Classification of Chest Lesions with Using Fuzzy C-Means Algorithm and Support Vector Machines

Donia Ben Hassen[1], Hassen Taleb[1], Ismahen Ben Yaacoub[2], and Najla Mnif[2]

[1] LARODEC Laboratory, Higher Institute of Management, University of Tunis, Tunisia
donia_ben_hassen@yahoo.fr
[2]Medical Imaging Department, University Hospital Charles Nicolle, Tunisia

Abstract. The specification of the nature of the lesion detected is a hard task for chest radiologists. While there are several studies reported in developing a Computer Aided Diagnostic system (CAD), they are limited to the distinction between the cancerous lesions from the non-cancerous. However, physicians need a system which is significantly analogous to a human judgment in the process of analysis and decision making. They need a classifier which can give an idea about the nature of the lesion. This paper presents a comparative analysis between the classification results of the Fuzzy C Means (FCM) and the Support Vector Machines (SVM) algorithms. It discusses also the possibility to increase the interpretability of SVM classifier by its hybridization with the Fuzzy C method.

Keywords: Chest lesions, Clustering, Features, FCM, SVM.

1 Introduction

Radiography continues to be the most widely used imaging technique of the initial detection of chest diseases because of its low cost, simplicity, and low radiation dose. Though, uncertainty is widely present in data in this modality. Computer-assisted approaches may be helpful for handling this vagueness and as a support to diagnosis in this field. Therefore, the development of a reliable computer aided diagnosis (CAD) system for lung diseases is one of the most important research topics. Despite, lesion classification systems provide the fundation for lesions diagnosis and patient cure, the studies reported in developing a CAD application was limited to the distinction between the cancerous lesions from the non-cancerous. Physicians need a system which is significantly analogous to a human judgment in the process of analysis and decision making. The design of a classifier which can give an idea about the nature of the lesion, for example the lesion is of 50% an infection, 10% a cancer and 30% a tuberculosis etc… can help the radiologist to be suitable for handling a decision making process concerning.

To reach this goal, we propose a comparison study for chest lesions classification based on Fuzzy C Means (FCM) and Support Vector Machines (SVM) methods.

Á. Herrero et al. (eds.), *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*, Advances in Intelligent Systems and Computing 239,
DOI: 10.1007/978-3-319-01854-6_33,

The paper is organized as follows: after considering related works in section 1, section 2 describes the different classification systems and section 3 computerized schemes of our CAD system. Section 4 presents the results obtained on a real datasets.

2 Related Works

Many methods have been proposed in the literature for chest lesions classification and diagnosis utilizing a wide variety of algorithms. The majority of researches include the classification process under a description of whole Computer Aided Diagnosis systems. We can discern two main classes of studies concerning the classification of lesions in chest radiographs. The first class considers the classification process as distinction between true lesions and normal tissues in order improve radiologists' accuracy in detecting lung nodules. The work of [1] is an example. The second class adds to the first type of classification another one which distinguishes between the benign lesions and the malign ones. [2] proposed a system which automatically detects lung lesions from chest radiographs. The system extracts a set of candidate regions by applying to the radiograph three different multi-scale schemes. Support Vector Machines (SVMs), using as input different sets of features, has been successfully applied for the classification of chest lesions to benign and malign. [3] has used image processing algorithms for nodule classification to cancerous and non-cancerous tumors. We found that the majority of work in computer aided diagnosis systems in chest radiography has focused on lung cancerous nodule detection. Considering the load of lung diseases and the position of chest radiography in the diagnostic workflow of these diseases, we could argue that the classification of other type of lesions such as tuberculosis should receive more attention. We think also that an adapted framework for extraction of the adopted features describing the different kinds of lesions is the missing part of the methods presented in the literature.

