Cluster Analysis for Breast Cancer Patterns Identification*

Beatriz Flamia Azevedo[1,2][0000−0002−8527−7409], Filipe Alves[1,2][0000−0002−8387−391X], Ana Maria A. C. Rocha[2][0000−0001−8679−2886], and Ana I. Pereira[1,2][0000−0003−3803−2043]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
{beatrizflamia,filipealves,apereira}@ipb.pt
[2] ALGORITMI Center, University of Minho, 4710-057 Braga, Portugal
arocha@dps.uminho.pt

Abstract. Safety in patient decision-making is one of the major health care challenges. Computational support in establishing diagnoses and preventing errors will contribute to an enhancement in doctor-patient communication. This work performs a three-dimensional cluster analysis, using k-means algorithm, to identify patterns in a breast cancer database. The methodology proposed can be useful to identify patterns in the database that are normally difficult to be noted by classical methods, such as statistical methods. The three-dimensional cluster approach was explored combining three variables at once. The k-means algorithm is used to recognize the hidden patterns on the database. Sub-clusters are used to separate the benign and malignant tumors inside the global cluster. The results present effective analyses of three different clusters based on different combinations between variables. Thus, health professionals can obtain a better understanding of the properties of different types of tumor, identifying the mined abstract tumor features, through the cluster data analysis.

Keywords: cluster analysis · disease diagnosis · breast cancer

1 Introduction

Breast cancer, among the various type of cancer, is one of the most common and deadly around the world [5, 8, 18]. Like any other disease early diagnosis and treatment initiation are determining factors to control the development of the pathology.

The use of digital technologies, mainly data mining algorithms, has been widely used in medicine, providing remarkable advances in the early diagnosis of several diseases and analysis of patterns related to treatments, symptoms,

* This work has been supported by FCT – Fundação para a Ciência e a Tecnologia within the R&D Units Projects Scope: UIDB/00319/2020 and UIDB/05757/2020. Filipe Alves is supported by FCT Grant Reference SFRH/BD/143745/2019.

patient particularities and of course, characteristic of the disease itself. According to [13], machine learning tools are swiftly infiltrating many medicine areas, with significant potential to transform the medical landscape. Some practical applications of machine learning or data mining algorithms in the medical area can be found in [1, 2, 5, 17, 18]. According to [5], the detection of the pattern of symptoms using data mining is an important technique for the correct understanding of hidden patterns. In the case of breast cancer, the use of machine learning techniques is essential to reduce the disease diagnosis time and improve the accuracy of diagnosis for the patient [15].

In this work the breast cancer Wisconsin database [4] is explored in order to identify patterns in the cancer diagnosis, using the k-means cluster algorithm. First, a pre-processing procedure is applied to eliminate the outliers and provide a clear data classification. Thereafter, the correlations between the mean values of the ten features, whose database is composed, are evaluated to define the variables selection and features extraction for carrying out the cluster analysis.

Most of the work that applies the cluster analysis, create a two-dimensional cluster representation/visualization. However, in this work the cluster analysis relied on a three-dimensional representation/visualization, for a more intuitive analysis of the results. Besides, to separate the type of tumor inside the cluster, sub-clusters are considered to identify the benign or malignant tumor inside the cluster provided by the k-means algorithm.

This paper is organized as follow: after the introduction, Sect. 2 presents the methodology applied in this work, it is based on the cluster classification and the k-means algorithm. Thereafter, Sect. 3 describes the database used in this work. The data pre-processing, as well as the results, are presented in Sect. 4. The conclusion and future perspectives are shown on Sect. 5.

2 Classification Methods

The classification of individuals can be supervised or unsupervised. In the supervised classification, the class that generates each pattern in the modelling sample, is known *a priori*, and the classifier is trained to replicate the knowledge acquired to classify unknown samples. On the other hand, in unsupervised classification, the classes are not known, so the algorithm must find a structure in the data that allows it to divide the data into groups [3].

Clustering technique is an unsupervised method, which is appropriate for exploring relationships between data and the underlying structures. Cluster analysis techniques can be broadly categorized as hierarchical and non-hierarchical, in which the latter including partitional methods [10].

