methods, the results are not good because the tests with RBF and sigmoid kernels have misclassification problems (cf. above); for the polynomial kernel, the false positive rates for some images (e.g image 20) have very high value that nullify our goals. In the same way, note that for

TABLE III
RESULTS OBTAINED FOR SOME REPRESENTATIVE IMAGES FOR THE SKFCM
COMBINATIONS

Image		SVM Approaches					
		OA0 -RBF	OA0 -SIG.	OA0 -POLY.	MSVM -RBF	MSVM -SIG.	MSVM -POLY.
5	TP	83.33	81.25	93.75	83.33	70.83	97.92
	FP	24.43	24.05	28.19	24.39	23.58	28.87
7	TP	4.29	1.43	22.89	4.29	54.29	28.57
	FP	6.74	6.53	14.56	6.9	6.47	15.64
13	TP	55.56	51.58	77.78	55.56	29.63	88.89
	FP	8.31	7.99	14.23	8.38	7.74	15.11
18	TP	26.54	24.07	43.21	26.54	61.11	48.77
	FP	11.79	11.14	25.76	11.55	10.34	27.99
20	TP	82.86	82.86	82.86	82.86	65.71	82.86
	FP	33.14	30.16	47.77	30.53	14.2	45.81
21	TP	8.45	5.63	26.76	7.04	71.83	35.21
	FP	9.61	9.02	21.4	9.43	6.1	23.78

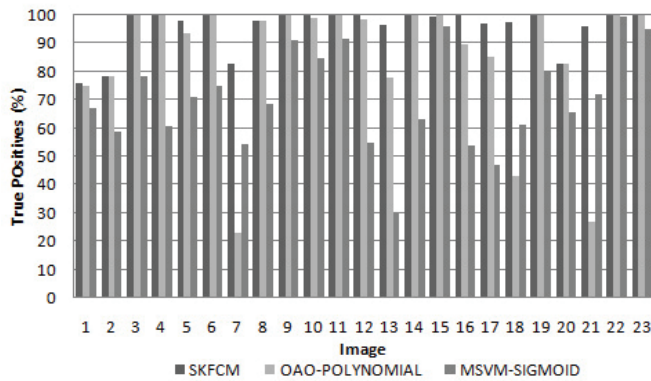


Fig. 3. True positive rates for the SKFCM algorithm and the best combinations.

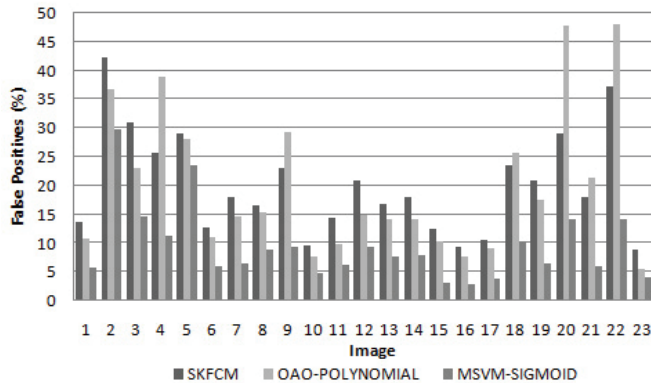


Fig. 4. False positive rates for the SKFCM algorithm and the best combinations.

this last approach, image 20 has very high false positive rates except for the sigmoid kernel; hence we choose this combination instead of others with better true positive rates.

The results obtained for the Weston approach with sigmoid kernel (Fig. 3 and Fig. 4) show the success rates for the selected variability images. These decrease in a high percentage, 54% and 29% being the true positive rate of images 7 and 13, respectively. The remaining images have rates up to 60%. In Fig. 4, the false positive rate experiments a high decrease, reaching 23% in the case of image 22, and 5%-15% for the remaining images.

In essence, the SKFCM algorithm does not function well for data sets with high variability because of the misclassification problems; it is therefore not a good election for the CAD system.

Finally, the results obtained by the MKSFCM algorithm, shown in Fig. 5 and Fig. 6, reveal success rates of approximately 100% for most images; even though in this case the worst rate is 68% rather than 76% for the previous methods, this reduction is not critical. It compensates for the decrease of the false positive ratio in all images, most of them falling into the 5%-15% rates and none above 21%.

