TEI CRETE

SCHOOL OF APPLIED TECHNOLOGY

DEPARTMENT OF APPLIED INFORMATICS AND MULTIMEDIA

PATTTERN RECOGNITION

Professor: George Papadourakis, Ph.D.

Supervised versus unsupervised learning methods

An application that summarizes the results of learning methods

Author:

Jonasz Kulpinski

EP1418

1. Introduction

In classification problem, there exists two kinds of algorithms: supervised and unsupervised learning. The distinction is drawn from the way the algorithm classifies the data. In supervised learning, the classification is done by training the model on prelabelled data. With a trained mode, it is then able to make inference for unseen instance. On the other hand, unsupervised learning seeks out similarity between pieces of data in order to determinate whether they can be characterized as forming a group without training on the data first. It is able to discover hidden structure of the data.

Advantage of supervised learning is that because it is trained on the data, it is better than unsupervised learning in terms of accuracy. However, it requires training data and is prone to overfitting. Unsupervised learning is useful in practise, where labels do not necessarily come with the data. They are often hard and expensive to obtain. Learning directly from data is one of the strengths of unsupervised learning.

2. Description of the application

The application is designed to allow user to compare supervised and unsupervised learning methods and make it easier to decide which algorithm is better in a given task.

Very important for pattern recognition system essentially involves the following three steps [1][2]:

1) Data acquisition and pre-processing: Input data is prepared and passed to the pattern recognition system. Raw data is later pre-processed by removing the erroneous instances or the extraction pattern of interest from the background so that the input data is readable by the system.

2) Extracting the function: Then decide which functions of the data set will be used for classification. Important features together form a set that is to be recognized or classified.

3) Making decisions: the classification is performed here.

2.1. Choosing Data

The Haberman’s dataset was used in the application. The application was made in Octave. The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Dataset contains the following attributes [3]:

- age of patient at time of operation,
- patient's year of operation,
- number of positive axillary nodes detected,
- survival status (class attribute)
  - the patient survived 5 years or longer
  - the patient died within 5 year

2.2. Statistical analysis

Haberman's dataset was sorted by class. There were 225 instances of class 1 and 81 of class 2. Next step was to made boxplots every of the three features to see the differences between data for different classes and to choose the best features what was done in the next chapter. Box chart, box chart, box-plot chart - a form of graphical presentation of statistical distribution distribution, often found in computer packages supporting the process of analysis and interpretation of statistical data. It allows to include in a single drawing messages regarding the location, dispersion and shape of the empirical distribution of the examined feature.

A box chart (presented horizontally) is created by placing some of the distribution parameters on a horizontal axis. A rectangle (box) is placed above the axis, the left side of which is designated by the first quartile, and the right side by the third quartile. The width of the box corresponds to the value of the quarter interval. Inside the rectangle there is a vertical line, which determines the median value.

The drawing of the box is completed on the right and left side with sections called "mustaches". In the simplest variant, the left end of the left section determines the smallest value in the set, while the right end of the right segment is the largest. In another variant, the "mustache" has a length of one and a half of the quarter interval, and values lying outside this range are represented by points [4].

