Multiple Sclerosis Classification via Random Forest Distances Robust to Missing Data

Antonella Mensi[1(✉)], Vincenzo di Maria[2], Elena Barusolo[1],
Roberta Magliozzi[1,3], and Manuele Bicego[2]

[1] Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy
{antonella.mensi,elena.barusolo,roberta.magliozzi}@univr.it
[2] Department of Computer Science, University of Verona, Verona, Italy
vincenzo.dimaria_02@studenti.univr.it, manuele.bicego@univr.it
[3] Department of Brain Sciences, Faculty of Medicine, Imperial College London, London, UK

Abstract. Random Forest (RF) distances represent powerful data-dependent measures, which usefulness has been shown in many different contexts. Very recently, they have been extended to deal with missing data, a crucial problem which affects many biomedical domains. However, their characterization has been mainly theoretical, with simple experiments aimed at showing that the resulting pairwise distances, in the presence of missing data, are nearly equivalent to those computed with complete data, without focusing on a specific task like classification or clustering. In this paper, we take a fundamental step forward in their evaluation, showing their usefulness in a challenging real-world scenario: the classification of Multiple Sclerosis (MS) according to the levels of various cerebrospinal fluid (CSF) protein markers. We based our experiments on MissRatioRF, a state-of-the-art RF-distance adapted for missing data. A thorough experimental evaluation on real data shows that this RF-distance outperforms all other state-of-the-art distances, for many–even severe–degrees of missingness.

Keywords: Random Forest Distances · Missing data · Multiple Sclerosis

1 Introduction

Missing data, i.e., unknown values for a particular feature in one or more subjects, is a pervasive issue across many different fields. However, its gravity is highly dependent on the specific context: for example, in the healthcare domain, it is crucial to adequately manage this issue. Indeed, in healthcare, missing data are common and may arise from several causes [17]. These, among others, include: i) manual data collection, e.g., a physician while visiting patients might

E. Rodolà et al. (Eds.): ICIAP 2025, LNCS 16168, pp. 90–102, 2026.
https://doi.org/10.1007/978-3-032-10192-1_8

not be able to collect all required data; ii) data digitalization, e.g., nowadays clinical reports are digitalized, but human errors can result in missing entries, or some parts of the report might be unreadable; iii) sensor-related issues, e.g., measured values might be missing or incorrect, requiring their removal. Furthermore, healthcare datasets are often small and sensitive in nature, and even a few missing values may have an impact on the reliability of downstream tasks such as diagnosis, prognosis, or treatment recommendation.

From the above example, it is clear that missing data must be dealt with carefully. There exist several approaches to do so, the most well-known being data removal and data imputation [22]. Both methodologies are widely employed, but can be disadvantageous if the adoption choice is not adequately pondered. In detail, data removal decreases the overall size of the dataset, which may result in an inaccurate representation of the original data, whereas data imputation, by estimating missing values, can introduce bias in the data distribution. Clearly, a better alternative would be to devise techniques that are robust to missing data, i.e., which can be employed without resorting to imputation or data deletion. A clear example is represented by the adaptation of Support Vector Machines and Naive Bayes to work with missing data [21,23]. In these approaches, missing data are handled via an adaptation of the training process. With respect to the task of distance computation, some approaches have been presented which are robust to missing data [1–3,5,10,11,26,27]: in detail, even if an object has a large number of missing values, its distance to other instances can be computed. Most of these techniques work by setting the distance between two objects to 0 along the feature(s) with at least one missing value, that is, the feature does not contribute to the final distance value. This is rather common for data-independent metrics, e.g. variants of the Euclidean distance [2,3,26,27]. More recent data-dependent approaches [5] treat missing values differently, i.e., they do not simply compare the values of the two objects along each feature.

