

A New Approach on Density-Based Algorithm for Clustering Dense Areas

Paola Perchinunno(✉) and Samuela L'Abbate

Department of Economics, Management and Business Law, University of Bari "Aldo Moro", Bari, Italy
{paola.perchinunno,samuela.labbate}@uniba.it

Abstract. This paper presents a new approach to density-based clustering for the identification of dense areas. In particular, the focus is on identification of breast masses in the X-ray imaging of a mammography. The idea was to apply cluster analysis by identifying breast masses as clusters, understood as dense regions of space separated by areas of lower density. Attention was focused on a particular method of clustering based on density, the DBSCAN, proposing a new approach by applying it to a real dataset: a supervised approach, based on ROC curves and a weighted distance, for the choice of input parameters.

Keywords: Cluster analysis · DBSCAN · ROC curves

1 Introduction

The rapid developments in the availability and access to spatially referenced information in a variety of areas, has induced the need for better analysis techniques to understand the various phenomena. Spatial clustering algorithms, which groups similar spatial objects into classes, can be used for the identification of areas sharing common characteristics. DBSCAN is the first density-based algorithm. In fact, these algorithms identify clusters as dense regions of space, separated by areas of lower density (noise). These algorithms therefore possess the ability to detect clusters of arbitrary shape as concave zones. The aim of this paper is to presents a new approach to density-based clustering for the identification of dense areas. Special attention was paid to the identification of breast masses in the X-ray imaging of a mammography.

2 DBSCAN Method

2.1 Description of Algorithm

DBSCAN (Density Based Spatial Clustering of Application with Noise) [1] is the first density-based algorithm. The key idea of this algorithm lies in the concept of proximity

The contribution is the result of joint reflections by the authors, with the following contributions attributed to S. L'Abbate (chapter 2, 3, 4) and to P. Perchinunno (chapter 1, 5).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13377, pp. 530–542, 2022.
https://doi.org/10.1007/978-3-031-10536-4_35

between patterns. A pattern of a cluster must have other patterns in its vicinity, that is, in the vicinity of a pattern the density of the points must exceed a certain threshold. Conversely, the density calculated in the noise zones does not exceed a certain threshold. So each point, to belong to a cluster, must have, within a certain radius, at least a certain number of other patterns. This intuitive idea was formalized with the DBSCAN algorithm.

DBSCAN is one of the cluster analysis algorithms that allows to identify clusters of different shapes from the ellipsoidal one. In fact, the shape around the pattern depends on the type of distance function chosen for two patterns p and q denoted by dist(p; q). There is no predetermined distance function to use, but it can be chosen according to the application.

The algorithm needs two input parameters ε and *MinPts* with which to distinguish the points present in the dataset into core points, border points and noise points.

The formal description of the algorithm proposed in [2] and reformulated in [3–5], requires the following definitions:

Definition 1: (ε-neighborhood)
Fixed a point p of the dataset and a parameter ε, the ε -neighborhood, that is, the $\varepsilon-$ neighborhood of p with respect to ε, is the set of points of the dataset with distance from p less than ε:

$$N_{\varepsilon}(p) = \{x | dist(x, p) \leq \varepsilon\}$$

Indicated *MinPts* the minimum number of points around the point p it is possible to define a core point that differs from a border point and a noise point.

A point p is called a core point or inner point if there are at least *MinPts* points in its ε radius circle: $|N_{\varepsilon}(p)| \geq MinPts$.

A point p is called a border point if there are fewer than *MinPts* points in its ε radius circle: $|N_{\varepsilon}(p)| < MinPts$, that is, if it is not a core point, but is within a neighborhood of radius ε and q center for some core point q.

A point p is called a noise point if it is not a core neither a border point. See Fig. 1.