3 Description of the Classification System

The classification problem has been addressed in many computing applications such as medical diagnosis. In the literature we can find the term grouping or clustering instead of classification. Although these two terms are similar in terms of language, they are technically different. In fact, classification is a discriminant analysis that uses a supervised approach where we are provided with a collection of labeled data; the problem is to label a newly unlabeled data. Usually, the given training data is used to learn the descriptions of classes which in turn are used to label a new data. Among these methods are bayes classifier, artificial neural networks, deformable models and support vector machines. Normally, clustering always refers to unsupervised framework which its problem is to group a given collection of unlabeled patterns into meaningful clusters. There is a whole family of unsupervised methods, including probabilistic ones, fuzzy ones, evidential ones. Especially, there are two variants of unsupervised

classifiers: Classifiers are known as hard methods and fuzzy methods. The commonly used fuzzy clustering methods are: FCM Fuzzy C- Means and its variants.

3.1 Fuzzy C-Means Algorithm

In this context, the fuzzy diagnosis concept is widely applied [4]. Fuzzy classifiers have been proposed to deal with classification tasks in presence of uncertainty. Fuzzy c-means (FCM) is one of these methods. It is an algorithm of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn [5] and improved by Bezdek[6]) is frequently used in pattern recognition.

The fuzzy c-means algorithm is an extension of the classic k-means algorithm [7] based on the minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \ , \ 1 \le m \le \infty \tag{1}$$

Where N is the number of data points, C represents the number of cluster center, m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{K=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i c_k \right\|} \right)^{\frac{2}{m-1}}} \qquad , \ C_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{2}$$

This iteration will stop when $max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k is the iteration number. This procedure converges to a local minimum or a saddle point of J_m.

With the work of Bezdek, fuzzy clustering methods attain a certain maturity [8]. Other variants of these algorithms were then developed in order to increase performance. These improved versions are often dedicated to a particular application, and still the FCM are generally useful in many situations. Some variants of FCM have been proposed to reduce the influence of data points which do not belong to a cluster.

Recently, many researchers have brought forward new methods to improve the FCM algorithm [9]. The most popular is: Possibilistic C-Means algorithm.

The Fuzzy C means algorithm can provide faster approximate solutions that are suitable for the treatment of issues related to understandability of models and incomplete and noisy data. This makes the technique effective in classification and very close to reasoning of physicians.

3.2 SVM Classification

Support vector machines (SVM) represent a classifier that has been successfully used for chest lesions classification. Moreover, we possess labeled data and it is expected that a supervised classifier achieves a good accuracy.

Let a set of data $(x_1, y_1), \ldots, (x_m, y_m) \in \Re^d \times \{\pm 1\}$ where X={ x_1, x_m} a dataset in $\Re^d$ where each x_i is the feature vector of an image. In the nonlinear case, the idea is to use a kernel function $k(x_i, x_j)$, where $k(x_i, x_j)$ satisfies the Mercer conditions [10]. Here, we used a Gaussian RBF kernel whose formula is:

$$k(x, x') = \exp\left[-\left\|x - x'\right\|^2 \Big/ 2\gamma^2 \right] \tag{3}$$

Where $\|.\|$ indicates the Euclidean norm in $\Re^d$.

Let Ω be a nonlinear function which transforms the space of entry $\Re^d$ to an intern H called a feature space. Ω allows to perform a mapping to a large space in which the linear separation of data is possible [11].

$$\begin{gathered} \Omega: \Re^d \longrightarrow H \\ (x_i, x_{ji}) \longmapsto \Omega(\mathrm{x_i})\Omega(\mathrm{x_j}) = \mathrm{k}(\mathrm{x_i}, \mathrm{x_j}) \end{gathered} \tag{4}$$

The H space is reproducing Kernel Hilbert space (RKHS). Thus, the dual problem is presented by a Lagrangian formulation as follows:

$$\max \mathrm{W}(\alpha) = \sum_{\mathrm{i}=0}^{\mathrm{m}} \alpha_\mathrm{i} - \frac{1}{2} \sum_{\mathrm{i,j}=1}^{\mathrm{m}} \mathrm{y_i y_j} \alpha_\mathrm{i} \alpha_\mathrm{j} \mathrm{k}(\mathrm{x_i}, \mathrm{x_j}), \mathrm{i} = 1, \ldots, \mathrm{m} \tag{5}$$