Table IV introduces the results of the combinations for this fuzzy algorithm. These manifest a stable behavior of the

TABLE IV
RESULTS OBTAINED FOR SOME REPRESENTATIVE IMAGES FOR THE MKSFCM COMBINATIONS

Image		SVM Approaches					
		OAO -RBF	OAO -SIG.	OAO -POLY.	MSVM -RBF	MSVM -SIG.	MSVM -POLY.
5	TP	62.5	68.75	72.92	64.58	62.5	77.08
	FP	17.11	17.37	17.92	17.24	16.81	17.99
7	TP	60	62.86	64.29	62.86	58.57	64.29
	FP	10.25	10.52	11.13	10.41	9.83	11.04
13	TP	81.48	81.48	92.59	81.48	74.07	81.48
	FP	10.2	10.51	11.04	10.36	9.87	10.98
18	TP	82.72	84.57	87.04	82.72	79.63	85.8
	FP	11.42	11.81	12.51	11.65	11.03	12.39
20	TP	80	81.43	82.86	81.43	80	81.43
	FP	7.14	7.41	8.19	7.25	6.87	8.06
21	TP	66.2	70.42	77.46	70.42	61.97	77.46
	FP	8.89	9.23	9.81	9.07	8.85	9.73

different SVM approaches, highlighting those obtained by the polynomial kernel with a slightly better success rate and a restrained false positive ratio. Fig. 5 and Fig. 6 analyse the two SVM approaches with polynomial kernel.

Fig 5 indicates a maintenance of the true positive ratio in concordance with the initial fuzzy segmentation; in some

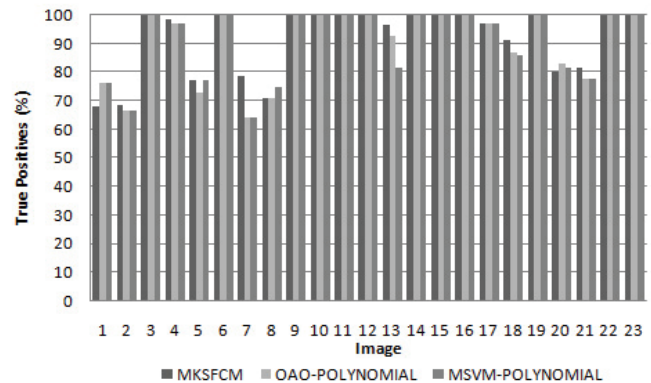


Fig. 5. True positive rates for the MKSFCM algorithm and the best combinations

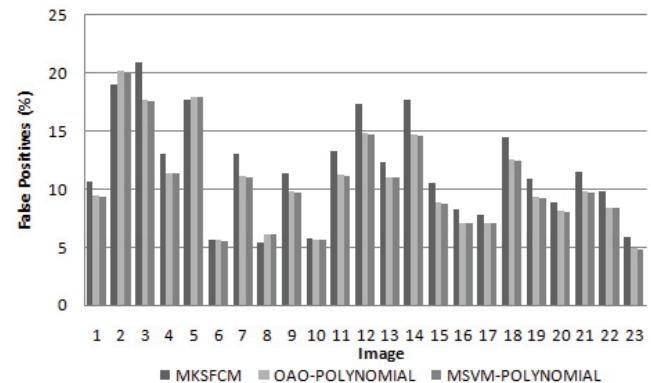


Fig. 6. False positive rates for the MKSFCM algorithm and the best combinations.

images (e.g. images 1,8, and 20) this rate is improved with an increase of up to 8% for the first image. The false positive ratio, in Fig. 6, decreases for most images in terms of a stable 2%. There are some images (e.g 2 and 8) for which this rate increases, but less than 0.5%, so the worsening is slightly marked.

The two analysed polynomial combinations of this algorithm present a stable behaviour with good rates for true and false positives, although the decrease of this final rate is not high because the initial fuzzy segmentation presented low values and it is difficult to realize a better classification.

VII. CONCLUSIONS

This paper presents an analysis of the combination of various fuzzy clustering algorithms with different approaches of the SVM technique. The principal aim of the study is to improve the defuzzification phase of the analysed fuzzy algorithms, obtaining better segmentations with fewer rates of false positives, by applying the SVM technique to the membership factors obtained in the initial ROIs classification.