The proposal made in [5] is framed within the context of RF-distances. Historically, RFs have been widely and successfully used for a variety of learning tasks, e.g., classification and clustering [9,28]. However, in recent years, they have also been explored as tools for computing data-dependent distances [6,7,29,31], i.e., functions embedding contextual information into the computation. In detail, RF-distances are based on the concept that RFs are space partitioners, and the distance between two objects can be computed by comparing the paths they traverse in each tree of the forest. The simplest approach [7,29] defines two objects as similar only if they reach the same leaf, i.e., in any other case their distance is maximum. Instead, state-of-the-art RF-distances, such as RatioRF [6], compare the entire paths traversed by the objects. In detail, RatioRF evaluates how the objects answer to the tests associated with the traversed nodes, i.e., the similarity between two objects is proportional to the number of tests on which they agree relative to the overall number of traversed nodes.

Variants of RF-distances robust to missing data have recently been proposed in [5]. In detail, the authors generalize the distance functions and propose adaptations of tree construction and traversal processes. Specifically, they define

MissRatioRF, an extension of RatioRF, and demonstrate its robustness by showing the equivalence between distance matrices computed on complete data and those computed on data with missing values, and by comparing this information to other metrics. However, the authors do not explore the possibility of using MissRatioRF to perform a learning task, which could further support the effectiveness of the methodology they propose.

In this paper, our aim is to fill this gap by using MissRatioRF in a real-life scenario: the classification of 37 patients as healthy or MS, according to 75 CSF markers. This task is not only clinically relevant, but also challenging. Indeed, several CSF markers are significantly altered in MS subjects compared to healthy ones, with some, such as oligoclonal bands (OCBs) and kappa free light chains (kFLCs), included in the McDonald diagnostic criteria [18,30]. However, the related classification task is challenging due to several factors: limited data availability, the inherent characteristics of the dataset–typically small sized and permeated with missing data–, and the lack of machine learning-based approaches in the MS domain that employ CSF data–further details will be provided in Sect. 3. Given this context, we compare the classification performance of MissRatioRF with several other distance functions robust to missing data. In addition, we study the effect of various missing rates by introducing artificial missing values using the Missing Completely at Random (MCAR) protocol [25].

The rest of the paper is organized into four sections. In Sect. 2 we provide an overview of RF-distances and their extension robust to missing data, and also briefly present the proposed pipeline. Then in Sect. 3 we describe our dataset and the related classification problem. Sect. 4 is dedicated to the experimental evaluation, and lastly in Sect. 5 we draw some conclusions.

2 Robust RF-distances

This section is divided into three subsections: Sect. 2.1 provides an overview of RF-distances, Sect. 2.2 describes extensions that handle missing data, and Sect. 2.3 briefly outlines the pipeline used in this work.

2.1 RF-distances

As briefly mentioned in Sect. 1, RF-distances [6,29,31,34] are part of a wider class of distance functions known as data-dependent distance functions [32], which differently from data-independent ones take into account the context, e.g., the density of the surrounding space. For example, two pairs of objects which Euclidean distance is equal, may have a different distance according to data-dependent functions. In detail, if one pair is surrounded by fewer objects it will have a higher similarity. In addition to being theoretically robust and highly descriptive, RF-distances have been proven to be extremely successful in a variety of contexts [16,24,28]. The reason behind their goodness is also linked to the innate nature of a Random Forest as a flexible data description tool: each forest comprises multiple decision trees, where each tree consists of tests on features that result in data partitioning, since each object can answer either true or false to the test. In essence, RFs are able to capture the relationship between two objects, i.e., compute their distance, by: i) having both objects walk down all trees, root to leaf, and collect their paths; ii) comparing the two paths in each tree using a specific distance function. Typically, most RF-distances consider two objects to be closer, relative to how the tree has partitioned the input space, if the traversed paths are more similar. Indeed, the simplest RF-distance [7,29] extremizes this concept by assigning a similarity of 1 to two objects if they end up in the same leaf, 0 otherwise. The final value is the average across all trees.

Over the years, more complex functions have been proposed, such as **RatioRF** [6], a very recent and well-performing RF-distance. In its computation, RatioRF considers all nodes of the traversed paths, and it also integrates Tversky's idea [32] that the similarity between two objects is characterized by both their commonalities and their differences. More in detail, to compute the RatioRF distance between two objects, two elements must be considered: i) the number of tests associated to the nodes traversed by the two objects in their root to leaf paths; ii) the number of tests, among those identified by i), to which the two objects answer in the same way. The ratio between ii) and i) is the RatioRF similarity within a tree, which, in accordance with Tversky, takes into account both the similarities and the differences between the two objects. The distance at forest level is obtained by first computing the average similarity across all trees and then converting it into a distance. RatioRF, however, is not able to handle missing data, i.e., missing values should either be removed or imputed prior to distance computation.

