SAPIENZA
UNIVERSITÀ DI ROMA

Multilevel mixed-type data analysis for validating partitions of scrapie isolates

Scuola di dottorato in Scienze Statistiche

Dottorato di Ricerca in Statistica Metodologica – XXIX Ciclo

Candidate

Giorgia Rocco
ID number 1432661

Thesis Advisor

Prof. Luca Tardella

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Methodological Statistics

2017

Thesis defended on September 2017
in front of a Board of Examiners composed by:

Caterina Conigliani (chairman)
Pietro Coretto
Maria Giovanna Ranalli
REFEREE: Marilena Barbieri
REFEREE: Alessandra Nardi

Multilevel mixed-type data analysis for validating partitions of scrapie isolates
Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: September 11, 2017

Author's email: g.rocco@uniroma1.it

Acknowledgments

This work is within the research project: *"The zoonotic potential of Transmisible Spongiform Encephalopathies"* supported by the Italian Ministry of Health.

RF-2009-1474624

The PhD fellowship was sponsored by the Italian **Istituto Superiore di Sanità**.

The author wishes to thank **Romolo Nonno** (Scientific Coordinator) and **Michele Angelo Di Bari** from the Department of Food Safety and Veterinary Public Health; **Maria Puopolo** from the Department of Cell Biology and Neuroscience.

Abstract

This dissertation arises from a joint study with the Department of Food Safety and Veterinary Public Health of the Istituto Superiore di Sanità. The aim is to investigate and validate the existence of distinct strains of the scrapie disease taking into account the availability of a priori benchmark partition formulated by researchers. Scrapie of small ruminants is caused by prions, which are unconventional infectious agents of proteinaceous nature affecting humans and animals. Due to the absence of nucleic acids, which precludes direct analysis of strain variation by molecular methods, the presence of different sheep scrapie strains is usually investigated by bioassay in laboratory rodents. Data are collected by an experimental study on scrapie conducted at the Istituto Superiore di Sanità by experimental transmission of scrapie isolates to bank voles.

We aim to discuss the validation of a given partition in a statistical classification framework using a multi-step procedure. Firstly, we use unsupervised classification to see how alternative clustering results match researchers' understanding of the heterogeneity of the isolates. We discuss whether and how clustering results can be eventually exploited to extend the preliminary partition elicited by researchers. Then we motivate the subsequent partition validation based on the predictive performance of several supervised classifiers.

Our data-driven approach contains two main methodological original contributions. We advocate the use of partition validation measures to investigate a given benchmark partition: firstly we discuss the issue of how the data can be used to evaluate a preliminary benchmark partition and eventually modify it with statistical results to find a conclusive partition that could be used as a “gold standard” in future studies. Moreover, collected data have a multilevel structure and for each lower-level unit, mixed-type data are available. Each step in the procedure is then adapted to deal with multilevel mixed-type data. We extend distance-based clustering algorithms to deal with multilevel mixed-type data. Whereas in supervised classification we propose a two-step approach to classify the higher-level units starting from the lower-level observations. In this framework, we also need to define an ad-hoc cross validation algorithm.

Contents

Introduction

This work analyzes data from an experimental study on prion diseases, focusing on the scrapie of small ruminants. Prions are unconventional infectious agents of proteinaceous nature affecting humans and animals. They lack nucleic acids, then direct analysis of strain variation by molecular methods is not possible. The studies on these diseases use data from the phenotypes observed by bioassay in laboratory rodents.

Scrapie is historically known to occur as different strains (Bruce et al. 2002), a fact which has renewed interest due to the zoonotic risk posed by animal prions (Cassard et al. 2014).

The scrapie strain characterization is routinely performed evaluating the phenotype of each rodent, and taking the average of collected variables from all animals inoculated with the same sheep scrapie disease. A large experimental study on strain characterization in bank vole has been carried out at Istituto Superiore di Sanità (Italian National Institute of Health). The resulting experimental data can be conceived as multilevel data with multiple bank voles (lower-level units) inoculated with the same sheep scrapie (higher-level unit). Each lower-level unit presents mixed-type because both continuous (survival time) and ordinal (vacuolation scores of brain areas) variables are assessed for each experimental animal.

We address the general scientific issue of investigating how the available experimental data can be used to evaluate the appropriateness of a partition, identifying strains or groups of strains. This is done, taking into account a possible preliminary benchmark partition proposed by researchers. Our aim is to validate a final, possibly finer, partition.

The identification of natural partitions underlying available data is usually achieved by means of unsupervised classification. It is common practice to employ a variety of different clustering techniques and use visual inspection and prior knowledge to select what is considered the most appropriate results (Handl et al. 2005, Hennig 2015). Several clustering validation techniques are also used (see Handl et al. (2005) for a review) and we usually distinguish between internal and external validation measures (Halkidi et al. 2001). Internal measures are based on different notions of clustering quality, as in intra-cluster homogeneity or separation between clusters (Liu et al. 2010), but also on predictive power as proposed for example in Yeung et al. (2001) Tibshirani & Walther (2005) and in Volkovich et al. (2009). External measures take a known set of class labels (the "gold standard") and compare the cluster results with the known labels.

In our study, the "gold standard" is not known and the aim is to define it with a statistically well-grounded partition which may account for some preliminary knowledge elicited by researchers in terms of a prior partition. This partition will be shortened hereafter with RP. Statistical analyses of the experimental data are performed to question the a priori partition given by the researchers that could undergo an extension or a reformulation.

Hence, we propose a sequential procedure based on both unsupervised and supervised classification. The unsupervised classification is used to have alternative clustering results to compare with RP in order to confirm or extend this partition. Once the partition is refined, a final validation step can be based on the predictive performance of supervised classification. We deviate somehow from the mainstream of cluster validation techniques proposed in the literature, which often use a distance-based index such as the silhouette index. We believe that a solid approach should rely also on measuring the predictive performance via flexible supervised classification, possibly combining multiple classifiers. In our experimental context, a sufficiently good predictive performance, based on probability statements, could be better interpreted by researchers and enable researchers to understand to what extent new experimental data could be correctly identified within the validated group structure so that the new partition could be used in future studies as the "gold standard".

The experimental data have a multilevel structure with mixed-type data measured for the lower-level units, so the statistical methods are extended to deal with this challenging data structure.

The analysis of grouped data is often approached by using probabilistic mixed-effects models (Goldstein 2011). Model-based clustering for this type of data relies on finite mixture modeling as pioneered in Vermunt & Magidson (2005), Vermunt (2008) and more recently extended in Calò et al. (2014). However, with our mixed-type data, we have decided to avoid the use of probabilistic assumptions on not yet well understood ordinal data for which the use of finite mixture models would require the distributional elicitation of suitable mixture components. Indeed, the model-based approaches for clustering mixed-type data consider each group of variable's type separately and then a model for the joint distribution of the continuous latent variables is assumed (Everitt & Merette 1990, McParland et al. 2014, Ranalli & Rocci 2017, Carmona et al. 2016). Hence, we stick to distance-based methods.

Distance-based methods for clustering repeated measures are proposed in Yeung et al. (2003) to deal with gene-expression data. They propose to employ hierarchical clustering algorithms with an ad-hoc adjustment named FITSS (forcing into the same subtrees) to account for the hierarchical nature of the data. Starting from this work, we extend and refine distance-based methods for clustering hierarchical mixed-type data. In our context, there is a relevant heterogeneity within observations belonging to the same higher-level unit, differently from the typical case in Yeung et al. (2003) where repeated measurements are considered. The substantial impact of such heterogeneity in our hierarchical data can undermine the stability of the results of a specific cluster analysis relying on a particular choice of distance and algorithm. In order to account for this aspect, we propose a consensus clustering of hierarchical algorithms based on alternative sensible choices of distances with FITSS adjustment. We also discuss the results obtained using partitioning strategies relying on a modified Partition Around Medoids (PAM) algorithm (Kaufman & Rousseeuw 2009).

We also propose to measure the distance between two higher-level units with the resolution of the optimal transport plan problem, such as a Wasserstein distance between two higher-level units is defined and classical distance-based algorithms can be used to perform the cluster analysis.

Another statistical challenge for our data analysis is related to the supervised classification methods for hierarchical data. This problem has been addressed in Yamal et al. (2011) and Yamal et al. (2015). They identify three possible approaches to the problem: (i) extract higher-level features from the lower-level data, (ii) use a statistical model that accounts for the hierarchical data structure, and (iii) classify at the lower-level and use an ad hoc approach to classify at the higher-level. We provide a contribution within the third approach using a model based combination of classifiers (Kakourou et al. 2014) on the average of lower-level predicted probabilities, to enhance the performance of the supervised classification and have a unique class label prediction for higher-level units. Comparative evaluation of supervised classification is done in order to find the best classifier and hence validate the group structure. In this process, overfitting issues may arise hence we need a cross-validation (CV) strategy to obtain unbiased estimates of the classifiers performance. In this framework, we propose a suitably modified double cross-validation (Mertens et al. 2006) for higher-level units which takes into account the unbalanced nature of the data.

The thesis is organized as follows. In Chapter 1 we discuss the two different approaches to deal with partitions: unsupervised and supervised classification with attention to the validation problem. We also discuss the Istituto Superiore di Sanità motivating example. We explicit our proposal to deal with the scientific issue proposed by experimental researchers and we motivate it with a simulated example. Chapter 2 starts with the description of the dataset and the multilevel mixed-type data challenge. The central part is devoted to a review of the current methodologies for the multilevel data and mixed-type data separately. We also address some of the dataset's limits. Clustering methods, both for multilevel data and mixed-type data, are reviewed in Chapter 3 and our distance-based method proposals are presented. In Chapter 4 we firstly review the current methods to perform supervised classification with multilevel data. Secondly, we present a two-step approach to performing classification and within this framework we propose a model that simultaneously combines classifiers and aggregates the lower-level units using the average predictive probabilities. At the end, a double cross-validation algorithm is proposed for multilevel data. Finally, results on the Scrapie dataset and final remarks are discussed in Chapter 5.

Chapter 1

A different perspective to deal with partitions

1.1 Learning from data with group structures

The recognition of patterns and regularities in data is a mainstream topic in statistical literature. Pattern recognition, machine learning or data mining are different fields dealing with this problem. Following the distinction described by Bishop (2006) pattern recognition has its origins in engineering, whereas machine learning grew out of computer science. Duda et al. (2012), Friedman et al. (2009) and Bishop (2006) are three of the most important books dealing with this problem and they talk about “learning from the data”. Focusing on the problem of classification for data with group structures, they distinguish when the learning process is trained from labeled “training” data (supervised learning or classification), from the cases where no labeled data are available and the aim is to discover previously unknown patterns (unsupervised learning or classification).

In supervised classification, class labels are available for a set of data (the training set) related to several variables called covariates or predictors. A supervised algorithm analyzes the training data to learn an inferred function, which can be used to classify new units in which a set of predictors is given, but the class label is unknown.

Unsupervised classification, or cluster analysis, has a different aim: the learning algorithm is used to find unknown groups in data. Various algorithms are used to cluster data and they differ in the notion of what constitutes a cluster: in distance-based methods, clusters are formed by units for which the relative distances are smaller for units belonging to the same group than for units belonging to different groups; while in model-based methods, clusters are composed of units belonging most likely to the same distribution. It is common practice to employ a variety of different clustering techniques and use visual inspection and prior knowledge to select what is considered the most appropriate result, but as it was noted by Estivill-Castro (2002), "clustering is in the eye of the beholder". Hennig (2015) offers a philosophical point of view about the "true clusters", analyzing the problem of choosing and assessing clustering methods and results. A critical point of his discussion is that what the "true cluster" are, depends on the context and clustering aim. Therefore researchers should specify what problem-specific "truth" they are interested in.

The existence of a rich variety of different algorithms, both in unsupervised and in supervised classification, highlights the need for an objective criterion to assess the performance of different methods. Validation techniques and indexes are available to evaluate the alternative algorithms' results. Jain & Dubes (1988) suggest "The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage".

1.2 Clustering validation techniques

Clustering techniques are used in the identification of group structures in data, but alternative algorithms can lead to different results. The clustering-validation techniques have been widely developed in the literature and several reviews are available: Handl et al. (2005) give a detailed review in post-genomic data analysis; Halkidi et al. (2001) distinguish between internal and external validation measures whereas Liu et al. (2010) focus on internal clustering validation measures. Brock et al. (2011) provide an R package named `clValid` with the most popular validation indexes, and more recently a new package, `clusterCrit` was developed by Desgraupes (2013*a*) with a joint clustering validation techniques review provided by Desgraupes (2013*b*).

1.2.1 Internal validation measures

Internal measures of cluster validity use only quantities and features related to the dataset. Following Handl et al. (2005) we can distinguish six different types of internal measures. Three types are related to notions of clustering quality as compactness, connectedness, and separation, while another type is based on the combination of these notions. Other validation techniques are based on predictive power/stability and compliance between partitioning and distance information.

Type I: Compactness. The measures of this type are based on the topological concept of compactness: units of each cluster should be as close to each other as possible. The most popular index is the intra-cluster variance and another related index is the sum-of-squared errors minimum variance criterion (Halkidi & Vazirgiannis 2001). Lower intra-cluster variance indicates better compactness. Also, there are other indexes based on distance, such as maximum or average pairwise distance, and maximum or average center-based distance. The variance based indexes can be calculated only for quantitative data, while the indexes based on distance can be adapted on other types of data (as categorical) with an appropriate choice of the distance index.

Type II: Connectedness. The second type of internal validation technique is based on the local concept of connectedness: data items should be on the same cluster with their nearest neighbors. Ding & He (2004) introduce the K-nearest-neighbor consistency concept requiring that for any data point in a cluster, its k-nearest neighbors should also be in the same cluster. Handl et al. (2005) propose a synthetic index called Connectivity defined using an indicator $x_{i,nn_{i(j)}}$ which is equal to 0 if the i-th observation and its j-th nearest-neighbor are in the same cluster of a partition $\mathscr{C}$ and $1/j$ otherwise:

$$Conn(\mathscr{C}) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}} \tag{1.1}$$

this index varies from 0 to ∞, but it depends on the choice of L, the number of nearest-neighbors to be considered.

Type III: Separation. The third type of internal measure evaluates the degree of separation between clusters. For example, an average of the distances between cluster centroids or the minimum distance between data items belonging to different clusters can be used. Separation is measured by the *between cluster sum of squares* in contrast with the *within cluster sum of squares* used for the compactness. Separation indexes are mostly defined for quantitative data and are widely used for the number of groups' choice in the K-means algorithm.

Type IV: Combinations. These types of internal indexes simultaneously allow for multiple concepts introduced above. Combinations of type one and type three gives new indexes: the compactness improves with an increasing number of clusters, while the separation is better suited for a lower number of clusters. The most popular validation indexes are linear or non-linear combinations of the two measures to validate clusters, balancing the two properties. A linear index is the SD-validity index introduced by Halkidi et al. (2000) as a weighted average of the *scattering* of clusters and total separation of clusters. The scattering, used for the compactness, is calculated by the ratio of the clusters variance and variance of the dataset. The total separation of clusters is based on the distance between cluster center points. This index must be used only for quantitative data. Examples of non-linear combinations are the Dunn Index (Dunn 1974), and the Silhouette Width (Rousseeuw 1987). The Dunn Index measures the ratio between the smallest cluster distance and the largest intra-cluster distance in a partition. Let C_m denote a cluster and $dist(C_k, C_l)$ is the minimal distance between pairs of data items i and j with $i \in C_k$ and $j \in C_l$, while $diam(C_m)$ is the maximum intra-cluster distance within cluster C_m, the Dunn index is defined as:

$$D(C) = \min_{C_k \in C} \left(\min_{C_l \in C} \frac{dist(C_k, C_l)}{\max\limits_{C_m \in C} diam(C)} \right). \tag{1.2}$$

The Silhouette width index is an evaluation of how well a unit fits within its cluster compared to the other clusters. Let $a(j)$ be the average dissimilarity between j and all objects in its cluster $C_j \ni j$ and let $d(j; C)$ be the average dissimilarity of j to all objects in C, for a fixed $C \neq C_j$. Denote with $b(j)$ the smallest distance $d(j; C)$ found among all clusters $C \neq C_j$. An evaluation of how well object j is classified in

C_j or in the neighbor cluster C is given by the following *silhouette width index*:

$$s(j) = \frac{b(j) - a(j)}{\max\{a(j), b(j)\}}. \tag{1.3}$$

$s(j)$ assumes values into the interval $[-1; 1]$. Observations with a value of $s(j)$ close to 1 are very well clustered, while a small value of $s(j)$ means that the observation can be assigned to two clusters, and observations with a negative $s(j)$ are misplaced. The silhouette index may help to select the number of clusters using the average silhouette width $\tilde{s}(K)$, which is the average of $s(K)$ over all objects of any possible clustering with K groups (Kaufman & Rousseeuw 2009).

Type V: Predictive power. These indexes are not related to cluster properties, but use the ability of an algorithm to find the same groups even when the data are perturbed. This ability is called *predictive power* and it is related to the stability of an algorithm. These indexes are not external since they do not make use of a class label but rely on the chosen clustering algorithm. The idea is re-sampling or perturbing the original dataset and re-clustering these data. The consistency of the corresponding results provides an estimate of the significance of the clusters obtained from the original dataset.

Breckenridge (1989), Fridlyand & Dudoit (2001), Lange et al. (2004), Tibshirani & Walther (2005) take this kind of approach: the data are repeatedly split into a training and a test set (typically of equal sizes and with no overlap), and both sets are clustered. The training set is then employed to derive a classifier to predict all class labels for the test set. The cluster results on the test set are then compared with the predicted class label of the classifier using a binary index of agreement. We have to note that the classifier used for prediction has a significant impact on the performance of this method. Lange et al. (2004) recommend the use of a nearest-neighbor classifier for the single link, and of centroid-based classifiers for algorithms such as k-means that assume spherically shaped clusters.

The stability of a clustering result can also be assessed by comparing the partitions obtained from perturbed data (Bittner et al. 2000, Kerr & Churchill 2001, Li & Wong 2001). For this purpose, a number of bootstrap datasets are generated from the original data: a noise component is added to each data item using a simple error model (Bittner et al. 2000) or more advanced methods such as ANOVA (Kerr & Churchill 2001). The resulting datasets are subjected to a cluster analysis and the partitioning results of different datasets are compared using external binary indexes.

Yeung et al. (2001) use the jackknife approach to validate clustering from gene expression data, and define an index which they call “Figure of Merit”.

Type VI: Compliance between a partitioning and distance information. Methods in this category are based on the cophenetic matrix C: a symmetric binary matrix of size $N \times N$ where N is the size of the dataset, the element $C(i,j)$ is equal to one, only if the unit i and the unit j have been assigned to the same cluster, while for hierarchical clustering, $C(i,j)$ represents the continuous value of the level within the dendrogram in which the two data units i and j were first assigned in the same cluster. The cophenetic matrix is compared to the original dissimilarity matrix with a measure of correlation (Halkidi et al. 2001) like the Pearson or the Spearman rank correlation.

1.2.2 External validation measures

External measures take a known set of class labels (the “gold standard”) and compare the cluster results with the known labels to assess the degree of consensus between the two partitions. Data mining literature provides a high number of indexes based on the contingency table built between the true classes and the estimated one.

Given a set of n objects $X = \{X_1, \ldots, X_n\}$, suppose $U = \{u_1, \ldots, u_R\}$ and $V = \{v_1, \ldots, v_P\}$ represent two different partitions of X such that $u_i \subset X$, $v_j \subset X$ and $\bigcup_{i=1}^{R} u_i = X = \bigcup_{j=1}^{P} v_j$ and $u_i \cap u_{i'} = \emptyset = v_j \cap v_{j'}$ for $1 \leq i \neq i' \leq R$ and $1 \leq j \neq j' \leq P$. Suppose U is the clustering resulting partition and V the gold standard partition, let a be the number of pairs of elements in the dataset that are in the same subset in U and in the same subset in V; b be the number of pairs of elements in the dataset that are in different subsets in U and in different subsets in V; c be the number of pairs of elements in the dataset that are in the same subset in U and in different subsets in V; d be the number of pairs of elements in the dataset that are in different subsets in U and in the same subset in V. The quantities a and b can be interpreted as agreements, while c and d as disagreements. The Rand index (Rand 1971) R is

$$R = \frac{a+b}{a+b+c+d}. \tag{1.4}$$

The Rand index has a value between 0 and 1, with 0 indicating that the two data clusterings do not agree on any pair of points and 1 indicating that the data clusterings are exactly the same.

The problem with the Rand index is that the expected value of the Rand index of two random partitions does not have a constant value on average and depends on the number of units . Hubert & Arabie (1985) propose an adjusted form of the Rand index assuming the generalized hypergeometric distribution as the model of randomness. Considering a contingency table in comparing the two partitions U and V, let n_{ij} be the number of objects that are in both class u_i and class v_j; then let $n_{i.}$ and $n_{.j}$ be the number of objects in class u_i and in class v_j respectively, the Adjusted Rand index (ARI) is then defined as follows:

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2} \sum_i \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_i \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2} \sum_i \binom{n_{.j}}{2}\right] / \binom{n}{2}}. \tag{1.5}$$

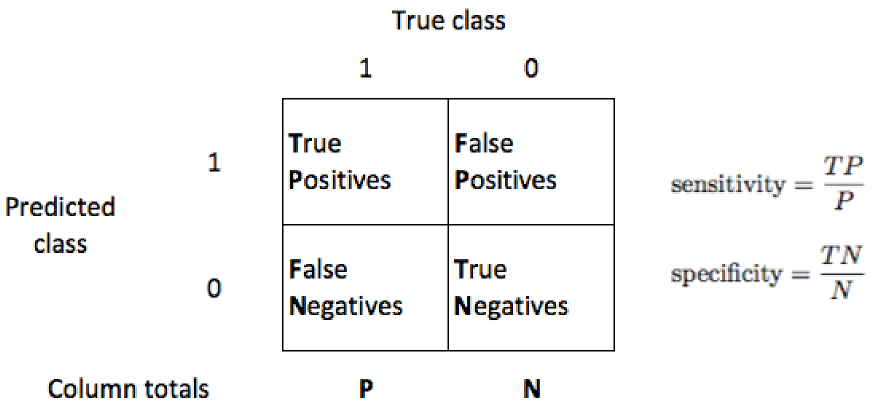
Milligan & Cooper (1986) evaluate many different indexes in measuring the agreement between two partitions in clustering analysis with different numbers of clusters, and they recommended the adjusted Rand index as the index of choice. Another index is the Jaccard coefficient (Jaccard 1908), which can be defined using the above notation of the Rand Index:

$$J = \frac{a}{a+b+c}. \tag{1.6}$$

Hubert & Arabie (1985) and Halkidi & Vazirgiannis (2001) highlight that external validation techniques suffer from biases with respect to the number of clusters, the distribution of resulting cluster sizes and the distribution of true class sizes in a partition: for a one-cluster partition, indexes obtain maximum possible value and tend to decrease with an increasing number of clusters. Moreover, the Rand Index tends to be overly optimistic in situations where relatively small clusters have been overlooked.

1.3 Supervised classification validation techniques

Let $\mathcal{D} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ be a dataset of n labelled observations with labels $y_i \in \mathcal{Y} = \{1, \ldots, G\}$, a classifier $\mathcal{L}$ maps a rule to assign an unlabelled observation to a label in $\mathcal{Y}$. The most used misclassification measure in a supervised problem is the 0/1 loss function which is related to the probability of correctly classifying a new observation called *accuracy* of a classifier $\mathcal{L}$ (Kohavi et al. 1995). The accuracy index is then estimated as the number of the observations correctly classified over the total number of observations. In medical classification problems, the *Sensitivity* and *Specificity* indexes are used to characterize a rule for the disease (Friedman et al. 2009). These indexes can be defined by elements of a confusion matrix. Fawcett (2006) shows a confusion matrix and use it to calculate several common metrics by considering classification problems using only two classes. Following this idea, we can define a confusion matrix in Figure 1.1 for each class in $\mathcal{Y}$ when the problem has more than two classes: the value 1 indicates that the observations belong to the class and otherwise the value 0 is assigned.



		True class	
		1	0
Predicted class	1	True Positives	False Positives
	0	False Negatives	True Negatives
Column totals		P	N

$$\text{sensitivity} = \frac{TP}{P}$$

$$\text{specificity} = \frac{TN}{N}$$

Figure 1.1. Confusion matrix for a classification problem, given a specified class. Definition of Sensitivity and Specificity indexes for a class.

Estimating the accuracy of a classifier $\mathcal{L}$ induced by supervised learning algorithms is important, not only to assess its future prediction accuracy but also for model selection (Wolpert 1992). Typically supervised classification algorithms have one or more tuning parameters, for example, the number of neighbors in a k-nearest neighbor classifier or the lasso/ridge coefficients in a lasso/ridge regression. The idea is to choose the optimal parameters to minimize the predictive error criterion, but at the same time, we want to evaluate the predictive performance of the model in assessing how the results of a supervised classifier could be generalized to an independent data set. As suggested by Friedman et al. (2009) if we have a large number of observations, the best approach for both problems is to randomly divide the dataset into three parts: a training set, a validation set, and a test set. This is the *holdout method*, the simplest kind of cross-validation. The training set is used to fit the models; the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final chosen model. Since data are often scarce, this is usually not possible.

K-fold cross-validation is one way to improve the holdout method. The data set is randomly divided into k subsets; each time, one of the k subsets is used as the test set and the other $k - 1$ subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that how the data gets divided is less important. Every observation is included in a test set only once, while $k - 1$ times it is used in a training set. The disadvantage of this method is that it is computationally heavy because the algorithm is run k times. When $k = n$, where n is the number of observations in the dataset, we have the *leave-one-out cross-validation.* Each observation, in turn, is left out, and the learning method is trained on all the remaining units. Witten et al. (2016) highlight two reasons to use this method: using all the possible data for the training set presumably increases the chance that the classifier is an accurate one and the procedure is deterministic since no random sampling is involved.