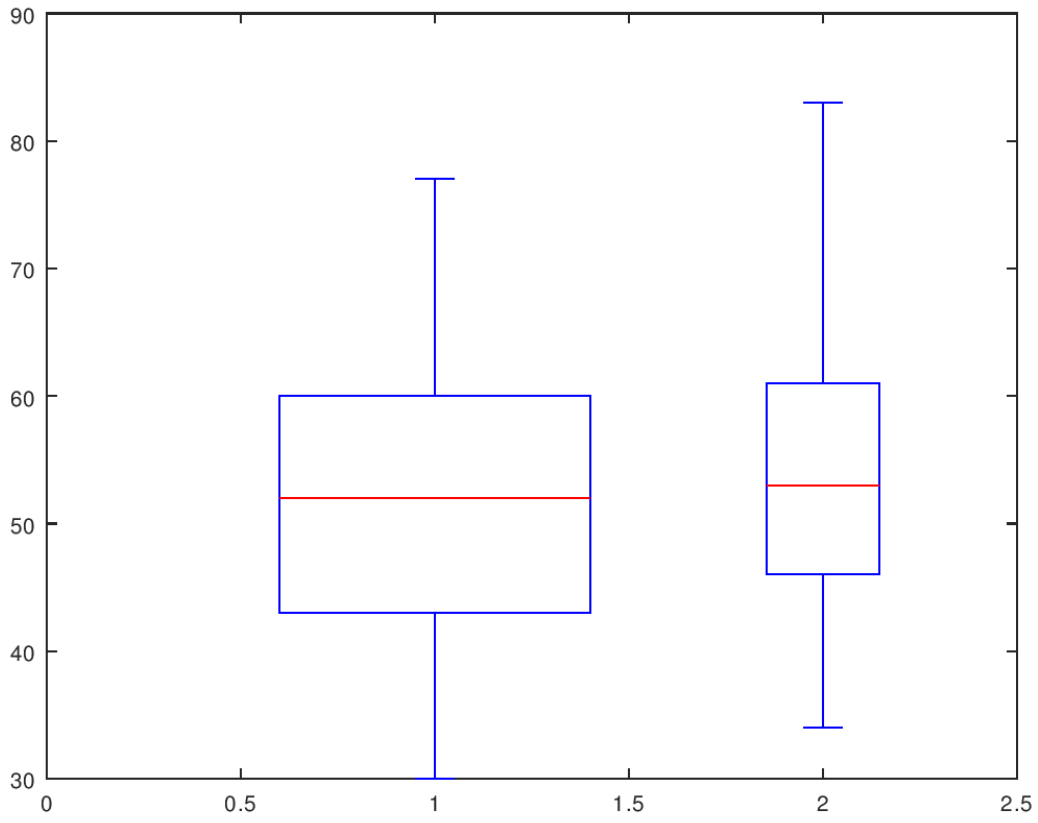

Figure 1. Age of patient at time of operation

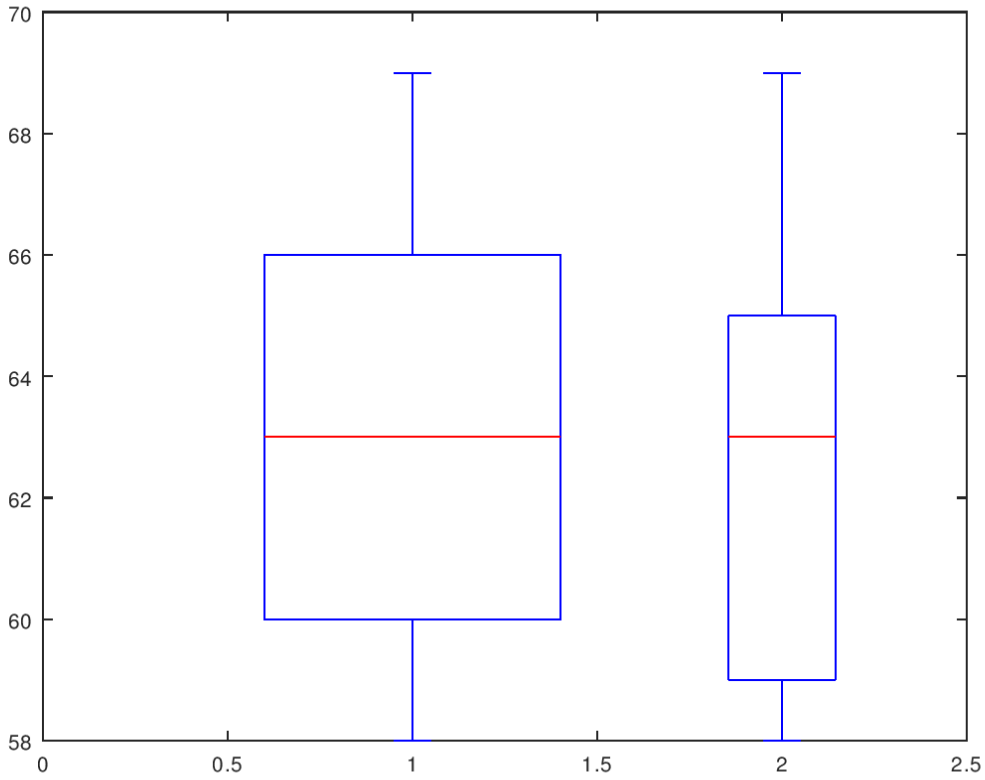

Figure 2. Patient's year of operation

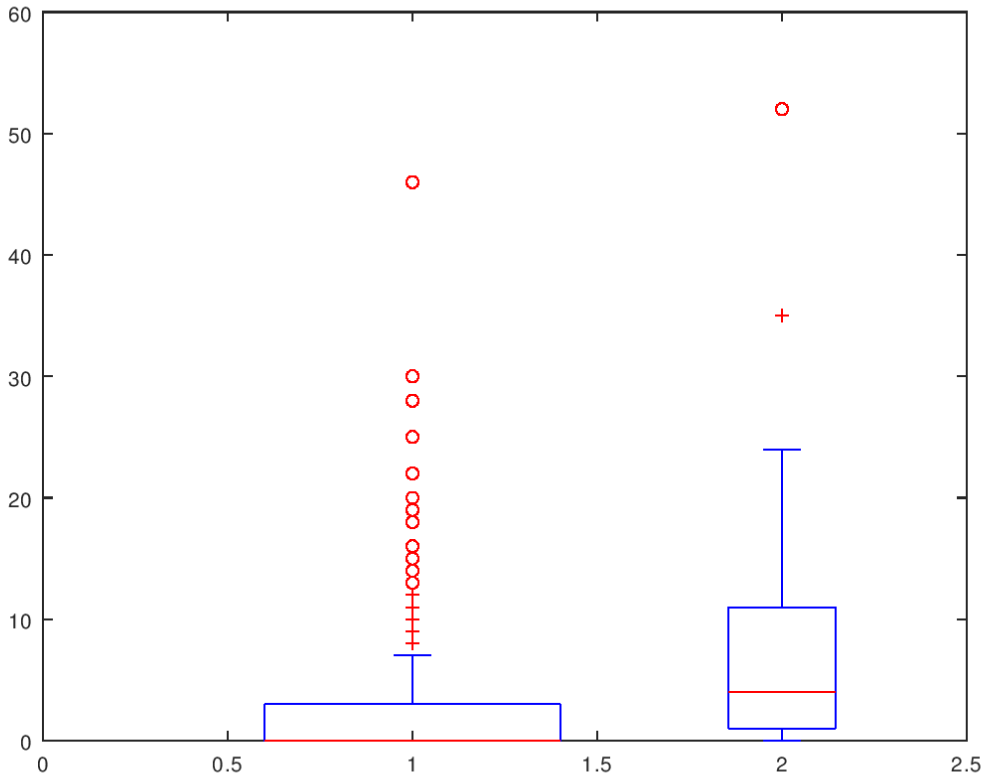

Figure 3. Number of positive axillary nodes detected

2.3. Preprocessing data

The sets of data we have collected need to be analyzed to identify the type of information they contain. This is one of the most important steps for the correct classification. We need to have enough data for each group. In cases of lack of data, the best way is to produce artificial data or to reuse data in different ways of separation.

Was made separated matrix with informations about instances class and deleted class column from dataset.

Later was found 5 outliers, removed them and transformed data into [0 1]. Outlier is an observation having an unusual value of an independent (explanatory) variable or unusual values of both variables - dependent (explained) and explanatory (explaining in the multiple regression analysis). This means that the relationship between Xi and Yi for a given observation is different than for the rest of the observations in the data set.

Outlier observations are generally caused by errors in the data, as a result of erroneous measurement, mistakes in entering information into the database, etc. A large number of outliers can also be a signal of choosing the wrong model.

Outlier observations resulting from errors in the data make it difficult and in extreme cases make analysis impossible. The methods and coefficients based on normal distribution and linear relationships, such as Pearson correlation, linear regression, classical correspondence analysis, etc., are particularly resistant. One outliers can completely change the value and the correlation sign, even from 0.9 to - 0.9 [5].

Then was splited dataset into training and validation set : 80-20%, 70-30%, 90-10%, the best accuracy was for 80-20% so this split percentage was used in later measurements. The biggest error was for 70-30%.