2.2 RF-distances Robust to Missing Data

As briefly mentioned in Sect. 1, data removal and data imputation are the most adopted solutions to manage missing data [22]. Data removal consists of either removing the samples that present at least one missing value, or the features which measurement is missing for at least one instance. Both choices lead to a decrease in the dataset size, even if along different directions, which then may lead to an incomplete dataset representation, e.g., if a highly discriminant feature gets removed. The second solution consists of imputing data, that is, replacing unknown values with an estimate. A very simple example is to populate missing values along one feature using a summary statistic, e.g., mean, median or mode, but more complex methodologies exist [14,20]. Data imputation might be a good choice if the mechanisms causing missing data are known, otherwise these approaches will not only yield poor estimates, but will also introduce unwanted bias in the distribution of the affected feature [26]. It is therefore evident that the disadvantages behind these approaches are highly impacting, and better alternatives should be adopted.

One such alternative consists of the recent approach proposed in [5] in which they provide a generalization of several RF-distances robust to missing data, with a focus on **MissRatioRF**, an extension of RatioRF. In detail, to adequately extend RF-distances, they: i) propose alternative tree building and traversal mechanisms; ii) modify the notion of traversed path; iii) define an alternative aggregation function to combine the similarities at forest level.

As to the tree building procedure, consider the test associated to a node within a tree and an object that has a missing value along the feature used within the test: it must be established if and how the object will be used to build the rest of the tree. The authors design two alternative approaches: *NanBoth* and *NanStop*. *NanBoth* makes the object go down both the left and right paths, the rationale being that the information is not sufficient to choose only one of them, i.e., to separate the object from the rest. The second approach, *NanStop*, stops the propagation of the object, the principle being that a test should be defined only on existing values. Whereas the first approach will likely produce bigger trees, the latter will do the converse; however, neither technique will affect the rest of the tree building process. Instead, in relation to the traversal of an existing tree, the authors propose using only *NanBoth*. Indeed, the *NanStop* technique would likely produce a path too short for objects with missing values, making the distance computation less reliable. In addition to these modifications, the authors revise the formulation of RatioRF and other RF-distances by replacing the concept of traversed paths with that of traversed subtrees, i.e., multiple root to leaf paths where each test is considered only once. Indeed if an object encounters a test on a feature for which its value is unknown, it will reach more than one leaf. Lastly, the authors propose a novel way, other than the average, to combine the tree level similarities at forest level based on the key concept that each tree contributes differently depending on the amount of missing values used during its construction, i.e., they propose a weighted average. In detail, they define the information content of a tree with respect to a given pair of objects as the ratio between the number of tests that define the subtrees traversed by the object, and the total number of tests summed across all trees. The information content is then used as weight within the aggregation function.

2.3 Classification with RF-distances Robust to Missing Data

The pipeline we propose to solve the classification task consists of two steps: i) compute pairwise (PW) distances between all subjects, independently of the missing rate, using MissRatioRF; ii) use the output from step i), i.e., the distance matrix, as input to a classifier that works with PW distances. As to the latter step, in our scenario, we decided to employ a simple but straightforward technique, the Nearest Neighbor (NN) classifier [8]. Given a novel subject, for which PW distances to the rest of the dataset are known, NN labels it with the class of its most similar instance, i.e., the one at the minimum distance.

3 Classification of Multiple Sclerosis: a Challenging Task

Multiple Sclerosis is a chronic, inflammatory, demyelinating disease of the central nervous system, which progression causes neurodegeneration [13]. Various heterogeneous factors are involved in its onset, although the exact process is still unknown. To diagnose MS, specific guidelines are usually followed, known as McDonald criteria [18,30]. In detail, the latest revisions include CSF-based evidence, alongside clinical and Magnetic Resonance Imaging (MRI)-based criteria. Specifically, the presence of OCBs and kFLCs must be evaluated. Both markers are linked to active and past inflammation, potentially reducing the need for comparative MRIs to detect damage at different timepoints, crucial for an accurate diagnosis. The collection of CSF is therefore essential for a smoother diagnostic process. In addition, several markers can help differentiate not only MS from healthy controls but also between various subtypes of MS, while others have prognostic value [12].