Stone (1974) introduces the *double cross-validation approach*: each observation is removed in turn from the data, and in the remaining training set another cross-validation step is done to recalibrate the classifier $\mathcal{L}$. The resulting classification rule is then applied to the left-out observation to obtain an unbiased allocation of this sample. The procedure is then repeated for all observations, after which misclassification rates are calculated using the results for all the left-out observations. Even if this method is computationally heavy, we are able to combine predictive optimization and predictive unbiased validation in the same procedure, without loss of data (Mertens et al. 2006). So it is a good approach for datasets with a small number of observations. Mertens (1998, 2001, 2003) gives further details on the computational burden.

1.4 A motivating example

Our work is motivated by an experimental study on scrapie, a prion disease, conducted at the Istituto Superiore di Sanità by experimental transmission of scrapie's diseases to bank voles.

Scrapie of small ruminants is caused by prions, which are unconventional infectious agents of proteinaceous nature affecting humans and animals. Scrapie is historically known to occur in different strains (Bruce et al. 2002), a fact which has renewed interest due to the zoonotic risk posed by animal prions (Cassard et al. 2014). Due to the absence of nucleic acids, which precludes direct analysis of strain variation by molecular methods, the presence of different sheep scrapie strains is usually investigated by bioassay in laboratory rodents. Mice and bank voles are the animal models of the election. Bioassay studies are performed inoculating a brain homogenate of sheep scrapie intracerebrally into the recipient rodent. Complete isolation, stabilization, and characterization of a scrapie agent requires serial passages of a sheep scrapie sample, hereafter referred to as *isolate*, in congenic mouse or vole lines (Bruce et al. 2002, Di Bari et al. 2012). Within this experimental setting, scrapie strain characterization is routinely performed evaluating, at least, the survival time and the spongiform change (as vacuolization ordinal scores in specific brain areas) of each injected animal, and taking the average of collected variables from all animals inoculated with the same sheep scrapie isolate.

A large experimental study on strain characterization in bank vole model of 32 sheep scrapie isolates from several European countries has been carried out at Istituto Superiore di Sanità (Italian National Institute of Health). The resulting experimental data can be conceived as multilevel data with multiple bank voles (lower-level units) inoculated with the same sheep scrapie isolate (higher-level unit). Both continuous (survival time) and ordinal (vacuolation scores of brain areas) variables are assessed for each experimental animal.

The general scientific issue is investigating how the available experimental data can be used to evaluate the appropriateness of an isolate partition, identifying strains or groups of strains. This is done by taking into account a preliminary benchmark partition proposed by researchers. Our aim is then to validate a final, possibly finer, partition that could be used as “gold standard” in future studies.

Remark 1. *We believe that this is a relevant scientific question for which there is not an off-the-shelf statistical methodology. Indeed unsupervised classification is used to find an unknown partition of units, while supervised classification algorithms are used, starting from a well-known or objectively defined partition, in order to find the* *relationship between the classes and the data. Researchers' prior knowledge has a key role, not only to understand the statistical results, but to also set up a statistical strategy to answer their scientific questions.*

1.5 A procedure to validate an a priori partition: a proposal

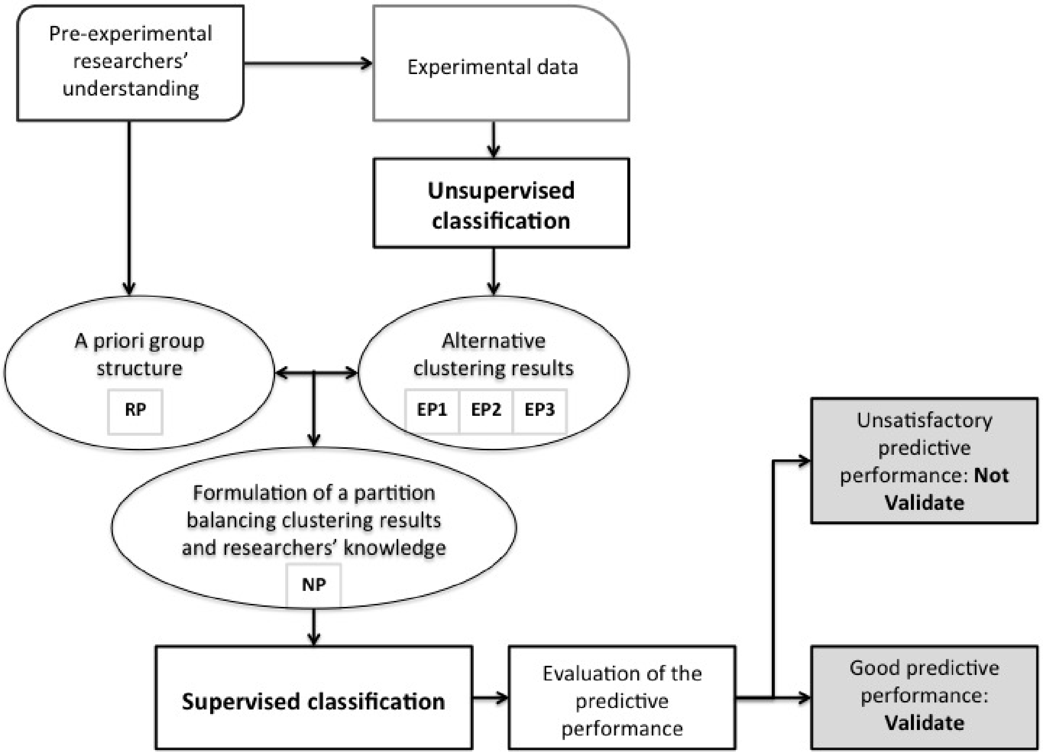

Pre-experimental researchers' understanding
Experimental data
Unsupervised classification
A priori group structure
RP
Alternative clustering results
EP1
EP2
EP3
Formulation of a partition balancing clustering results and researchers' knowledge
NP
Supervised classification
Evaluation of the predictive performance
Unsatisfactory predictive performance: Not Validate
Good predictive performance: Validate

Figure 1.2. Diagram to illustrate the procedure proposed in this work.

Based on their pre-experimental understanding, researchers formulate a preliminary partition (RP) of isolates. The attribution of each isolate to the corresponding group in the partition is not based on direct evidence. Cluster analysis is used in the attempt to find evidence in support of the partition using the experimental data, but it can also lead to new interesting results. Hence the proposed procedure may provide a formulation of a new candidate partition (denoted with NP) taking into account alternative clustering results (generically denoted with the acronym EP) and RP. EP results are chosen using internal clustering validation methods and are compared with RP using external validation clustering measures. NP does not necessarily reflect the best EP outcome, instead, it is formulated by researchers accounting for both RP and EP.

The new partition is then used to define labels for the isolates and a supervised classification step is performed using these labels.

In order to validate the NP, an evaluation of the predictive performance of several models ensues, crucially, with a cross-validation that properly accounts for the data structure and avoids overfitting.

Figure 1.2 outlines the procedure: the rectangles indicate the steps using statistical tools adapted to deal with multilevel mixed-type data.

1.6 Why not just use supervised classification?

Since the aim of the applied problem is predicting new further discovered isolates within the strains, the question could be why not just use a supervised classification to validate the researchers' partition? We discuss this question using a simple example composed of two similar simulation studies within the same framework: we simulate $N = 200$ samples with $n = 300$ observations from two i.i.d. $Uniform(0, 1000)$, then we fix arbitrary thresholds to create 5 classes. In the *Simulation 1* the groups are defined by the thresholds but they are not well separated, while in the *Simulation 2* we use hard thresholds to better separate the groups.

Firstly, we perform supervised classification on the simulated samples. We use 4 different supervised classification algorithms for the supervised classification: the Gradient Boosting Method (GBM) and the Random Forest (RF) as classification tree-based methods; the Generalized Linear Elastic Net regression (GLMNET) and the K-Nearest Neighbor (KNN). We split the original dataset into a training set (75% of the observations) and in a testing set (25%); then use a 10-fold cross-validation within the training set to calibrate the algorithms. Therefore we use the testing set to predict the classes using the best models found within the cross-validation. On the test set, we measure the indexes introduced in Section 1.3.

Finally, we use unsupervised classification on the same samples. We use distance-based methods for clustering: hierarchical clustering with an average linkage, and two partitioning algorithms, the K-means and the Partition Around Medoids (PAM). For each dataset, we choose the best clustering results using three different validation

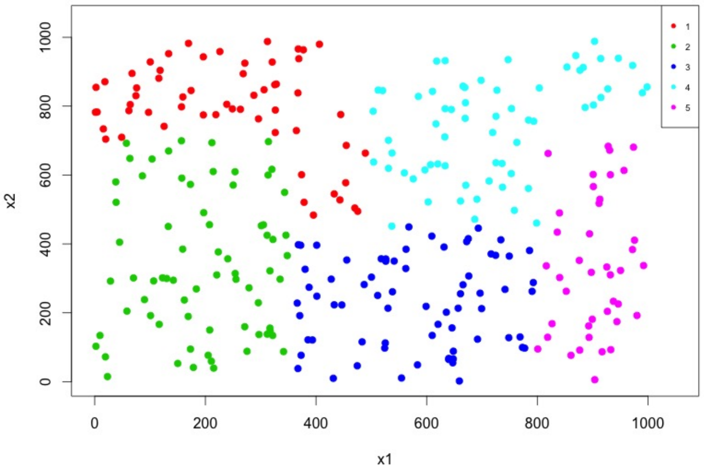


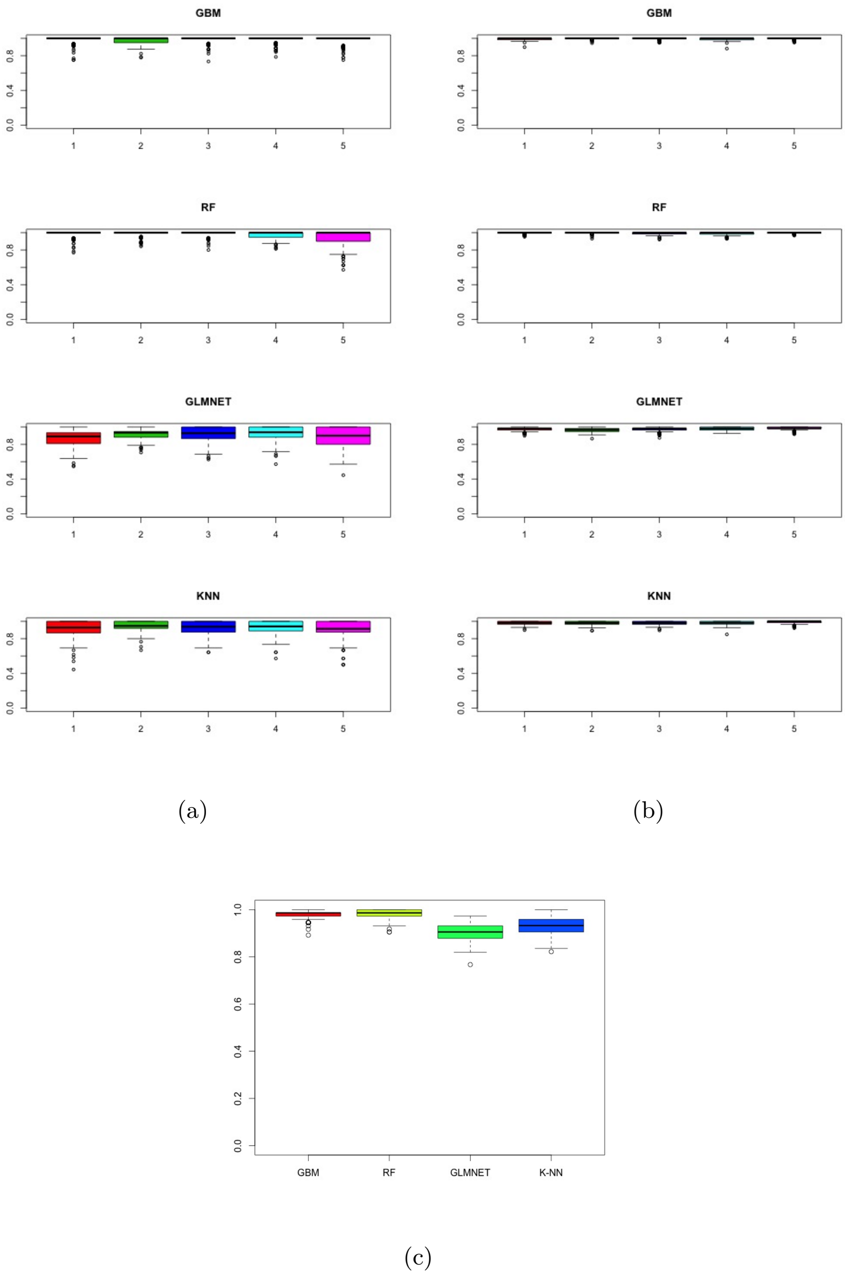
Figure 1.3. *Simulation 1* example: two variables generated from two i.i.d. uniform distributions. The observations are assigned to 5 classes created with arbitrary thresholds.

indexes: the Connectivity index (Equation 1.1 with $L = 5$), the Dunn Index (Equation 1.2) and the Silhouette index (Equation 1.3). The highest index value allows us to select the best method combined with the best number of groups for each dataset and for each index.

Simulation 1: not separated groups. Figure 1.3 shows one simulated sample with the assigned classes. In Figure 1.4, we report all the supervised validation indexes measured over the 200 samples: both the Sensitivity and the Specificity indexes in all the classes are very high for the two tree-based algorithms (from 0.8 to 1.0). The sensitivity index distributions of GLMNET and KNN are more variable, even if the distributions are concentrated over 0.8, there are samples with lower values in at least one group, while the Specificity indexes are always over 0.85. The Accuracy indexes range from 0.8 to 1.0 for all the samples even if the GLMNET and the KNN have more variable distributions than the GBM and the RF, which show very concentrated values around 1.0. We can conclude that the supervised algorithm will predict further new observations in those classes very well even though the underlying classes are not well separated. Note that the samples did not have a structured statistical distribution and the groups' definition is based on arbitrary thresholds. The supervised classification algorithms would be powerful if the classes were exogenously recognized as “gold standard”. Given they are not in this example, we are not able to conclude that the supervised classification on these data provide empirical evidence in support of the classes.

Therefore we rely also on cluster analysis to find if the classes are consistent with the data. Table 1.1 summarizes the best choices: the Connectivity index, based on the 5 nearest-neighbor observations, present maximum values only for a number of groups equal to 2. The Dunn index mainly allows choosing the hierarchical method with the highest number of groups that we fix at 8. The Silhouette index favors the partition algorithms. These contradictory and heterogeneous results hint that data do not have a well-defined partition structure.

Finally, the best clustering results are compared to the partition considered in the supervised classification (Figure 1.5) using the Rand Index (Equation 1.4), the Adjusted Rand Index (Equation 1.5) and the Jaccard index (Equation 1.6). Only the Rand index (which is optimistic as discussed in Section 1.2.2) has high values but the clustering results do not seem to agree with the defined partition.



(a)

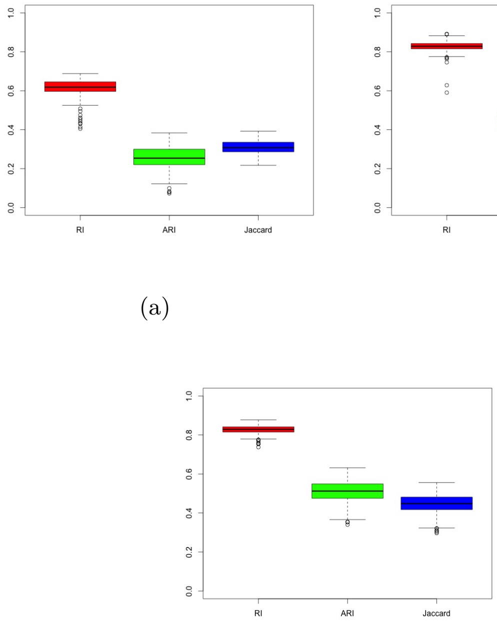
(b)

(c)

Figure 1.4. Supervised classification validation indexes for the *Simulation 1* on the 200 simulated samples using the 4 supervised algorithms for the 5 classes: (a) Sensitivity index distributions (b) Specificity index distributions (c) Accuracy index distributions.

Index	Number of groups	Methods		
		Hierarchical	*K-means*	*PAM*
Connectivity	2	180	12	8
Dunn	2	2	0	0
	3	0	0	0
	4	5	1	2
	5	8	1	0
	6	18	0	2
	7	47	2	0
	8	110	2	0
Silhouette	2	0	0	0
	3	0	9	5
	4	0	120	39
	5	0	6	1
	6	0	5	2
	7	0	4	2
	8	0	3	4

Table 1.1. Cluster analysis: internal validation indexes for the *Simulation 1* example.



(a) (b)

(c)

Figure 1.5. External validation indexes (RI: Rand Index; ARI: Adjusted Rand Index; Jaccard: Jaccard index) on the best clustering results for the *Simulation 1*. The clustering results are selected by the (a) Connectivity index (b) Dunn index (c) Silhouette index.

Simulation 2: well-separated groups We repeat the simulation within the same framework, but the classes are now well separated as shown in Figure 1.6. Now the partition is well defined and we expect to have better results than the first simulation in term of validation indexes.

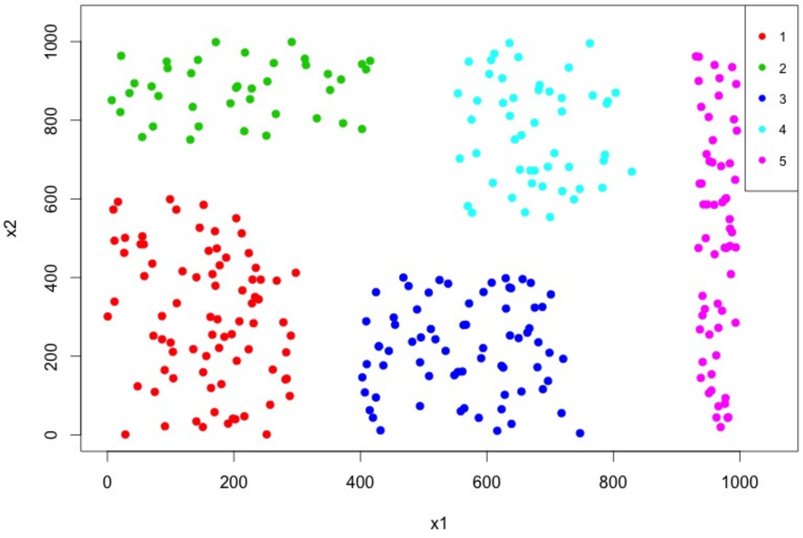


Figure 1.6. *Simulation 2* example: two variables generated from two i.i.d. uniform distributions. The observations are assigned to 5 classes created with arbitrary thresholds and the groups are well separated.

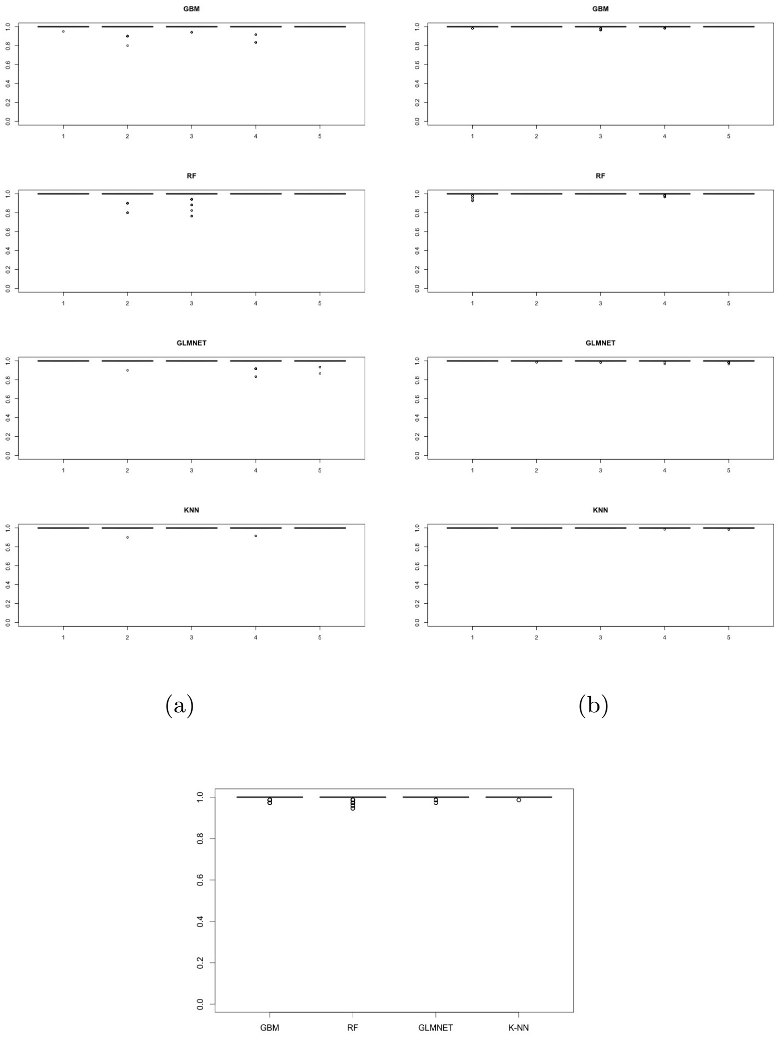
Figure 1.7 shows that all the supervised classification algorithms have a good performance with all the indexes concentrated around 1.

The unsupervised classification results are reported in Table 1.2 and the best choices are very heterogeneous: the Connectivity index select always a number of groups equal to 2; the Dunn index and the Silhouette index select different best combination of methods and number of groups. Regarding to the number of groups, the Silhouette index in most cases chooses the right number of groups (5), while the Dunn index allows choosing the highest number of groups (8) in most cases. Even if the founded groups are not the same, the external validation indexes on the best partitions in Figure 1.8 show a good agreement between the resulting and the initial partitions. So, even if we use a thresholds-based rule as in *Simulation 1*, the well-separated groups are detected by unsupervised and supervised classification.

From this example we highlight some peculiarity from the validation indexes: the

Connectivity index has the highest values always with a number of groups equal to 2, while the Dunn index in most cases has the highest values with the highest number of groups. The Silhouette index is better calibrated than the other internal indexes, even if this index fits better with the partition algorithms than the hierarchical ones. Regarding the external validation indexes, we highlight that the Rand Index is more optimistic than the Adjusted Rand Index and the Jaccard index. These two indexes also properly account for the different number of groups in the two partitions.

Using this toy example, we can conclude that supervised classification algorithms find the relation between data and groups even if no distributional rules are used to define the classes. The predictive performances are really good when the groups are well separated and the supervised classification allows to find the relation between the defined groups and the covariates. Therefore, if the aim of the statistical analysis is finding empirical evidence in support of a certain partition, we also need to rely on clustering algorithms. This example also shows that in partition analysis, we have to compare several validation indexes to understand what we can learn from the data. Sometimes the validation indexes involve contradictory results and the evaluation of these results have to be contextualized to the dataset and the problems. Therefore we can conclude that in applied problems, results must be validated by experts, particularly when statistical analysis provides heterogeneous results.



(a)

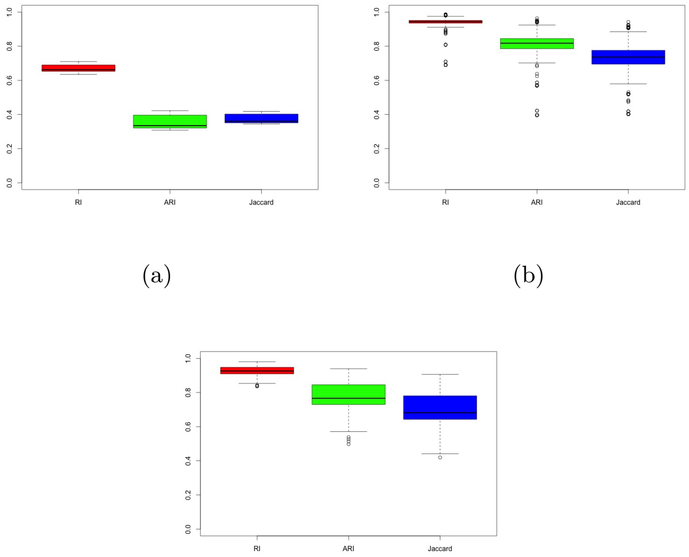
(b)

(c)

Figure 1.7. Supervised classification validation indexes for the *Simulation 2* on the 200 simulated samples using the 4 supervised algorithms for the 5 separated classes: (a) Sensitivity index distributions (b) Specificity index distributions (c) Accuracy index distributions.

Index	Number of groups	Methods		
		Hierarchical	*K-means*	*PAM*
Connectivity	2	169	17	14
Dunn	2	18	0	0
	3	10	0	0
	4	1	2	0
	5	35	4	1
	6	17	2	0
	7	40	1	2
	8	59	8	0
Silhouette	2	0	0	0
	3	0	0	0
	4	18	3	8
	5	35	40	11
	6	7	39	12
	7	1	6	1
	8	0	16	3

Table 1.2. Cluster analysis: internal validation indexes for the *Simulation 2* example with separated groups.



(a)

(b)

(c)

Figure 1.8. External validation indexes (RI: Rand Index; ARI: Adjusted Rand Index; Jaccard: Jaccard index) on the best clustering results for the *Simulation 2*. The clustering results are selected by the (a) Connectivity index (b) Dunn index (c) Silhouette index.

Chapter 2

Scrapie dataset and multilevel mixed-type data

2.1 Data description

Sheep scrapie strains can be characterized by experimental transmission in laboratory rodents, by evaluating the strain-specific differences in survival time, neuropathology, type and distribution of misfolded prion proteins in the brain. Transmission studies are performed by inoculating laboratory rodents with brain homogenates from affected sheep. These experiments, usually referred to as primary transmissions, typically result in long and variable incubation times, because the infecting prion needs to cross a so-called species barrier, (it derives from sheep and needs to infect rodents). Subsequent sub-passages in congenic laboratory rodents result in shortening and stabilizing of the incubation time, a phenomenon referred to as adaptation, during which a given prion agent adapts to the new species. Therefore, full isolation of prion agents requires serial passages of natural isolates, which also leads to the stabilization of the neuropathological phenotype in the new species.