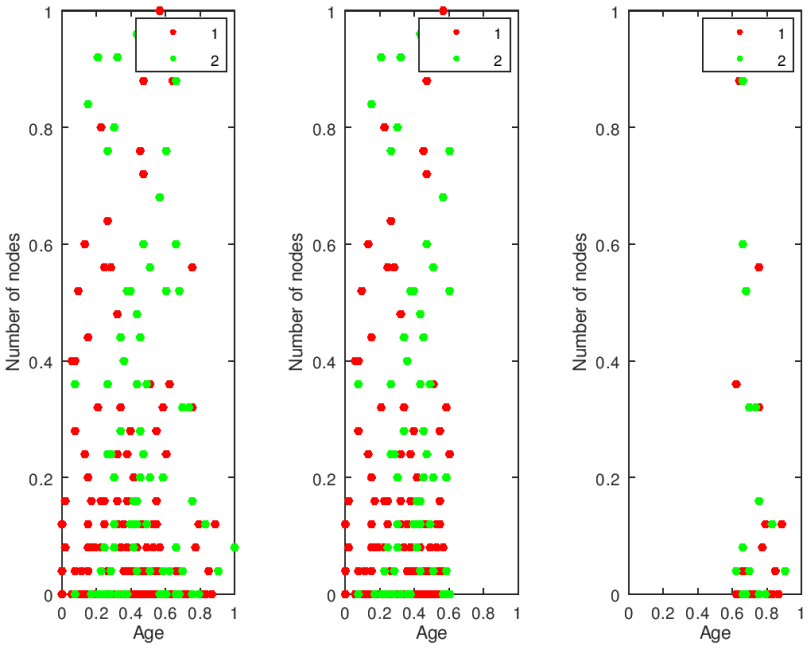

Figure 4. Gscatter plots of features: patients age, number of nodes. First gscatter – instances of dataset, second – training set, third - validation set. Legend: 1 – survived more than 5 years , 2 – died within 5 years

2.4. Classifiers

Linear Classifier (LDA) and Quadratic Classifier (QDA) were chose as supervised learning methods. After classification application draw gscatter plots to compare the results.

Discrimination (Latin discriminatio - discrimination) - a form of unjustified marginalization (social exclusion), manifested by treating a person less favorably than another in a comparable situation due to a feature, such as developmental period, disability, sexual orientation, sex, professed religion, worldview, nationality or race [6].

Referring to the problem of statistical methods, we find the method that is called the discriminant analysis. The method used to find a linear combination of features that best distinguishes between two or more object classes or events is *linear discriminant analysis* (LDA). Discriminant analysis to ensure that every observation of $x \in X$ is assigned to a class, the observation is completed. The discriminating (classifying) rule divides the set of X into g discrete subsets (classes) and gives the record as a mapping: $d(x) : X \rightarrow G$, where G is a completed g-element set of class labels, to which observations belong, while what $g \geq 2$ [6].

Function:

```
class = classify(sample,training,group,'LDA')
```

In its basic form, the function classifies each row of the 'sample' matrix, having first been trained from the 'training' matrix.

- 'group' shows which group corresponds to each instance of the 'training' matrix.
- 'class' contains the predicted group of each instance of the 'sample'.
- 'group' contains the classes for the training.
- for Quadratic classifier must be 'QDA2' instead of 'LDA' which is linear classifier

The matrices named 'sample' and 'training' must have the same column number. 'group' contains the classes for the training.

[class,err] = classify(...)

Another syntax of the function returning an estimate of the error rate based on training data set.

The QDA algorithm is useful in machine learning and statistical classification to separate measurements of two or more object or event classes using a square. It can be said that this is a general version of the linear classifier. QDA is a parametric approach in supervised learning, involving the use of Gaussian distribution. Gaussian parameters for all classes can be found on the basis of training points with an estimate of the maximum probability. This simple Gauss model is best suited for cases where there is not much information, i.e. if there are too few training samples to infer the class distributions. Especially when the number of training samples is small compared to the number of dimensions of each training sample, the covariance estimate can be poorly done [8].

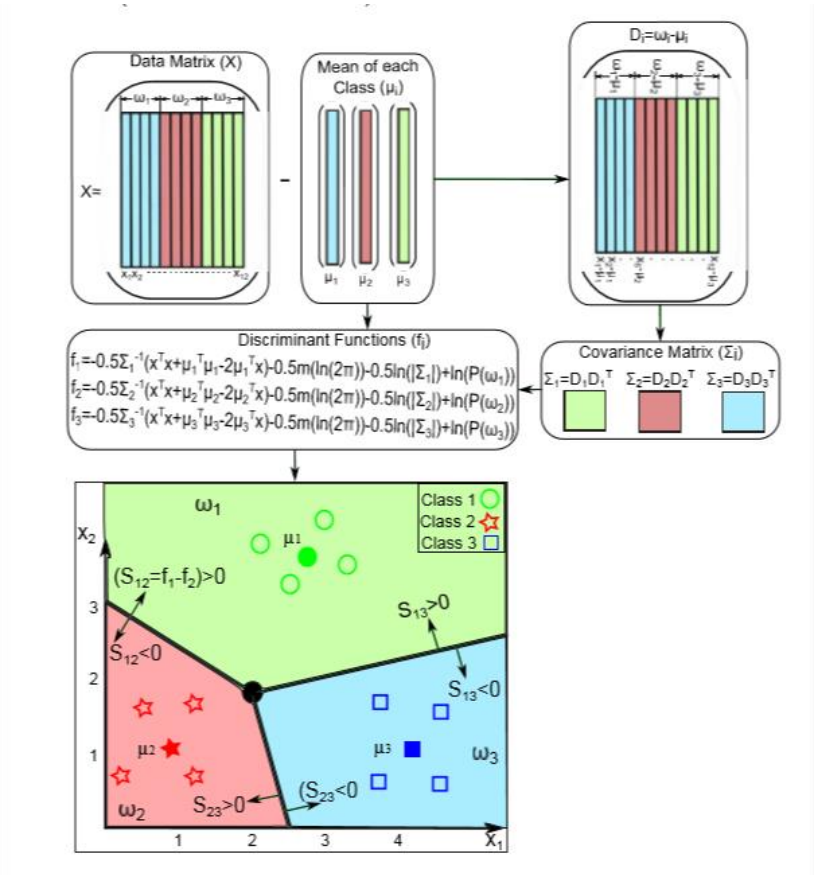

Figure 5. Steps of calculating example of discriminant analysis classifier given three classes, each class has four samples

2.5. Kmeans and Hierarchical Grouping

Kmeans and Hierarchical Grouping were chose as unsupervised learning methods. After classification application draw plots to compare the results and dendrograms.

K-mean concentration is a cluster analysis method that divides n different observations into k different clusters, and each observation belongs to the cluster with the nearest average. This problem is difficult, but there are effective heuristic algorithms that are often used and reach the local minimum quickly. These algorithms are similar to the algorithm of maximizing expectations for mixtures of Gaussian distributions using the iterative approach to improvement used by both algorithms.

Octave function [9]:

`[idx, centers, sumd, dist] = kmeans (data, k, 'distance', value1, …)`

'idx': contains the cluster number that each instance belongs to, 'centers': contains information about the centers, 'sumd': contains the sums of the distances of the instances from the center, 'dist': contains the distances of each instance from each center

'distance' - the distance measure used for partitioning and calculating centroids. Possible values are:

- sqeuclidean
- cityblock
- cosine
- correlation
- hamming

The Linkage algorithm creates a hierarchy of groups in which groups on one level connect to the next highest level. Thanks to this, we can decide which level is the best for our data.