Our dataset consists of 37 post-mortem subjects, 27 of whom suffer from MS and the remaining 10 are healthy controls (HC). All MS subjects, at the time of death, were classified as secondary progressive, the most advanced stage of the disease. In addition to clinical and neuropathological data, which are available only for MS subjects, CSF data were measured for the entire cohort. Therefore, we will use the latter to build a classifier that distinguishes MS from HC. The task, as mentioned in Sect. 1, is highly challenging: i) this type of data is highly unlikely to be publicly available, thus making this classification task rare; ii) the existing literature on learning tasks related to MS focuses mainly on other types of data, such as MRI or clinical data [4,19,33], making CSF-based approaches often overlooked; iii) MS-related datasets are often unbalanced, small, high-dimensional, and permeated by missing values.

This is also evident in our dataset: originally, it consisted of 83 immunological markers, but the final version contains only 75. In detail, we had to exclude 8 markers that were measured only in MS subjects. This is a common source of missing data: certain features are often measured only for a specific group, MS in our case, because for example they were not considered relevant at the time of collection. In general, recording CSF data is a lengthy process subjected to various factors that can cause missing data. Another example concerns strictly technical issues resulting in incorrect measurements that need to be removed. Often, the experiment may not be repeatable, leading to missing values. Moreover, even if the experiment succeeds, measurements must be adequately recorded. However, data are often managed by individual researchers and are rarely stored in proper databases, increasing the probability of losing the measurement.

4 Experimental Evaluation

This section is divided into two parts. In Sect. 4.1, we describe the details of the experiments we carried out, whereas in Sect. 4.2 we present our results, comparing MissRatioRF with other distance functions robust to missing data.

4.1 Experimental Setup

Given the dataset described in Sect. 3, we performed several experiments: first, we build a RF using Extremely Randomized Trees [15], then we compute pairwise distances using MissRatioRF and lastly, we use the distances to classify the subjects as MS or HC. In detail, each tree was built to its maximum depth and with 50% of the features, and we varied the following parameters: i) propagation variant: NanStop and NanBoth–both described in Sect. 2.2; ii) number of trees in a forest T; iii) ratio of training samples S used to build each tree. Given the latter, we defined the following configurations (cf):

1. **cf1**: NanStop, $T = 100, S = 50\%$
2. **cf2**: NanBoth, $T = 100, S = 50\%$
3. **cf3**: NanStop, $T = 500, S = 50\%$
4. **cf4**: NanBoth, $T = 500, S = 50\%$
5. **cf5**: NanStop, $T = 200, S = 80\%$
6. **cf6**: NanBoth, $T = 200, S = 80\%$
7. **cf7**: NanStop, $T = 400, S = 30\%$
8. **cf8**: NanBoth, $T = 400, S = 30\%$
9. **cf9**: NanStop, $T = 200, S = 100\%$
10. **cf10**: NanBoth, $T = 200, S = 100\%$

Please note that given a cf, distances were aggregated at forest level in two different ways, **V1** and **V2**, which are respectively the average and the weighted average as briefly described in Sect. 2.2, for a total of 20 different variants. Additionally, given the random nature of RFs, for each configuration, we iterated the whole procedure 100 times to increase robustness. The obtained matrices were then used to train and test a NN classifier using the Leave-One-Out protocol. Performances were evaluated by computing the average classification error across the 100 iterations.

MissRatioRF was then compared to the following state-of-the-art distances for missing data: **HEOM**, **HVDM**, **HEOM-Redef**, **HVDM-Redef** [26,27] and **Mean ED** [2,3], which are all variants of the Euclidean distance; **Bhattacharyya** [1] which extends the Mahalonobis distance; **FWPD-Clas** and **FWPD-Clus** which are both extension of the FWPD distance and that differ only in the value of the parameter α [10,11]. The same classification protocol used for MissRatioRF is adopted, except for the 100 iterations.