The hierarchical clusters methods build the clusters in such that given two clusters, these are either disjoint or one of them is contained in the other, thus obtaining a hierarchy of clusters. Basically, these methods try to grouping into classes proceeds in stages, generally identifying from n subgroups (of a single individual each) successive mergers of subgroups considered to be more "similar". Each merger reduces the number of subgroups. The results of the hierarchical

methods are usually presented in dendrograms, which contain the relationships between the clusters and the order in which the clusters were put together (agglomerative methods) or divided (divisive methods).

In turn, the partitional methods aim to obtain partitions on the set to be classified. In general, the partition method applied to a set of n elements, starts from a subset of k points, considered the centers of the aggregation classes or poles. Through transfers of individuals from one class to another, an attempt is made to determine the best classification, in order to make the classes more internally homogeneous and externally heterogeneous. The method is iterated, recalculating the centers at each stage, and the elements of the set to be classified are reallocated according to their dissimilarity to the centers. Usually the stopping criterion is defined as no modifications after two successive iterations. In practice, the algorithm is typically run multiple times with different initial states, and the best configuration obtained from all stages is used as the final cluster. These types of methods are usually based on a central point (average of the attributes of the "k-means" objects).

The k-means method is the simplest, most popular and generally the most widely used partitioning algorithm, which employs the square error criterion [16]. The k-means method assigns the elements to the class with the nearest centroid. At the beginning of this algorithm, a set of k points is selected, representatives of the classes or centroids [10]. In this sense, the k-means clustering algorithm existing in MatLab® library, was trained to apply and analyse the data present in the following section.

3 Breast Cancer Wisconsin Diagnostic Database

The database used in this work, provided by Wisconsin University [4], contains 10 features for each cell nucleus for the breast cancer diagnosis.

This database provides information about 569 patients (357 patients who were diagnosed with a benign nodule and 212 who were diagnosed with a malignant nodule). The features are computed from a digitized image of a fine needle aspirate of a breast mass. The description of each feature, which describe the characteristics of the cell nucleus found for each patient, follows:

- **feature 1:** radius (distance, in mm, from the center to points on the perimeter)
- **feature 2:** texture (standard deviation of gray-scale values)
- **feature 3:** perimeter, in mm
- **feature 4:** area, in mm^2
- **feature 5:** smoothness (local variation in radius lengths)
- **feature 6:** compactness, given by $\frac{(Perimeter)^2}{Area-1}$
- **feature 7:** concavity (severity of concave portions of the contour)
- **feature 8:** concave points (number of concave portions of the contour)
- **feature 9:** symmetry
- **feature 10:** fractal dimension, given by $CoastlineApproximation - 1$.

In the database, for each feature, three indicators are given: mean value, standard error, and "worst" or maximum value, resulting in 30 attributes. Those different measurements are treated as different features in the data set. Since these values have different scales (standard error values), in this study the mean values of each feature will be used.

In the identification of breast cancer nodules, two diagnoses can be obtained, which are represented by 0 and 1, benign and malignant nodules, respectively. Therefore, the general idea is to apply the k-means algorithm on breast cancer Wisconsin diagnostic database. The objective is to analyse the clusters and separate the data into benign and malignant tumors, to infer the impacts of the features on the cancer diagnosis.

4 Clustering Application

4.1 Database Pre-processing

The data pre-processing used in this paper is divided into two stages: first, an outlier analysis is done aiming to obtain a clear database to perform the cluster analysis. Then, the correlation between the features is analysed in order to reduce the features dimension and select the optimal cluster combination.

The outliers analysis was done through the feature box plot evaluation. Outliers are unusual or noise values in the dataset that can distort statistical analysis [9]. Although, there are some techniques to deal with outliers, here it is intend to generalize the patterns (similarities and dissimilarities), so it is important to eliminate all divergence values to provide a more accurate analysis. Thus, all the outliers were removed from the malign group and also from the benign group of patients. Thereby, the database used to perform the cluster analysis considers 248 data from patients with benign tumor and 172 from patients with a malignant tumor.