While laboratory mice have historically been used in studies of prion diversity, they are not fully susceptible to natural scrapie isolates. Bank voles have been shown to be a species remarkably susceptible to different prion sources (Nonno et al. 2006, Agrimi et al. 2008, Di Bari et al. 2013, Pirisinu et al. 2016) and have been exploited to isolate and discriminate different scrapie strains (Di Bari et al. 2008). Strain characterization was performed using 32 selected sheep scrapie isolates from seven European countries. A variable number of bank voles were inoculated with each isolate. After two subsequent vole-to-vole sub-passages (third passage), survival times and brain lesion profiles were registered for every individual vole. When different pathological phenotypes were observed in animals inoculated with the same inoculum at first passage, more than one second passage was initiated. This was the case for 4 natural isolates so that 36 second passages were performed from the initial 32 primary transmissions.

Due to these reasons we use the stabilized data of the third passage. The experimental data has a multilevel structure: the higher-level units, indexed by $j = 1, \ldots, 36$, are the isolates; the lower-level observations, indexed by $i = 1, \ldots, n_j$, are the bank voles. The number of bank voles in each isolate n_j (also reported in Figure 2.1) varies from 3 to 11 to better characterize the unbalanced clustered data. Note that $\sum_{j=1}^{36} n_j = 279$. For each lower-level observation the mixed-type outcome variables are $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp}, \ldots, x_{ijP})$ where $p = 1$ corresponds to the survival time measured in days (continuous), while $p = 2, \ldots, 10$ correspond to the nine lesion profile ordinal variables. In the supervised classification, we will also use the class label $y_{ij} = 1, \ldots, g, \ldots, G$ defined by a putative partition.

The individual lesion profile is a method (Fraser & Dickinson 1968, Di Bari et al. 2012) in which vacuolar change is assessed by assigning a non-negative integer score from 0 to 5, depending on the severity of vacuolation, in nine brain areas, including (1) medulla, (2) cerebellum, (3) superior colliculus, (4) hypothalamus, (5) thalamus, (6) hippocampus, (7) septum, (8) retrosplenial and adjacent motor cortex, and (9) cingulate and adjacent motor cortex.

Variability is observed both within and between scrapie isolates. The violin plots of survival times for each isolate in Figure 2.1 show asymmetric distributions with heterogeneous degrees of variability and tendency. Only 10 isolates (28.2%) have an average survival time greater than 100. Lesion profiles also differ between and within scrapie isolates: differences may be in modal values or in distributional frequencies, in one or more brain areas. A partial representation of heterogeneity of lesion severity profiles within isolates is displayed in Figure 2.2. A remarkable heterogeneity between isolates can be seen for It4, Uk4 and Fr4-A (Figure 2.2(a), 2.2(c) and 2.2(d)). On the other hand, many other pairs of isolates are more homogeneous as is clearly the case for It4 and Uk14 (Figure 2.2(a) vs 2.2(b)).

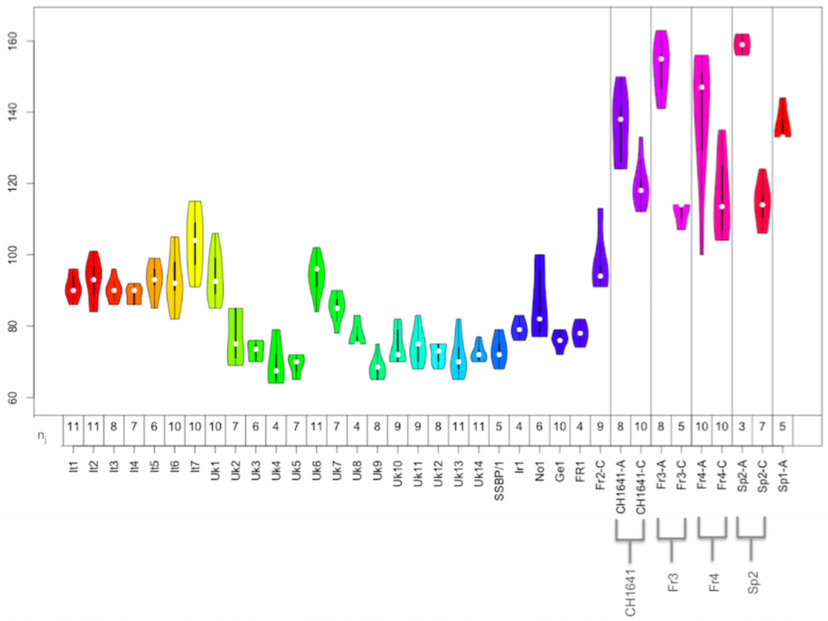
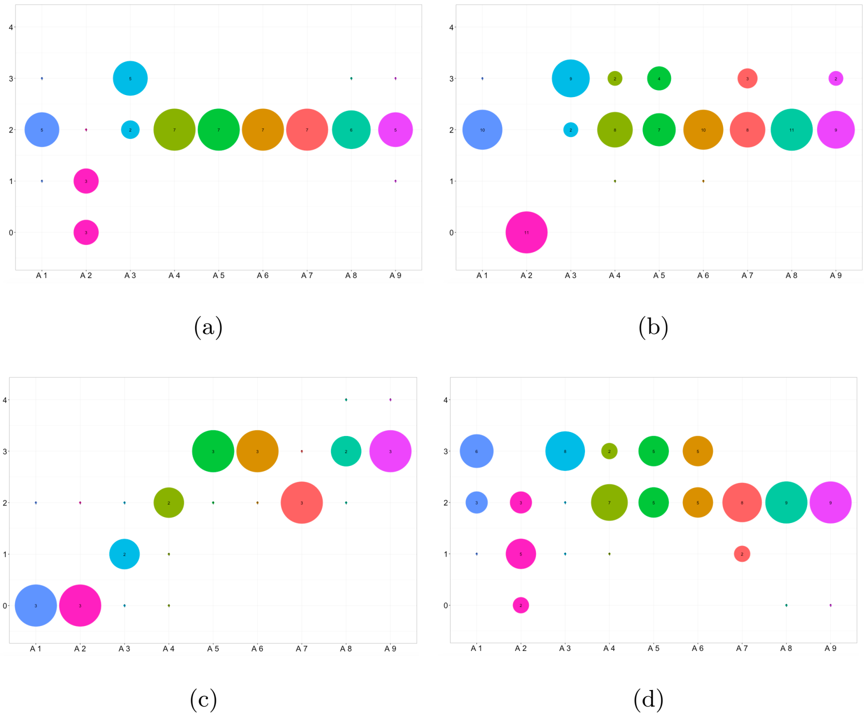


Figure 2.1. Violin plots for Survival Time for each experimental isolate. n_j are reported for each isolate. The 4 inocula, split into 2 isolates are highlighted in the labels.

2.2 Multilevel data

Many kinds of data, including experimental data in biological sciences and observational data in social sciences, have a hierarchical, nested, or clustered structure (Goldstein 2011). Generally, we refer to these data structures as multilevel data, because each observation belongs to more than a single level. For example, individuals from the same family, or students in the same school tend to show similar behaviors influenced by the same environment. Multilevel data structures also exist in medical studies like e.g. when measuring cell belonging to the same patients or different patients exposed to the same treatment but in different hospitals. Our data have a multilevel structure because the same isolate is inoculated into multiple bank voles. We are interested in the higher-level unit, but data can not be obtained for the isolate because it is not directly observable. The phenotype is thus indirectly measured by only inoculating multiple bank voles, our lower-level units, with the same isolate. The lower-level observations are then correlated within the same isolate.

Latent variable models are widely used for the analysis of multilevel data especially by using means of mixed effects models. A linear model that incorporates both fixed effects (for the covariates $\mathbf{x}_{ij}$) and random effects u_j can be used in a two-level



(a)

(b)

(c)

(d)

Figure 2.2. Distributional plots for lesion profile in 9 brain areas (A1-A9) for isolates (a) It4 (b) Uk14 (c) Uk4 (d) Fr4-A. A1=medulla, A2=cerebellum, A3=superior colliculus, A4=hypothalamus, A5=thalamus, A6=hippocampus, A7=septum, A8=retrosplenial and adjacent motor cortex, A9=cingulate and adjacent motor cortex.

framework to model the dependent variable y_{ij}. u_j are latent variables used to account for the heterogeneity of lower-level observations within the same higher-level unit. A mixed effects model can be represented as

$$y_{ij} = \alpha + \beta \mathbf{x}_{ij} + u_j + \epsilon_{ij} \tag{2.1}$$

where

$$\begin{aligned} u_j &\sim N(0, \sigma_u^2) \qquad i.i.d \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2) \qquad i.i.d \end{aligned}$$

are independent from each other.

The increasing popularity of mixed effects models is explained by the flexibility they offer in modeling the within-group correlation often present in grouped data and by the availability of reliable and efficient software for fitting them (Fox 2002, Gelman & Hill 2006).

An alternative way to model the multilevel data is analyzing distribution functions or probability density functions (pdf). Calò et al. (2014) shows in Figure 2.3 how the two-level dataset has an intrinsic two-level hierarchical structure, in which pdf-objects correspond to higher-level units and the random realizations from the different pdfs represent lower-level units. Some examples of the use of this approach are the study of age distributions across countries in a given year (Delicado 2011), the characterization of computer images by the distribution of the respective gray-scale pixel values (Spellman et al. 2005), the comparison of the distributions of collagen fibril diameters observed in different mice strains (Chervoneva et al. 2012), and the idea of performing customer segmentation after each customer has been represented by the distribution of the item unit price across his purchases (Sakurai et al. 2008). In this context, another way to look at the two-level data is considering histogram data (Irpino & Romano 2007). Histogram data are introduced in the context of Symbolic Data Analysis by Bock & Diday (2012) and they are defined by a set of contiguous intervals of the real domain, which represent the support of each histogram, with an associated a system of weights (frequencies, densities).

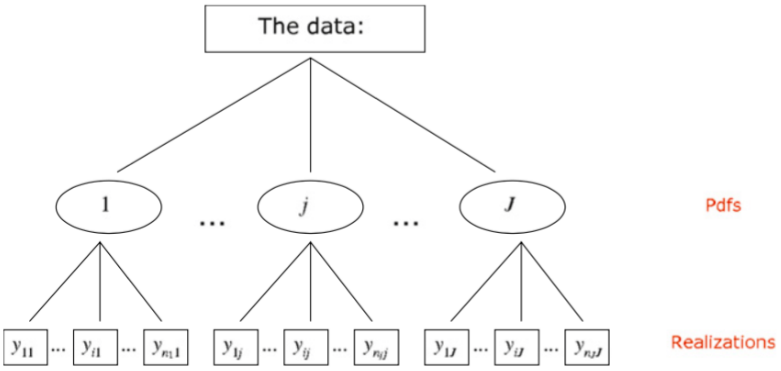

The data:
1
...
j
...
J
Pdfs
y_{11} ... y_{i1} ... $y_{n_1 1}$
y_{1j} ... y_{ij} ... $y_{n_j j}$
y_{1J} ... y_{iJ} ... $y_{n_J J}$
Realizations

Figure 2.3. Hierarchical structure of a two-level dataset (Calò et al. 2014)

Several approaches are available for clustering multilevel data, both using latent variable models and distributional analysis. These models are built to manage continuous data. To the best of our knowledge, we have not found an approach to clustering multilevel ordinal or mixed-type data. Learning from data approaches with distance-based methods are less common (Yeung et al. 2003), but they can be used more easily with any data type because they do not require the distributional elicitation of suitable mixture components.

Supervised classification of multilevel data has gained increasing importance in many applied problems, but it has not received the same attention in the statistical literature as cluster analysis for multilevel datasets or multilevel modeling.

2.3 Mixed-type data

Mixed-type data modeling has received increasing attention over time, especially in model-based clustering approaches. Latent variable models are widely used. Consider a vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ih}, \ldots, x_{iH})$ of mixed-type variables, in latent variable models $\mathbf{x}_i$ are a manifestation of an underlying latent continuous vector, $\mathbf{z}_i$ (for $i = 1, \ldots, N$). The basic idea is to consider each group of variable type separately and then modeling the joint distribution of the latent $\mathbf{z}_i$. Arminger & Küsters (1988) assume $\mathbf{z}_i$ to have multivariate Gaussian distribution with expectation potentially dependent on covariates. Sammel et al. (1997) assume that each manifest variable follows a one-parameter exponential family model, and they use a generalized linear model to connect the manifest variables to fixed covariates and a specific continuous latent variable. Dunson (2000) generalizes the framework of Arminger & Küsters (1988) to accommodate non-normal latent variables and clustered data. The author also manages non-linear relationships between the underlying and latent variables, multiple latent variables for each outcome type and covariate-dependent modifications of the relationship between the latent and underlying variables. The joint distribution of the underlying variables is described by a mixture of generalized linear models to detect heterogeneity. Furthermore, a distinct link function is used for each underlying variable.

Focusing on the modeling of each type of variable separately, two latent variable approaches are used to model ordinal variables in statistical literature: Item Response Theory (IRT) and Underlying Response Variable (URV). IRT was first introduced by Thurstone (1925) and is widely used in psychological and educational tests. Extensions are given by Rasch (1960) with the Rasch model for binary outcomes, while for data with K ordinal responses Samejima (1969) proposes the graded response model, and Masters (1982) introduces the partial credit model. IRT assumes that each observed ordinal response is a manifestation of a latent continuous variable. IRT is based on the local independence assumption on variables x_{ih}, where i denotes the individual and h is the item. The dependence between the manifest variables x_{ih} is explained by the ability factor θ_i using a generic link function f that specifies the link between the ordinal variables and the quantitative unobserved variables (Equation 2.2):

$$f(x_{ih}) = \lambda_h \theta_i - b_h. \tag{2.2}$$

The latent trait θ_i can be unidimensional or multidimensional and it allows a dimensionality reduction.

The URV approach was first introduced by Muthén (1984) within the Structural Equation Modeling (SEM) framework and developed by Jöreskog (1990) and Lee et al. (1990). The idea of URV is to consider ordinal variables $x_{ih} = \{0, \ldots, k, \ldots K\}$ as a categorization of the underlying continuous variables z_{ih} through a threshold model (Equation 2.3). For each item a vector of threshold parameters $\gamma_h = (\gamma_{h,0}, \ldots, \gamma_{h,k}, \ldots \gamma_{h,K_h})$ exists. This vector is subject to the constraint:

$$-\infty = \gamma_{h,0} \leq \gamma_{h,1} \leq \ldots \leq \ldots \gamma_{h,K_h} = +\infty,$$

and the relation between the observed variables x_{ih} and the latent z_{ih} is:

$$\gamma_{h,k-1} < z_{ih} \leq \gamma_{h,k} \qquad \Longleftrightarrow \qquad x_{ih} = k. \tag{2.3}$$

Models are therefore developed for the latent variables, for example assuming for $(z_{i1}, \ldots z_{ih}, \ldots z_{iH})$ a multivariate Gaussian distribution. The thresholds $\gamma_{h,k}$ can be arbitrarily fixed or they have to be estimated. The URV thresholds method is also used as link function between x_{ih} and z_{ih} in Equation 2.2 (McParland & Gormley 2013), so that the IRT model becomes $z_{ih} = \lambda_h \theta_i - b_h$.

Also binary data are modeled as ordinal variables, with one unique threshold generally fixed at 0:

$$x_{ih} = \begin{cases} 0 & \text{if} \quad z_{ih} < 0 \\ 1 & \text{if} \quad z_{ih} \geq 0. \end{cases} \tag{2.4}$$

Nominal variables are categorical variables with unordered responses. For each nominal variable x_{ih} with K_h possible responses, we need $K_h - 1$ underlying continuous variables z_{ih} i.e. $\mathbf{z}_{ih} = (z_{ih}^1, \ldots, z_{ih}^{K_h - 1})$. The observed nominal response x_{ih} is related to the z_{ih}^s as follows:

$$x_{ih} = \begin{cases} 1 & \text{if } \max \{z_{ih}^s\} \\ k & \text{if} \quad z_{ih}^{k-1} = \max_s \{z_{ih}^s\} \quad \text{and} \quad z_{ih}^{k-1} > 0 \quad \text{for} \quad s = 2, \ldots, K_h. \end{cases} \tag{2.5}$$

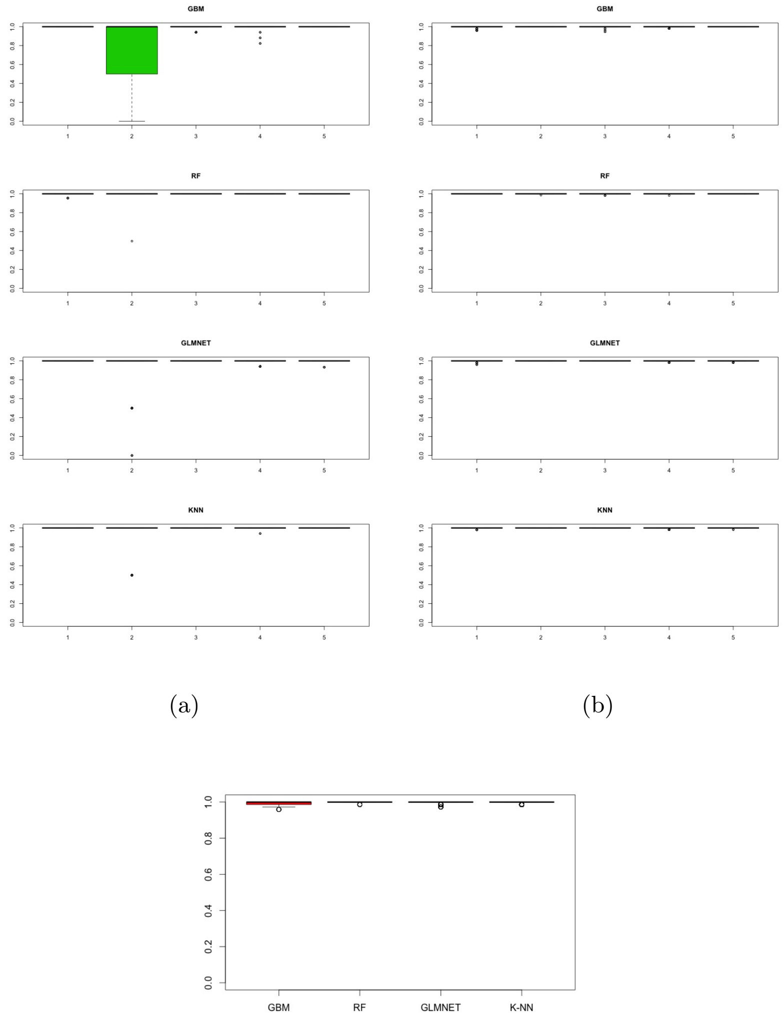
Let X denote a data matrix with N rows and P columns. Suppose that the continuous variables are in the first C columns, the ordinal and binary variables are in the following O columns and the nominal variables are in the final $P-(C+O)$ columns. We associate at X the matrix Z where the first C columns are equal to the same of X and the other columns are built with the latent variables. Finally, we have to model the joint distribution for Z.

2.4 Addressing some of the dataset's limits

Even if the multilevel mixed-type data modeling is a methodological challenge, we have to deal with some limits due to the dataset. The dataset has unbalanced sample sizes between classes and also within isolates (Table 2.1). The lowest number of bank voles within an isolate is equal to 3, whereas the smallest group is SSBP/1like for which the phenotypes are measured on 9 bank voles nested in 2 isolates. It is not always easy to detect a small group within a supervised classification analysis because of the splitting into the training and the testing sets. For example, we perform the supervised classification for the *Simulation 2* presented in the Section 1.6 forcing the group 2 to be "small" with 10 units. Figure 2.4 shows the resulting supervised classification index. Note that for the GBM algorithm the Sensitivity index is very heterogeneous, assuming also values equal to 0.

The problem of size discourages the use of a model-based clustering approach: using a URV model for ordinal data with the integration of the survival time, the number of parameters to estimate is at least 10 μ and $10*(10+1)/2=55$ for the Σ covariance matrix with 65 parameters over 279 total observations. If we also want to estimate those parameters for each class, we end up requiring more parameters than observations. We have implement a Bayesian hierarchical model with uninformative prior distributions to model our dataset. We use a Markov chain Monte Carlo (MCMC) sampling process to perform the statistical inference procedures necessary for estimation of the parameters, but we have experienced convergence problems for MCMC and more improperly unstable parameter estimates. For these reasons we opted for distance-based methods to cluster our data.

Looking at the lower-level units and focusing only on the ordinal data (lesion profile), the observations are similar and they do not show high heterogeneity. We calculate some descriptive statistics on the lesion profile (Figure 2.5). The median



(a) (b)

(c)

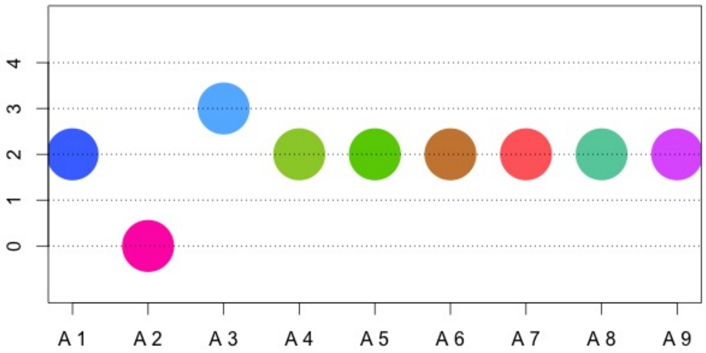
Figure 2.4. Supervised classification validation indexes for the *Simulation 2* on the 200 simulated samples using the 4 supervised algorithms for the 5 separated classes and with a “small class” (group 2): (a) Sensitivity index distributions (b) Specificity index distributions (c) Accuracy index distributions.

lesion profile corresponds to the modal one (Figure 2.5(a)) and all the areas show median scores equal to 3, except from the cerebellum (1) and the superior colliculus (4). Looking at the absolute difference from this profile, we find 15 observations (5.3% of the dataset) with the same values and 37 observations with a difference of one point in one area (Figure 2.5(b)). Since the lesion profiles are not highly heterogeneous, we expect that the statistical results, both for the clustering and the classification, will be more influenced by the survival time variable than by the ordinal scores.

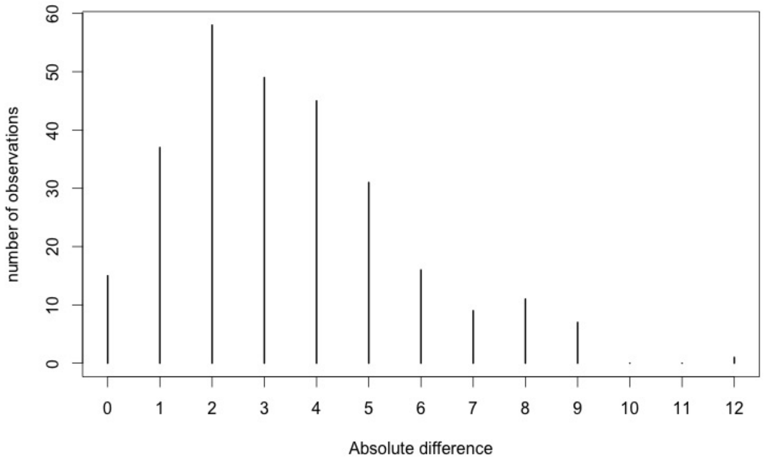
Inocula	Isolates	Number of bank voles	Researchers' partition (RC)
It1	It1	11	It93
It2	It2	11	It93
It3	It3	8	It93
It4	It4	7	It93
It5	It5	6	It93
It6	It6	10	It93
It7	It7	10	It93
Uk1	Uk1	10	It93
Uk2	Uk2	7	Uk85
Uk3	Uk3	6	Uk85
Uk4	Uk4	4	SSBP/1like
Uk5	Uk5	7	Uk85
Uk6	Uk6	11	It93
Uk7	Uk7	7	It93
Uk8	Uk8	4	Uk85
Uk9	Uk9	9	Uk85
Uk10	Uk10	9	Uk85
Uk11	Uk11	9	Uk85
Uk12	Uk12	8	Uk85
Uk13	Uk13	11	Uk85
Uk14	Uk14	11	Uk85
SSBP/1	SSBP/1	5	SSBP/1like
Ir1	Ir1	4	Uk85
No1	No1	6	It93
Ge1	Ge1	10	Uk85
Fr1	Fr1	4	Uk85
Fr2	Fr2-C	9	TypeC
CH1641	CH1641-A	8	TypeA
	CH1641-C	10	TypeC
Fr3	Fr3-A	8	TypeA
	Fr3-C	5	TypeC
Fr4	Fr4-A	10	TypeA
	Fr4-C	10	TypeC
Sp2	Sp2-A	3	TypeA
	Sp2-C	7	TypeC
Sp1	Sp1-A	5	TypeA

Class	Number of isolates	Number of observations
It93	11	97
Uk85	13	98
SSBP/1like	2	9
TypeA	5	34
TypeC	5	41

Table 2.1. Inocula used to perform the primary transmission and the relative isolates selected for the second passages. Researchers' a priori classification in 5 groups.



(a)



(b)

Figure 2.5. (a) Median and modal lesion profile in 9 brain areas (A1-A9) for isolates. (b) Absolute distances from the median profile: frequency distribution on the dataset A1=medulla, A2=cerebellum, A3=superior colliculus, A4=hypothalamus, A5=thalamus, A6=hippocampus, A7=septum, A8=retrosplenial and adjacent motor cortex, A9=cingulate and adjacent motor cortex.