Procedure is that we use the pdist function to calculate the distance between instances. Then we use the linkage function. Thanks to this function, observations that have a smaller distance (information from the previous step) are grouped into clusters. The next connection is formed between the observations and / or the group with the only criterion of the shortest distance. In this way, clusters are created continuously until one cluster contains all instances.

Functions [10]:

```
D = pdist(X,distance)
```

It calculates the Euclidean distance of a matrix X (m x n), by using the method defined to the parameter 'distance' according to the next table. The default value is 'Euclidean'.

```
y = linkage (D, method)
```

The method defines the way the distance between two clusters is calculated and how they are indexed when two clusters are merged.

3. Results

3.1. Boxplots results:

People who have lived for more than 5 years (1st class) are 2,7 times more than those from the 2nd class. The first boxplot shows that the median age of people who underwent surgery is similar for both classes (about 52). However, people representing class 2 were a little older. The year of operation for both classes has the same median (1963). The difference can be seen in the first and third quartiles, they have lower values for the 2nd class. Differences occurred on the boxplot of number of positive axillary nodes detected. The median for 1st class is 0, although there have been cases of even tens nodes for individual results. For patients who lived less than 5 years after surgery, the average number of nodes is definitely higher. The conclusion is the higher result for this feature is characterized by class 2.

The best feature is number of nodes (feature 3) because we can see clearly the difference between the results for classes. In feature 1 and 2 there are also differences, however smaller than in feature 3. Features **patients age, number of nodes** were chosen to use in next learning methods.

3.2. Linear and Quadratic Classifier

Error (errn):

Linear: 0,25

Quadratic: 0,2375

Visualization classification results (1 subplot) and Validation Set before (2 subplot) and after classification (3 plot):

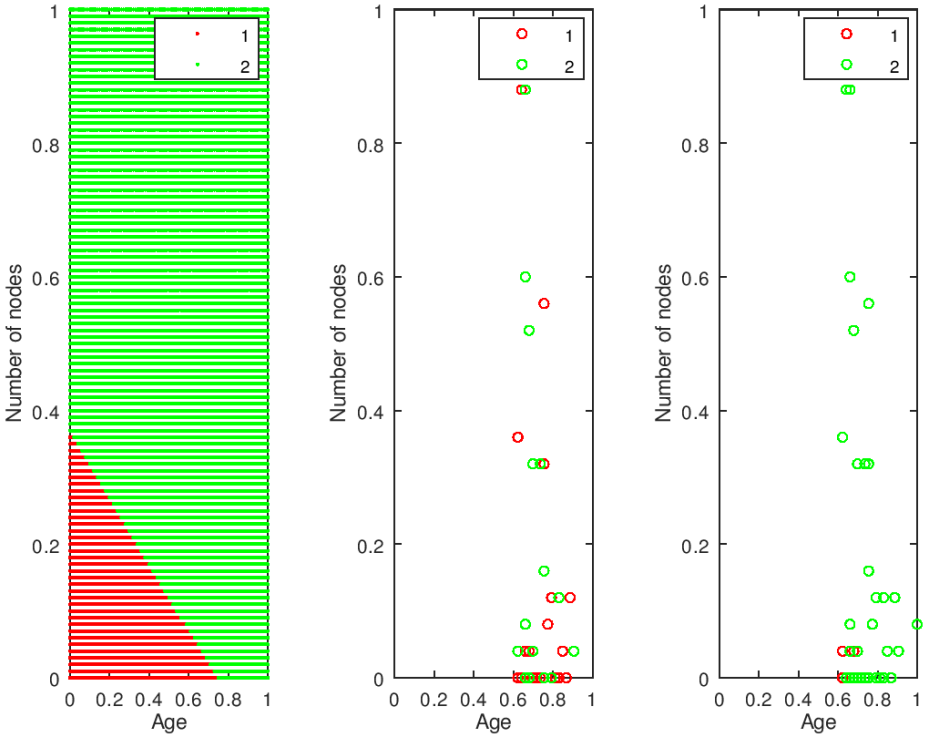

Figure 6. Visualization results of linear classify

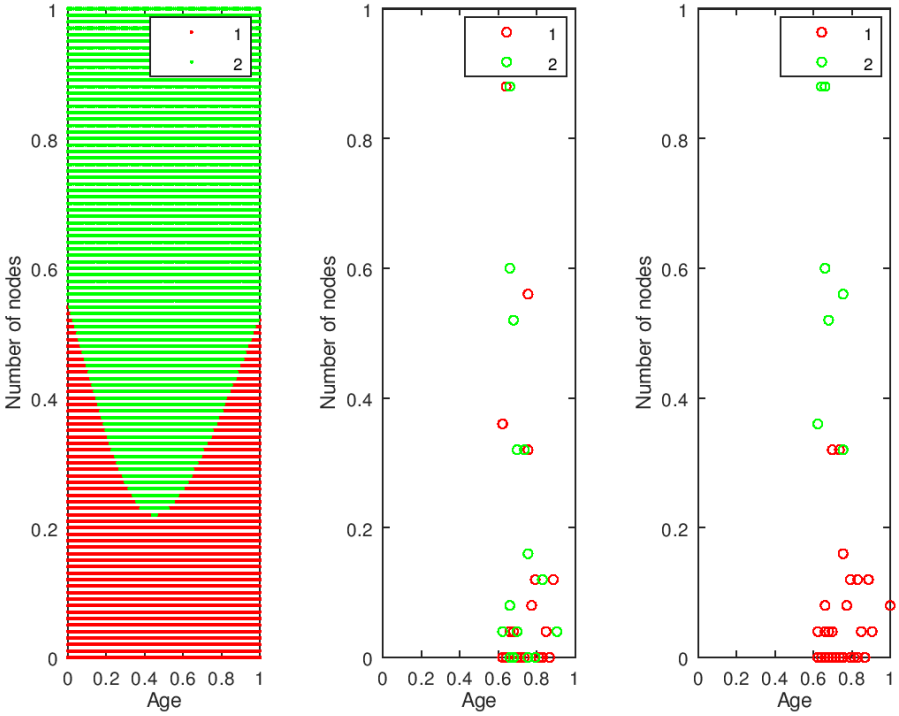

Figure 7. Visualization results of quadratic classify

There were differences between the results of classifiers. For the linear, only a few instances, with a small number of nodes Validation Set after classification belongs to class 1. Most is in class 2, unlike Quadratic classifier.
Quadratic classifier has lower error and the result is better for this classifier as we can see on plots. To class 1 belong instances with a small number of nodes, and to class 2 those with a larger number, what agrees with gscatters.
Both classify methods were used with another percentage splits of dataset. The results were similar but more instances belongs to class 1 in validation set after using linear classifier:

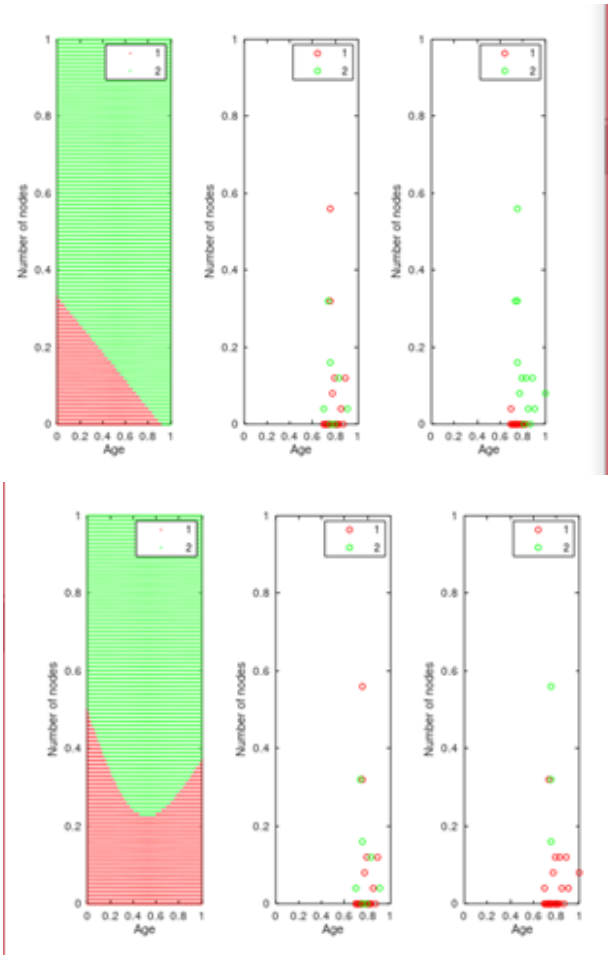

Figure 8. 90-10% split data percentage results visualization

3.3. Kmeans

Firstly, was changed number of **clusters to 2** and distance measure setting to **'hamming'.**
Legend: 1-class1 (live more than 5 years), 2- died in 5 years.

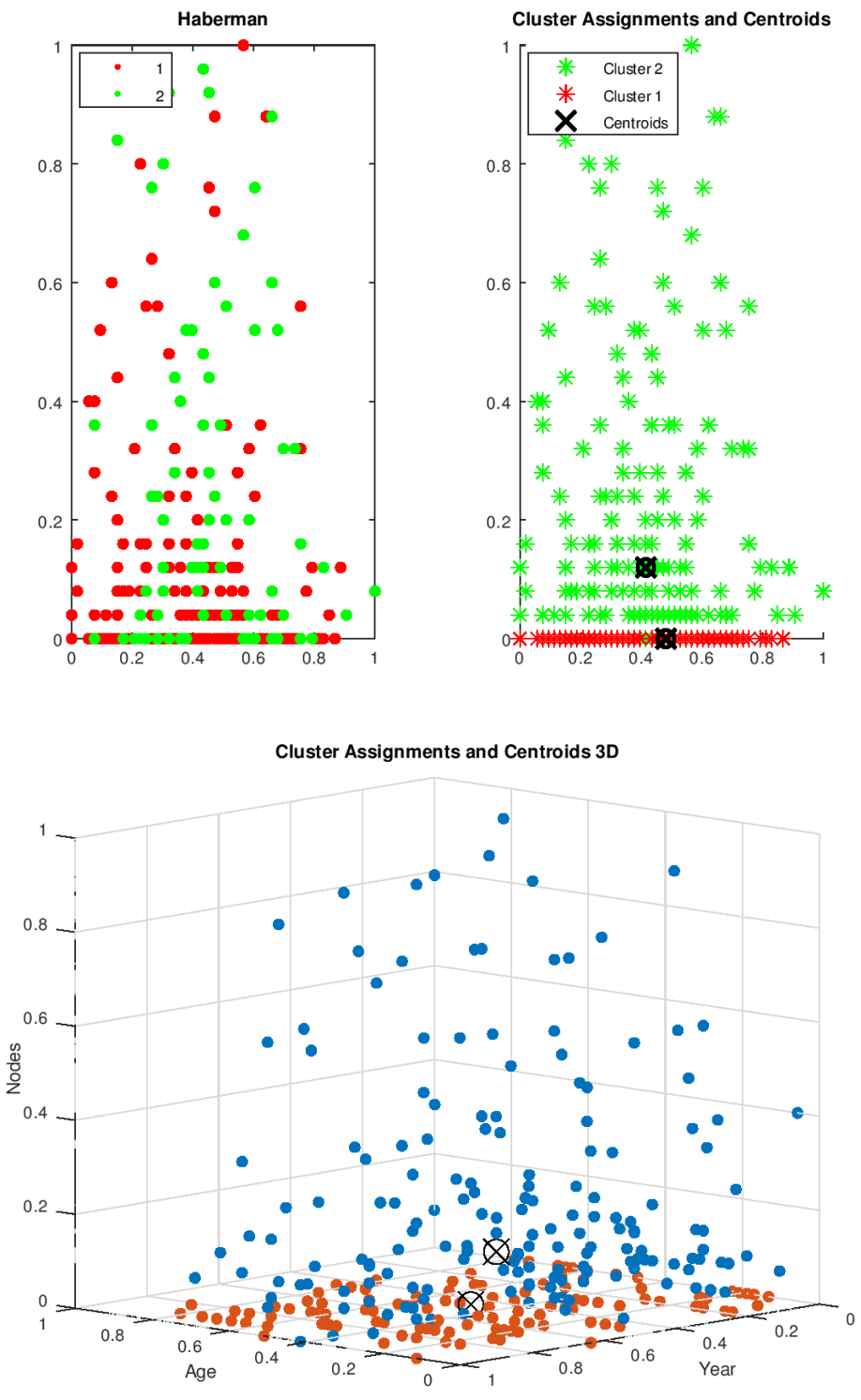

Figure 9. Visualization results of kmeans clustering (2 clusters, ‘hamming’) in 2D and 3D.