Before starting the cluster analysis the Pearson correlation between the features is analysed. The correlation coefficients are used to assess the strength and direction of the linear relationships between pairs of features [7]. To analyse the correlation between the features, the Pearson coefficient will be used to identify the less correlated features, as presented in Fig. 1.

It is important to mention that there are other measures, sometimes more efficient, to analyse the dependence, causality and other statistical relationship between the features that can infer similarities and dissimilarities between features. As it is intended here to perform a clusters analysis, which is a machine learning technique, the features that have the least correlation with each other were chosen, since, in general, they are the most difficult to find patterns by statistical techniques. Then, only the features with correlation smaller than 0.80 will be considered, due to fact that superior values will imply features with very strong correlation, as suggested in [6].

From Fig. 1, it possible to note that the features f1, f3, f4, f6 and f7 have very strong correlation with the feature f8. Therefore, the feature f8 is kept and the

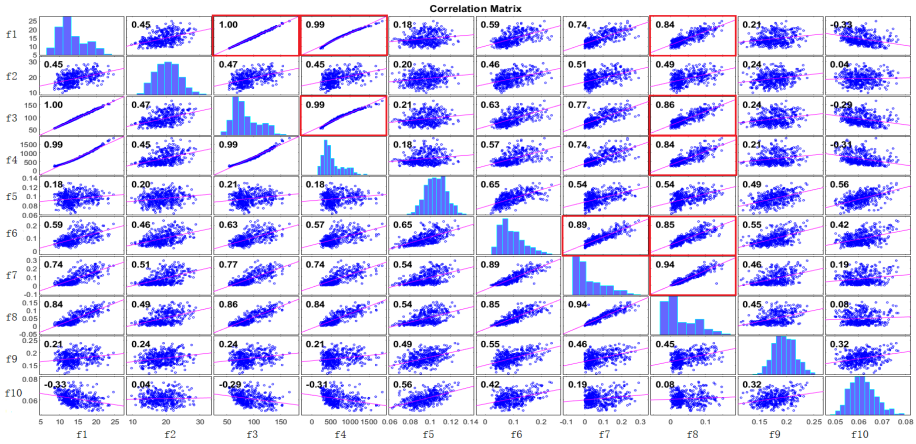

Fig. 1. Pearson Correlation Matrix.

remaining ones are discarded from the cluster analysis. Thus, 5 features remain to be analysed (f2, f5, f8, f9, f10), resulting in 10 possible combinations (5 features, 3 to 3). All the 10 combinations between the features were analysed, however only the three more interesting combinations stand out to be presented in this work, considering the identified patterns, the distances between the centroids and the distribution of the sub-clusters.

4.2 Cluster Results

To identify the optimum number of clusters in the data set is a fundamental issue in clustering partitioning. As already stated before, the k-means algorithm requires the specification *a priori* of the number of k clusters to be generated [11]. In this work, the Silhouette method [14], a similarity measurement, is adopted to estimate the k value.

The Silhouette method, used for the interpretation and validation of consistency within clusters, displays a measure of how close are the data points in the database, the distance within the cluster and the distance between clusters [12]. For the database under study, the Silhouette index value assigned $k = 2$, as the optimal number of cluster partition for all feature combinations considered. The following results present the 3 most relevant combinations between the features, 3 cases. As an example, only the combination of features for case 1 will be represented by a three-dimensional figure.

Case 1: This clustering is provided by the features texture, concave points and fractal dimension. The clusters generated can be visualized in the three-dimensional space in Fig. 2. On the left side of the figure, the 2 clusters obtained by the k-means algorithm are represented (in black and light blue), whereas on the right side, for each cluster, the malignant and benign points are marked. The