Chapter 3

Unsupervised classification for multilevel mixed-type data

Unsupervised classification algorithms are here used in order to find similar groups of the higher-level units (the isolates). One very common (and simple) approach for dealing with multilevel data is to compute the average of variables over all lower-level units belonging to each higher-level unit and use these averages to build up a new data matrix for the higher-level units only. On this matrix, the pairwise dissimilarities can be computed and standard distance-based algorithms can be implemented. However, in the presence of mixed-type data, with also ordinal variables, the average is either impossible or misleading. Moreover, the average is not advised also for continuous data because of the possible loss of useful information, due to the presence of varying degrees of heterogeneity in lower-level units as well as because of different sample sizes for each isolate. The average is also not a good summary statistic where the data present asymmetric distributions.

To the best of our knowledge, methods for clustering multilevel mixed-type data have not been proposed so far; while model-based approaches to clustering multilevel data exist as well as clustering mixed-type data models. However, with our mixed-type data, we have decided to avoid the use of probabilistic assumptions on not yet well understood ordinal data for which the use of finite mixture models would require the distributional elicitation of suitable mixture components. Hence we stick to distance-based methods. As explained in Section 2.4 our dataset also presents unbalanced and often small sample sizes, this issue discourages us from the use of a model with a too rich parameterization.

3.1 Clustering algorithms for multilevel data

Looking at the problem of cluster analysis for multilevel data, we can classify four different ways to approach the problem: (i) model-based clustering relying on finite mixture modeling; (ii) distance-based methods extended to multilevel data; (iii) cluster analysis of cumulative distribution functions or probability density functions (pdf); (iv) functional cluster analysis. The way to look at the data hierarchies produces the different approaches: methods in (i) and (ii) extend the classical cluster analysis methodologies to the multilevel data by accounting of specific features of higher and lower-level units; methods in (iii) look at the lower-level units as realizations of the pdf objects (higher-level objects); methods in (iv) look at the higher-level unit as a functional object and the lower-level observations are used to build the functional data. Methods (i) and (iii) can also overlap in some definitions. All these methods are mostly applied to continuous data.

3.1.1 Clustering multilevel data via mixture models

Within the first approach, Celeux et al. (2005) propose a mixture of linear mixed-effects models (Equation 2.1) in order to cluster gene expression repeated data.

They choose linear mixed-effects models to take into account data variability. Authors provide a maximum likelihood estimation approach through the EM algorithm. Their model requires the strong independence assumption for the genes, so that Ng et al. (2006) consider the extension of normal mixture models to correlated and replicated data for gene expression data.

Higher-level units can also be classified through non-parametric maximum likelihood (NPML) (Aitkin 1999) or via multilevel latent class (or mixture) models (Vermunt & Magidson 2005, Vermunt 2008, Asparouhov & Muthen 2008) where random effects occur as discrete latent variables. Vermunt & Magidson (2005) propose the hierarchical mixture model which consists of two parts. The first part allow to classify the higher-level unit $\mathbf{x_j}$ in one of the K classes with the discrete random effect u_j:

$$\begin{aligned} f(\mathbf{x}_j) &= \sum_{k=1}^{K} \pi(u_j = k) f(\mathbf{x}_j | u_j = k) \\ f(\mathbf{x}_j | u_j = k) &= \prod_{i=1}^{n_j} f(\mathbf{x}_{ij} | u_j = k). \end{aligned} \tag{3.1}$$

Note that the lower-level observations $\mathbf{x}_{ij}$, given the class membership for the higher-level unit, are assumed conditionally i.i.d. within a higher-level unit.

The second mixture is used to model the $f(\mathbf{x}_{ij}|u_j = k)$, at the lower-level with the aim of detecting the heterogeneity within a higher-level unit. The random effect at this level is denoted by w_{ij}; given T components for the mixture, the model at this level is defined as follows:

$$f(\mathbf{x}_{ij} | u_j = k) = \sum_{t=1}^{T} \pi(w_{ij} = t | u_j = k) f(\mathbf{x}_{ij} | w_{ij} = t). \tag{3.2}$$

The use of a mixture at the lower-level allows modeling the data variability with discrete random effects instead of the continuous ones used by Celeux et al. (2005). The hierarchical mixture model also permits to classify both the higher-level units and the lower-level observations at the same time in the two different levels. If we compare the standard mixture model with the hierarchical mixture model, we see two important differences: (1) we obtain information on class membership for both higher-level units and lower-level units; (2) groups are assumed to differ with respect to the prior distribution of their members across lower-level latent classes.

These models are estimated using variants of the EM algorithms, but external methods are used to choose the number of components. Azzimonti et al. (2013) extend it with an EM-based approach where the number of mass points is automatically selected depending on some problem-driven tuning parameters.

3.1.2 Distance-based methods for multilevel data

The challenge with the distance-based methods for multilevel data is how to measure the distance between a pair of higher-level units (Figure 3.1). Methods proposed in literature do not take into account the distances between lower-level units. Usually, the information within the same object is previously summarized and then the distance-based algorithms are performed.

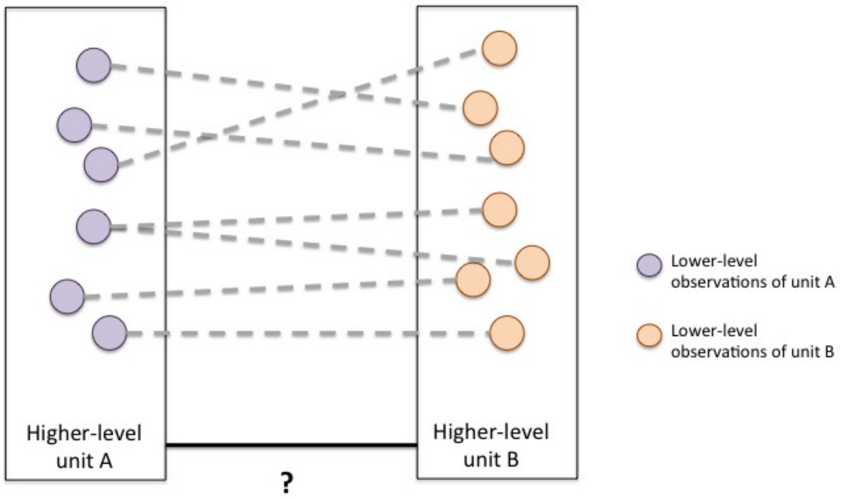

Lower-level observations of unit A
Lower-level observations of unit B
Higher-level unit A
Higher-level unit B
?

Figure 3.1. Distance between two higher-level units.

In this framework, Yeung et al. (2003) compare several clustering algorithms that incorporate repeated measurements. They perform an empirical study on real and synthetic gene expression datasets with the aim of evaluating alternative clustering methods.

Yeung et al. (2003) compare four different approaches to clustering repeated measurements: (i) average over repeated measurements; (ii) variability-weighted similarity measures; (iii) hierarchical clustering of repeated measurements; (iv) IMM-based approach.

The simplest and widely used approach is to compute the average over all the lower-level units for each higher-level unit. Then, the pairwise similarities can be computed on these average values and the classical distance-based algorithms are used. The averaging approach does not take into account the variability in repeated measurements within the same higher-level object.

Hughes et al. (2000) propose an error-weighted clustering approach. The idea is use error estimates to weigh values in pairwise similarities such that higher-level units with high variability are down-weighted. These error-weighted pairwise similarities are then used as inputs to clustering algorithms. Yeung et al. (2003) use directly variability's estimates as a weight for the pairwise dissimilarities, in particular, they use the standard deviation and the coefficient of variability for continuous data only.

Yeung et al. (2003) propose to cluster the repeated measurements as an individual object in hierarchical clustering algorithms. The idea is to employ hierarchical clustering algorithms with an ad-hoc adjustment named FITSS (forcing into the same subtrees) to account for the multilevel nature of the data: lower-level units of each higher-level object are assigned to the same subtree in the dendrogram and then the analysis continues as usual to build the final dendrogram. The initial rules to force lower-level units in the same dendrogram are not explicitly illustrated in the paper.

Medvedovic & Sivaganesan (2002) propose an infinite Gaussian mixture model (IMM model) for gene-expression data which incorporates repeated data. Within a higher-level unit, the lower-level observations are assumed to follow a multivariate normal distribution. They use a Gibbs sampler to estimate the posterior pairwise probabilities. These posterior pairwise probabilities are treated as pairwise similarities, which are finally used as inputs to hierarchical clustering algorithms.

We will focus on these methods proposing methodologies for measuring the distance between higher-level units as questioned in Figure 3.1. We avoid the use of a summary statistic for the data, but we summarize all the possible distances between pairs of lower-level units.

3.1.3 Cluster analysis of probability density function

The third way to consider multilevel data is by using pdf objects. Gibbs & Su (2002) provide a review of the metrics used to measure the distance between probability measures. Symbolic Data Analysis (SDA) is widely used in this field. SDA aims to analyze complex and structured data with higher-level units modeled by symbolic objects described by multivalued variables (Bock & Diday 2012, Billard & Diday 2003). Irpino & Verde (2006) present a new distance, based on the Wasserstein metric, in order to cluster a set of data described by distributions with a finite continuous support (histograms as called in SDA). A Wasserstein-based distance is also used by Irpino & Verde (2008) for interval data, that are statistical units described by means of intervals of values.

Terada & Yadohisa (2010) have proposed a non-hierarchical clustering method that works with empirical cumulative distribution functions, in order to avoid that the number of histogram bins (or the range of bins) affects the results of clustering. Kim & Billard (2013) introduces various dissimilarity measures for histogram data and they also develop a cumulative distribution measure for histograms. These approaches have been developed for both univariate and multivariate settings, even if in the last case, a large number of observations is required due to the curse of dimensionality.

Vrac et al. (2012) propose a model to cluster a set of estimated multivariate distribution functions based on the notions of a function of distributions and of multi-dimensional copulas.

Calò et al. (2014) introduce a hierarchical mixture modeling for the pdf estimation in the univariate context; then they extend the model to multivariate densities by means of a factorial model performing dimension reduction. EM algorithm is used to fit the model. This approach is classified by the authors as pdf unsupervised classification, even if it is an extension of the Vermunt & Magidson (2005) approach.

We have attempted to build a model-based clustering for multilevel mixed-type data following the model proposed by Calò et al. (2014). We use the mixed-type model proposed by McParland et al. (2014) for modeling the pdf objects, but we run into identification and overparameterization issues because of the use of several latent variables at the different levels.

3.1.4 Functional cluster analysis

The multilevel data can occur also when the lower-level observations are associated with a particular domain as, for example, time or space. In this case, the higher-level object can be seen as a function of the lower-level data. In functional data analysis (FDA) (Ramsay 2006) data can be seen as a realization of continuous functions on a given domain $X = \{X(t), t \in T\}$. The lower-level units are used to build the functional data (curves) for the higher-level object. The choice of an FDA approach is motivated by two main aspects, (1) data are collected at an almost continuous rate in space and time; (2) FDA allows for dimensional reduction and it is then suitable for very large datasets. Several standard statistical techniques have been adapted in FDA in order to capture features of functions and of their derivatives.

FDA is widely used with spatiotemporal data and high-dimensional data especially in environmental studies (see for instance Ignaccolo et al. 2015, Secchi et al. 2015, Ranalli et al. 2016).

Clustering functional data is generally a difficult task. The lack of a definition for the probability density of a functional random variable, the definition of distances or estimation from noisy data are some examples of such difficulties. Different approaches have been proposed. Jacques & Preda (2014) provide a comprehensive review for the functional cluster analysis (FCA). They suggest to distinguish four main approaches to the FCA: (*i*) methods working directly on the evaluation points of the curves; (*ii*) filtering methods that first approximate the curves into a finite basis of functions and then perform clustering using the basis expansion coefficients (two-stage methods); (*iii*) methods that perform simultaneously dimensional reduction of the curves and clustering, leading to functional representation of data depending on clusters; (*iv*) distance-based methods using clustering algorithms based on specific distances for functional data.

3.2 Clustering mixed-type data

The challenge of clustering mixed-type data is essentially due to how one can model the joint distribution of multivariate data measured by continuous, binary, categorical and ordinal data. In particular, the challenge is how to model categorical, binary and ordinal variables as explained in Section 2.3. Latent variable models for clustering categorical or mixed-type data have been proposed by many authors. On the other hand, only two dissimilarity indexes are proposed for mixed-type data to properly account for ordinal data.

3.2.1 Model-based clustering for mixed-type data

As pioneered by Everitt (1988) the URV approach (Equation 2.3) is widely used for ordinal data to establish thresholds in the categorical data and then using the underlying continuous latent variables in a joint distribution with the observed continuous variables. Assuming a mixture for the joint distribution for Z (see Section 2.3) is the way to perform cluster analysis.

Everitt & Merette (1990), Muthén & Shedden (1999) are the first to provide the intuition of clustering mixed-type data with the use of latent variable models. Everitt (1988) assumes a multivariate homoscedastic normal mixture density to model both observed and latent variables, while Lubke & Neale (2006) assume a heteroscedastic one. The estimation is done by maximum likelihood, but due to computational reasons, they are able to include only a few ordinal variables (Everitt & Merette 1990).

Cai et al. (2011) propose a mixture of generalized latent variable models to handle mixed-type heterogeneous data using different link functions to model data of multiple types. They estimate the model within the Bayesian framework. Browne & McNicholas (2012) use a mixture of latent variables model for the model-based clustering and also to perform discriminant analysis of mixed-type data. The estimates are carried out within the expectation-maximization (EM) framework. Mixed binary and continuous variables are analyzed by Morlini (2012) with an approach where each binary attribute is generated by a latent continuous variable that is dichotomized with a suitable threshold value, following the URV approach.

Cagnone & Viroli (2012) propose a latent variables model for binary data by a finite mixture of multivariate Gaussians in order to achieve both a dimension reduction and model-based clustering. They use a generalized version of the EM algorithm for the model estimation. Gollini & Murphy (2014) extend latent class analysis with a mixture of latent trait analyzers model by assuming a model for the categorical response variables that depends on both a categorical latent class (for finding the group structure) and a continuous latent trait variable (dependence within the unknown groups).

Ranalli & Rocci (2016) use a latent Gaussian mixture model to classify ordinal data following the URV approach: the observed categorical variables are considered as a discretization of an underlying finite mixture of Gaussians and the model is estimated within the EM framework maximizing a pairwise likelihood. Recently Ranalli & Rocci (2017) incorporate also continuous variables in the model: both the continuous and the ordinal variables are assumed to follow a heteroscedastic Gaussian mixture model, where the ordinal variables are modeled by the URV approach and the threshold parameters are estimated within the model.

The mixture of factor analyzers model for mixed-type data is proposed for ordinal data by McParland & Gormley (2013), using the IRT approach (Equation 2.2) with a URV link function. Successively McParland et al. (2014) extend their previous work to the mixed-type data, with a finite mixture model (MFA-MD) based on a combination of factor models for quantitative data; IRT models for ordinal data (Equation 2.2) and ideas from the multinomial probit model for categorical data (Equation 2.5). The MFA-MD model can explicitly model the nature of each variable type directly, as we have explained in Section 2.1. Even if McParland et al. (2014) achieve a dimensionality reduction with the IRT, they assume unknown thresholds and the model is computationally expensive. Therefore, the authors introduce also the model clustMD (McParland & Gormley 2016) that employs a parsimonious covariance structure for the latent variables. They use an EM algorithm combined with a Monte Carlo EM algorithm to estimate parameters.

All these models need the estimation of a high number of parameters: in addition to the mixture components' parameters, we need to estimate also the thresholds within the URV approach or the IRT parameters. More recently Canale & Dunson (2011) and Carmona et al. (2016) propose another parsimonuis model within the Bayesian framework, also extending the problem to complex design setting. Carmona et al. (2016) propose an infinite mixture and they model the ordinal variables within the URV approach. They do not reduce the dimensionality in the parameterization of the mean of the latent variables, but they fix thresholds' values.

Model-based approaches with latent variables present some difficulties: firstly all the models are computationally expensive, secondly, the models are deeply associated with the underline assumptions for categorical and ordinal variables. To our knowledge all the models reviewed in this subsection have not been extended to deal with multilevel data structures.

3.2.2 Measuring the dissimilarity between mixed-type data

On the other hand, distance-based clustering methods rely on the choice of an index to measure the distances between units, but they are computationally less demanding. We account for two different types of dissimilarities used in literature, the Gower index (Gower 1971) and the Gower index with Podani correction (Podani 1999). While the Gower index treats the ordinal variables as quantitative variables rescaled for their range, Podani extends the Gower index to ordinal variables considering a distance-based on their rank.

Gower dissimilarity for two units l and i is defined, in absence of missing data and binary variables, as follows:

$$d(l, i) = \sum_{p=1}^{P} d_p(l, i); \tag{3.3}$$

where the p-th variable contribution to the total distance, denoted by $d_p(l, i)$, is a distance between $x_{l,p}$ and $x_{i,p}$.

The contribution $d_p(l, i)$ for quantitative and ordinal variables is

$$d_p(l, i) = \frac{|x_{l,p} - x_{i,p}|}{range(x_p)}, \tag{3.4}$$

where the $x_{i,p}$ is the observed numeric value in the ordinal scale. For ordinal variables, this is the same as using them as quantitative variables. When the dataset has only quantitative and ordinal variables the Gower index is a rescaled Manhattan distance.

Podani (Podani 1999) extends the Gower index to ordinal variables considering the ranks:

$$d_p(l, i) = \frac{|r_{l,p} - r_{i,p}| - (T_{l,p} - 1)/2 - (T_{i,p} - 1)/2}{r_{max,p} - r_{min,p} - (T_{max,p} - 1)/2 - (T_{min,p} - 1)/2}, \tag{3.5}$$

where $T_{i,p}$ ($T_{l,p}$) is the number of units which have the same rank score for variable p as unit i (l) including i (l) itself, $r_{max,p}$ and $r_{min,p}$ are the maximum and minimum observed ranks for variable p respectively, $T_{max,p}$ is the number of units with the maximum rank, and $T_{min,p}$ is the number of units with the minimum rank.

3.3 Unsupervised classification via consensus clustering of distance-based methods

Starting from the FITSS proposed by Yeung et al. (2003), we extend and refine distance-based methods for clustering multilevel mixed-type data. In our context, there is a relevant irregular heterogeneity, with possibly asymmetric features, within observations belonging to the same higher-level unit differently from the typical case in Yeung et al. (2003) where repeated measurements are considered. The substantial impact of such heterogeneity in our multilevel data can undermine the stability of the results of a specific cluster analysis, relying on a particular choice of distance and algorithm. In order to account for this aspect, we propose a consensus clustering of hierarchical algorithms based on alternative sensible choices of distances with FITSS adjustment. With a similar idea we use partitioning strategies relying on a modified Partition Around Medoids (PAM) algorithm (Kaufman & Rousseeuw 2009) to deal with the multilevel data challenge.

We choose to implement both a hierarchical clustering and a partitioning method in order to compare different results: we consider hierarchical methods in order to build alternative hierarchies of clusters; and we use a non-hierarchical PAM algorithm (Kaufman & Rousseeuw 2009). Both methods are adapted to the multilevel data using the consensus clustering (Gordon 1999, Hornik 2005) illustrated below.

Starting from different partitions/hierarchies the general idea of consensus clustering is finding the partition/hierarchy z that minimize the following objective function over all the $b = 1, \dots, B$ partitions/hierarchies,

$$L(z) = \sum_{b=1}^{B} w_b d(z_b, z) \tag{3.6}$$

where z_b is the partition/hierarchy of the $b-$th method, the weights w_b are the weights for the partitions and d is the distance between partitions. Consensus clustering will be used both in the hierarchical algorithms and in the PAM to have a unique, more stable and robust clustering result.

3.3.1 Cluster analysis via hierarchical algorithms: extending the FITSS

Our aim is to propose a general method accounting for the heterogeneity of lower-level units belonging to the same higher-level observation. Starting from the FITSS (forcing into the same subtrees) introduced by Yeung et al. (2003) we initialize the agglomerative algorithms by assigning units of the same higher-level unit to the same subtree in the dendrogram. The idea is to use different agglomerative algorithms managing distances between higher-level units starting from different summary statistics (minimum, average, quantiles...) of the dissimilarities computed on multivariate outcomes observed on each lower-level unit. After this step, we use a linkage method with distance update between subgroups relying on the same (or possibly more similar) summary statistics. For example in a single linkage hierarchical algorithm, the cluster dissimilarity of two clusters is the minimum dissimilarity between two pairs of units belonging to the clusters, while in a complete linkage hierarchical algorithm the cluster dissimilarity is the maximum distance. However, with data which can be substantially more heterogeneous than repeated measures, we believe that other summary statistics of distances are better suited. We have also to deal with the mixed-type data using the dissimilarities presented in Section 3.2.2.

Since both the Gower index and the Gower index with Podani correction are not metrics we restrict the attention to the following common hierarchical algorithms (Kaufman & Rousseeuw 2009): single linkage, complete linkage, average linkage and McQuitty linkage.

Method Names	Linkage method	Distance update between higher-level units	FITSS distance (Distances between couples units belonging to j and v)
SINGLE	single linkage	$d(jv,l) = min\{d(j,l), d(v,l)\}$	Minimum
SINGLE_Q1	single linkage	$d(jv,l) = min\{d(j,l), d(v,l)\}$	First Quartile
COMPLETE	complete linkage	$d(jv,l) = max\{d(j,l), d(v,l)\}$	Maximum
COMPLETE_Q3	complete linkage	$d(jv,l) = max\{d(j,l), d(v,l)\}$	Third quartile
AVERAGE	average linkage	$d(jv,l) = \frac{n_j d(j,l)+n_v d(v,l)}{n_j+n_v}$	Average
MCQUITTY	McQuitty	$d(jv,l) = \frac{d(j,l)+d(v,l)}{2}$	Average

Table 3.1. Hierarchical methods used for consensus clustering.

Table 3.1 shows the basic ingredients of these four agglomerative methods with FITSS distance where, in the case of single and complete linkage, the FITSS has been diversified using two different summary statistics for the initialization: minimum (SINGLE) and first quartile(SINGLE_Q1) for the single linkage and maximum (COMPLETE) and third quartile (COMPLETE_Q3) for the complete linkage.

With the hierarchical methods, we show how the use of consensus clustering improves the quality and robustness of our analysis yielding a unique dendrogram. We remind that there is a bijection between dendrograms and ultrametrics so that a generic dendrogram can be denoted by $u = (u_{lj})$ where u_{lj} is the ultrametric distance between the l-th unit and the j-th unit.

In our consensus approach we use the Manhattan distance for d and the weights w_b are considered all equal so that the objective function (3.6) becomes:

$$L(u) = \sum_{b=1}^{B} \sum_{i<j} |h_{ljb} - u_{lj}| \tag{3.7}$$

where h_{ljb} is the ultrametric distance between the l-th unit and the j-th unit corresponding to the b-th dendrogram. The minimization problem for (3.7) is solved via the Sequential Unconstrained Minimization Technique (SUMT) (De Soete 1984). We use visual inspection for the final choice of the number of clusters because validation indexes can not be used for the consensus final partition. In fact the indexes are calculated using distances between units: in the consensus clustering of hierarchical algorithms initialized by different distances we are not able to rely on final distances between units. Otherwise the notions of compactness, connectedness and separation are not extendible to the multilevel data because also observations far away from each other could be measured in the same higher-level unit.

3.3.2 Cluster analysis via partition algorithms: extending PAM for hierarchical data

Despite hierarchical clustering is more suitable for exploring and visualizing heterogeneity in our experimental data, we decided to challenge our results implementing also a non-hierarchical partitioning method. PAM (Kaufman & Rousseeuw 2009) looks for k representative units $m_1, \dots, m_k, \dots, m_K$, called medoids so that each medoid m_k represents a cluster C_k. Each unit j belongs to the cluster whose medoid is closest to that unit. The optimal search of K medoids is based on the following objective function to be minimized:

$$\sum_{k=1}^{K} \sum_{j \in C_k} d(j, m_k).$$

The PAM algorithm used for minimization has two phases: (i) the build/initialization phase and (ii) the swap phase. It first looks for a good initial set of medoids and then at each swap step, one iteratively tries to swap in turn each unit with a medoid and the swap is accepted provided that the new choice of medoids improves the objective function. The algorithm stops when no swap can improve the objective function. Compared to the k-means approach, PAM is more flexible and robust since it is not limited to use euclidean distances and it accepts any kind of dissimilarities. The Silhouette index (Equation 1.3) may help to select the number of clusters using the average silhouette width $\tilde{s}(K)$, which is the mean of $s(K)$ over all objects of any possible clustering with K groups (Kaufman & Rousseeuw 2009).

In order to implement the algorithm with our data we need to input a dissimilarity/distance matrix for higher-level units. Since PAM is a partitioning method there is no natural linkage distance to drive the choice of a summarizing distance between lower-level units belonging to different higher-level units. Hence we found it more appropriate to try alternative starting summaries. We use the minimum, the first quartile, the median, the third quartile and the maximum. Once the distance is provided, the PAM algorithm is implemented with a fixed number of clusters. Then, one could approach the selection of a suitable number of clusters relying on some criterion. The most used is the maximization of $\tilde{s}(K)$ provided there is reasonable evidence of well-formed partition, which is verified, as suggested by Kaufman & Rousseeuw (2009), when $\tilde{s}(k) \geq 0.51$. In order to get a unique partition taking into account the possibly different results, we use again a consensus clustering approach. Suppose we have B alternative partitions, $\{P_1, \ldots, P_b, \ldots, P_B\}$ of the same set of n units. A partition can be represented as a graph whose nodes correspond to the units and two units in the same class are connected by an arc. Hence a partition P_b can be represented in terms of the corresponding adjacency matrix $\mathbf{C}_b = (c_{lj}^{(b)})$, where $c_{lj}^{(b)} = 1$ if the l-th and the j-th units belong to the same class and 0 if they do not. Similarly to the consensus clustering for hierarchies we opted for d as the Manhattan distance and equal weights w_b so that (3.6) can be rewritten as

$$L(z) = \sum_{b=1}^{B} \sum_{l<j} |c_{ljb} - z_{lj}| \tag{3.8}$$

where $z_{lj} = 1$ if the l-th and the j-th units belong to the same class in the optimal partition P and 0 if they do not. Note that the minimization is carried out with suitable conditions as in Gordon (1999) to ensure that z identifies a partition. In order to find a unique resulting partition, we include in the consensus clustering algorithm all partitions in K clusters obtained from the different starting distances. The optimization in (3.8), solved via the SUMT, automatically selects a suitable number of clusters within the range of groups which have been given in input.