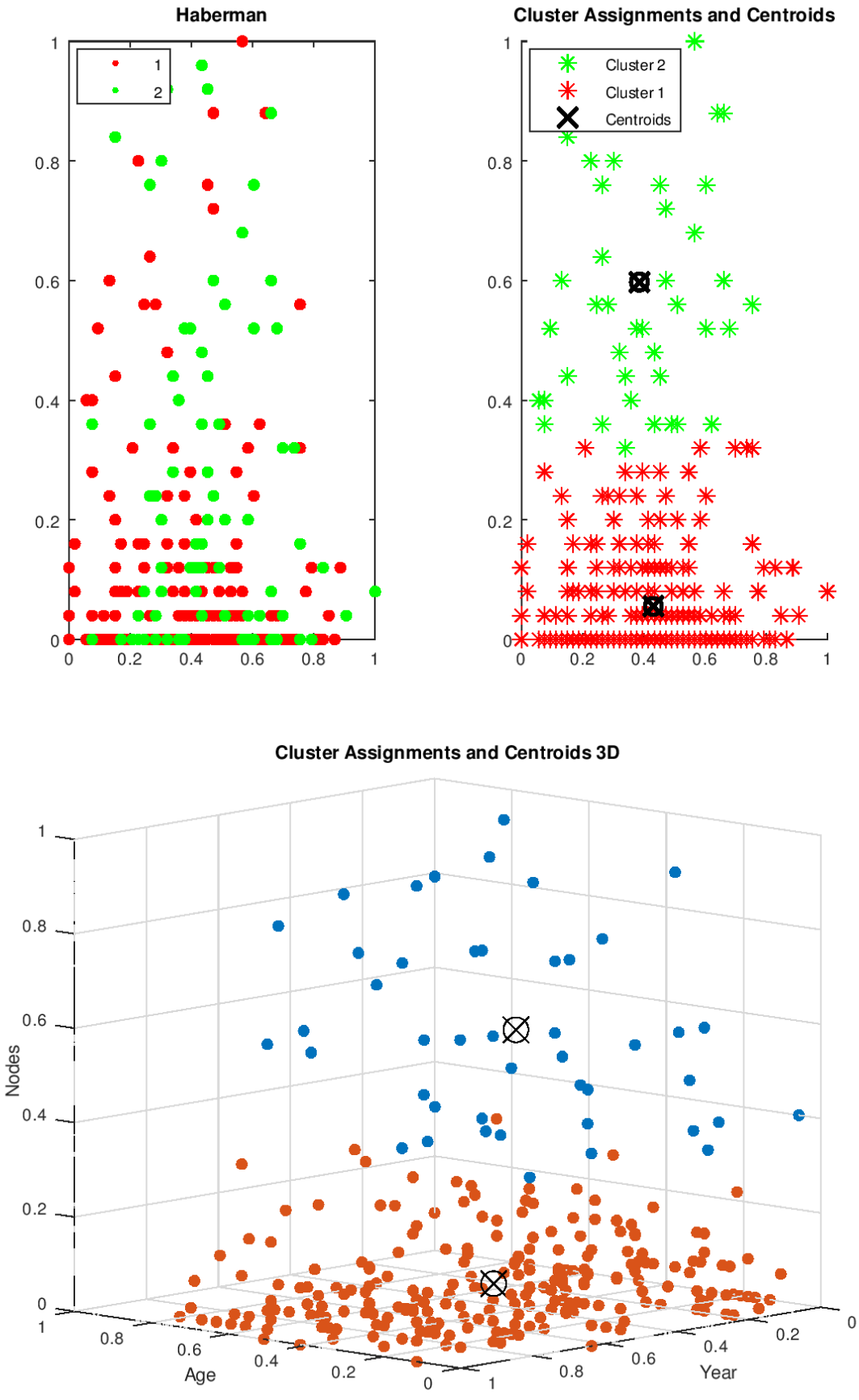

Figure 10. Visualization results of kmeans clustering (2 clusters, 'Euclidean')

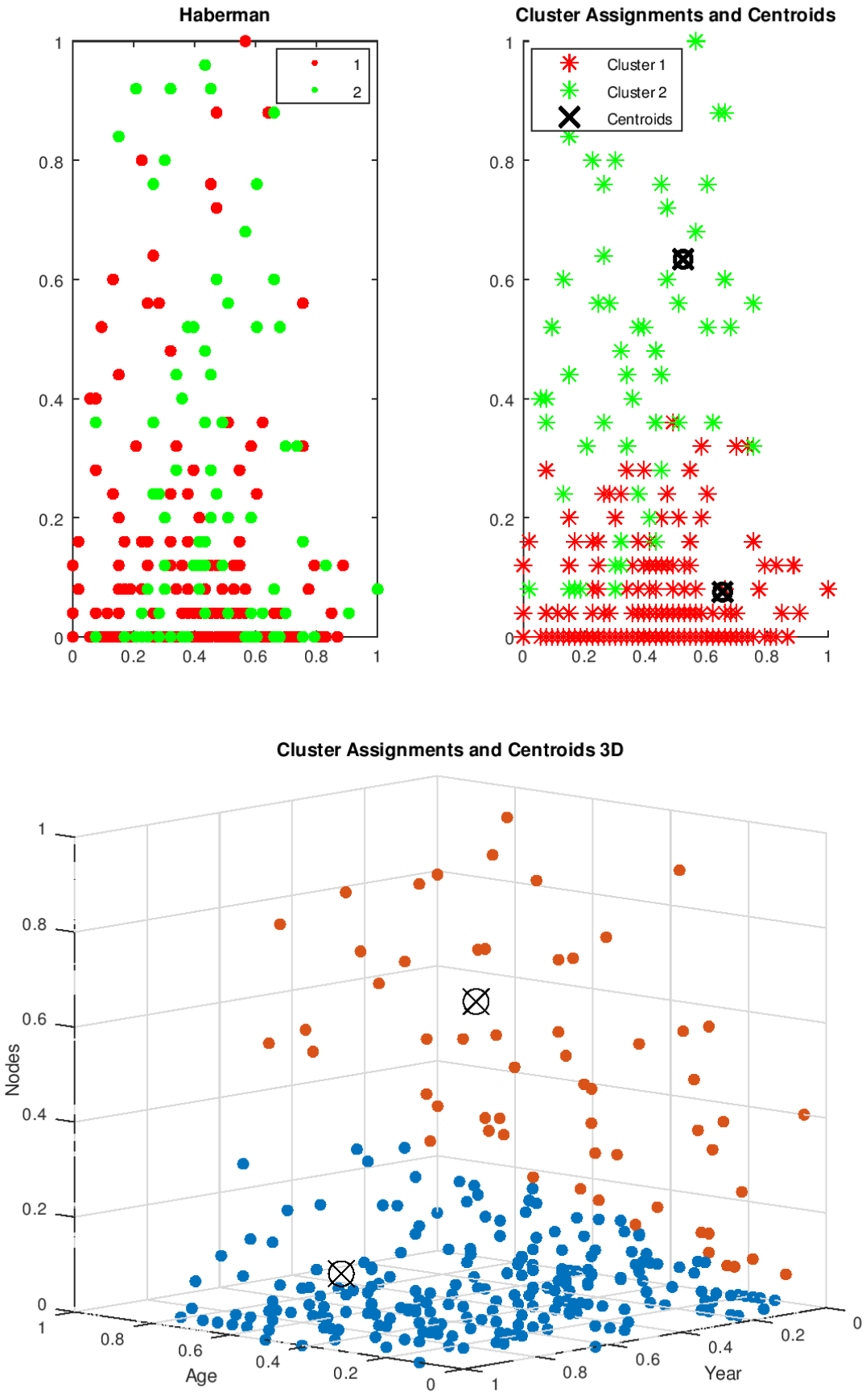

Figure 11. Visualization results of kmeans clustering (2 clusters, ‘cosine’)