The entire protocol, applied to both MissRatioRF and the other distance functions, is repeated for 8 different missing rates, starting from the original (0.0011). This rate is then increased by synthetically generating missing values using MCAR [25], reaching up to 50.09% of missing values. Briefly, MCAR works by removing data without considering the possible existing relations among the features. However, we ensure that each subject retains at least one known value.

4.2 Results

In Tables 1 and 2 we report the (average) classification error for each configuration and aggregation variant of MissRatioRF, along with the results obtained with the extensions of geometric-based distance functions. Each column indicates a different missing rate (MR) with Tables 1 and 2 containing, respectively, results

Table 1. Comparison of MissRatioRF with state-of-the-art distances (low MR).

Distance	MR = 0.0011 (Original)	MR = 0.0112	MR = 0.0512	MR = 0.1013
MissRatioRFv1-cf1	0	0	0	0
MissRatioRFv2-cf1	0	0	0	0
MissRatioRFv1-cf2	0	0	0	0
MissRatioRFv2-cf2	0	0	0	0
MissRatioRFv1-cf3	0	0	0	0
MissRatioRFv2-cf3	0	0	0	0
MissRatioRFv1-cf4	0	0	0	0
MissRatioRFv2-cf4	0	0	0	0
MissRatioRFv1-cf5	0	0	0	0
MissRatioRFv2-cf5	0	0	0	0
MissRatioRFv1-cf6	0	0	0	0
MissRatioRFv2-cf6	0	0	0	0
MissRatioRFv1-cf7	0	0	0	0
MissRatioRFv2-cf7	0	0	0	0
MissRatioRFv1-cf8	0	0	0	0
MissRatioRFv2-cf8	0	0	0	0
MissRatioRFv1-cf9	0	0	0	0
MissRatioRFv2-cf9	0	0	0	0
MissRatioRFv1-cf10	0	0	0	0
MissRatioRFv2-cf10	0	0	0	0
HEOM	0.027	0.054	0.027	0.108
HVDM	0.027	0.054	0.054	0.108
HEOM-R	0.027	0.054	0.027	0.108
HVDM-R	0.027	0.054	0.054	0.108
MeanED	0.081	0.081	0.081	0.081
Bhattacharyya	0.027	0.054	0.027	0.054
FWPD-Clas	0.081	0.081	0.108	0.054
FWPD-Clus	0.081	0.054	0.108	0.054
Best MissRatioRF	0	0	0	0
Best Alternative	0.027	0.054	0.027	0.054

for low (from 0.0011 up to 0.1013) and high (from 0.2027 up to 0.5059) degrees of missingness. Further, for each missing rate, we also report in separate rows both the best configuration of MissRatioRF and the best alternative distance measure. We can observe that MissRatioRF is very robust, showing almost perfect classification results independently of the chosen configuration and aggregation variant. As evident from both tables, MissRatioRF is highly reliable also when the missing rate reaches very high values, showing performances that decrease only slightly and only for specific configurations. Indeed, the highest classification error is 0.0620–an average of 2.3 misclassified subjects–for MissRatioRFv1-cf1 and the highest degree of missingness. The other distance functions also perform

Table 2. Comparison of MissRatioRF with state-of-the-art distances (high MR).