The best combinations obtained for the SFCM and MKSFSCM algorithms provide good results, being able to maintain and increase in some cases the fuzzy success rate and decrease the false positive ratio in a significant value for most of the images.

In the case of SFCM, the better combination was that provided by the one-against-one approach with sigmoid kernel obtaining a general true positive ratio above 80% for most images and reducing the false positives rate up to 10% for various images. Fig. 7 shows a comparative of the improvements introduced by this approach applied to an image of the dataset. The left upper image represents the preprocessed image, where the lungs are isolated, the right upper image is the output obtained by the fuzzy algorithm without any combination, and finally, the image below them is the defuzzification result provided by the best combination.

The best combination for MKFSCM was the one-against-one approach, as for SFCM, but in this case with the polynomial kernel. This obtains the best and most robust results, improving even the success rates for some images up to 8% and preserving it for the rest of images. In the case of false positive rate, the combination reduces the good rates of false positives (already obtained by previous results of the fuzzy algorithm) for all images but in a minor way if compared with the better SFCM combination. As in the previous case, Fig. 8 indicates the results obtained by this approach.

Once again, the SKFCM fuzzy combinations have classifications problems with data sets with high variability, due to the presence of images with nodules in early stages or that are hard to differentiate from the other lung structures. In this case, the results are not good for our problem, because the low true positive rate is less than 10% in some images and the high true positive rate is above 40%.

As a result, the SFCM and MKFSCM fuzzy clustering algorithms with their optimal SVM combinations can serve to improve the defuzzification phase in the task of ROIs detection.

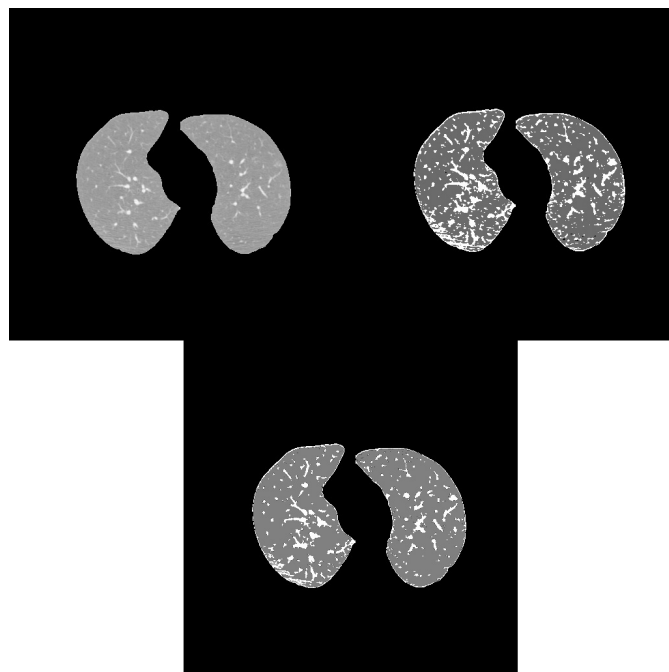


Fig. 7. Example of result obtained with the best combination of SFCM with SVM.

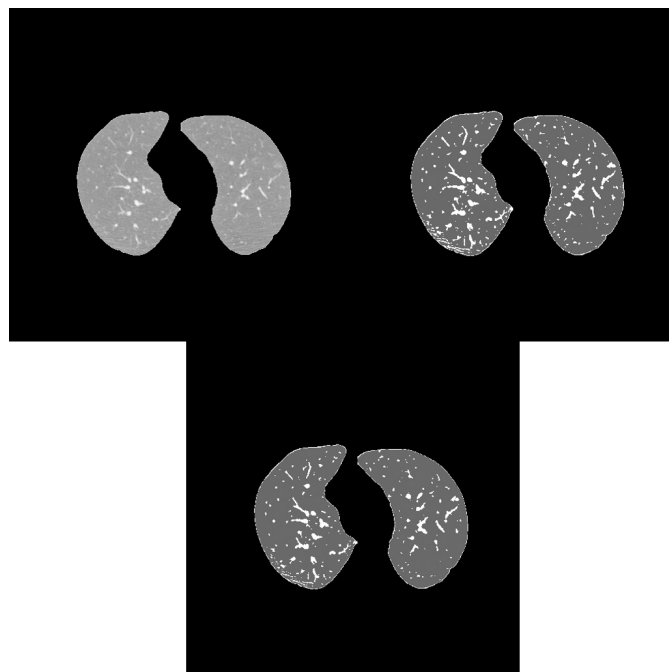


Fig. 8. Example of result obtained with the best combination of MKSFSCM with SVM.