3.4 Distance between higher-level units: a proposal

Using the FITSS methods, proposed in Section 3.3, we measure the distance between two higher-level units (Figure 3.1) relying on a summary statistics of the all possible distances between pairs of lower-level units. Now we want address the issue of measuring the distance between two higher-level objects considering all the distances between all the possible pairs of lower-level units.

We rely on the optimal transport problem introduced by Monge (1781). We formulate the problem in the general form as follows: given two objects A and B composed of a set of objects $\{A_1, \ldots, A_i, \ldots, A_m\}$ and $\{B_1, \ldots, B_j, \ldots, B_n\}$. At each object A_i is associated a mass $a_i \geq 0$, such as at each object B_j is associated a mass b_j. Let x_{ij} be the non negative mass transported from A_i to B_j. Hence $x_{ij} \geq 0$. The transport plan $\gamma = \{x_{11}, \ldots, x_{ij}, \ldots, x_{mn}\}$ (Figure 3.2) for moving masses from A to B is a bijection between $\{A_1, \ldots, A_i, \ldots, A_m\}$ and $\{B_1, \ldots, B_j, \ldots, B_n\}$. Let c_{ij} be the unit cost of transporting a mass from A_i to B_j, we want to find the transport plan γ for moving masses from A to B that minimizes the total cost, subject to the following constraints:

- *Necessary and sufficient condition*: the total mass transported from A must be equal to the total mass received by B:

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j; \tag{3.9}$$

- All the mass of A_i must be transported to the object B:

$$\sum_{j=1}^{n} x_{ij} = a_i \quad i = 1, \ldots, m; \tag{3.10}$$

- All the mass of B_j must be received from the object A:

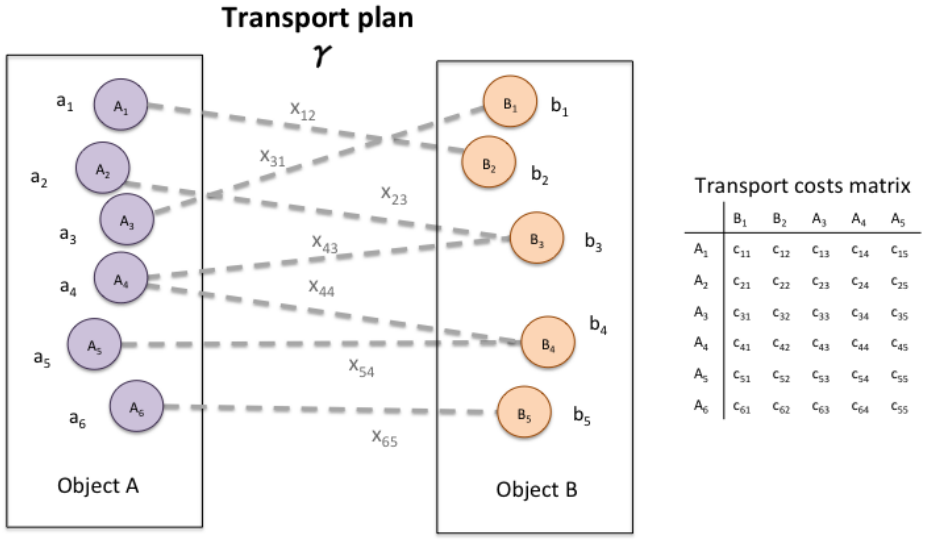
$$\sum_{i=1}^{m} x_{ij} = b_j \quad j = 1, \ldots, n. \tag{3.11}$$

The subset Γ represents all the possible transportation plans γ for which (3.10) and (3.11) hold true. The optimal transport plan is then defined as the minimization problem:

$$\gamma^* : \arg\min_{\gamma \in \Gamma} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{3.12}$$

subject to the constraints (3.9), (3.10) and (3.11).

When the cost is defined in terms of a distance, we end up formalizing the *Wasserstein distance* between two objects A and B, which is in fact related to the optimal transport plan. As suggested by Villani (2008), the terminology of Wasserstein distance is very questionable since the explicit definition of the distance



Transport costs matrix

	B1	B2	A3	A4	A5
A1	c11	c12	c13	c14	c15
A2	c21	c22	c23	c24	c25
A3	c31	c32	c33	c34	c35
A4	c41	c42	c43	c44	c45
A5	c51	c52	c53	c54	c55
A6	c61	c62	c63	c64	c55

Figure 3.2. An example of a transport plan between two objects A and B with a given matrix for the costs of transport.

is not so easy to find in Wasserstein's work. Moreover, these distances are "discovered and rediscovered" by several authors over time, including Gini (1921), Kantorovitch (1958), Vaserstein (1969), Mallows (1972), Tanaka (1973), so that the distance occurs with different names as the Mallows distance, the optimal transport distance, the Earth Movers' distance etc...

Villani (2008) define the Wasserstein distance of order p between two probability measures μ and ν, using the notion of transport plan γ, as following.

Definition 1 (Wasserstein Distance)**.** *Let $(\mathcal{X}, d)$ be a Polish metric space (complete separable metric space), and let $p \in [1, \infty)$. For any two probability measures μ, ν on $\mathcal{X}$, the Wasserstein distance of order p between μ and ν is defined by the formula:*

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}} d(x, y)^p d\gamma(x, y) \right)^{1/p}. \tag{3.13}$$

Given $\mu \in P(\mathcal{X})$ and $\nu \in P(\mathcal{Y})$, then $\Gamma(\mu, \nu)$ is the set of all joint probability measures on $\mathcal{X} \times \mathcal{Y}$ whose marginals are μ and ν respectively.

For example, when we want to calculate the Wasserstein distance between two objects $A = \{A_1, \dots, A_k\}$ and $B = \{A_1, \dots, B_k\}$ with the same masses equal to 1 and with the Euclidean distance as cost function we have the simplified formulation:

$$W_p(A,B) = \sum_i \parallel a_i - b_{\gamma^*(i)} \parallel^p . \tag{3.14}$$

The Wasserstein distance has no closed form solution and can only be solved by linear programming algorithms. The R-package `transport` (Schuhmacher et al. 2017) propose several algorithms to solve the problem for the optimal transport plan:

- `shortsimplex`: The shortlist method, based an a revised simplex algorithm, as described in Gottschlich & Schuhmacher (2014);
- `revsimplex:` The revised simplex algorithm with various speed improvements, including a multiscale approach, as described in Luenberger & Ye (2015);
- `primaldual:` The primal-dual algorithm as described in Luenberger & Ye (2015);
- `aha`: The Aurenhammer et al. (1998) method with the multiscale approach presented in Mérigot (2011);
- `auction:` The auction algorithm by Bertsekas (1988) with epsilon-scaling, see Bertsekas (1992);
- `auctionbf:` A refined auction algorithm that combines forward and revers auction, see Bertsekas (1992).

Finally, our proposal is use the optimal transport plan problem to measure the distance between two higher-level units A and B by taking a dissimilarity measure (Gower or Podani for mixed-type data) as cost function. The masses are fixed as follows: $a_j = 1/n_a$ and $b_i = 1/n_b$, so that $\sum_{j=1}^{n_a} a_j = \sum_{i=1}^{n_b} b_i = 1$ to account for the unbalanced sizes of the higher-level units.

Chapter 4

Supervised classification for multilevel data

Supervised classification is used in order to estimate the predictive performance of several classifiers, given a classification of the isolates formulated by researchers. Our final goal is the partition validation, so we focus on the cross-validation algorithm to avoid the overfitting problem. We use several classifiers to have more indexes to evaluate the predictive performance. The challenge of mixed-type data is not relevant in supervised classification because the variables are used as covariates in the algorithms. On the other hand we need to define modified strategies for supervised classification of hierarchical data structure, using a model that simultaneously combine predictors and gives a unique prediction for a higher-level unit. In this way, with supervised classification, we have the possibility to classify new future discovered prion diseases within the defined partition.

4.1 Supervised classification for multilevel data

Supervised classification problems with a multilevel structure have not received the same attention in the statistical literature as the general classification problem, as evidenced also by Yamal et al. (2011) and Yamal et al. (2015). These are perhaps the only two papers published in statistical journals, while heuristics methods are widely proposed in medical literature where we can find some examples of that problem: classifying measurements of cell nuclei from fine needle aspirates to diagnose breast cancer (Mangasarian et al. 1995), measurements of mouthwashes for patient diagnosis of oral cancer and periodontal pathogens (Sciubba et al. 1999, Christian 2002, Boutaga et al. 2007) and flow cytometric measurements (Cadez et al. 1999). Yamal et al. (2011) identify three possible approaches to the problem: (i) extracting higher-level features from the lower-level data, (Bashashati & Brinkman 2009, Lugli et al. 2007, Thiran & Macq 1996) (ii) using a statistical model that accounts for the multilevel data structure (Swartz et al. 2005) and (iii) classifying at the lower-level and then use an ad hoc approach to classify at the higher-level (two-steps approach) (Cadez et al. 1999). Cadez et al. (1999) use a hierarchical Bayesian model with two levels, firstly modeling the distribution of the lower-level units with a mixture model, and then modeling the probability at the higher-level for each class.

Swartz et al. (2005) propose the cumulative log-odds (CLO) approach. They deal with the multilevel classification problem by obtaining estimates for the posterior log-odds of diseases, given the observed data at the lower-level and then combining those estimates to calculate the posterior log-odds of disease at the higher-level. The model is applied to classify patients (higher-level unit) with adenocarcinoma of the cervix, while the lower-level data are cellular measurements within each patient. CLO assumes that given the disease state, the cellular measurements share an identical distribution and are independent of each other, modeled by the posterior log-odds of disease. In other words, the assumption is that conditional on the class, the feature vectors for the members of the same population (the lower-level units) are i.i.d. The density functions are approximated using a kernel density estimate. CLO is similar to the naïve Bayes. The only difference is that in the naïve Bayes the unit of measurement and unit of classification are the same, whereas in the CLO lower-level units are for the measurement and higher-level units are for the classification. An extension of this method was proposed using latent-class cumulative log-odds (LACLO) by (Yamal et al. 2011), allowing for heterogeneity of the distributions of the data within a disease state. They assume the existence of an unobserved latent variable and that the features are i.i.d., given the class and the latent variable. The densities are estimated with a functional clustering method: firstly they compute a density estimate for each higher-level unit and then apply a clustering algorithm to the densities. The proposal methods only deal with the binary classification problem; in the model-based methods, only quantitative variables that can be modeled with density estimation are considered.

There is not evidence of the best approach to the multilevel supervised classification problem: Tsybrovskyy & Berghold (1999) show that ignoring the higher-level as the unit of analysis, leads to an increase in bias. Therefore a multilevel method could be the best approach to this problem. Yamal et al. (2015) apply several methods of the three approaches to the same dataset and compare the predictive performance to find the best one. Results highlight good performances of the two steps approach using Elastic Net and Tree-based methods for the lower-level classification.

These papers do not deal with the problem of adapting cross-validation algorithms to consider the multilevel structure of the dataset. Only Yamal et al. (2015) explain their cross-validation strategy. They divide the data into three sets: training, validation, and test sets, following the basic rules of Friedman et al. (2009). They use the training set to estimate the parameters of a classifier, either by using the whole training set to fit models with no "free" parameters (e.g., logistic regression) or by using five-fold cross-validation to choose the model parameters (e.g., the penalization parameter of L1-regularized logistic regression), relying on double cross-validation even if they do not formalize the algorithm. The validation set is used to obtain estimates of the trained classifier's performance using the parameters estimated in the training set and to select a classifier to apply to the test set. The test set is used to obtain an unbiased estimate of the chosen classifier's performance, re-estimating the classifier's parameters using fivefold cross-validation within the combined training and validation sets.

4.2 A classification algorithm for multilevel data

Our problem is different from the typical one already dealt with in the literature. Firstly we do not have a binary partition, but a multi-class partition where a class label $y_{ij} = 1, \ldots, g, \ldots, G$ is defined for each higher-level unit. Secondly, our predictors or covariate are the mixed-data $\mathbf{x}_{ij}$ and we can not use methodologies to model the joint density. We propose a method within the "two-steps approach": firstly we classify the lower-level units and then we use an ad hoc approach to classify at the higher-level.

We assume that the probability for a lower-level unit to belong to a group, given the class label, does not depend on the higher level unit. These probabilities can be then used to classify the higher level unit. We start from predicting the class probabilities at the lower-level using several classification algorithms and then use a suitable approach to classify at the higher-level. In this way, we can also evaluate alternative standard classifiers and use a combination of classifiers. We now introduce methodological details of our supervised classification analysis.

4.2.1 Classification for lower-level units

We use classifiers for the i-th lower-level observation nested within the j-th higher-level unit. The probability of belonging to the g-th class, derived from the k-th classifier, is denoted by $\hat{p}^k_{ijg}(\mathbf{x})$. Classification methods dealing with mixed-type data used for the analysis are Random Forest (RF, $k = 1$), Gradient Boosting Method (GBM, $k = 2$) and Neural Network (NNET, $k = 3$) (Friedman et al. 2009). The best tuning parameters are chosen using a 10-fold cross-validation, for each classifier, within a step of the higher-level leave-one-out procedure as described in Section 4.3. The double cross-validation approach of Mertens et al. (2006) is adjusted for multilevel data. Moreover, in order to enhance the predictive performance, combination approaches at the lower-level may be considered (Kakourou et al. 2014). As a first approach, we consider a simple convex combination. We take the average of predicted probabilities fitted by each classifier k:

$$p^{AVE}_{ijg}(\mathbf{x}) = \sum_{k=1}^{K} w_k p^k_{ijg}(\mathbf{x}) \tag{4.1}$$

where w_k are the weights for the k-th classifier with $\sum_{k=1}^{K} w_k = 1$. In the absence of a priori information, we set equal weights: $w_k = \frac{1}{K}$.

Alternatively, we use a model based combination approach (Kakourou et al. 2014) by fitting a multinomial logistic regression to the set of posterior class probabilities functions:

$$\log\left(\frac{P(y_{ij} = g)}{P(y_{ij} \neq g)}\right) = \beta_0 + \sum_{k=1}^{K}\sum_{g=1}^{G} \beta_{kg}\text{logit}\left(p^k_{ijg}(\mathbf{x})\right) \tag{4.2}$$

There are some non trivial issues with the model in Equation 4.2. Using the predictive probabilities $p^k_{ijg}(\mathbf{x})$, calibrated on the training set, can lead to both overfitting and bias (LeBlanc & Tibshirani 1996). Kakourou et al. (2014) choose to use leave-one-out “double cross-validation” (Mertens et al. 2006) to calibrate the predicted class probabilities of the classifiers to be combined. On the other hand

$\text{logit}(p^k_{ijg}(\mathbf{x}))$ are correlated predictors, so we use the Elastic Net Regression (Zou & Hastie 2005) that performs both regression and variable selection. The penalty index $\alpha \in (0,1)$ appearing in the term $\sum_{j=1}^{p}(1-\alpha)\frac{1}{2}\|\beta_j\|_2^2 + \alpha\|\beta_j\|_2$, is also chosen via cross-validation. Note that in this elastic net framework with $\alpha = 0$, we get the lasso penalty and with $\alpha = 1$, we get the ridge penalty.

4.2.2 Prediction for the higher-level unit

Given a classifier k, for each lower-level unit, probabilities $\hat{p}^k_{ijg}(\mathbf{x})$ are derived as explained in Section 4.2.1. However, it may happen that different lower-level units within the same higher-level unit may result as predicted in different classes. Indeed we need to elaborate predictions from lower-level units nested within a higher-level unit to get a single higher-level prediction. Hence we consider the following three different methods for exploiting lower-level predictions.

Modal value method. Firstly, we assign each lower-level unit to the most probable class:

$$\hat{y}^k_{ij} = \underset{g=1,\ldots,G}{\text{argmax}}\, \hat{p}^k_{ijg}(\mathbf{x}) \tag{4.3}$$

and obtain a vector of labels: $\hat{\mathbf{y}}^k_j = \left(\hat{y}^k_{1j} \ldots, \hat{y}^k_{n_j j}\right)$ for the higher-level unit. Then,for each j-th higher-level unit, the class predicted is the modal class,

$$\hat{y}^k_j = \text{Mode}\{\hat{\mathbf{y}}^k_j\} \tag{4.4}$$

A possible issue occurs when the subset of class labels is multi-modal, in that case, there is no unique assignment of class label.

Average method. For each class g, the predictive probability for the j-th higher-level unit is obtained as the average of predicted probabilities for the n_j lower-level units nested in the j-th higher-level unit,

$$\hat{\overline{p}}^k_{jg} = \frac{1}{n_j}\sum_{i=1}^{n_j}\hat{p}^k_{ijg} \tag{4.5}$$

and we finally use the class with the highest value of the average probability for classification:

$$\hat{y}^k_j = \underset{g=1,\ldots,G}{\text{argmax}}\, \hat{\overline{p}}^k_{jg} \tag{4.6}$$

Higher-level model-based combination (HLMBC method). The basic idea is to aggregate the information of the standard classifiers using the average of the predicted probabilities $\left(\overline{p}_{jg}^{k}(\mathbf{x})\right)$ as the predictors on the model-based combination approach as follows:

$$\log\left(\frac{P(y_j = g)}{P(y_j \neq g)}\right) = \beta_0 + \sum_{k=1}^{K}\sum_{g=1}^{G}\beta_{kg}\text{logit}\left(\overline{p}_{jg}^{k}(\mathbf{x})\right). \tag{4.7}$$

However, issues related to the estimation of the predicted probabilities $\overline{p}_{jg}^{k}(\mathbf{x})$ arise. We decide to estimate the probabilities $\overline{p}_{jg}^{k}(\mathbf{x})$ exploit the 10-fold cross-validation step within the leave-one-out. For each fold, we calculate the estimated probabilities $\hat{p}_{ijgf}^{k}(\mathbf{x})$. Finally, we take the average within each fold to have a probability for the higher-level unit, denoted by $\widehat{\overline{p}}_{jgf}^{k}(\mathbf{x})$. We now have a new training set composed only of higher-level units repetitions (indexed with j for the higher-level and f for the different folds) and each repetition is composed of different lower-level units:

$$\mathcal{T} = \left\{\left(y_j, \widehat{\overline{p}}_{jgf}^{k}(x)\right); g = 1, \ldots, G; f = 1, \ldots, F\right\} \tag{4.8}$$

On this training set, we perform another 10-fold cross-validation to tune the best parameters of an Elastic Net Regression based on Equation 4.7. This way we avoid overfitting problems since the $\widehat{\overline{p}}_{jgf}^{k}(\mathbf{x})$ are calibrated in an external way during the 10-fold cross-validation on the standard classifiers and the performance is evaluated in an external testing set as described in Section 4.3.

4.3 Cross Validation for multilevel data: higher-level unit leave-one-out

We use a double cross-validation approach each training set of the overall cross-validation an embedded cross-validation is performed in order to find the best tuning parameters and estimate the predictive class probabilities.

We modify the cross-validation algorithm in order to take into account the multilevel structure of the dataset. With this algorithm, we want to avoid the overfitting problems and also have an evaluation of the predictive performance of the classifiers using different samples of lower-level units for the same higher-level unit. We also aim at avoiding a possible bias due to a specific testing set. We have to work both in the training set, to correctly calibrate the predictive probabilities for the model-based combination of classifiers, and in the testing set in order to have alternative samples of lower-level units for the same higher-level unit to predict the class which the higher-level unit belongs to. We first use a leave-one-out strategy for the higher-level units so that the training set is composed of all lower-level units not belonging to the j-th higher-level unit. With this training set, we use the 10-fold cross-validation in order to find the best tuning parameters, as well as model parameters. Hence we get the calibrated estimates of lower-level class probabilities in (4.2) and (4.7).

Let $\mathcal{D}_j = \{(y_{ij}, x_{ij}); i = 1, \ldots, n_j\}$ denote the subset of data corresponding to the j-th higher-level unit. We use all possible samples of the lower-level units in $\mathcal{D}_j$ to create alternative testing sets. In this way, we can evaluate the predictive performance of the same higher-level unit. In fact, we start from different combinations of lower-level units because we want to avoid results influenced by a specific choice of lower-level units. Another benefit of this sampling in $\mathcal{D}_j$ is to use a fixed minimum number m of lower-level units. Since the goal of classifiers is the potential prediction of new strains, a possible future testing set will be new experiments with m bank voles for each new strain. If we fix $m = min\,\{n_j\} = 3$ and get a good predictive performance then we can safely plan triplicate experiments. Indeed we note that m should be as low as possible for ethical reasons. The notation for the leave-one-out training datasets is the following: $\mathcal{D}_{-1}, ..., \mathcal{D}_{-j}, ..., \mathcal{D}_{-J}$, where $\mathcal{D}_{-j} = \bigcup_{l \neq j} \mathcal{D}_l$.

We specify the pseudo code for the adjusted higher-level “leave-one-out” modified strategy in Algorithm 1 and in Figure 4.1

Note that for the HLMBC method we use the predicted probabilities $\widehat{\widehat{p}}^{k}_{jgf}(\mathbf{x})$ for the folds created in the first cycle and then we use the training set defined in (4.8) with another 10-fold cross-validation to calibrate the model (4.7).

```
for $j$ = $1$ to $J$ do
    Remove all the lower-level units belonging to the $j$-th higher-level from $\mathcal{D}$: $\mathcal{D}_{-j}$;
    for $k$= $1$ to $K$ do
        Train classification algorithms on $\mathcal{D}_{-j}$;
        Do a 10-Fold cross-validation to find best tuning parameters;
    end
    Use $\mathcal{D}_j$ to form different testing sets $\mathcal{S}_l^j \quad l = 1, \ldots, L_j$ with a fixed number $m$ of lower-level units with $L_j = \binom{n_j}{m}$ ;
    for $k$= $1$ to $K$ do
        for $l$= $1$ to $L_j$ do
            Determine the predictive probability $\hat{p}_{ijg}^k(\mathbf{x})$ of the $i$-th lower-level unit in the $l$-th testing set $\mathcal{S}_l^j$ of the $j$-th higher-level unit to belong to the $g$-th class;
            Assign the $j$-th higher-level unit to one of the $G$ classes properly combining $\hat{p}_{ijg}^k(\mathbf{x})$;
        end
    end
end
```

Algorithm 1: Higher-level unit “leave-one-out”

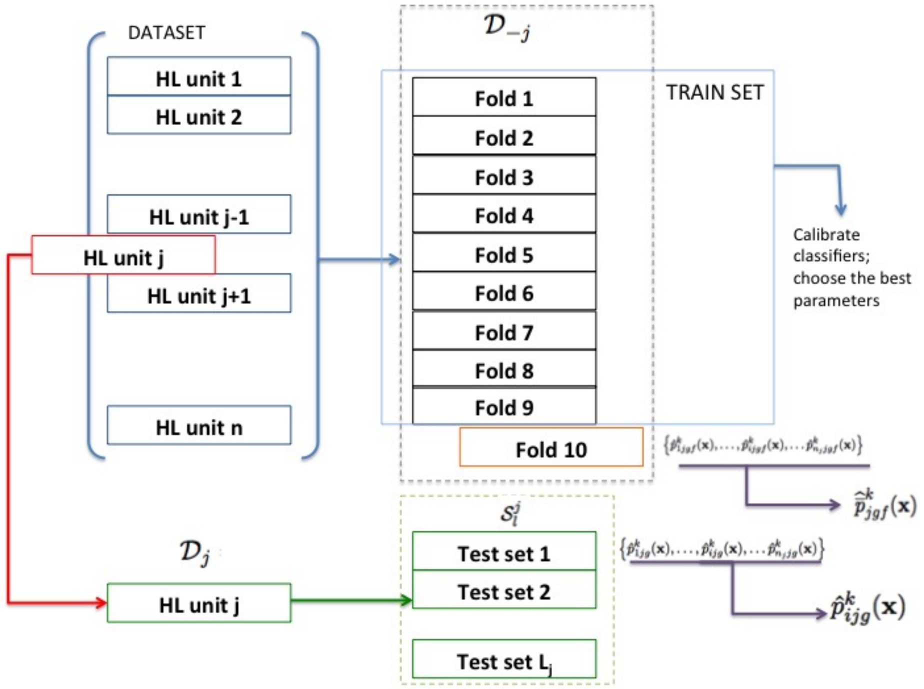

DATASET
HL unit 1
HL unit 2
HL unit j-1
HL unit j
HL unit j+1
HL unit n
$\mathcal{D}_{-j}$
TRAIN SET
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10
Calibrate classifiers; choose the best parameters
$\{\hat{p}^k_{1jgf}(\mathbf{x}),\ldots,\hat{p}^k_{ijgf}(\mathbf{x}),\ldots\hat{p}^k_{n_jjgf}(\mathbf{x})\}$
$\widehat{\overline{p}}^k_{jgf}(\mathbf{x})$
$\mathcal{S}^j_l$
$\mathcal{D}_j$
HL unit j
Test set 1
Test set 2
Test set L$_j$
$\{\hat{p}^k_{1jg}(\mathbf{x}),\ldots,\hat{p}^k_{ijg}(\mathbf{x}),\ldots\hat{p}^k_{n_jjg}(\mathbf{x})\}$
$\hat{p}^k_{ijg}(\mathbf{x})$

Figure 4.1. A schematic representation of the Algorithm 1 .

Chapter 5

Application to the scrapie disease data

In this Chapter, we apply the proposed methods to the scrapie dataset, described in Section 2.1. The aim is the validation of the a priori partition elicited by researchers. They have originally formulated 5 groups for the 36 isolates (RP partition): (1) the SSBP/1 like group composed only by 2 isolates, (2) the UK35 with 13 isolates, (3) the It93 groups with 11 isolates, (4) the Type-C with 5 isolates and the (5) Type-A with 5 isolates. The researchers' external evidence for eliciting this group partition have been essentially based on the geographic sources, while the Type-A and the Type-C groups have occurred as new variants when different pathological phenotypes were observed in animals inoculated with the same inoculum at first passage.