Distance	MR = 0.2027	MR = 0.3305	MR = 0.4048	MR = 0.5069
MissRatioRFv1-cf1	0.001	0.003	0.028	0.062
MissRatioRFv2-cf1	0.001	0	0.020	0.024
MissRatioRFv1-cf2	0	0	0	0.009
MissRatioRFv2-cf2	0	0	0	0.013
MissRatioRFv1-cf3	0	0	0.017	0.046
MissRatioRFv2-cf3	0	0	0.011	0.018
MissRatioRFv1-cf4	0	0	0	0.004
MissRatioRFv2-cf4	0	0	0	0.011
MissRatioRFv1-cf5	0	0.001	0.022	0.044
MissRatioRFv2-cf5	0	0.001	0.018	0.021
MissRatioRFv1-cf6	0	0	0	0.011
MissRatioRFv2-cf6	0	0	0	0.017
MissRatioRFv1-cf7	0	0.001	0.023	0.054
MissRatioRFv2-cf7	0	0	0.009	0.014
MissRatioRFv1-cf8	0	0	0	0.008
MissRatioRFv2-cf8	0	0	0	0.011
MissRatioRFv1-cf9	0	0.001	0.021	0.039
MissRatioRFv2-cf9	0	0	0.023	0.019
MissRatioRFv1-cf10	0	0	0	0.017
MissRatioRFv2-cf10	0	0	0	0.021
HEOM	0.351	0.243	0.135	0.324
HVDM	0.378	0.297	0.270	0.324
HEOM-R	0.216	0.405	0.405	0.405
HVDM-R	0.297	0.405	0.378	0.459
MeanED	0.135	0.162	0.162	0.378
Bhattacharyya	0.054	0.081	0.108	0.189
FWPD-Clas	0.135	0.162	0.216	0.243
FWPD-Clus	0.270	0.216	0.216	0.216
Best MissRatioRF	0	0	0	0.004
Best Alternative	0.054	0.081	0.108	0.189

well, with Bhattacharyya being, on average, the best one; however, MissRatioRF consistently outperforms them all. Further, these measures are more affected by the increase in missing rate, i.e., the average classification errors reported in Table 2 are considerably worse than those in Table 1. To provide a clearer view of the latter point, Fig. 1 shows the trend of classification error for all missing rates for each distance function. As for MissRatioRF, for the sake of clarity, we only report one specific configuration, which is the worst one on average.

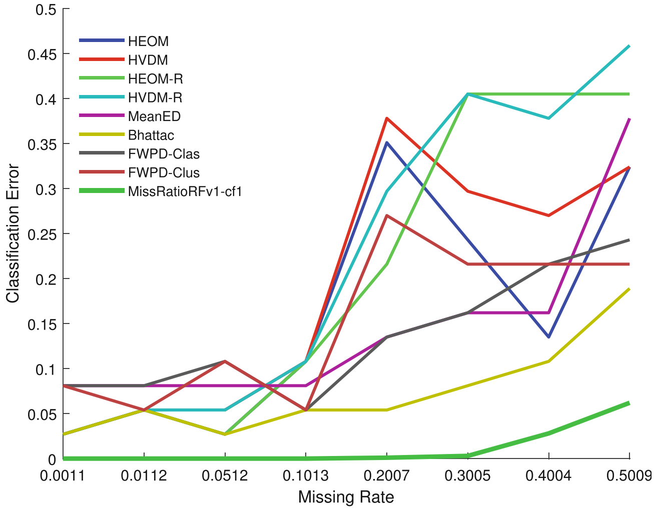

Fig. 1. Comparison of the distance functions across varying missing rates.

Observing Fig. 1, previously drawn conclusions are even more striking: whereas the performance of MissRatioRF begins to worsen at 30% missing rate, all other distance measures exhibit an increasing error as soon as the missingness degree exceeds 10%. In conclusion, even if the classification problem is highly challenging, MissRatioRF performs extremely well independently of the degree of missingness and also compared to other distance functions robust to missing data. Regarding interpretability, perfect classification indicates that some markers have enough predictive power to avoid any misclassification. However, due to the small size of the dataset, we should test the pipeline on a larger cohort to identify a reliable set of markers with potential diagnostic value.

5 Conclusions

In this paper, we applied MissRatioRF, a novel methodology that efficiently computes distances in the presence of missing data, to a complex real-world problem, the diagnosis of MS based on CSF data. In detail, missing data are neither removed nor imputed, instead they are actively used in the distance computation. Although the original paper proved the theoretical robustness of MissRatioRF, it was not used for any learning task. In this work, we demonstrate its effectiveness in the context of classification. Indeed, MissRatioRF not only achieves better results than other distance functions, but its performance remains stable as the missing rate increases. As mentioned, future research will focus on testing the pipeline on a more complex dataset from mass spectrometry that is currently being generated. This dataset will have more samples–addressing the limitation of the current study–over 10000 markers, and, as is typical of omic datasets, higher amounts of missing data.