Under the following constraints:

$$\textstyle\sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_\mathrm{i} \mathrm{y_i} = 0, 0 \leq \alpha_\mathrm{i} \leq \mathrm{C} \tag{6}$$

They α_i are called Lagrange multipliers and C is a regularization parameter which is used to allow classification errors. The decision function will be formulated as follows:

$$\mathrm{f(x)} = \mathrm{sgn}(\textstyle\sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_\mathrm{i}\, \mathrm{y_i k(x, x_i)} + \mathrm{b}) \tag{7}$$

We hence adopted one approach of multiclass classification: One-against-One. This method consists of creating a binary classification of each possible combination of classes, the result for K classes K (K -1) / 2 .

4 Computerized Scheme for Classification of Lung Lesions

Here, we only describe the final stage of our CAD system: The segmentation and the feature selection are discussed in our previous works [12] and [13] that are briefly presented in the next sections.

4.1 Preprocessing

Preprocessing may lead to better results because it allows more flexibility. For example, the contrast feature is almost meaningless in discriminating normal and lesion pixels, since lesions can occur in regions of both high and low contrast. However, combining this feature with image enhanced intensity can be used for much more effective discrimination.

4.2 Segmentation

We closely follow our previous work described in [12] which is an automatic chest radiography segmentation framework with integration of spatial relations. The algorithm developed as the initial step of our system works under no assumption. The results obtained proving that this is an excellent initialization step for a CAD system aimed at lung lesions detection and recognition of their sites. The segmented lung area includes even those parts of the lungs which are usually excluded from the methods presented in the literature. This decision is motivated by the recognition of the site of the lesion which can help in deducing its nature for example, "lesion is in the right apical" that means that the lesion is localized in the right upper lobe it's an infection. The segmented area is processed with a method that enhances the visibility of the lesions, and an extraction scheme is then applied to select potential features.

4.3 Feature Extraction and Selection

The purpose of features extraction and selection is to reduce the original data set by measuring certain properties that distinguish one input pattern from another pattern. The extracted feature should provide the characteristics of the input type to the classifier by considering the description of relevant properties of the image into a feature space. We believe that the calculation of features is a primordial step to well perform the task of classification. In fact, each pixel should have characteristics used for the differentiation between the lesion and the normal pixels nevertheless for the discrimination between malign and benign lesions. A great variety of features can be computed for each image such that intensities and textures depending on the nature of problem. The description of features in works cited above is not very detailed. Almost features cited in the literature are classical. Between them those of Haralick, were used without really giving details about their meaning. However, it is difficult to interpret what these features are. Based on characteristics given by service of medical imaging of CHU Charles Nicolle (described in table 1), we selected 8 features (size, circularity, x-fraction, y-fraction, skewness, kurtosis, homogeneity, correlation) that we believe are able to specify the nature of the lesion.

Table 1. Characteristics of principals lesions for radiologists

LESION	METASTASIS	BENIGN TUMORS	MALIGN TUMORS	TUBERCULOSIS	INFECTION
UNIQUE	Exceptional	indifferent	Very common		
MULTIPLES	Very common	indifferent	rare	Common	
LOCALISATION	Basal			Apical	
CONTOURS	sharp	sharp	sharp	blurred	blurred
FORME	Rounded, ovalaire	Rounded, ovalaire polylobed	spiculated	Ill-defined	Ill-defined or fissural limit
OPACITY	homogeneous	homogeneous, dense	heterogene-ous +/- excavated	heterogeneous +/- excavated	homogene-ous
CALCIFICATIONS	rare except metastasis of sarcoma	frequent central lobulated	rare	Frequent in sequelae stage	Very rare
SIZE		< 1cm	> 1 cm		

5 Experimental Study

The chest radiographs are taken from the JSRT database [14]. This is a publicly available database with 247 Posterior Anterior chest radiographs. 154 images contain exactly one pulmonary lung lesion. The other 93 images contain no lung lesions.