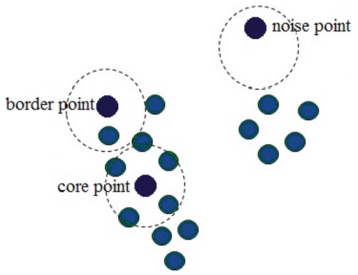


Fig. 1. Core point, border point and noise, with $\varepsilon = 1$ and $MinPts = 5$

Definition 2: (directly density-reachable)
A point p is *directly density-reachable* from point q with respect to ε and *MinPts* in a set of points D if.

1) $N_\varepsilon(q)$, where $N_\varepsilon(q)$ is the subset of D contained in the ε-neighborhood of q, i.e. $N_\varepsilon(q) = \{p \in D | dist(p, q) \leq \varepsilon\}$.
2) $|N_\varepsilon(q)| \geq MinPts$.

The second condition means that q is a *core point*. Only from core points, other points can be directly density-reachable. (See Fig. 2a).

Definition 3: (density-reachable)
A point p is *density-reachable* from a point q with respect to ε and *MinPts* in the set of points D if there is a chain of objects $p_1, \ldots, p_n$ with $p_1 = q$ and $p_n = p$, such that $p_i \in D$ and p_{i+1} is directly density-reachable from p_i with respect to ε and *MinPts.*

Density-reachability is the transitive hull of direct density reachability and this relation is not symmetric in general; only core points can be mutually density-reachable. (See Fig. 2 b).

Definition 4: (density-connected)
A point p is *density-connected* to point q with respect to ε and *MinPts* in the set of points D if there is a point $o \in D$ such that both p and q are density-reachable from o. (See Figure c).

A density-based *cluster* is now defined as a set of density-connected points which is maximal with respect to density-reachability and the *noise* is the set of points not contained in any cluster.

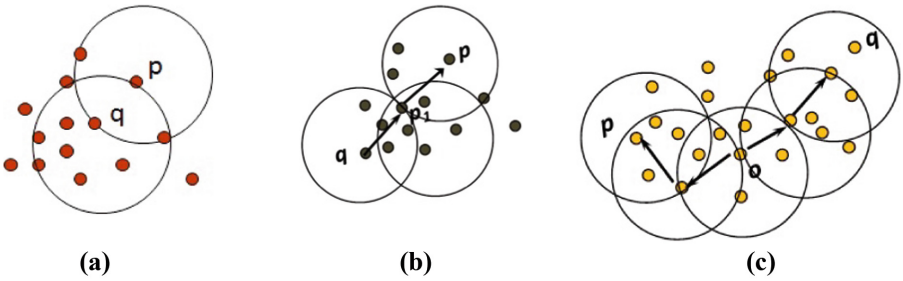


Fig. 2. a) Directly density-reachable, b) Density-reachable, c) Density connected

Definition 5: (cluster and noise)
Let D be a set of points. A *cluster* C with respect to ε and *MinPts* in $D$is a non-empty subset of D satisfying the following conditions:

1) Maximality: $\forall p, q \in D$: if $p \in C$ and q is density-reachable from p with respect to ε and *MinPts*, then also $q \in C$.
2) Connectivity: $\forall p, q \in C$: p is density-connected to q with respect to ε and *MinPts* in D then.

Every point not contained in any cluster is *noise.* Note that a cluster contains not only core points but also points that do not satisfy the core point condition: border points.

These points of the cluster are, however, directly density-reachable from at least one core point of the cluster (in contrast to noise points).

The algorithm DBSCAN [1], which discovers the clusters and the noise in a database according to the above definitions, is based on the fact that a cluster is equivalent to the set of all points in D which are density-reachable from an arbitrary core point in the cluster. The retrieval of density-reachable points is performed by iteratively collecting directly density-reachable points. DBSCAN checks the ε -neighborhood of each point in the database. If the ε -neighborhood $N_\varepsilon(p)$ of a point p has more than *MinPts* points, a new cluster C containing the objects in $N_\varepsilon(p)$ is created. Then, the ε -neighborhood of all points q in C which have not yet been processed is checked. If $N_\varepsilon(q)$ contains more than *MinPts* points, the neighbors of q which are not already contained in C are added to the cluster and their ε-neighborhood is checked in the next step. This procedure is repeated until no new point can be added to the current cluster C.