Following the outline in Figure 1.2, we discuss the cluster analysis with several methods and then we compare the resulting EP partitions with the a priori RP partition using the Adjusted Rand Index. Finally, researchers elicit the new partition (NP) taking advantage of the results of the cluster analysis elicited a new partition based on the EP and the RP. Then we use supervised classification to validate the NP partition using the evaluation of the predictive performance of the Accuracy index, the Sensitivity and Specificity index for the groups (see Section 1.3). Good predictive performances of the indexes validate our group structure and we can use the best classifier for predicting new further discovered isolates within this group structure.

5.1 Clustering analysis

We use distance-based algorithms proposed in Chapter 3 to cluster the scrapie disease dataset. We first need to choose the metrics for measuring the distance between units for all the clustering methods. We focus on the two indexes introduced in the Section 3.2.2: the Gower index (Gower 1971) and the Gower index with Podani correction (Podani 1999) (thereafter called “Podani index” to simplify). For example, in Figure 5.1 we compare the two indexes computed between all couples of lower-level units, one belonging to Uk1 and the other belonging to a possibly different isolate. On the left, in red, we highlight the distribution of dissimilarities within the same isolate (Uk1). Podani index is better suited for detecting more dissimilarity between units belonging to different isolates rather than units within Uk1. This happens also when we replace Uk1 with a different isolate. For this reason, we choose Podani index as dissimilarity measure for our cluster analysis.

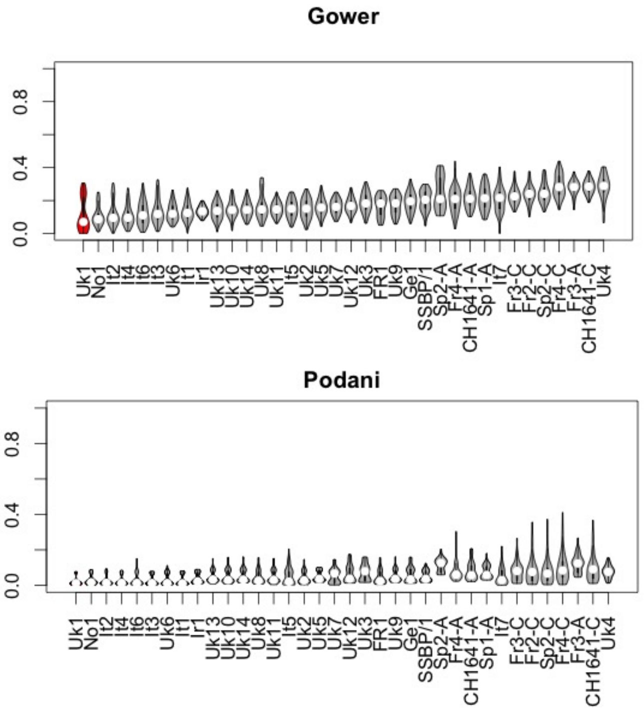


Figure 5.1. Distribution of Gower index and Podani index for the isolate Uk1. In red on the left the distribution of dissimilarities within the isolate Uk1. In grey the distributions of dissimilarities between all couples of lower-level units one belonging to Uk1 and the other belonging to a possibly different isolate.

5.1.1 Hierarchical FITSS clustering results

We apply the FITSS algorithms to the scrapie dataset using the alternative six methods listed in Table 3.1. We initialize each algorithm with a summary statistics for distances which are associated with the linkage method.

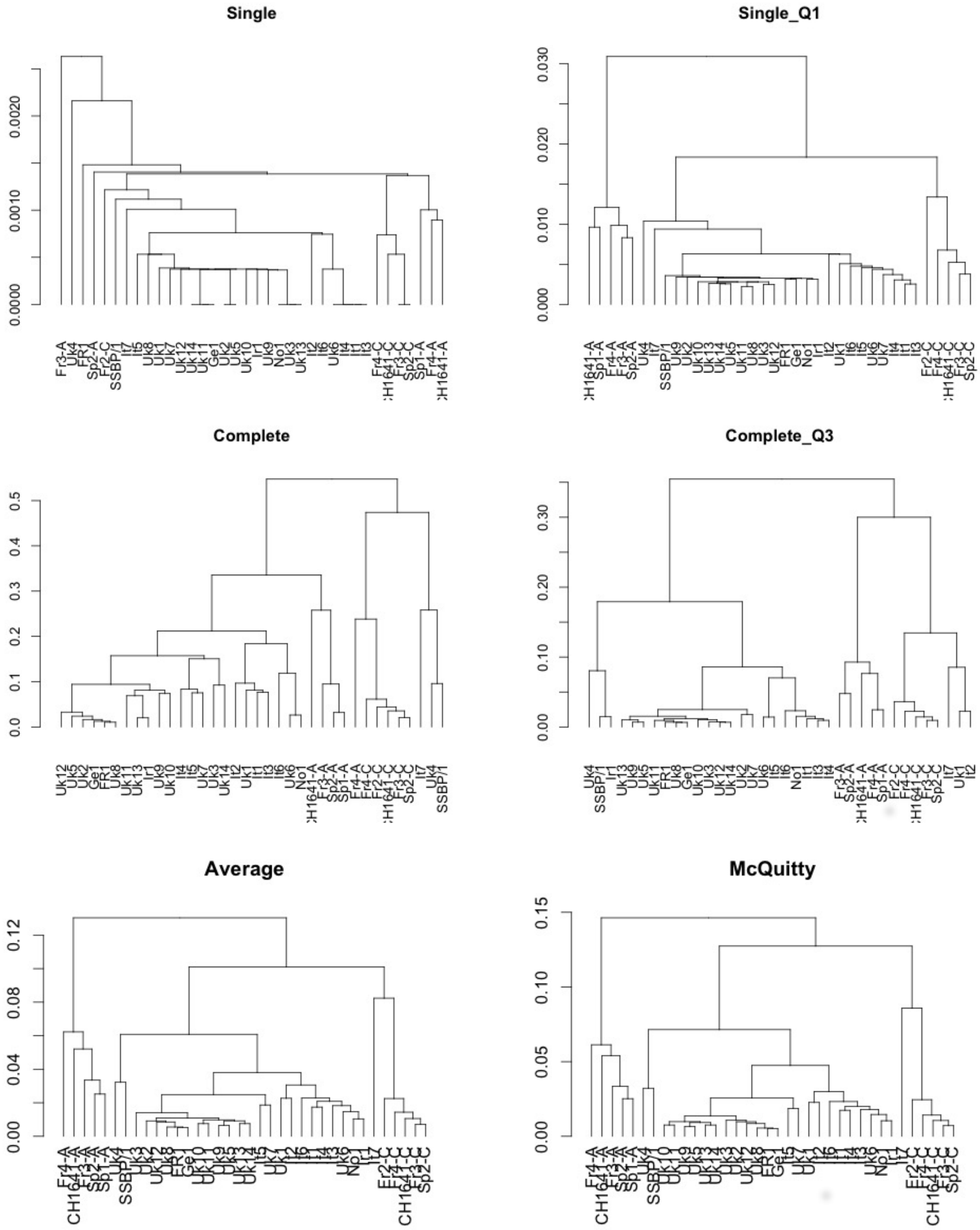


Figure 5.2. Dendrogram of hierarchical algorithms in Table 3.1

Figure 5.2 shows the resulting six dendrograms for all the considered methods. Despite some of the dendrograms appear to provide more similar evidence, others show some peculiar features: a chaining effect (units merging at every step in one unique elongated cluster) is apparent for the single linkage (SINGLE) whereas the same is less apparent for the complete linkage (COMPLETE) method. The chaining effect is not present when the FITSS initial distances are the first quartile for the single linkage (SINGLE_Q1) and the third quartile for the complete linkage (COMPLETE_Q3). Looking at each dendrogram separately we observe that the isolates TypeA and TypeC merge the other isolates at high distances. The isolates SSBP/1 and Uk4 are neighbors in all the dendrograms except for the single linkage methods (SINGLE and SINGLE_Q1).

Validation indexes lose importance with the multilevel data because observations which are different from each other are forced to be in the same cluster by the membership to the same higher-level unit. However, we calculate the internal validation indexes within each dendrogram: the Connectivity index is always minimized for a number of groups equal to 2; the Dunn index does not show homogeneous results, but it shows very low values for the SINGLE algorithm; the Silhouette index yields negative values for the SINGLE algorithm, while in the other methods presents the highest value for a number of groups equal to 3.

Even if the hierarchical methods show overall similarity, finally, we use consensus clustering to improve the quality and robustness of our analysis yielding a unique dendrogram. Figure 5.3 shows the resulting dendrogram for the consensus clustering obtained from combining the alternative hierarchical methods. We can not calculate the internal validation indexes for the consensus clustering results, so we consider to cut the dendrogram at different levels and compare the results with the RP. There is visual evidence of at least $k = 3$ groups (H3 partition in Table 5.1) with the possible further partitioning of two of them for a maximum of $k = 5$ subgroups (H5 partition in Table 5.1). In the latter finer partition there are two groups made of few isolates: one consists of a single isolate It7, whereas the other contains only Uk4 and SSBP/1. As we have observed in the singular results, isolates with “A” suffix aggregate each other at a distance of 0.06 and are in the last cluster aggregated at a distance of 0.14. Also, the isolates with “C” suffix are well isolated from the other units. The consensus clustering of hierarchical methods seems to aggregate the two groups Uk85 and It93 in a unique big group. These are also the two groups with the highest number of isolates and with the big sizes also in terms of lower-level units. The small sizes in isolates do not seem to affect the hierarchical clustering method.

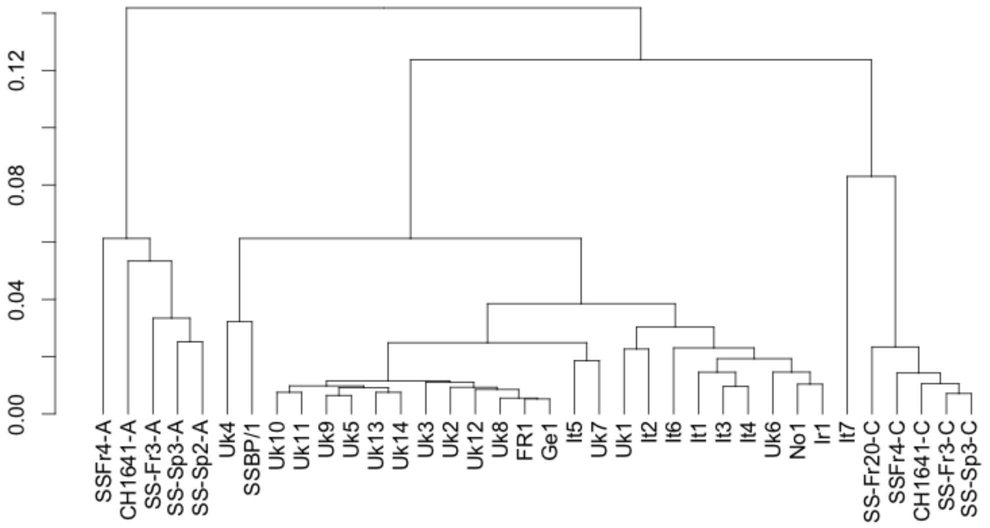


Figure 5.3. Dendrogram for consensus clustering of hierarchical methods.

5.1.2 Partitioning clustering results

We now rely on the Partitioning algorithm as explained in Section 3.3.2. To measure the distances between higher-level units we rely on several summaries of the distances between all the couples of lower-level units. We use the same summaries of the hierarchical algorithms: the minimum, the first quartile, the third quartile and the maximum and we consider the median instead of the average. Then we apply the PAM algorithm starting from all the distances. We compute the algorithm for all the partitions with a number of groups from 2 to 8. We include in the consensus clustering all partitions with $\tilde{s}(k) \geq 0.51$ (Figure 5.4) regardless of their number of groups. The partition with the highest number of groups has 6 clusters and is obtained using the First Quartile distance. Other partitions included for consensus clustering have from 2 up to 5 clusters for the other distances. The algorithm with initial maximum distance yields partitions for which the condition $\tilde{s}(k) \geq 0.51$ is not met for any k. The consensus clustering yields a final partition of 4 groups, denoted with PAM4 in Table 5.1.

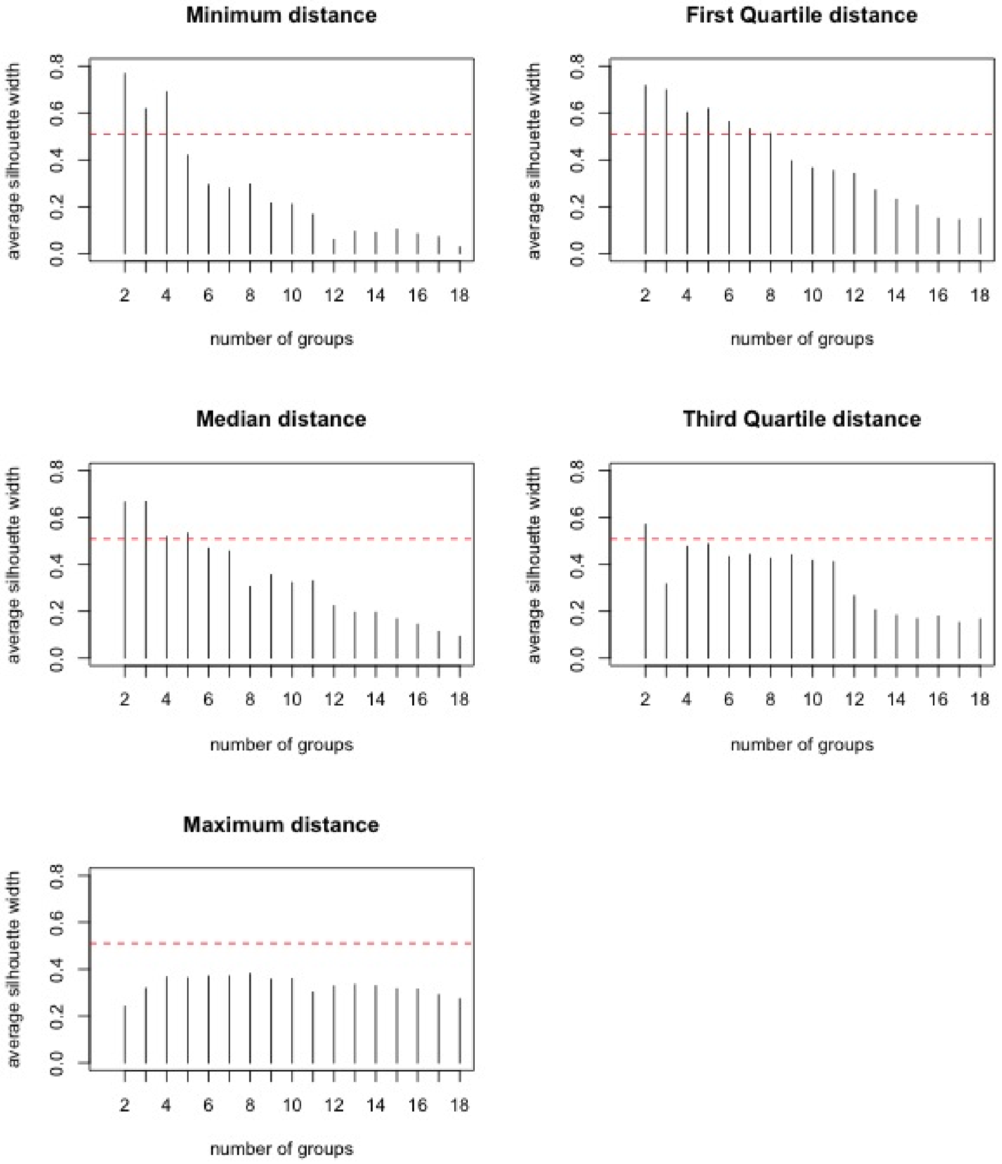


Figure 5.4. Silhouette indexes for the PAM results starting from different summaries of the distances.

5.1.3 Results for the distance-based methods with the Wasserstein distance

We apply distance-based algorithms using the Wasserstein distance introduced in Section 3.4 within the optimal transport plan paradigm. Firstly, we rely on hierarchical algorithms with the single, the complete, the average and the McQuitty linkages. We use consensus clustering to have a unique result. In Figure 5.5 we report the dendrogram for each of the 4 methods. The groups TypeA and TypeC are well separated from the other isolates by all the methods. In all dendrograms, we note that the isolate Uk4, which belongs to the SSPB1like group in RP, merges the other isolates at a higher distance than the other isolates without TypeA and TypeC suffixes.

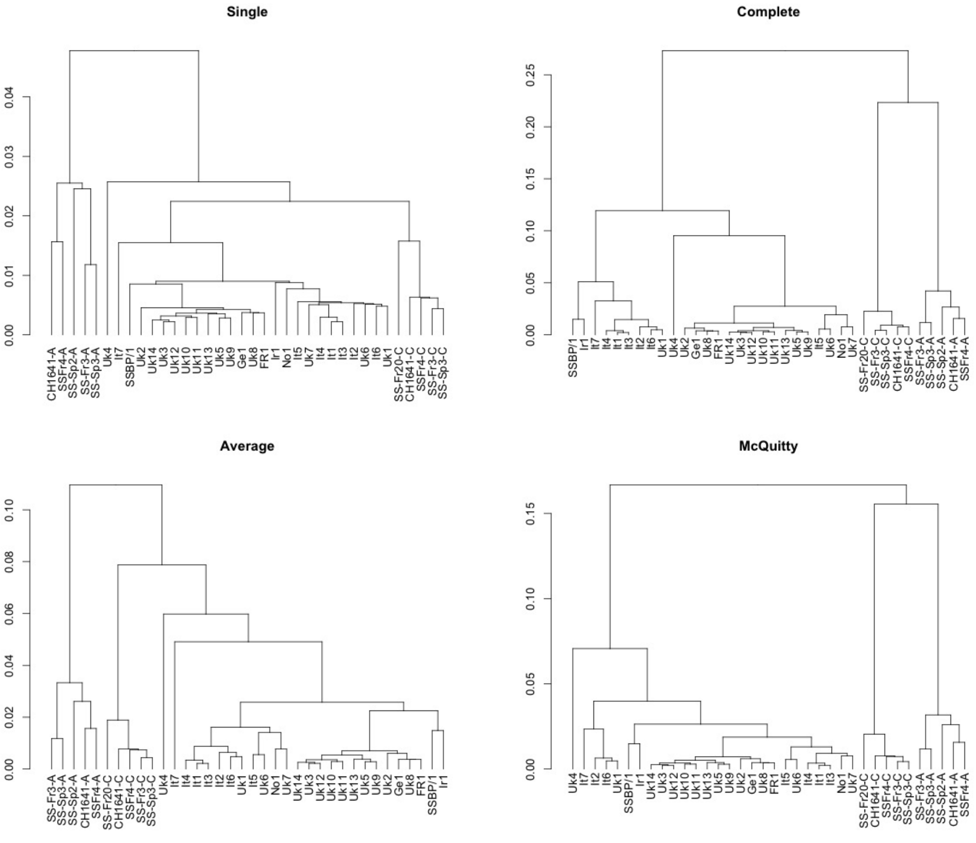


Figure 5.5. Dendrogram of hierarchical algorithms using the Wasserstein distance.

Figure 5.6 shows the final dendrogram obtained with the consensus clustering. The isolates with suffix TypeA and TypeC aggregates the other isolates only at the last step. The isolates Uk4 and the It7 seem to be different from the majority of the other units. The Uk4 forms the SSBP1like group with the SSBP/1 isolate in the RP partition, but in the dendrogram, the isolate Uk4 shows a relevant difference from all the other isolates. Also, It7 is far away from the others. Due to a visual inspection, the best number of groups is $k = 3$ (H3.W partition in Table 5.1). If we choose $k = 4$ we will have a group formed just by the Uk4 isolate, while if we choose $k = 5$ we will have also a group composed just of the It7 isolate.

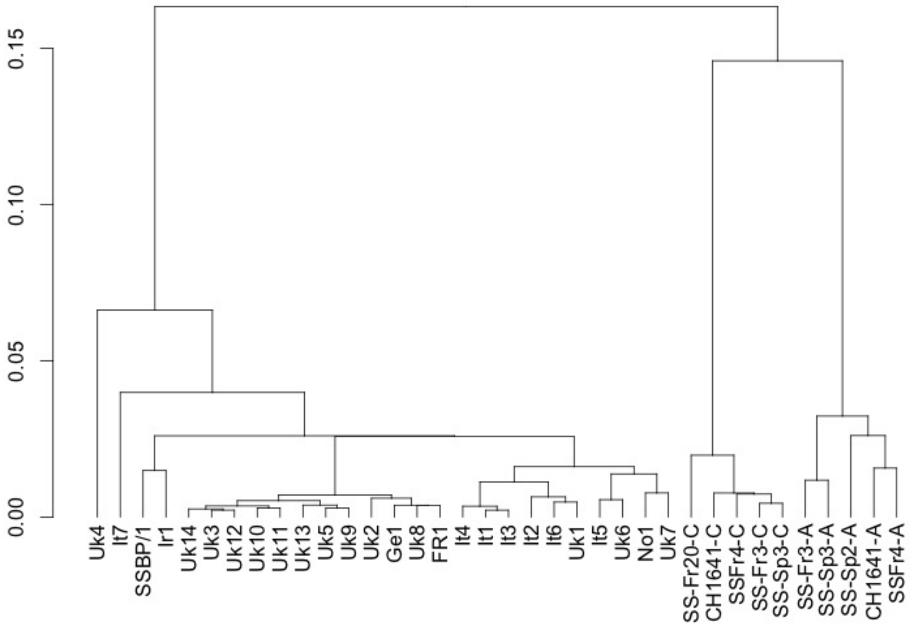


Figure 5.6. Dendrogram for consensus clustering of hierarchical methods with the Wasserstein distance.

Secondly, we use the Wasserstein distance also with the PAM algorithm. We compute the average silhouette index $\tilde{s}(k)$ (Figure 5.7) to choose the best number of groups. We choose to compare with the RP partition both the $k = 3$ and the $k = 4$ PAM results (PAM3.W and PAM4.W partitions in Table 5.1). Also, this results highlight the peculiar features of the TypeA and the TypeC isolates which form two distinct classes.

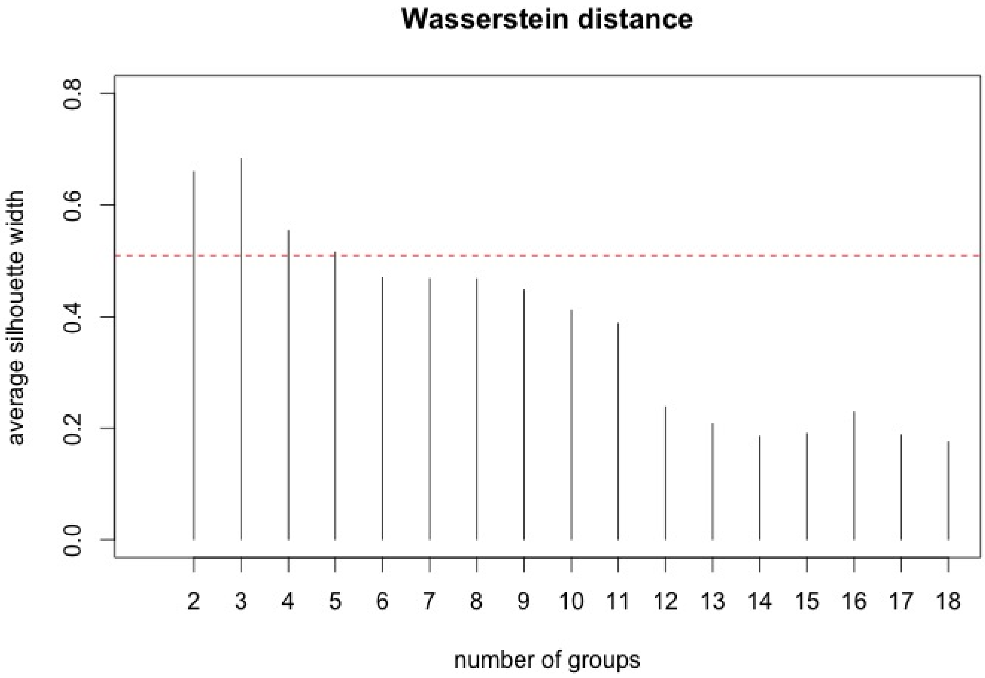


Figure 5.7. Silhouette indexes for the PAM algorithm with the Wasserstein distance.

5.2 Comparison of clustering results and the a priori researchers' partition

In Table 5.1 we compare all six different partitions (H3, H5, PAM4 and H3.W, PAM3.W, PAM4.W for Wasserstein distance), obtained with our cluster analyses, to a benchmark partition in 5 groups, that has been elicited a priori by researchers (RC: 1- SSPB1like, 2- Uk85, 3- It93, 4- TypeC, 5- TypeA). The H3, H3.W, and PAM3.W have obtained the same clustering results. All the clustering results distinguish TypeC and TypeA groups. The single isolate It7 is allocated in different groups over all the clustering results. The group SSPB1like is found only in the hierarchical method with 5 groups. The groups labeled as Uk85 and It93 are collapsed in the hierarchical methods, while in the consensus PAM with 4 groups they are partially found because Uk85 integrate also the SSPB1like and the isolate No1 belonging to It93; while in the PAM4.W No1 is correctly assigned to It93.