Acknowledgments. We thank the UK MS Society Tissue Bank at Imperial College for the supply of post-mortem MS samples/data; the Laboratory of Neuropathology at LURM (University Laboratory of Medical Research), University of Verona. Antonella Mensi and Roberta Magliozzi received support from the Horizon Europe project BEHIND-MS (grant n. 101137235) and the Italian MS Foundation (grant FISM 2023/R-Single/038).

References

1. AbdAllah, L., Shimshoni, I.: A distance function for data with missing values and its application. Int. J. Comput. Sci. Eng. 7 (2013)
2. AbdAllah, L., Shimshoni, I.: Mean shift clustering algorithm for data with missing values. In: Data Warehousing and Knowledge Discovery: 16th International Conference DaWaK 2014, Proceeding, vol. 16, pp. 426–438. Springer (2014). https://doi.org/10.1007/978-3-319-10160-6_38
3. AbdAllah, L., Shimshoni, I.: K-means over incomplete datasets using mean Euclidean distance. In: International Conference Machine Learning Data Mining in Pattern Recognition, pp. 113–127. Springer (2016). https://doi.org/10.1007/978-3-319-41920-6_9
4. Barile, B., Ashtari, P., Stamile, C., Marzullo, A., Maes, F., Durand-Dubief, F., Van Huffel, S., Sappey-Marinier, D.: Classification of multiple sclerosis clinical profiles using machine learning and grey matter connectome. Front. Rob. AI **9**, 926255 (2022). https://doi.org/10.3389/frobt.2022.926255
5. Bicego, M., Cicalese, F.: Computing random forest-distances in the presence of missing data. ACM Trans. Knowl. Discov. Data **18**(7), 1–18 (2024). https://doi.org/10.1145/3656345
6. Bicego, M., Cicalese, F., Mensi, A.: Ratiorf: a novel measure for random forest clustering based on the Tversky's ratio model. IEEE Trans. Knowl. Data Eng. **35**(1), 830–841 (2021). https://doi.org/10.1109/TKDE.2021.3086147
7. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
8. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (1967). https://doi.org/10.1109/TIT.1967.1053964
9. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found. Trends Comput. Graph. Vis. **7**(2–3), 81–227 (2012). https://doi.org/10.1561/0600000035
10. Datta, S., Bhattacharjee, S., Das, S.: Clustering with missing features: a penalized dissimilarity measure based approach. Mach. Learn. **107**(12), 1987–2025 (2018). https://doi.org/10.1007/s10994-018-5722-4
11. Datta, S., Misra, D., Das, S.: A feature weighted penalty based dissimilarity measure for k-nearest neighbor classification with missing features. Pattern Recognit. Lett. **80**, 231–237 (2016). https://doi.org/10.1016/j.patrec.2016.06.023
12. Deisenhammer, F., Zetterberg, H., Fitzner, B., Zettl, U.K.: The cerebrospinal fluid in multiple sclerosis. Front. Immunol. **10**, 726 (2019). https://doi.org/10.3389/fimmu.2019.00726
13. Dobson, R., Giovannoni, G.: Multiple sclerosis-a review. Eur. J. Neurol. **26**(1), 27–40 (2019). https://doi.org/10.1111/ene.13819