centroids of each cluster are marked by the symbol x. It is possible to note that the texture feature allows a clear division between clusters 1 and 2, in which all the tumors with texture smaller than 20 belong to cluster 1 and greater than 20 belong to cluster 2. The concave feature allows leveraging a separation between sub-clusters (benign and malignant tumors). It is clear that data from benign tumors are usually located below the value of 0.03, while the data with values above correspond to malignant tumors. The feature of fractal dimension shows that the dispersion in cluster 2 is higher than in cluster 1, so cluster 1 is more compact than cluster 2. In this way, the within-cluster sums of point-to centroid distances, in cluster 1 are smaller than cluster 2 distances, being equal to 867.39 and 955.96, respectively. The centroid 1 is located at $(x, y, z) = (22.097, 0.062, 0.061)$ and centroid 2 at $(x, y, z) = (16.179, 0.031, 0.061)$. In general, with these 3 features, the texture was decisive for the representation of the patterns and the separation of the 2 clusters. On the other hand, when the sub-clusters are analysed, the feature concave points is decisive in the division of the sub-clusters, since the benign tumors have the smallest concave points (between 0 and 0.03), and the data with malignant tumors have concave points greater than 0.03.

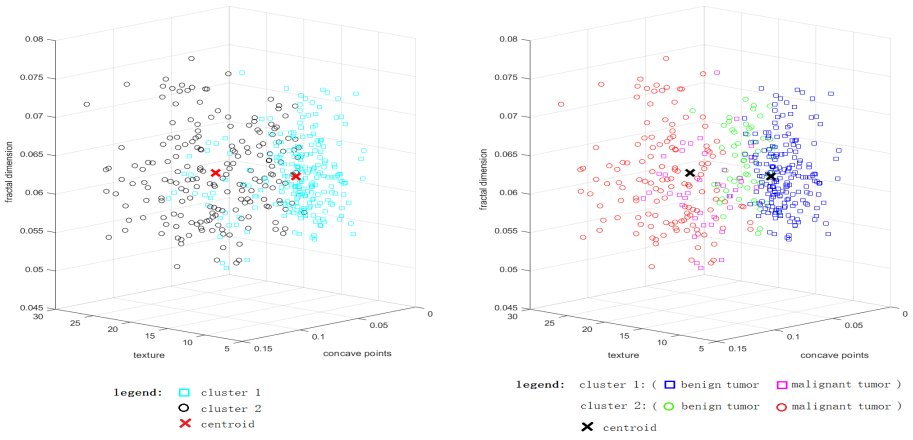

Fig. 2. Clustering with the features combination: texture, concave points and fractal dimension.

Case 2: This clustering is obtained combining the features smoothness, symmetry and fractal dimension. In this case, it is possible to verify a similar dispersion in both clusters, besides the point-to-centroids distance sum in both clusters is closer to 0.06. The centroid 1 is located at $(x, y, z) = (0.101, 0.196, 0.063)$ and centroid 2 at $(x, y, z) = (0.089, 0.160, 0.059)$. The symmetry feature provides a clear division between clusters 1 and 2, showing that below 0.18 there are more data in the cluster 1 and above this value refers to data of cluster 2. In this

case, an intriguing pattern stands out, benign tumors and malignant tumors are separated by a hyperplane that passes through the highest value of the fractal dimension and the lowest value of smoothness. By this way, it is possible to observe there is a greater agglomeration of benign tumors above the hyperplane, while below it is possible to observe the situations of malignant tumors.

Case 3: The third case presents the clustering obtained combining the features concave points, symmetry and fractal dimension. By the cluster analysis, it is possible to note that the cluster 1 is less compact than cluster 2, since the point-to-centroids distance sum is 0.16 for cluster 1 and 0.12 for cluster 2. The centroid 1 is located at $(x, y, z) = (0.169, 0.023, 0.060)$ and centroid 2 at $(x, y, z) = (0.192, 0.089, 0.061)$. There is a clear perception that the feature concave points define the cluster division. Cluster 1 has all data with concave point above the 0.05, whereas the cluster 2 has the most data below this value. The cluster analysis reveals that in cluster 1 there are no groups of sub-clusters, so, all data in this cluster, refer to malignant tumors, revealing this particularity in the combination of these 3 variables. In turn, in cluster 2, the presentation of sub-clusters between benign and malignant tumors is maintained, as in the previous cases. Analysing the cluster 2, it is possible to identify, when the feature concave points is smaller than the value of 0.04 the data belong to benign tumors, and the malignant tumors data are found on the concave points values between 0.04 and 0.06. Regarding the feature symmetry, the data reveal a higher concentration of cluster 1 above the symmetry value of 0.18, while in cluster 2 the data have a more homogeneous distribution, considering the feature symmetry. Besides, when the sub-cluster of cluster 2 are analysed, the malignant tumors are more concentrated on the values of 0.15 and 0.18 of symmetry.