2.2 DBSCAN: Advantages and Disadvantages

The DBSCAN algorithm has several advantages: it detects clusters of arbitrary size and shape; it does not require a priori the number of clusters to be searched for and the initial patterns with which to start the algorithm; automatically determines noise points; is applicable to any metric space.

The greatest advantages of DBSCAN are that it can follow the shape of the clusters and that it requires only one distance function and two input parameters. Their choice is crucial because they determine whether a group is a cluster of points or a simple noise.

In [2] the authors provide a heuristic that they consider sufficiently valid for the determination of ε and to choose as a threshold (ε) the point at the first depression that can be noticed in the trend of the graph; and as a graph they recommend choosing the sorted k-dist graph for $k = 4$ as experimental results have shown a very similar trend for values above 4, instead of a considerable increase in time required for the construction of the graph itself. However, this determination of ε has ambiguities within it since the graph does not have a net depression but often oscillations are evident so the choice of the ε is reduced to a choice of attempts of the most appropriate.

2.3 Supervised Approach to Choosing MinPts

ROC curves were obtained by plotting the fraction of true positives (sensitivity) along the y axis and the fraction of false positives (1 – specificity) along the x axis. The area A_z under the ROC curve was obtained to serve as a measure of the diagnostic accuracy of the feature.

R.O.C.- Receiver Operating Characteristic curves are used to evaluate the validity of a diagnostic test (see [6–9]). A ROC curve is drawn by placing on the ordinate axis the values of sensitivity (the fraction of true-positive):

$$Se = \frac{TP}{TP + FN}$$

where TP means true positive, FN means false negatives; and on that of abscissa the values of 1-specificity (the fraction of false-positives):

$$Sp = \frac{TN}{FP + TN}$$

where TN means true negative.

The A_z area under the ROC curve is a measure of accuracy (performance) of a diagnostic test, [10, 11]. The value of A_z is between 0 and 1; the greater the better the method is, that is, the more correct the classification. In this work the DBSCAN algorithm was modified for the choice of the optimal value to be assigned to *MinPts* with a supervised approach that uses ROC curves to measure the effectiveness of the classification obtained with respect to a reference ideal. Calculating the sensitivity and specificity, for different values of *MinPts*, we have chosen as the optimal value to be assigned to *MinPts* the one corresponding to the curve with the highest area under the curve (AUC-area under curve), see Fig. 3.

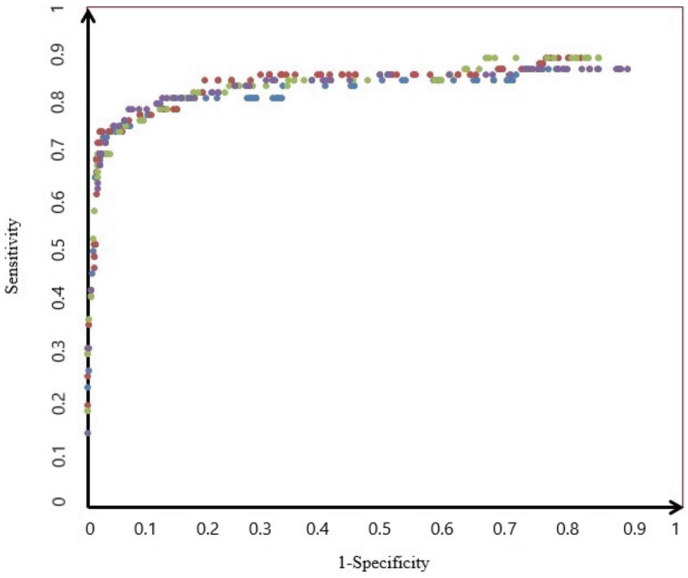


Fig. 3. ROC curves for different MinPts values

2.4 Choice of Epsilon

Being able to attribute a certain value to ε is functional to limit the arbitrariness of its choice and thus reduce the time of formation of clusters. For each dataset, the distribution of ε, is considered, and for a fixed value of *MinPts* a histogram is constructed. The first relative minimum of the histogram is then chosen as the value of ε optimal, as shown in Fig. 4.