Overall these results have a different number of groups from the RP. The results seem to be in good agreement with the a priori partition although they suggest that the presence of a separated new group It7 could be appropriate. It7 segregation might be explained by the biological properties of the original isolate, composed of a pool of more brains of sheep affected with scrapie, instead of single brain. In Table 5.1 we report also the Adjusted Rand index. The different number of groups implies lower values of the index but they indicate a good agreement between the partition conceived by researchers and the clustering results. Finally, we propose a new partition denoted with NP which integrates the It7 class. Groups in H3 and H5 agree with the new proposal NP better than RC. While the PAM algorithms with 4 groups show a worse agreement in terms of the Adjusted Rand index. The NP classification will be validated via predictive performance using the supervised classification.

5.3 Supervised classification for the NP partition

After labeling each isolate with the class denoted by the NP partition, we evaluate the performance of alternative classifiers and combination methods using the accuracy index as a quantitative key indicator. Accuracy and its confidence interval for modal value, average probability, and HLMBC method are compared in Table 5.2. There is no substantial difference in term of accuracy in using the modal value methods rather than the average probability for each specific classifier. On the other hand the distribution of predictive groups for each isolate can be multi-modal and in this case, there is still some degree of uncertainty in classification. Indeed we consider a correct classification for the j-th isolate if the right group is within the modal values. This means that the accuracy in Table 5.2 is in fact slightly overstated as far as the modal value method is concerned. We report in Table 5.2 the mean percentage of times (%SMV) that there is a single modal value and the overall average number of modal values in each isolate (AMV).

Regardless of the way of exploiting group probabilities at lower-level the best single classifier is the Random Forest. On the other hand, all combination methods relying on group probabilities at lower-level are outperformed by the HLMBC method (Equation 4.7) which first aggregates at higher-level the lower-level group probability and then fits a model based on such higher-level predictors. The good accuracy consistently achieved by all methods can be regarded as a first solid validation of the NP partition. However, we could look at a finer level the predictive performance

Isolates	RP	H3	H5	PAM4	H3.W	PAM3.W	PAM4.W	NP
Uk4	1	1	1	1	1	1	1	1
SSBP/1	1	1	1	1	1	1	1	1
Uk12	2	1	2	1	1	1	1	2
Uk10	2	1	2	1	1	1	1	2
Uk11	2	1	2	1	1	1	1	2
Uk13	2	1	2	1	1	1	1	2
Uk14	2	1	2	1	1	1	1	2
Uk5	2	1	2	1	1	1	1	2
Uk8	2	1	2	1	1	1	1	2
Uk9	2	1	2	1	1	1	1	2
Fr1	2	1	2	1	1	1	1	2
Ge1	2	1	2	1	1	1	1	2
Ir1	2	1	2	1	1	1	1	2
Uk2	2	1	2	1	1	1	1	2
Uk3	2	1	2	1	1	1	1	2
No1	3	1	2	1	1	1	2	3
Uk6	3	1	2	2	1	1	2	3
Uk1	3	1	2	2	1	1	2	3
Uk7	3	1	2	2	1	1	2	3
It4	3	1	2	2	1	1	2	3
It5	3	1	2	2	1	1	2	3
It6	3	1	2	2	1	1	2	3
It1	3	1	2	2	1	1	2	3
It2	3	1	2	2	1	1	2	3
It3	3	1	2	2	1	1	2	3
It7	3	2	3	2	2	2	2	6
Sp2-C	4	2	4	3	2	2	3	4
Fr4-C	4	2	4	3	2	2	3	4
Fr3-C	4	2	4	3	2	2	3	4
Fr2-C	4	2	4	3	2	2	3	4
CH1641-C	4	2	4	3	2	2	3	4
CH1641-A	5	3	5	4	3	3	4	5
Fr3-A	5	3	5	4	3	3	4	5
Fr4-A	5	3	5	4	3	3	4	5
Sp1-A	5	3	5	4	3	3	4	5
Sp2-A	5	3	5	4	3	3	4	5
Adjusted Rand Index								
RP		0.403	0.524	0.795	0.403	0.403	0.894	0.956
NC	0.956	0.435	0.556	0.759	0.435	0.435	0.851	

Table 5.1. A priori classification of researchers (NP) compared with clustering results. A new proposal of partition (NP) is then compared with the others. The Adjusted Rand indexes are presented to investigate the similarity between groups. The (RP)groups are: 1- SSPB1like, 2- Uk85, 3- It93, 4- TypeC, 5- TypeA.

evaluation for each group. We consider the predictive performance for each group using Balanced Accuracy (BA), Sensitivity (SE) and Specificity (SP), (Table 5.3). These indexes help to validate the proposed NP partition and assess whether or not our proposed NP partition with a separated It7 group is supported. In fact, both Sensitivity and Specificity are more than satisfactory thus providing additional support to the partition and confirming the usefulness of our integrated approach. On the other hand, the class featuring a less impressive performance is the SSBP1like with SI$\leq$ 0.5 for all classifiers, except for the GBM with modal value method and the HLMBC method. However, the HLMBC method with an SI= 78.57% and the highest values of SP for all classifiers still allow to specifically validate this cluster.

Aggregation method for higher-level unit	Method for lower-level unit	Accuracy (95% CI)	%SMV	AMV
Average	Random Forest	98.11% (97.48%-98.62%)		
	Gradient Boosting	97.39% (96.67%-98.00%)		
	Neural Network	95.71% (94.82%-96.49%)		
	Convex combination	98.07% (97.43%-98.58%)		
	Model Combination	97.56% (96.86%-98.14%)		
Modal value	Random Forest	98.07% (97.43%-98.58%)	94.59%	1.02
	Gradient Boosting	97.35% (96.62%-97.96%)	89.18%	1.06
	Neural Network	95.58% (94.67%-96.37%)	91.89%	1.03
	Convex combination	97.35% (96.62%-97.96%)	94.59%	1.04
	Model Combination	97.27% (96.53%-97.88%)	94.59%	1.04
HLMBC method (4.7)	RF-GBM-NNET	98.87% (98.35%-99.25%)		

Table 5.2. Evaluation of classifiers' performance for the NP partition: Accuracy. %SMV is the mean percentage of times that there is a single modal value and AMV is the overall average number of modal values in each isolate.

5.4 Characterization of the final partition

Finally, we want to analyze the data belonging to each group of the final partition. The aim is to understand how the variables allow characterizing the groups. Especially we look at the data in relation with the clustering and supervised classification results. We have decided to rely on descriptive graphs to describe the distributions

Aggregation method for higher-level unit	Method for lower-level unit	Groups	Balanced Accuracy	Sensitivity	Specificity
Average	Random Forest	It93	98.63%	98.34%	98.91%
		SSBP1like	60.71%	21.43%	100.00%
		It7	75.00%	50.00%	100.00%
		TypeA	100.00%	100.00%	100.00%
		TypeC	99.93%	100.00%	99.85%
		Uk85	98.62%	98.91%	98.32%
	Gradient Boosting	It93	98.22%	98.34%	98.10%
		SSBP1like	64.28%	28.57%	100.00%
		It7	82.50%	65.00%	100.00%
		TypeA	99.81%	100.00%	99.63%
		TypeC	98.84%	97.83%	99.85%
		Uk85	97.93%	97.34%	98.52%
	Neural Network	It93	96.40%	94.70%	98.10%
		SSBP1like	50.00%	0.00%	100.00%
		It7	50.00%	0.00%	100.00%
		TypeA	95.37%	91.77%	98.97%
		TypeC	99.93%	100.00%	99.85%
		Uk85	98.42%	100.00%	96.84%
	Convex combination	It93	98.50%	98.01%	98.98%
		SSBP1like	60.71%	21.42%	100.00%
		It7	72.50%	45.00%	100.00%
		TypeA	100.00%	100.00%	100.00%
		TypeC	99.93%	100.00%	99.85%
		Uk85	98.74%	99.28%	98.19%
	Model Combination	It93	98.33%	98.56%	98.10%
		SSBP1like	50.00%	0.00%	100.00%
		It7	77.50%	55.00%	100.00%
		TypeA	100.00%	100.00%	100.00%
		TypeC	99.93%	100.00%	99.85%
		Uk85	97.80%	97.34%	98.26%
Modal value	Random Forest	It93	98.74%	99.45%	98.03%
		SSBP1like	67.85%	35.71%	100.00%
		It7	75.00%	50.00%	100.00%
		TypeC	100.00%	100.00%	100.00%
		TypeC	99.93%	100.00%	99.85%
		Uk85	98.22%	97.34%	99.10%
	Gradient Boosting	It93	97.94%	97.57%	98.30%
		SSBP1like	82.14%	64.29%	100.00%
		It7	85.00%	70.00%	100.00%
		TypeA	99.81%	100.00%	99.63%
		TypeC	98.84%	97.83%	99.85%
		Uk85	97.80%	97.34%	98.26%
	Neural Network	It93	96.35%	94.80%	97.89%
		SSBP1like	50.00%	0.00%	100.00%
		It7	50.00%	0.00%	100.00%
		TypeA	94.93%	90.95%	98.92%
		TypeC	99.93%	100.00%	99.85%
		Uk85	98.15%	99.40%	96.90%
	Convex combination	It93	97.80%	97.57%	98.03%
		SSBP1like	67.86%	35.71%	100.00%
		It7	75.00%	50.00%	100.00%
		TypeA	100.00%	100.00%	100.00%
		TypeC	99.93%	100.00%	99.85%
		Uk85	97.67%	97.34%	98.00%
	Model Combination	It93	97.80%	97.57%	98.03%
		SSBP1like	60.71%	21.43%	100.00%
		It7	75.00%	50.00%	100.00%
		TypeA	100.00%	100.00%	100.00%
		TypeC	99.93%	100.00%	99.85%
		Uk85	97.35%	97.34%	97.35%
HLMBC method (4.7)	RF-GBM-NNET	It93	99.02%	98.78%	99.25%
		SSBP1like	89.29%	78.57%	100.00%
		It7	92.50%	85.00%	100.00%
		TypeA	100.00%	100.00%	100.00%
		TypeC	99.95%	100.00%	99.90%
		Uk85	98.94%	98.79%	99.10%

Table 5.3. Evaluation of performance of the classifiers for the NP partition: Sensitivity, Specificity and Balanced Accuracy.

for both the survival time and the lesion profiles in 9 brain areas. Figure 5.8 shows the violin plots for the survival times of the bank voles in the 6 final groups. The TypeA class shows an asymmetric distribution which essentially overcomes the other distributions, with a median survival time equal to 145. Also, bank voles inoculated with TypeC isolates show higher survival times than the other remaining groups. The It7 isolate shows higher survival times than the It93 that is the group in which researchers classified the It7 in the RP. Finally the Uk85 and the SSPB1like shows similar distributions for the survival times with the same median equal to 72. Looking at these evidence for the survival time we can also find some relation with the clustering results: all the methods have identified clearly the TypeA and the TypeC groups, while many methods have not been able to detect the differences between It93 and Uk85.

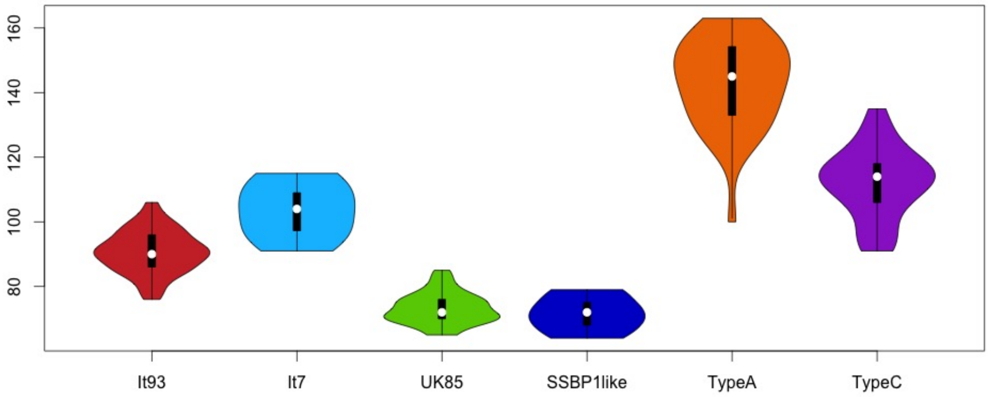


Figure 5.8. Violin plots for the Survival time in the 6 classes of the NP partition.

We study the discrete distributions of the lesion profile in the 9 areas with bubble plots (Figure 5.9) for the six final groups. The It93 (Figure 5.9(a)) shows a similar distribution to the Uk85 group (Figure 5.9(c)), and the modal value for each area are equal to the modal values of all the dataset as shown in Section 2.1. However, the It93 is more heterogeneous than the Uk85. The It7 (Figure 5.9(b)) shows a more variable distribution, and also low values for the A6 (hippocampus). The SSBP1like (Figure 5.9(d)) shows very heterogeneous results deviating from the dataset modal values especially for the A8 and A9 areas (retrosplenial and adjacent motor cortex;

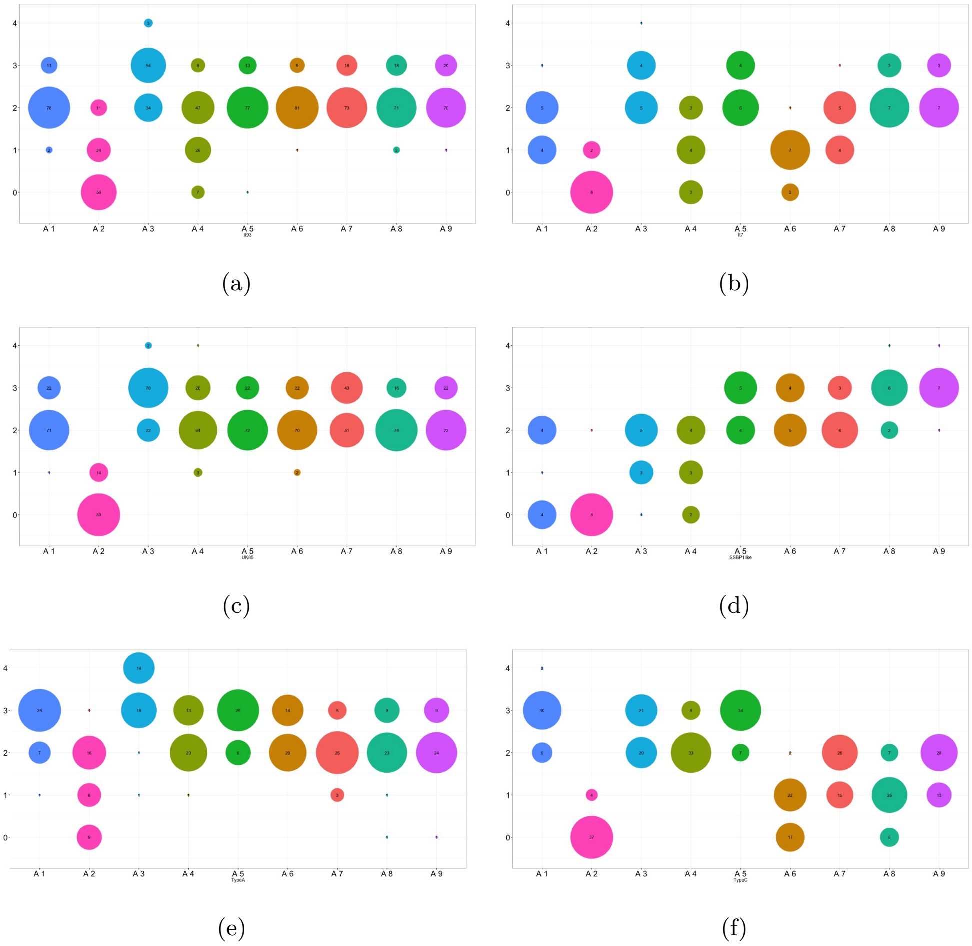
cingulate and adjacent motor cortex). The TypeA (Figure 5.9(e)) shows more bank voles with high values in all the areas with respect to the overall modal values. The difference is highlighted especially for the A2 area (cerebellum) which is not infected in the majority of the other isolates. The TypeC (Figure 5.9(f)) shows high values in the A1 (medulla), but lower values for the A6 and the A8 areas (hippocampus; retrosplenial and adjacent motor cortex). Again the descriptive analysis highlights the difference of the TypeA and the TypeC isolates which are variants arisen in the second passages of the experiment. The lesion profile seems to be different also for the It7 and the SSBP1like, the two groups which are not always separated in the cluster analysis and which also show relatively lower values of Sensitivity index in the supervised classification. These groups are also affected by the size problem in term both of isolates (higher-level units: 1 and 2 respectively) and bank voles (lower-level units: 10 and 9 respectively).

Finally, we study the correlations between the survival times and the ordinal variables of the lesion profile areas using the Spearman correlation index. Using the correlogram plot (Figure 5.10) we are also able to detect some "spatial" information. In fact, the areas are enumerated following a proximity criteria. The It93 group does not show significant correlations between the survival times and the scores evaluated in the nine areas. Significative correlations are highlighted in the quadrant composed of the areas from A3 to A9 in Figure 5.10(a). The A2 present significative negative correlations with no adjacent areas A4, A6, and A7. The It7 correlations (Figure 5.10(b)) show a significant negative correlation between the survival time and the scores in the A7 area. From A3 to A9 there are some significative positive correlations, especially for the two areas A8 and A9. The Uk85 group (Figure 5.10(c)) shows a high significative positive correlation only between the A8 and A9 areas. The survival time is positively correlated with the A3 and the A8 areas. SSBP1like group (Figure 5.10(d)) present a high positive correlation between the A1 and the A3 areas and also between A5 and A6. The survival time for the TypeA group (Figure 5.10(d)) is positively correlated with A3, A4, and A9, while the other correlations appear substantially a block excluding the A2 and the A6 areas. TypeC (Figure 5.10(e)) shows positive correlations between the survival times and the A5 area, while adjacent correlations appear from A7 to A9 areas.

In conclusion, the TypeA and TypeC groups stand out from the other groups both in term of survival time and in lesion profiles. For TypeA the highest survival times are correlated with the superior colliculus which is very lesioned with scores greater than 3. Also the hypothalamus and cingulate and adjacent motor cortex influence the survival time even if they do not present peculiar values in the distribution.

The It7 isolate group shows middle values for the survival time, which is negatively correlated with the lesions in the septum area, where several bank voles show lower values than the others belonging to the other different groups. Low values of survival time in the SSBP1like are relatively lower uncorrelated with scores in the lesion profiles, but in a short time, many bank voles are very damaged in the areas from A5 to A9.

It93 is similar to Uk85 in lesion profile values, but It93 has higher survival times than the Uk85. the correlation structures of these two groups are also different: the survival time is uncorrelated with the areas for It93, while in Uk85 is correlated with the superior colliculus and retrosplenial and adjacent motor cortex scores. Those two groups are merged into the hierarchical cluster, while are well separated from the partition method PAM.



(a)

(b)

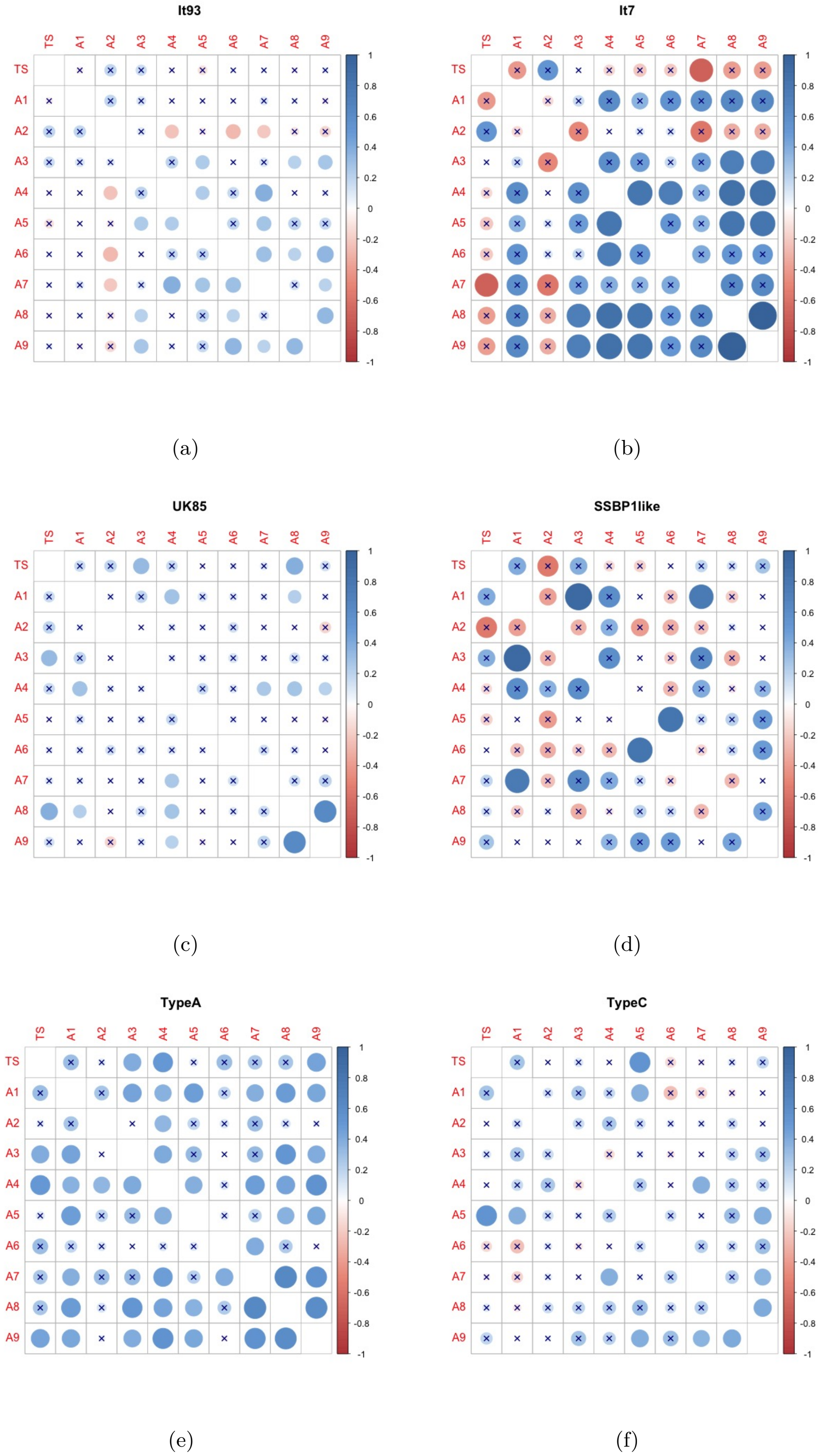
(c)

(d)

(e)

(f)

Figure 5.9. Distributional plots for lesion profile in 9 brain areas (A1-A9) for classes (a) It93 (b) It7 (c) Uk85 (d) SSBP1like (e) TypeA, (f) TypeC. A1=medulla, A2=cerebellum, A3=superior colliculus, A4=hypothalamus, A5=thalamus, A6=hippocampus, A7=septum, A8=retrosplenial and adjacent motor cortex, A9=cingulate and adjacent motor cortex.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5.10. Spearman correlation plots for survival times and lesion profile in 9 brain areas (A1-A9) for classes (a) It93 (b) It7 (c) Uk85 (d) SSBP1like (e) TypeA, (f) TypeC. Crosses indicate not-significative coefficients. TS=survival time A1=medulla, A2=cerebellum, A3=superior colliculus, A4=hypothalamus, A5=thalamus, A6=hippocampus, A7=septum, A8=retrosplenial and adjacent motor cortex, A9=cingulate and adjacent motor cortex.

Conclusions and future developments

The aim of this work is to provide appropriate tools for guiding researchers in understanding how data can be used to validate and possibly improve the understanding of an isolate partition, identifying strains or groups of strains. We discuss the problem of identifying and validating a putative "true partition" combining results from both unsupervised and supervised classification.

In the field of prion strains characterization, this study introduces methodologies that properly account for multilevel structures of transmission studies, mixed-type data, and heterogeneity. We have considered a multiple-step approach suggesting appropriate adjustments on both unsupervised and supervised classification steps. The distance-based cluster analysis is used in an explorative way and allows also to allows to match and integrate the starting partition with new information. In fact, a comprehensive cluster analysis points out a peculiar cluster, comprising a single isolate, It7. Although this finding was originally unexpected and overlooked in the partition proposed by the researcher experimental classification, it has represented an interesting contribution by the proposed extensive statistical analysis. Indeed, It7 segregation might be explained by the biological properties of the original isolate, which represents one of the two samples in our dataset (the other is Uk7) composed by a pool of more brains of sheep affected with scrapie, instead of a single brain. This explains the motivation for formulating another partition integrating this information and creating a new group with It7. The supervised classification is used to validate the new proposed partition with the predictive performance evaluation using the Accuracy index for the overall predictive performance and the Sensitivity and Specificity indexes for the groups. Good predictive performance provides further validation of the newly proposed group structure and we can use the best classifier for predicting new further discovered isolates within this partition.

In this work, we use the combination of easily understandable statistical tools to answer experimental questions about the similarity in prion disease strains. The proposed methods make fine-tuned use of strategies to account for the multilevel structure of the data and the unbalanced experimental design. We advocate the use of a data-driven approach with a mix of exploratory/confirmatory methods avoiding explicit probabilistic assumptions. We use the consensus clustering methods with distance-based algorithms to manage the multilevel structure of the data and the heterogeneity within higher-level units.

The main methodological issue faced in unsupervised classification is how one can measure distance between higher-level units also accounting for the variables' nature, as for example with our mixed-type data. We propose to summarize the distance between all the couples of lower-level units belonging to two different higher-level units. Using hierarchical algorithm, with a linkage coherent with the summary statistic chosen for the distances, is the first proposed method. This is similar to the FITSS approach proposed by Yeung et al. (2003), even if they do not explain how to force data into the same dendrogram. We avoid using a singular method, especially with the purpose of not relying on a single summary statistic for the distances between two higher-level units. We propose to use consensus clustering of different methods because of the heterogeneity of lower-level units within the same higher-level unit and also because two lower-level units of different higher-level units can be very close, even if the other lower-level units are far away from each other. Starting from this idea we also look for another way to measure the distance between two higher-level units. Hence we present a distance based on the optimal transport plan problem as the Wasserstein distance (a.k.a. Kantorovitch, Fortet-Mourier, Mallows, Earth Mover's distances etc..). For mixed-type data with unbalanced higher-level unit sizes, we consider a Wasserstein distance where the costs are measured by Gower or Podani dissimilarities, while the masses for the lower-level units in the higher-level unit j are taken equal to $1/n_j$.