First of all, we defined the number of clusters. We discussed with our collaborators in the service of medical imaging of the university hospital Charles Nicolle and we concluded the five important clusters which are: Lung cancer, Metastasis, Tuberculosis, Infection, Benign tumors.

Table 2. Classes of diseases and number of samples in the database used for performance Evaluation

Classes	Train	Test	Total number
Cancer	61	30	91
Metastasis	6	3	9
Infection	16	8	24
Tuberculosis	12	6	18
Benign tumor	5	3	8

The obtained feature vectors passed for the classification phase by using FCM and SVM's.

In the experiments of FCM, we have used a fuzziness coefficient m = 2.

Step 1. Our dataset contains samples of features belonging to five diseases.

Step 2. The data to be clustered is 8-dimensional data and represents features cited above. From each of the Five groups (Lung cancer Metastasis Tuberculosis Infection Benign tumors), two characteristics (for example, X-fraction vs. Y-fraction as shown in fig.1) of the lesion are plotted in a 2-dimensional plot.

Step 3. Next, the parameters required for Fuzzy C-Means clustering such as number of clusters, exponent for the partition matrix, maximum number of iterations and minimum improvement are defined and set.

Step 4. Fuzzy C-Means clustering is an iterative process. First, the initial fuzzy partition matrix is generated and the initial fuzzy cluster centers are calculated (show the centers in magenta in Fig. 1). In each step of the iteration, the cluster centers and the membership grade point are updated and the objective function is minimized to find the best location for the clusters. The process stops when the maximum number of iterations is reached, or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified.

Table 3 presents the results obtained with fuzzy C Means algorithm.

In table 3 also, we present the results obtained with SVM classifier with parameters C,γ ($2^{(1)}$,$2^{(-7)}$) settings of Gaussian RBF kernel. We have used the grid search which searches the optimal parameters values using cross validation. After learning phase, we test the test data.

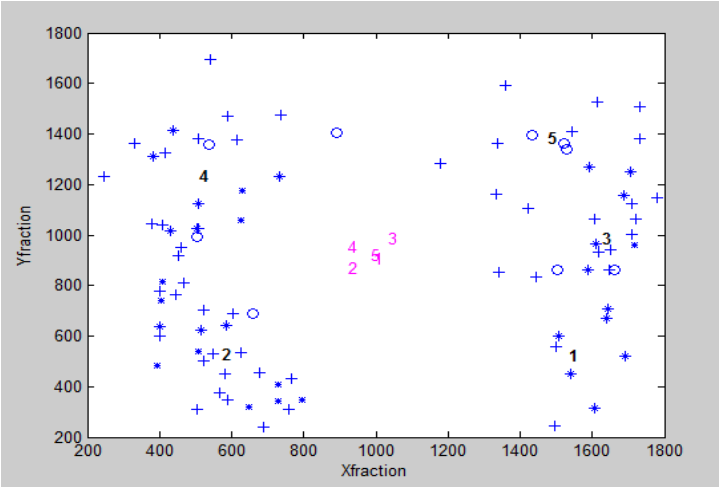

Fig. 1. The two characteristics (X-fraction vs. Y-fraction) of the lesion are plotted in a 2-dimensional plot

Table 3. Performances of FCM and SVM classifier

Classes	Acccuracy (%)		SE(%)		SP(%)	
	FCM	SVM	FCM	SVM	FCM	SVM
Cancer	70.33	74.19	73.00	80.00	71.33	60.00
Metastasis	60.25	50.94	63.88	66.66	61.56	50.00
Infection	55.66	51.72	52.76	50.00	53.00	52.00
Tuberculosis	60.23	56.56	55.85	53.33	57.00	51.88
Benign tumors	53.33	52.00	54.67	51.45	53.89	50.00