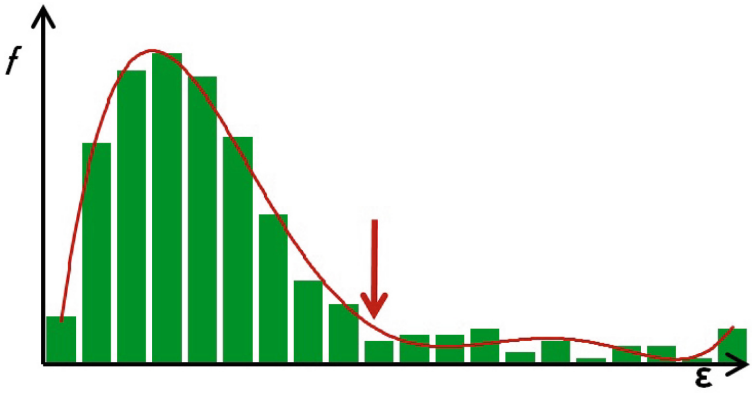


Fig. 4. Graphical representation of the distribution of ε

2.5 Choosing the Distance Function

In general, the function chosen as distance for the DBSCAN method depends on the application studied and must satisfy the typical properties of a distance.

In the epidemiological or environmental field, the analysis of clusters using a density-based approach aims to identify areas where a given phenomenon assumes values significantly higher than those recorded in other parts of the territory.

In the case study examined, starting from an X-ray image, a map was created by matching each pixel of the X-ray image with a color value based on a grayscale. Therefore it is necessary that the variable that describes the phenomenon is decisive in discriminating the neighboring points in the sense as well as topographical also under the analyzed aspect.

The presence of the phenomenon is understood as a weight that makes two points closer with high intensity of the phenomenon and two points distant spatially close but with low intensity of the phenomenon analyzed. For example, it is possible that two pairs of points: $(A; B)$ and $(C; D)$ are such that $d(A; B) = d(C; D)$, they are pairs of points close spatially, indicated by F the weight that evaluates the measure of the phenomenon for each point, we have that:

$$d(A; B) \cdot F(A; B) > d(C; D) \cdot F(C; D)$$

that is, the spatial distance is altered by this weight. The weighted distance can be extended to the case of n variables in which the weights can also be represented by several variables.

If the Euclidean distance were used, the DBSCAN would identify the clusters only on the basis of the geographical proximity of the points, see Fig. 5, without considering the intensity index represented by the third variable.

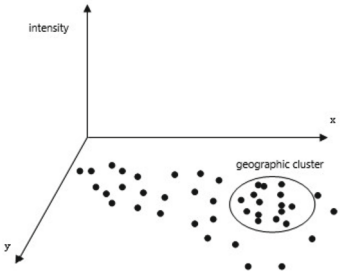


Fig. 5. Cluster identified with respect to Euclidean distance

The objective of our study is to identify dense areas in terms of intensity so instead of the Euclidean distance a function was chosen that deforms the geometric space in such a way that the geographically close points and that have similar intensities are even closer while geographically close points but with different intensities become more distant. The function that binds in the above terms two points:$A(x_A; y_A; w_A)$ and $B(x_B; y_B; w_B)$ is a weighted distance obtained by relating the Euclidean distance to the average power with negative exponent between the points under examination:

$$d(A, B) = \frac{\sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}}{\sqrt[-t]{\frac{w_A^{-t} + w_B^{-t}}{2}}}$$

This function satisfies the first three axioms of distance but does not verify the triangular inequality. In general, it is precisely the fourth and last property of distance that is discriminating and the fact that it is satisfied or not distinguishes metric measurements from so-called semimetric ones. Therefore the weighted distance that we propose is semimetric, however this restriction does not affect the Definitions 3 and 4 of Sect. 2.1 respectively of reachability and connection for which the properties of reflexivity and symmetry are necessary which remain satisfied. With this function at pairs of points with low intensity value the distance increases so that these are penalized in the formation of clusters, see Fig. 6. Empirically it has been verified that the most appropriate value of t is 5.