In our opinion, this result can be incorporated into the lung nodule detection CAD system.

REFERENCES

- [1] H. K. Weir et al., "Annual report to the nation on the status of cancer, 1975-2000," *J. Natl. Cancer Inst.*, vol. 95, no. 17, pp. 1276-1299, 2003.
- [2] X. Lu, G. Wei, J. Qian, and A. K. Jain, "Learning-based Pulmonary Nodule Detection from Multislice CT Data," in *Proc. CARS*, 2004.

- [3] Y. Lee, T. Hara, H. Fujita, S. Itoh, and T. Ishigaki, "Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique," *IEEE Trans. Med. Imaging*, vol. 27, no. 7, pp. 595-604, July 2001.
- [4] C. Lei, L. Xiaojian, Z. Jie, and C. Wufan, "Automated lung segmentation algorithm for CAD system of thoracic CT," *J. Med. of PLA*, vol. 23, no. 4, pp. 215-222, August 2008.
- [5] A. Castro, C. Boveda, A. Rey and B. Arcay, "An analysis of various clustering algorithms for ROI detection in high resolution CT lung images," *Lect. Notes Comput. Sc.*, vol. 6374, Warsaw, Poland, 2010.
- [6] A. Castro and B. Arcay, "Comparison of various fuzzy clustering algorithms in the detection of ROI in lung CT and a modified kernelized-spatial fuzzy c-means algorithm," *Proc. of 10th IEEE Int. Conf. on Inf. Tech. and Appl. in Biom.*, Corfu, Greece, 2010.
- [7] S.G. Armato III, G. McLennan, M.F. McNitt-Gray for the Lung Image Database Consortium Research Group, "Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community," *Radiology*, 232, pp. 739-748, 2004.
- [8] C.W. Hsu and C. J. Lin. (April, 2006). *BSVM: A Library for Support Vector Machines*. [Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/bsvm>.
- [9] K. Chuang, H. Tzeng, S. Chen, J. Wu, and T. Chen, "Fuzzy C-means clustering with Spatial information for image segmentation," *Comput. Med. Imaging Grap.*, no. 30, pp. 9-15, 2006.
- [10] S. Chen, D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure," *IEEE Trans. Syst. Man Cybern. Part B-Cybern.*, vol. 34, no. 4, pp. 1907-1916, Aug. 2004.
- [11] D. Q. Zhang and S. C. Chen, "A novel kernelized fuzzy C-means algorithm with application in medical image segmentation," *Artif. Intell. Med.*, vol. 32, no. 1, pp. 37-50, Sept. 2004.
- [12] T.M. Cover, "Geomeasure and statistical properties of systems of linear inequalities in pattern recognition," *IEEE Trans. Electron. Comput.* vol. EC-14, pp. 326-334, Mar. 1965.
- [13] W.D. Zhong, X.X. Wei, and Y.P. Jian, "Fuzzy C-Means clustering algorithm based on kernel method," *Proc. of 5th Int. Conf. on Comp. Intell. and Mult. Appl.*, Xi'an, China, 2003, pp. 49-54.
- [14] C. Hsu, and C. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. on Neural Netw.*, vol. 13, no. 2, pp. 415-425, March 2002.
- [15] U. Kreßel, "Pairwise classification and support vector machines," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C.J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, ch. 15, pp. 255-268.
- [16] J. Weston and C. Watkins, "Multi-class support vector machines," *Proc. 7th Eur. Symp. On Artif. Neural Netw.*, Burges, Belgium, 1999, pp. 219-224.
- [17] V. Vapnik, *Statistical Learning Theory*, New York: John Wiley & Sons, 1998.