We judge the performance of our classification approach using several evaluation criteria often used in the literature. For a five class clustering problem, one can distinguish true positive (TP) (sample correctly classified), false positive (FP) (false sample classified as true sample), false negative (FN) (false sample classifier as false sample), and true negative (TN) (false sample classified as true sample). From these values, measures such as accuracy, sensitivity (SE) and specificity (SP) can be computed given by the following equations.

$$SE = \frac{TP}{TP+FN} \tag{8}$$

.

$$SP = \frac{TN}{TN + FP} \tag{9}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

The results in table 3 show that the clusters are identified well by the FCM algorithm.

The clusters that contain more errors are those that contain fewer samples. The lung cancer is the cluster which is more clearly identified by the FCM algorithm (achieved an averaged accuracy rate of the order 70.33 % for cancer).

The obtained results by SVM are satisfactory. Indeed, we reached a recognition rate of the order 74.19% for the cancer. We remark that the classification rate is less in the other classes because the fewer number of the train and the test data.

We can conclude that the SVM can achieve better accuracy in our classification problem. However, the two methods cannot represent clusters of small size.

The Fuzzy C Means Algorithm uses a fuzzy clustering, in which the input vector x is pre-classified with different membership values. The outputs of the FCM algorithm may present the input vector to the SVM classifier. This last will be used for the automatic recognition of disease.

6 Conclusion and Future Works

The intelligibility is the motive force behind the use of FCM algorithm for this problem. However, a compromise between interpretability and accuracy is met. On the other hand, we focused on a more accurate solution by using SVM. Then we risk losing the linguistic sense defining the fuzzy models. Indeed, we have experiment also the possibility to increase the interpretability of SVM classifier by the hybridization with the clustering method Fuzzy C means.

References

1. Hardie, R., Rogers, S., Wilson, T., Rogers, A.: Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. Medical Image Analysis 12(3), 240–258 (2008)
2. Campadelli, P., Casiraghi, E., Valentini, G.: Lung nodules detection and classification. In: ICIP (1), pp. 1117–1120 (2005)
3. Nehemiah, H., Kannan, A.: An intelligent system for lung cancer diagnosis from chest radiographs. International Journal of Soft Computing, 133–136 (2006)
4. Masulli, F., Schenone, A.: A fuzzy clustering based segmentation system as support to diagnosis in medical imaging. Artificial Intelligence in Medicine 16, 129–147 (1999)
5. Dunn, J.C.: A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics 3, 32–57 (1973)
6. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
7. Lesot, M.J., Bouchon-Meunier, B.: Descriptive concept extraction with exceptions by hybrid clustering. In: Proc. of Fuzz-IEEE 2004, pp. 389–394. IEEE Comp. Intell. Society, Budapest (2004)
8. Khodja, L.: Contribution à la classification floue non supervisée. Thesis. Savoie University, France (1997)

9. Gomathi, M., Thangaraj, P.A.: New Approach to Lung Image Segmentation using Fuzzy Possibilistic C-Means Algorithm. International Journal of Computer Science and Information Security 7 (2010)
10. Vapnik, V., Chapelle, O.: Bounds on error expectation for support vector machines. Neural Computation 12 (2000)
11. Scholkopf, B., Smola, A.: Learning with Kernels. MIT Press (2001)
12. Ben Hassen, D., Taleb, H.: A fuzzy approach to chest radiography segmentation involving spatial relations. IJCA Special Issue on "Novel Aspects of Digital Imaging Applications", 40–47 (2011)
13. Ben Hassen, D., Taleb, H.: Automatic detection of lesions in lung regions that are segmented using spatial relations. Clinical Imaging (2012) (in press)
14. Ginneken, B.V., Stegmann, M.B., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Medical Image Analysis 10, 19–40 (2006)