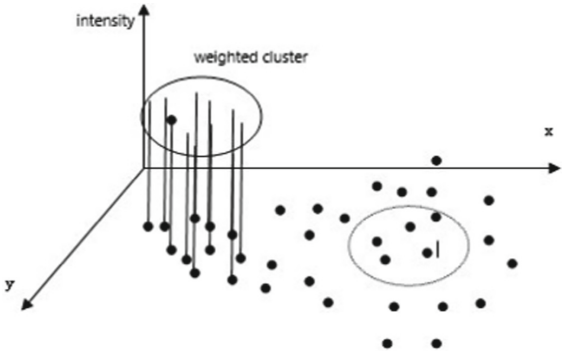
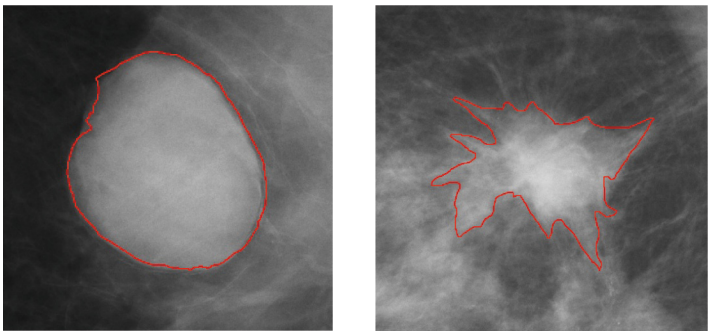


Fig. 6. Cluster detected with respect to weighted distance

3 Application

Database concerns digital X-ray images of mammograms related to benign and malignant lesions. Of each breast mass we know the diagnostic classification by biopsy and the outline drawn by hand by a radiologist specialized in mammograms, Fig. 7.



(a) (b)

Fig. 7. a) benign lesion, b) malignant lesion

Several maps were generated from the scanned X-ray images by matching each pixel of the X-ray image with a color value established based on a grayscale as shown in Fig. 8.

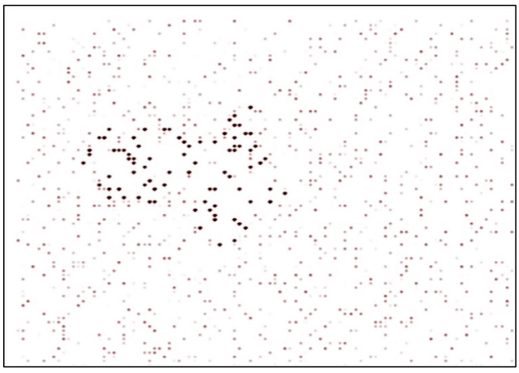


Fig. 8. Example of a map generated from an X-ray image

It was decided to apply, for the generated maps, the DBSCAN described in Sect. 2.1 because it allows to identify clusters as dense regions of space, separated by areas with lower density (noise) and can detect clusters of arbitrary shape as concave zones (the spicules in breast masses). In the application under consideration, therefore, dense areas are those areas whose points in addition to being spatially close have a high intensity of color, therefore a presence of the phenomenon. First of all, a distance function was appropriately chosen as seen in Sect. 2.5, in fact if the DBSCAN were applied considering the Euclidean distance function as a function, the algorithm would detect a single cluster corresponding to the rectangle containing all the points.

3.1 Input Parameters for Method Application

Regarding the input parameters ε and *MinPts* that the algorithm needs, the optimal values have been obtained as illustrated in Sects. 2.3 and 2.4. For the choice of ε, proceeding by trial and error would produce a high waste of time, see Fig. 3.3; as described in Sect. 2.4 to limit the arbitrariness of your choice and thus reduce training time of the clusters, for a fixed value of *MinPts* was considered the distribution of ε and was chosen as the value of optimal of ε the first relative minimum of the histogram.