14. Galán, C.O., Lasheras, F.S., de Cos Juez, F.J., Sánchez, A.B.: Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions. J. Comput. Appl. Math. **311**, 704–717 (2017). https://doi.org/10.1016/j.cam.2016.08.012
15. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1
16. Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D.: Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. Neuroimage **65**, 167–175 (2013). https://doi.org/10.1016/j.neuroimage.2012.09.065
17. Le, L.P., Nguyen, T., Riegler, M.A., Halvorsen, P., Nguyen, B.T.: Multimodal missing data in healthcare: a comprehensive review and future directions. Comput. Sci. Rev. **56**, 100720 (2025). https://doi.org/10.1016/j.cosrev.2024.100720
18. Levraut, M., Landes-Chateau, C., Mondot, L., Cohen, M., Lebrun-Frenay, C.: The kappa free light chains index and central vein sign: two new biomarkers for multiple sclerosis diagnosis. Neurol. Ther. 1–21 (2025)
19. Moazami, F., Lefevre-Utile, A., Papaloukas, C., Soumelis, V.: Machine learning approaches in study of multiple sclerosis disease through magnetic resonance images. Front. Immunol. **12**, 700582 (2021). https://doi.org/10.3389/fimmu.2021.700582
20. Nishanth, K.J., Ravi, V.: Probabilistic neural network based categorical data imputation. Neurocomputing **218**, 17–25 (2016). https://doi.org/10.1016/j.neucom.2016.08.044
21. Pelckmans, K., De Brabanter, J., Suykens, J.A., De Moor, B.: Handling missing values in support vector machine classifiers. Neural Netw. **18**(5–6), 684–692 (2005). https://doi.org/10.1016/j.neunet.2005.06.025
22. Pigott, T.D.: A review of methods for missing data. Educ. Res. Eval. **7**(4), 353–383 (2001). https://doi.org/10.1076/edre.7.4.353.8937
23. Ramoni, M., Sebastiani, P.: Robust Bayes classifiers. Artif. Intell. **125**(1–2), 209–226 (2001). https://doi.org/10.1016/S0004-3702(00)00085-0
24. Rennard, S.I., Locantore, N., Delafont, B., Tal-Singer, R., Silverman, E.K., Vestbo, J., Miller, B.E., Bakke, P., Celli, B., Calverley, P.M., Coxson, A., Crim, C., Edwards, L.A., Lomas, D.A., MacNee, W., Wouters, E.F.M., Yates, J.C., Coca, I., Agusti, A.: Identification of five chronic obstructive pulmonary disease subgroups with different prognoses in the eclipse cohort using cluster analysis. Ann. Am. Thorac. Soc. **12**(3), 303–312 (2015). https://doi.org/10.1513/AnnalsATS.201403-125OC
25. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976). https://doi.org/10.1093/biomet/63.3.581
26. Santos, M.S., Abreu, P.H., Fernández, A., Luengo, J., Santos, J.: The impact of heterogeneous distance functions on missing data imputation and classification performance. Eng. Appl. Artif. Intell. **111**, 104791 (2022). https://doi.org/10.1016/j.engappai.2022.104791
27. Santos, M.S., Abreu, P.H., Wilk, S., Santos, J.: How distance metrics influence missing data imputation with k-nearest neighbours. Pattern Recognit. Lett. **136**, 111–119 (2020). https://doi.org/10.1016/j.patrec.2020.05.032
28. Shi, T., Seligson, D., Belldegrun, A., Palotie, A., Horvath, S.: Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Modern Pathol. **18**, 547–557 (2005). https://doi.org/10.1038/modpathol.3800322

29. Shi, T., Horvath, S.: Unsupervised learning with random forest predictors. J. Comput. Graph. Stat. **15** (2005). https://doi.org/10.1198/106186006X94072
30. Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M.S., et al.: Diagnosis of multiple sclerosis: 2017 revisions of the Mcdonald criteria. Lancet Neurol. **17**(2), 162–173 (2018). https://doi.org/10.1016/S1474-4422(17)30470-2
31. Ting, K., Zhu, Y., Carman, M., Zhu, Y., Zhou, Z.H.: Overcoming key weaknesses of distance-based neighbourhood methods using a data dependent dissimilarity measure. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, pp. 1205–1214 (2016). https://doi.org/10.1145/2939672.2939779
32. Tversky, A.: Features of similarity. Psychol. Rev. **84**(4), 327 (1977). https://doi.org/10.1037/0033-295X.84.4.327
33. Vázquez-Marrufo, M., Sarrias-Arrabal, E., García-Torres, M., Martín-Clemente, R., Izquierdo, G.: A systematic review of the application of machine-learning algorithms in multiple sclerosis. Neurología (English Ed.) **38**(8), 577–590 (2023). https://doi.org/10.1016/j.nrleng.2020.10.013
34. Zhu, X., Loy, C., Gong, S.: Constructing robust affinity graphs for spectral clustering. In: Proceedings of the International Conference Computer Vision and Pattern Recognition, pp. 1450–1457 (2014). https://doi.org/10.1109/CVPR.2014.188