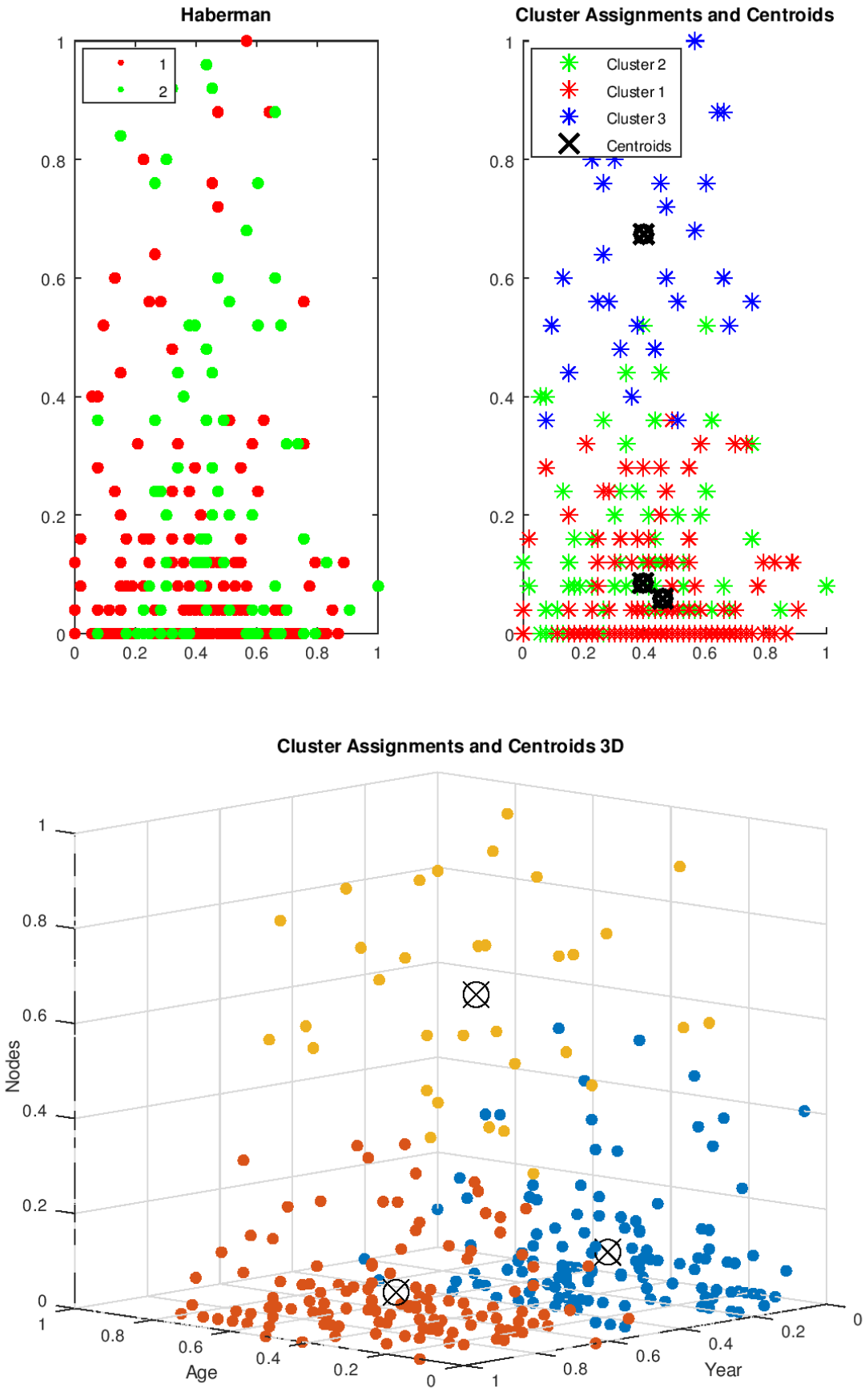

Figure 12. Visualization results of kmeans clustering (3 clusters, ‘Euclidean’)

In kmeans using the Hamming distance with creation of 2 clasters, you can see that the class 1 cluster was created close to zero nodes because many of the class 1 has very low values of these nodes, and class 2 above, because their values in this class are higher.
After changing the Euclidean distance, cluster 1 includes instances with a higher value of nodes feature than before, but it does not seem bad because up to 0.4 on the vertical axis there are still many instances of class 1.

In the experiment with the creation of 3 clusters, cluster 1 with a small node value, was divided into 2 clusters. The division is on the axis of the year of operation.

3.4. Hierarchical Grouping

Dendrograms:
Distance method: Euclidean, linkage method: ‘average’