Regarding the choice of the optimal value to be attributed to *MinPts*, we proceeded as described in Sect. 2.4 with a supervised approach. The sensitivities and specificities were calculated, for different *MinPts* values, comparing the contour drawn by the radiologist with the elaborated curve, meaning with true positives: the points of the cluster correctly classified, with true negatives: the correctly classified noise points, with false positives: the noise points classified as points of the cluster and finally with false negatives: cluster points classified as noise, see Fig. 9 (Fig. 10).

The optimal value of MinPts was chosen as the one corresponding to the curve with the highest area under the curve (AUC), see Fig. 11 and it was found that the optimal value to be attributed to MinPts is equal to 5.

By applying the DBSCAN algorithm, modified considering the optimal input parameters, the detection of the cluster is immediately obtained as in Fig. 3.6 whose contour coincides with that drawn by the radiologist also returning greater precision in the identification of the spicules, see Fig. 12.

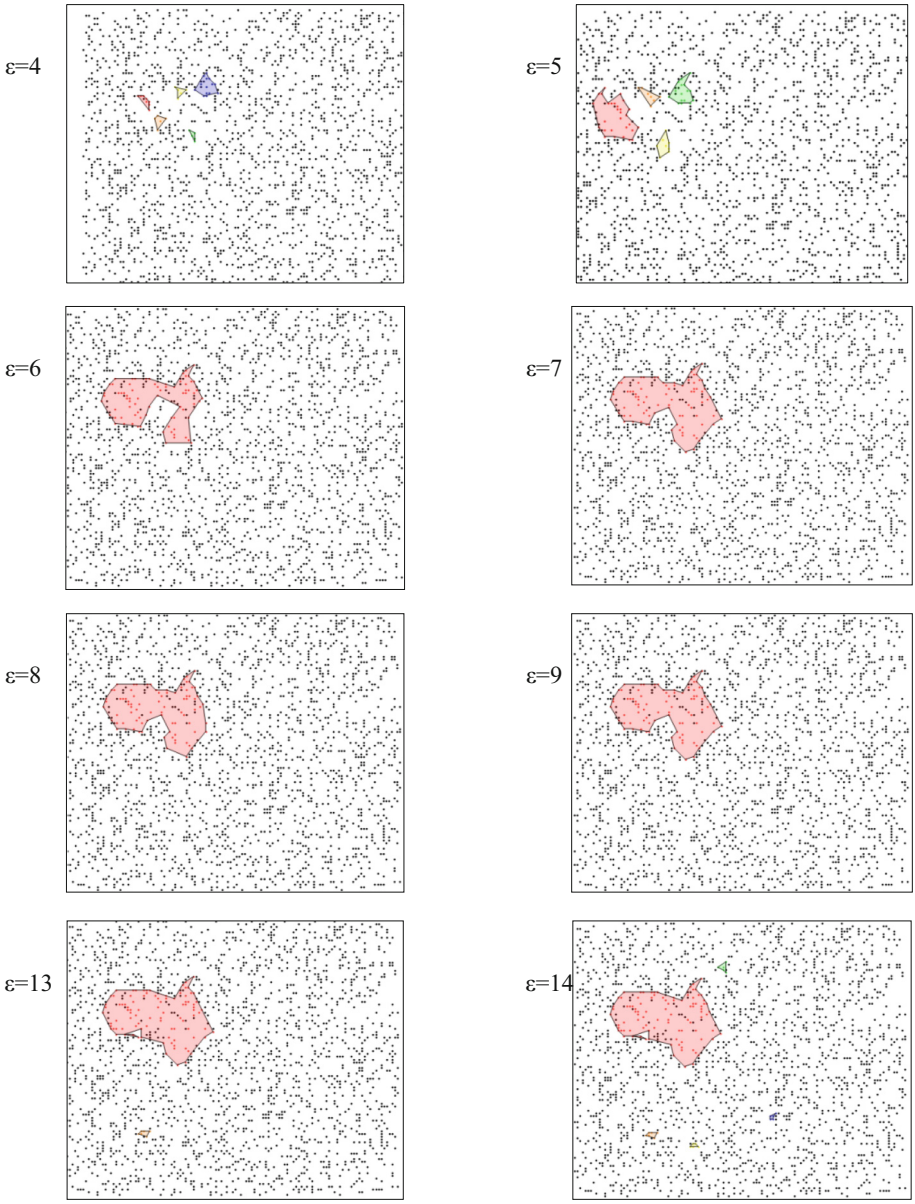


Fig. 9. Applying the DBSCAN algorithm for various ε values

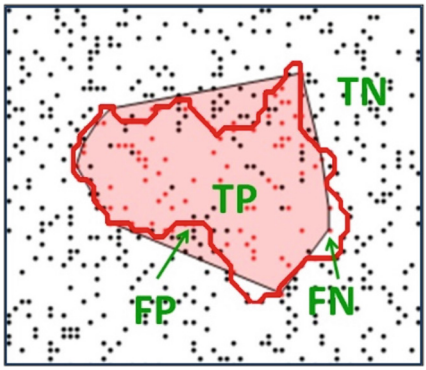


Fig. 10. Comparison between the radiologist's contour and the processed curve: TP: True Positive, TN: True Negative, FP: False Positive and FN: False Negative

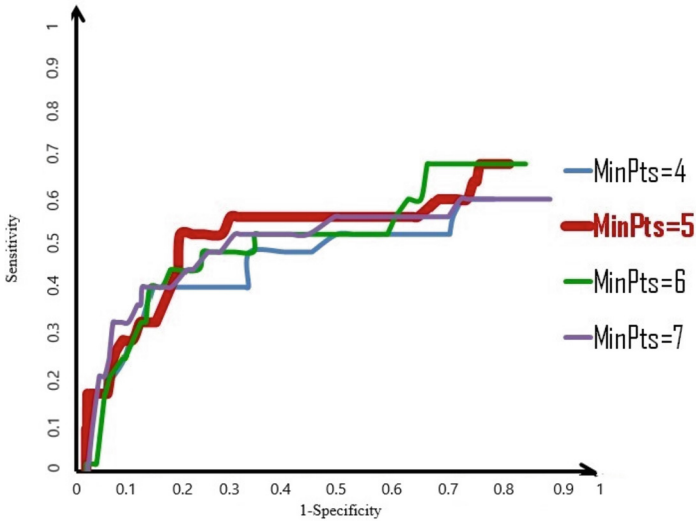


Fig. 11. ROC curves for different MinPts values

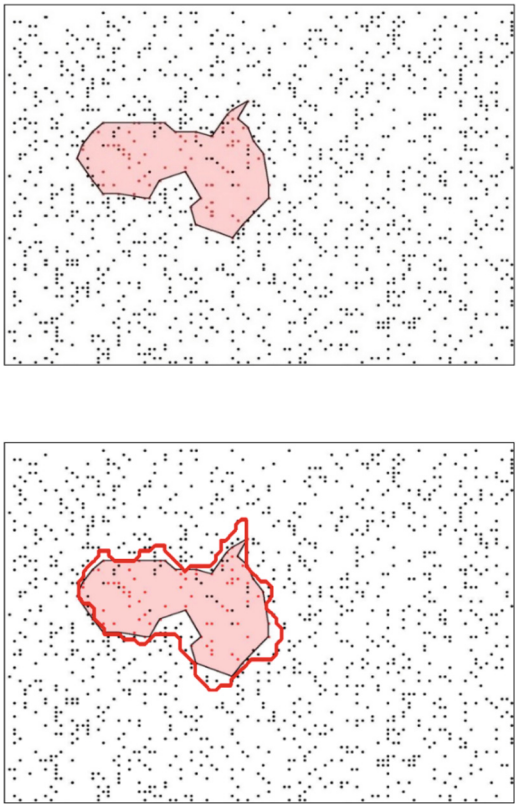


Fig. 12. Application of the DBSCAN algorithm with $\varepsilon = 8$ and MinPts $= 5$ and overlapping the contour extracted by the radiologist to the obtained cluster applying the DBSCAN algorithm with $\varepsilon = 8$ and MinPts $= 5$

4 Conclusions and Future Advancements

The basic idea that guided this work was to be able to apply a density-based clustering algorithm, DBSCAN, for the detection of dense areas such as breast masses.