We provide a contribution in supervised classification for multilevel data with the HLMBC method that simultaneously combines classifiers and aggregates the lower-level units using the average predictive probabilities. In this framework, we also propose the higher-level leave-one-out cross-validation that properly accounts for the multilevel nature of the dataset within the double cross-validation's paradigm. The HLMBC method, combined with this cross-validation, also allows for the problem of unbalanced data using the samples generated in the k-fold cross-validation. At the same time, we avoid overfitting because the testing set used is external to this k-fold cross-validation thanks to the double cross-validation's paradigm.

We are working on the development of a model-based clustering for multilevel mixed-type data within the Bayesian framework. Although we have tried to use the mixture of linear mixed-effect models combined with the latent variables models for mixed-type data, but we run into identification and overparameterization problems.

The methodologies proposed in this work are linked to the scrapie dataset. The small size of the data permits us to develop computationally expensive methods without bumping into the curse of dimensionality. In further developments, we want to challenge the Wasserstein distance with bigger datasets and compare it with the existing model-based methods. The supervised classification methods with the classifiers' combinations and the higher-level leave-one-out cross-validation algorithm are computationally expensive even with the scrapie dataset, so we would like to manage it to be suitable also for a bigger dataset.

Bibliography

Agrimi, U., Nonno, R., Dell'Omo, G., Di Bari, M. A., Conte, M., Chiappini, B., Esposito, E., Di Guardo, G., Windl, O., Vaccari, G. et al. (2008), 'Prion protein amino acid determinants of differential susceptibility and molecular feature of prion strains in mice and voles', *PLoS Pathog* **4**(7), e1000113.

Aitkin, M. (1999), 'A general maximum likelihood analysis of variance components in generalized linear models', *Biometrics* **55**(1), 117–128.

Arminger, G. & Küsters, U. (1988), Latent trait models with indicators of mixed measurement level, *in* 'Latent trait and latent class models', Springer, pp. 51–73.

Asparouhov, T. & Muthen, B. (2008), 'Multilevel mixture models', *Advances in latent variable mixture models* pp. 27–51.

Aurenhammer, F., Hoffmann, F. & Aronov, B. (1998), 'Minkowski-type theorems and least-squares clustering', *Algorithmica* **20**(1), 61–76.

Azzimonti, L., Ieva, F. & Paganoni, A. M. (2013), 'Nonlinear nonparametric mixed-effects models for unsupervised classification', *Computational Statistics* **28**(4), 1549–1570.

Bashashati, A. & Brinkman, R. R. (2009), 'A survey of flow cytometry data analysis methods', *Advances in bioinformatics* **2009**.

Bertsekas, D. P. (1988), 'The auction algorithm: A distributed relaxation method for the assignment problem', *Annals of operations research* **14**(1), 105–123.

Bertsekas, D. P. (1992), 'Auction algorithms for network flow problems: A tutorial introduction', *Computational optimization and applications* **1**(1), 7–66.

Billard, L. & Diday, E. (2003), 'From the statistics of data to the statistics of knowledge: symbolic data analysis', *Journal of the American Statistical Association* **98**(462), 470–487.

Bishop, C. M. (2006), 'Pattern recognition', *Machine Learning* **128**.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A. et al. (2000), 'Molecular classification of cutaneous malignant melanoma by gene expression profiling', *Nature* **406**(6795), 536–540.

Bock, H.-H. & Diday, E. (2012), *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*, Springer Science & Business Media.

Boutaga, K., Savelkoul, P. H., Winkel, E. G. & van Winkelhoff, A. J. (2007), 'Comparison of subgingival bacterial sampling with oral lavage for detection and quantification of periodontal pathogens by real-time polymerase chain reaction', *Journal of periodontology* **78**(1), 79–86.

Breckenridge, J. N. (1989), 'Replicating cluster analysis: Method, consistency, and validity', *Multivariate Behavioral Research* **24**(2), 147–161.

Brock, G., Pihur, V., Datta, S., Datta, S. et al. (2011), 'clvalid, an r package for cluster validation', *Journal of Statistical Software (Brock et al., March 2008)* .

Browne, R. P. & McNicholas, P. D. (2012), 'Model-based clustering, classification, and discriminant analysis of data with mixed type', *Journal of Statistical Planning and Inference* **142**(11), 2976–2984.

Bruce, M. E., Boyle, A., Cousens, S., McConnell, I., Foster, J., Goldmann, W. & Fraser, H. (2002), 'Strain characterization of natural sheep scrapie and comparison with bse', *Journal of General Virology* **83**(3), 695–704.

Cadez, I. V., McLaren, C. E., Smyth, P. & McLachlan, G. J. (1999), Hierarchical models for screening of iron deficiency anemia, *in* 'ICML', pp. 77–86.

Cagnone, S. & Viroli, C. (2012), 'A factor mixture analysis model for multivariate binary data', *Statistical Modelling* **12**(3), 257–277.

Cai, J.-H., Song, X.-Y., Lam, K.-H. & Ip, E. H.-S. (2011), 'A mixture of generalized latent variable models for mixed mode and heterogeneous data', *Computational Statistics & Data Analysis* **55**(11), 2889–2907.

Calò, D. G., Montanari, A. & Viroli, C. (2014), 'A hierarchical modeling approach for clustering probability density functions', *Computational Statistics & Data Analysis* **71**, 79–91.

Canale, A. & Dunson, D. B. (2011), 'Bayesian multivariate mixed-scale density estimation', *arXiv preprint arXiv:1110.1265* .

Carmona, C., Nieto-Barajas, L. & Canale, A. (2016), ‘Model based approach for household clustering with mixed scale variables’, *arXiv preprint arXiv:1612.00083* .

Cassard, H., Torres, J.-M., Lacroux, C., Douet, J.-Y., Benestad, S. L., Lantier, F., Lugan, S., Lantier, I., Costes, P., Aron, N. et al. (2014), ‘Evidence for zoonotic potential of ovine scrapie prions’, *Nature communications* **5**.

Celeux, G., Martin, O. & Lavergne, C. (2005), ‘Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments’, *Statistical Modelling* **5**(3), 243–267.

Chervoneva, I., Zhan, T., Iglewicz, B., Hauck, W. W. & Birk, D. E. (2012), ‘Two-stage hierarchical modeling for analysis of subpopulations in conditional distributions’, *Journal of applied statistics* **39**(2), 445–460.

Christian, D. C. (2002), ‘Computer-assisted analysis of oral brush biopsies at an oral cancer screening program’, *The Journal of the American Dental Association* **133**(3), 357–362.

De Soete, G. (1984), ‘A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix’, *Pattern Recognition Letters* **2**(3), 133–137.

Delicado, P. (2011), ‘Dimensionality reduction when data are density functions’, *Computational Statistics & Data Analysis* **55**(1), 401–420.

Desgraupes, B. (2013*a*), ‘clustercrit: Clustering indices’, *R package version* **1**(3).

Desgraupes, B. (2013*b*), ‘Clustering indices’, *University of Paris Ouest-Lab Modal’X* **1**, 34.

Di Bari, M. A., Chianini, F., Vaccari, G., Esposito, E., Conte, M., Eaton, S. L., Hamilton, S., Finlayson, J., Steele, P. J., Dagleish, M. P. et al. (2008), ‘The bank vole (myodes glareolus) as a sensitive bioassay for sheep scrapie’, *Journal of General Virology* **89**(12), 2975–2985.

Di Bari, M. A., Nonno, R. & Agrimi, U. (2012), ‘The mouse model for scrapie: inoculation, clinical scoring, and histopathological techniques’, *Amyloid Proteins: Methods and Protocols* pp. 453–471.

Di Bari, M. A., Nonno, R., Castilla, J., D’Agostino, C., Pirisinu, L., Riccardi, G., Conte, M., Richt, J., Kunkle, R., Langeveld, J. et al. (2013), ‘Chronic wasting disease in bank voles: characterisation of the shortest incubation time model for prion diseases’, *PLoS Pathog* **9**(3), e1003219.

Ding, C. & He, X. (2004), K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization, *in* 'Proceedings of the 2004 ACM symposium on Applied computing', ACM, pp. 584–589.

Duda, R. O., Hart, P. E. & Stork, D. G. (2012), *Pattern classification*, John Wiley & Sons.

Dunn, J. C. (1974), 'Well-separated clusters and optimal fuzzy partitions', *Journal of cybernetics* **4**(1), 95–104.

Dunson, D. B. (2000), 'Bayesian latent variable models for clustered mixed outcomes', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(2), 355–366.

Estivill-Castro, V. (2002), 'Why so many clustering algorithms: a position paper', *ACM SIGKDD explorations newsletter* **4**(1), 65–75.

Everitt, B. S. (1988), 'A finite mixture model for the clustering of mixed-mode data', *Statistics & probability letters* **6**(5), 305–309.

Everitt, B. S. & Merette, C. (1990), 'The clustering of mixed-mode data: a comparison of possible approaches', *Journal of Applied Statistics* **17**(3), 283–297.

Fawcett, T. (2006), 'An introduction to roc analysis', *Pattern recognition letters* **27**(8), 861–874.

Fox, J. (2002), 'Linear mixed models', *Appendix to An R and S-PLUS Companion to Applied Regression* .

Fraser, H. & Dickinson, A. (1968), 'The sequential development of the brain lesions of scrapie in three strains of mice', *Journal of comparative pathology* **78**(3), 301IN3311–310.

Fridlyand, J. & Dudoit, S. (2001), Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method, Technical report, Technical Report 600, Department of Statistics, UC Berkeley.

Friedman, J., Hastie, T. & Tibshirani, R. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer-Verlag New York.

Gelman, A. & Hill, J. (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge university press.

Gibbs, A. L. & Su, F. E. (2002), 'On choosing and bounding probability metrics', *International statistical review* **70**(3), 419–435.

Gini, C. (1921), 'Measurement of inequality of incomes', *The Economic Journal* **31**(121), 124–126.

Goldstein, H. (2011), *Multilevel statistical models*, Vol. 922, John Wiley & Sons.

Gollini, I. & Murphy, T. B. (2014), 'Mixture of latent trait analyzers for model-based clustering of categorical data', *Statistics and Computing* **24**(4), 569–588.

Gordon, A. D. (1999), *Classification*, 2nd edn, Chapman and Hall/CRC.

Gottschlich, C. & Schuhmacher, D. (2014), 'The shortlist method for fast computation of the earth mover's distance and finding optimal solutions to transportation problems', *PloS one* **9**(10), e110214.

Gower, J. C. (1971), 'A general coefficient of similarity and some of its properties', *Biometrics* pp. 857–871.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2001), 'On clustering validation techniques', *Journal of Intelligent Information Systems* **17**(2), 107–145.

Halkidi, M. & Vazirgiannis, M. (2001), Clustering validity assessment: Finding the optimal partitioning of a data set, *in* 'Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on', IEEE, pp. 187–194.

Halkidi, M., Vazirgiannis, M. & Batistakis, Y. (2000), Quality scheme assessment in the clustering process, *in* 'European Conference on Principles of Data Mining and Knowledge Discovery', Springer, pp. 265–276.

Handl, J., Knowles, J. & Kell, D. B. (2005), 'Computational cluster validation in post-genomic data analysis', *Bioinformatics* **21**(15), 3201–3212.

Hennig, C. (2015), 'What are the true clusters?', *Pattern Recognition Letters* .

Hornik, K. (2005), Cluster ensembles, *in* 'Classification—the Ubiquitous Challenge', Springer, pp. 65–72.

Hubert, L. & Arabie, P. (1985), 'Comparing partitions', *Journal of classification* **2**(1), 193–218.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D. et al. (2000), 'Functional discovery via a compendium of expression profiles', *Cell* **102**(1), 109–126.

Ignaccolo, R., Franco-Villoria, M. & Fassò, A. (2015), 'Modelling collocation uncertainty of 3d atmospheric profiles', *Stochastic environmental research and risk assessment* **29**(2), 417–429.

Irpino, A. & Romano, E. (2007), 'Optimal histogram representation of large data sets: Fisher vs piecewise linear approximations', *Revue des nouvelles technologies de l'information* **1**, 99–110.

Irpino, A. & Verde, R. (2006), A new wasserstein based distance for the hierarchical clustering of histogram symbolic data, *in* 'Data science and classification', Springer, pp. 185–192.

Irpino, A. & Verde, R. (2008), 'Dynamic clustering of interval data using a wasserstein-based distance', *Pattern Recognition Letters* **29**(11), 1648–1658.

Jaccard, P. (1908), *Nouvelles recherches sur la distribution florale.*

Jacques, J. & Preda, C. (2014), 'Functional data clustering: a survey', *Advances in Data Analysis and Classification* **8**(3), 231–255.

Jain, A. K. & Dubes, R. C. (1988), *Algorithms for clustering data*, Prentice-Hall, Inc.

Jöreskog, K. G. (1990), 'New developments in lisrel: Analysis of ordinal variables using polychoric correlations and weighted least squares', *Quality & Quantity* **24**(4), 387–404.

Kakourou, A., Vach, W. & Mertens, B. (2014), 'Combination approaches improve predictive performance of diagnostic rules for mass-spectrometry proteomic data', *Journal of Computational Biology* **21**(12), 898–914.

Kantorovitch, L. (1958), 'On the translocation of masses', *Management Science* **5**(1), 1–4.

Kaufman, L. & Rousseeuw, P. J. (2009), *Finding groups in data: an introduction to cluster analysis*, Vol. 344, John Wiley & Sons.

Kerr, M. K. & Churchill, G. A. (2001), 'Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments', *Proceedings of the national academy of sciences* **98**(16), 8961–8965.

Kim, J. & Billard, L. (2013), 'Dissimilarity measures for histogram-valued observations', *Communications in Statistics-Theory and Methods* **42**(2), 283–303.

Kohavi, R. et al. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection.

Lange, T., Roth, V., Braun, M. L. & Buhmann, J. M. (2004), 'Stability-based validation of clustering solutions', *Neural computation* **16**(6), 1299–1323.

LeBlanc, M. & Tibshirani, R. (1996), 'Combining estimates in regression and classification', *Journal of the American Statistical Association* **91**(436), 1641–1650.

Lee, S.-Y., Poon, W.-Y. & Bentler, P. (1990), 'Full maximum likelihood analysis of structural equation models with polytomous variables', *Statistics & probability letters* **9**(1), 91–97.

Li, C. & Wong, W. H. (2001), 'Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application', *Genome biology* **2**(8), research0032–1.

Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. (2010), Understanding of internal clustering validation measures, *in* '2010 IEEE International Conference on Data Mining', IEEE, pp. 911–916.

Lubke, G. & Neale, M. C. (2006), 'Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood?', *Multivariate Behavioral Research* **41**(4), 499–532.

Luenberger, D. G. & Ye, Y. (2015), *Linear and nonlinear programming*, Vol. 228, Springer.

Lugli, E., Pinti, M., Nasi, M., Troiano, L., Ferraresi, R., Mussi, C., Salvioli, G., Patsekin, V., Robinson, J. P., Durante, C. et al. (2007), 'Subject classification obtained by cluster analysis and principal component analysis applied to flow cytometric data', *Cytometry Part A* **71**(5), 334–344.

Mallows, C. (1972), 'A note on asymptotic joint normality', *The Annals of Mathematical Statistics* pp. 508–515.

Mangasarian, O. L., Street, W. N. & Wolberg, W. H. (1995), 'Breast cancer diagnosis and prognosis via linear programming', *Operations Research* **43**(4), 570–577.

Masters, G. N. (1982), 'A rasch model for partial credit scoring', *Psychometrika* **47**(2), 149–174.

McParland, D. & Gormley, I. C. (2013), Clustering ordinal data via latent variable models, *in* 'Algorithms from and for Nature and Life', Springer, pp. 127–135.

McParland, D. & Gormley, I. C. (2016), 'Model based clustering for mixed data: clustmd', *Advances in Data Analysis and Classification* **10**(2), 155–169.

McParland, D., Gormley, I. C., McCormick, T. H., Clark, S. J., Kabudula, C. W. & Collinson, M. A. (2014), 'Clustering south african households based on their asset status using latent variable models', *The annals of applied statistics* **8**(2), 747.

Medvedovic, M. & Sivaganesan, S. (2002), 'Bayesian infinite mixture model based clustering of gene expression profiles', *Bioinformatics* **18**(9), 1194–1206.

Mérigot, Q. (2011), A multiscale approach to optimal transport, *in* 'Computer Graphics Forum', Vol. 30, Wiley Online Library, pp. 1583–1592.

Mertens, B. (1998), 'Exact principal component influence measures applied to the analysis of spectroscopic data on rice', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **47**(4), 527–542.

Mertens, B. (2001), 'Downdating: interdisciplinary research between statistics and computing', *Statistica Neerlandica* **55**(3), 358–366.

Mertens, B. J. (2003), 'Microarrays, pattern recognition and exploratory data analysis', *Statistics in medicine* **22**(11), 1879–1899.

Mertens, B. J., Noo, M. D., Tollenaar, R. A. & Deelder, A. M. (2006), 'Mass spectrometry proteomic diagnosis: Enacting the double cross-validatory paradigm', *Journal of Computational Biology* **13**(9), 1591–1605.

Milligan, G. W. & Cooper, M. C. (1986), 'A study of the comparability of external criteria for hierarchical cluster analysis', *Multivariate Behavioral Research* **21**(4), 441–458.

Monge, G. (1781), *Mémoire sur la théorie des déblais et des remblais*, De l'Imprimerie Royale.

Morlini, I. (2012), 'A latent variables approach for clustering mixed binary and continuous variables within a gaussian mixture model', *Advances in Data Analysis and Classification* **6**(1), 5–28.

Muthén, B. (1984), 'A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators', *Psychometrika* **49**(1), 115–132.

Muthén, B. & Shedden, K. (1999), 'Finite mixture modeling with mixture outcomes using the em algorithm', *Biometrics* **55**(2), 463–469.

Ng, S.-K., McLachlan, G. J., Wang, K., Ben-Tovim Jones, L. & Ng, S.-W. (2006), 'A mixture model with random-effects components for clustering correlated gene-expression profiles', *Bioinformatics* **22**(14), 1745–1752.

Nonno, R., Di Bari, M. A., Cardone, F., Vaccari, G., Fazzi, P., Dell'Omo, G., Cartoni, C., Ingrosso, L., Boyle, A., Galeno, R. et al. (2006), 'Efficient transmission and characterization of creutzfeldt–jakob disease strains in bank voles', *PLoS Pathog* **2**(2), e12.

Pirisinu, L., Di Bari, M. A., D'Agostino, C., Marcon, S., Riccardi, G., Poleggi, A., Cohen, M. L., Appleby, B. S., Gambetti, P., Ghetti, B. et al. (2016), 'Gerstmann-sträussler-scheinker disease subtypes efficiently transmit in bank voles as genuine prion diseases', *Scientific reports* **6**.

Podani, J. (1999), 'Extending gower's general coefficient of similarity to ordinal characters', *Taxon* pp. 331–340.

Ramsay, J. O. (2006), *Functional data analysis*, Wiley Online Library.

Ranalli, M. G., Rocco, G., Jona Lasinio, G., Moroni, B., Castellini, S., Crocchianti, S. & Cappelletti, D. (2016), 'Functional exploratory data analysis for high-resolution measurements of urban particulate matter', *Biometrical Journal* **58**(5), 1229–1247.

Ranalli, M. & Rocci, R. (2016), 'Mixture models for ordinal data: a pairwise likelihood approach', *Statistics and Computing* **26**(1-2), 529–547.

Ranalli, M. & Rocci, R. (2017), 'Mixture models for mixed-type data through a composite likelihood approach', *Computational Statistics & Data Analysis* .

Rand, W. M. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical association* **66**(336), 846–850.

Rasch, G. (1960), 'Probabilistic models for some intelligence and achievement tests', *Copenhagen: Danish Institute for Educational Research* .

Rousseeuw, P. J. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65.

Sakurai, Y., Li, L., Chong, R. & Faloutsos, C. (2008), Efficient distribution mining and classification, *in* 'Proceedings of the 2008 SIAM international conference on data mining', SIAM, pp. 632–643.

Samejima, F. (1969), 'Estimation of latent ability using a response pattern of graded scores.', *Psychometrika monograph supplement* .

Sammel, M. D., Ryan, L. M. & Legler, J. M. (1997), 'Latent variable models for mixed discrete and continuous outcomes', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**(3), 667–678.

Schuhmacher, D., Bähre, B., Gottschlich, C., Heinemann, F. & Wilm, T. (2017), 'Package 'transport''.

Sciubba, J. J., Group, U. C. O. S. et al. (1999), 'Improving detection of precancerous and cancerous oral lesions: computer-assisted analysis of the oral brush biopsy', *The Journal of the American Dental Association* **130**(10), 1445–1457.

Secchi, P., Vantini, S. & Vitelli, V. (2015), 'Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of milan', *Statistical Methods & Applications* **24**(2), 279–300.

Spellman, E., Vemuri, B. C. & Rao, M. (2005), Using the kl-center for efficient and accurate retrieval of distributions arising from texture images, *in* 'Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on', Vol. 1, IEEE, pp. 111–116.

Stone, M. (1974), 'Cross-validatory choice and assessment of statistical predictions', *Journal of the royal statistical society. Series B (Methodological)* pp. 111–147.

Swartz, R. J., West, L. A., Boiko, I., Malpica, A., Guillaud, M., MacAulay, C., Follen, M., Atkinson, E. N. & Cox, D. D. (2005), 'Classification using the cumulative log-odds in the quantitative pathologic diagnosis of adenocarcinoma of the cervix', *Gynecologic oncology* **99**(3), S24–S31.

Tanaka, H. (1973), 'An inequality for a functional of probability distributions and its application to kac's one-dimensional model of a maxwellian gas', *Probability Theory and Related Fields* **27**(1), 47–52.

Terada, Y. & Yadohisa, H. (2010), Non-hierarchical clustering for distribution-valued data, *in* 'Proceedings of COMPSTAT', pp. 1653–1660.

Thiran, J.-P. & Macq, B. (1996), 'Morphological feature extraction for the classification of digital images of cancerous tissues', *IEEE Transactions on biomedical engineering* **43**(10), 1011–1020.

Thurstone, L. L. (1925), 'A method of scaling psychological and educational tests.', *Journal of educational psychology* **16**(7), 433.

Tibshirani, R. & Walther, G. (2005), 'Cluster validation by prediction strength', *Journal of Computational and Graphical Statistics* **14**(3), 511–528.

Tsybrovskyy, O. & Berghold, A. (1999), 'Primary unit for statistical analysis in morphometry: patient or cell?', *Analytical Cellular Pathology* **18**(4), 191–202.

Vaserstein, L. N. (1969), 'Markov processes over denumerable products of spaces, describing large systems of automata', *Problemy Peredachi Informatsii* **5**(3), 64–72.

Vermunt, J. K. (2008), 'Latent class and finite mixture models for multilevel data sets', *Statistical Methods in Medical Research* **17**(1), 33–51.

Vermunt, J. K. & Magidson, J. (2005), Hierarchical mixture models for nested data structures, *in* 'Classification—the Ubiquitous Challenge', Springer, pp. 240–247.

Villani, C. (2008), *Optimal transport: old and new*, Vol. 338, Springer Science & Business Media.

Volkovich, Z., Barzily, Z., Avros, R. & Toledano-Kitai, D. (2009), On application of the k-nearest neighbors approach for cluster validation, *in* 'Proceeding of the XIII International Conference Applied Stochastic Models and Data Analysis (ASMDA 2009), Vilnius', pp. 468–472.

Vrac, M., Billard, L., Diday, E. & Chédin, A. (2012), 'Copula analysis of mixture models', *Computational Statistics* pp. 1–31.

Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. (2016), *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Wolpert, D. H. (1992), 'Stacked generalization', *Neural networks* **5**(2), 241–259.

Yamal, J.-M., Follen, M., Guillaud, M. & Cox, D. D. (2011), 'Classifying tissue samples from measurements on cells with within-class tissue sample heterogeneity', *Biostatistics* **12**(4), 695–709.

Yamal, J.-M., Guillaud, M., Atkinson, E. N., Follen, M., MacAulay, C., Cantor, S. B. & Cox, D. D. (2015), 'Prediction using hierarchical data: Applications for automated detection of cervical cancer', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **8**(2), 65–74.

Yeung, K. Y., Haynor, D. R. & Ruzzo, W. L. (2001), 'Validating clustering for gene expression data', *Bioinformatics* **17**(4), 309–318.

Yeung, K. Y., Medvedovic, M. & Bumgarner, R. E. (2003), 'Clustering gene-expression data with repeated measurements', *Genome biology* **4**(5), R34.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.

List of Figures

List of Tables

Ringraziamenti

Il mio primo grazie va al Prof. Luca Tardella, supervisore di questa tesi. Grazie per l'aiuto, le mille idee ed i tanti dibattiti "filosofici" sulla statistica e per i preziosi consigli su come conciliare la metodologia con la realtà dei dati.

Ringrazio la Prof.ssa Giovanna Jona Lasinio per avermi insegnato a fare ricerca partendo dai dati e per avermi fatto scoprire che in fondo parlare davanti agli altri non è poi così male!

Ringrazio tutti i miei colleghi di dottorato con cui ho condiviso corsi, pranzi e tante lunghe giornate. In particolare grazie ad Alberto per avermi impedito di alzarmi al momento opportuno.

Un grazie speciale va ai miei genitori, sperando (al terzo ringraziamento) di averli ripagati per tutto quello che hanno fatto per me. Grazie anche a mia sorella e a Maurizio. Grazie a Viola e Tommaso, sperando che tra qualche anno saranno loro a ringraziare me!

L'ultimo grazie va a Claudio, il mio più grande sostenitore, che finalmente dal prossimo anno sarà parte della mia famiglia in modo ufficiale.