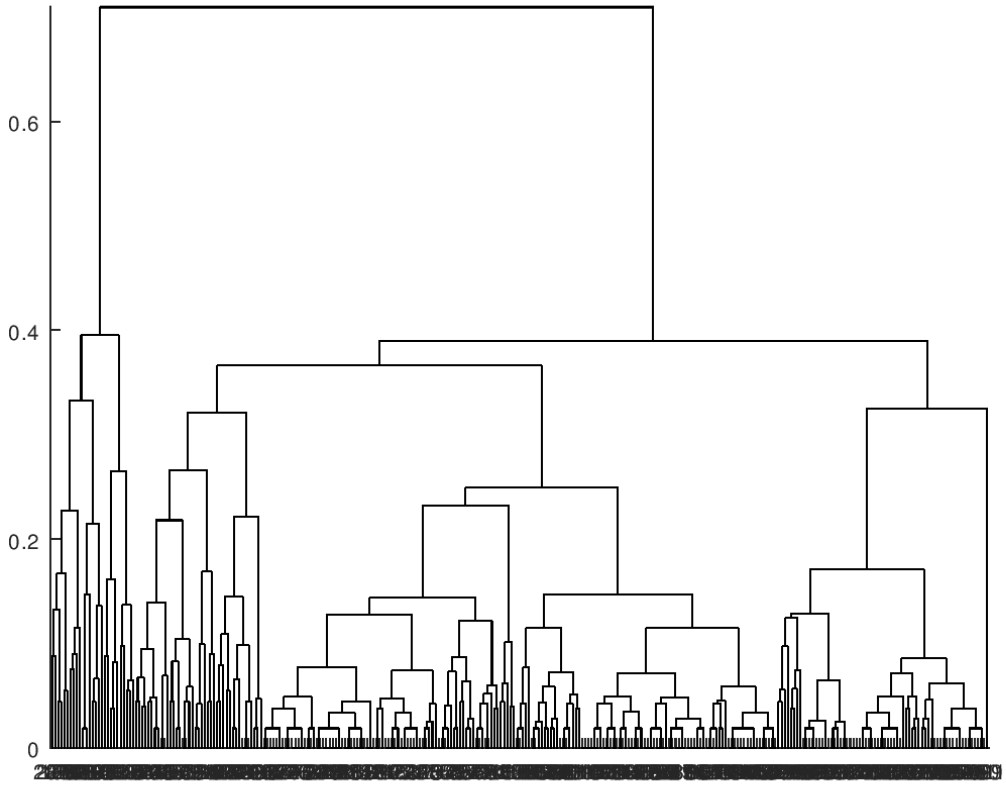

Figure 13. Visualization results of hierarchical grouping (‘average’,’hamming)

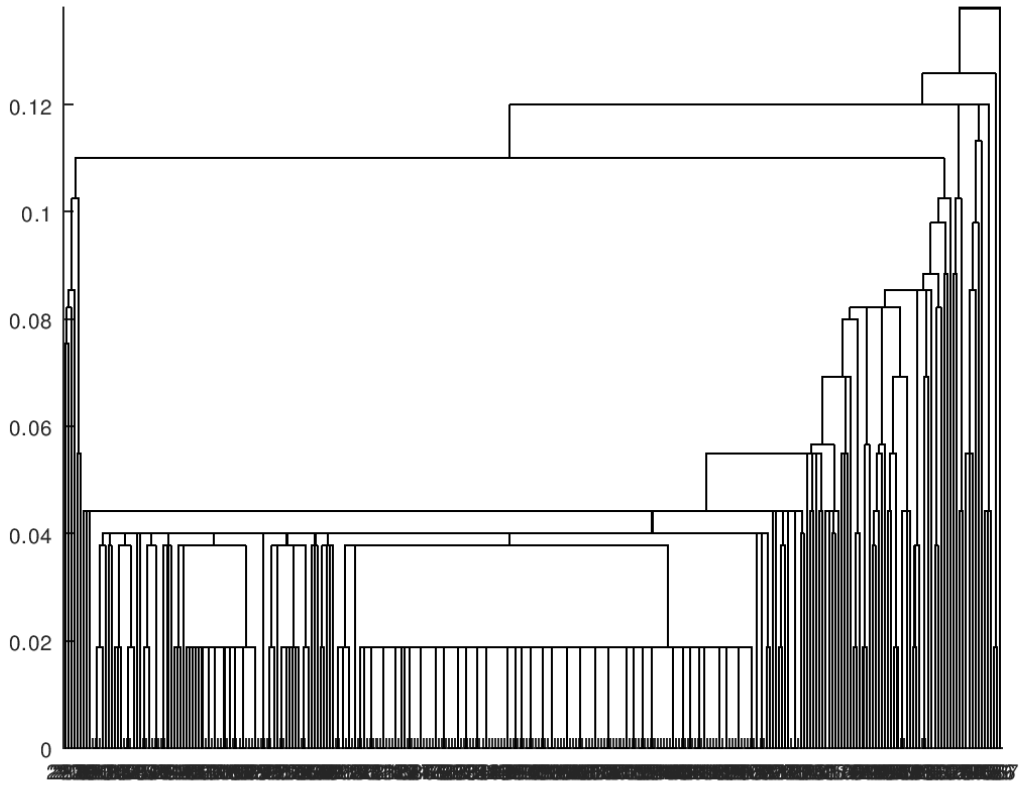

Figure 14. Visualization results of hierarchical grouping ('default','hamming)

The largest node in first dendrogram (method: average) tree is much higher than the other nodes, so there are two large groups. The first of the 2 major nodes which is on the left contains many instances in a small range of x axis values. Second major node which is on the right contains instances with a large range of values on the x-axis. As in the previous methods, there are 2 groups, one of which contains a lot of instances in a small area like class 1, the other one in a larger area like class 2.
On the second dendrogram (method:default) near to the beginning of coordinate system is one narrow high node. Then nodes high is growing with value on the x-axis.

4. Conclusions

Both supervised and unsupervised learning methods gave satisfied results of classification collections in the data set. The best results were given by quadratic classifier and kmeans. Both recognized a large number of instances with a small number of nodes as class 1, and those with a large number as class 2. Linkage dendrogram confirming also what was visible before: there are two large, separate groups. Linear classifier after classifying the majority of validation set as class 2, which was a less correct result compared to Quadratic.

Distinct differences occurred between results using similar methods. Linear and quadratic differed in results, similar to K-means (Euclid and Hamming methods).

In general, based on the results of learning algorithms, it can be concluded that the number of axillary nodes can affect the life expectancy of patients after breast cancer surgery. Those who survive longer have a small number of nodes in relation to those living shorter.

5. Bibliography

[1] M. Parasher, S. Sharma, A.K Sharma, and J.P Gupta, *Anatomy On Pattern Recognition*, Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, no. 3, Jun-Jul 2011.

[2] SeemaAsht and RajeshwarDass, *Pattern Recognition Techniques: A Review*, International Journal of Computer Science and Telecommunications, vol. 3, issue 8, August 2012.

[3] UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science.

[4] Frigge, Michael; Hoaglin, David C.; Iglewicz, Boris (February 1989). *Some Implementations of the Boxplot*. The American Statistician. 43 (1): 50–54.

[5] Jacek Koronacki, Jan Mielniczuk: *Statystyka dla studentów kierunków technicznych i przyrodniczych*. Warszawa: WNT, 2006, s. 289,304. ISBN 83-204-3242-1.

[6] *What is Discrimination?*. Canadian Human Rights Commission. Archived from the original on 2018-04-15. Retrieved 2018-04-15.

[7] Agnieszka Nowak-Brzezińska, *Analiza dyskryminacyjna* Konspekt do zajęć: Statystyczne metody analizy danych, 8 stycznia 2010.

[8] M. Parasher, S. Sharma, A.K Sharma, and J.P Gupta, *Anatomy On Pattern Recognition*, Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, no. 3, Jun-Jul 2011.

[9] https://octave.sourceforge.io/statistics/function/kmeans.html

[10] https://octave.sourceforge.io/statistics/function/pdist.html

[11] https://octave.sourceforge.io/statistics/function/linkage.html