5 Conclusions and Future Work

In a wide range of application domains, especially health, data analysis tasks heavily rely on clustering. The aim of this paper was to provide a renewed approach to a clustering application in the known Breast Cancer Wisconsin Diagnostic database, using the k-means algorithm. The results obtained were discussed considering three different feature combinations.

This work aimed at a more complete cluster analysis, integrating three features in a three-dimensional space. In this way, it was possible to demonstrate the relationship of the features in terms of clustering efficiency, similarities and dissimilarities, considering three at a time. Besides, to identify and also to provide new relevant knowledge about the clusters, an internal subdivision into sub-cluster was applied to facilitate the identification between benign and malignant tumors.

This approach, reveals excellent particularities of grouping and forecasting results for problems with unsupervised learning and with high interest, in this case, to support decision making in the diagnosis of malignant/benign tumors in breast cancer. The future research will focus on developing a more robust

clustering algorithm for evolving some dimension reduction methods to be used in high dimensional databases.

References

1. Bressan, G.M., Azevedo, B.C.F., Souza, R.M.: A fuzzy approach for diabetes mellitus type 2 classification. Brazilian Archives of Biology and Technology **63** (2020).
2. Bressan, G.M., Azevedo, B.F.A., Souza, R.M.: Métodos de classificação automática para predição do perfil clínico de pacientes portadores do diabetes mellitus. Revista Brasileira de Biometria **38**(2), 257–273 (2020).
3. Dougherty, G.: Pattern Recognition and Classification: An Introduction. Springer, 1 edn. (2013).
4. Dua, D., Graff, C.: Uci - machine learning repository (2017), http://archive.ics.uci.edu/ml, accessed on March, 2020
5. Dubey, A.K., Gupta, U. Jain, S.: Analysis of k-means clustering approach on the breast cancer wisconsin dataset. Int J Comput Assist Radiol Surg **11**(11), 2033–2047 (2016).
6. Evans, R.H.: An analysis of criterion variable reliability in conjoint analysis. Perceptual and Motor Skills **82**(3), 988–990 (1996).
7. Fisher, R.A.: Statistical methods for research workers. New York, Hafner (1958)
8. Gomes, I., Nunes, C.: Analysis of the breast cancer mortality rate in portugal over a decade: Spatiotemporal clustering analysis. Acta Médica Portuguesa **33**(5), 305–310 (2020).
9. Illowsky, B., Dean, S., Birmajer, D., Blount, B., Boyd, S., Helmreich, J., Keyon, L., Lee, S., Taub, J.: Introductory Statistics. https://openstax.org/details/books/introductory-statistics Last accessed 21 April, 2021. (2021)
10. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern recognition letters **31**(8), 651–666 (2010)
11. Kaufman, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis, vol. 344. John Wiley & Sons (2009)
12. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: 2010 IEEE international conference on data mining. pp. 911–916. IEEE (2010)
13. Materials, N.: Ascent of machine learning in medicine. Nature Materials **18**(407) (2019).
14. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
15. Rustam, Z., Hartini, S.: Classification of breast cancer using fast fuzzy clustering based on kernel. In: 9th Annual Basic Science International Conference 2019. Conf. Series: Materials Science and Engineering. vol. 546, pp. 1–9 (2019).
16. Steinley, D.: K-means clustering: a half-century synthesis. British Journal of Mathematical and Statistical Psychology **59**(1), 1–34 (2006)
17. Vidman, L., Källberg, D., Rydén, P.: Cluster analysis on high dimensional rna-seq data with applications to cancer research - an evaluation study. PLOS ONE **14**(12), 1–21 (2019).
18. Xie, J., Liu, R., Luttrell, J., Zhang, C.: Deep learning based analysis of histopathological images of breast cancer. Frontiers in Genetics **10**, 80 (2019).