The intuition that led to propose an optimization of the algorithm was to see the breast masses as 'spots'; therefore, clusters understood as dense regions of space, separated by areas with lower density. Hence the need to introduce an appropriate distance function that would identify the areas where the phenomenon examined assumed significantly higher values than in other areas.

The goal was to speed up the clustering algorithm, developing new methodologies to overcome the limits of typical clustering algorithms based on density, i.e. the arbitrariness of input parameters.

The ability of DBSCAN to detect clusters of arbitrary shape and in particular the concave areas has motivated the research to recognize also the spiculated breast masses. The results obtained were very satisfactory, the algorithm also identified clusters with particular concave or spicule areas. The comparison with the contour drawn by the

radiologist showed the high reliability in the recognition of the cluster that represented the breast masses. Automating the density-based clustering algorithm can support the radiologist in extracting the contour of breast masses.

In addition, being able to classify according to the shape of the contour means having information on the very nature of the pathology, because the morphology of the lesion is linked to its degree of malignancy. The shape is a fundamental characteristic to distinguish benign and malignant masses, so a correct clustering leads to the recognition of the actual form and therefore to a more likely exact distinction between malignant and benign mass. One of the future developments is in fact to define indices related to the shape of the cluster that based on the degree of irregularity can establish the pathology of the breast mass of which the contour is extracted, in order to provide a valid support to the verdict of the radiologist.

A next step will be to delve the cluster validity indices for data clustering with respect to internal and external criteria proposing a new inferential approach.

The future perspectives are many and are both methodological and applicative and demonstrate how statistics can also represent a valid investigative tool to quantify and rationalize notions and hypotheses formulated on the basis of experimental observations.

References

1. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland (1996)
2. Ester, M., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd ACM SIGKDD, Portland, OR, USA, pp. 226–231 (1996)
3. Montrone, S., et al.: A modified density-based algorithm for clustering. In: Supplemento ai Rendiconti del Circolo Matematico di Palermo. II 84, Palermo, Italy, pp. 231–239 (2012)
4. Montrone, S., et al.: Comparing SaTScan and DBSCAN methods inspatial phenomena. In: SPATIAL2, Spatial Data Methods for Environmental and Ecological Processes, 2nd edn. Baia delle Zagare, Foggia, Italy, pp. 116–119 (2011)
5. Montrone, S., et al.: Statistical methods for detecting geographical clustering of housing poverty. In: Rivista Italiana di Economia, Demografia e Statistica, vol. LXV. 3/4. SIEDS, Roma, Italy, pp. 157–164 (2011)
6. Metz, C.E.: Basic principles of ROC analysis. Seminars Nucl. Med. **VIII**(4), 283–298 (1978)
7. Metz, C.E.: ROC Methodology in radiologic imaging. Invest. Radiol. **21**, 720–733 (1986)
8. Pepe, M.S.: The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, New York (2003)
9. Zou, K.H.: Receiver operating characteristic (ROC) literature research (2002). http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html
10. Hanley, J.A., McNeil, B.I.: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology **148**(3), 839–843 (1983)
11. Hanley, J.A., McNeil, B.I.: Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med. Decis. Making **4**, 137–150 (1984)
12. Montrone, S., Massari, A., Perchinunno, P., Ligorio, C., L'Abbate, S.: Statistical Methods for Detecting Geographical Clustering of Housing poverty. In: Rivista Italiana di Economia, Demografia e Statistica, vol. LXV, no. 3/4 (2011). ISSN: 0035-6832
13. Efron, B.: Bootstrap methods: another look at the jackknife. In: The Annals of Statistics, vol. 7, pp. 1–26. Institute of Mathematical Statistics (1979)