ExACT Explainable Clustering: Unravelling the Intricacies of Cluster Formation

Federico Sabbatini[1],*, Roberta Calegari[2]

[1]Department of Pure and Applied Sciences, University of Urbino Carlo Bo
[2]Department of Computer Science and Engineering, Alma Mater Studiorum–University of Bologna

Abstract
Cluster assignments, in particular the deep clustering ones, are often hard to explain, partially because they depend on all the features of the data in a complicated way, so it is difficult to determine why a particular row of data is classified in a particular bucket. This opaqueness makes their predictions not trustable, as for many predictors based on black boxes. This paper aims to tackle the aforementioned issues by introducing the design and implementation of ExACT, a new explainable clustering algorithm based on the induction of decision trees and performing hypercubic approximations of the input feature space in order to provide output human-interpretable clusters. Furthermore, ExACT is versatile enough to perform explainable classification and regression as well, as demonstrated in this work, proving to be a competitive alternative to existing analogous algorithms.

Keywords
Explainable clustering, Explainable artificial intelligence, PSyKE

1. Introduction

Clustering is one of the most fundamental optimisation techniques constituting the heart of many applications in machine learning (ML) and data mining. However, in the past few years due to the increasing need for transparency [1, 2] – in particular in critical domains related to human health, safety, and wealth – people do not trust clusterings, or more generally learning models that are not interpretable by humans. Models lacking interpretability are defined *opaque* or *black boxes* (BBs), regardless of their nature (e.g., supervised neural network classifiers as well as unsupervised deep clustering techniques). ML models able to achieve the best predictive performance are generally the most complex and thus difficult to be inspected by humans and, therefore, the adoption of opaque models for high-stakes decisions is mandatorily subject to the derivation of some kind of human-intelligible knowledge.

To not renounce the impressive predictive capabilities of ML models, many strategies to obtain explainable behaviours have been proposed in the literature [3, 4]. When possible, interpretable ML predictors as decision trees are exploited [5]. Conversely, when these interpretable models do not have a satisfying performance (e.g., shallow decision trees) or their complexity hinders their actual readability (e.g., deep decision trees), it is possible to reverse-engineer the predictors'

KoDis'23: The 2nd International Workshop on Knowledge Diversity, September 3, 2023, Rhodes, Greece

*Corresponding author.

f.sabbatini1@campus.uniurb.it (F. Sabbatini); roberta.calegari@unibo.it (R. Calegari)
0000-0002-0532-6777 (F. Sabbatini); 0000-0003-3794-2942 (R. Calegari)

behaviour [6]. Symbolic knowledge-extraction (SKE) techniques are exploited to this end, acting in a post-processing phase to extract interpretable knowledge out of a BB predictor.

Inspired by the works on explainable clustering [7, 8] and the ones on SKE [9, 10, 11, 12], in this paper we provide a human-interpretable model blending both topics and resulting in the development of ExACT, a new explainable clustering technique also suitable to perform explainable classification and regression tasks. ExACT may be applied to continuous input features and both categorical and numerical output data. The technique is able to describe clusters of instances in terms of human-comprehensible rules derived from a binary tree built according to a hierarchical hypercubic partitioning of the input feature space.

The paper is organised as follows: Section 2 introduces background information on the topics discussed here and related works present in the literature. Section 3 describes the ExACT algorithm. Experiments and benchmark comparisons are discussed in Section 4. Finally, conclusions are drawn in Section 5.

2. Related Works

2.1. Traditional Clustering

A large amount of different clustering techniques have been proposed in the literature during the decades, each one providing some peculiar advantages but at the same time bound to specific limitations, usually regarding the properties of the clusters they can identify (e.g., shape, density). For such a reason, there is no widely acknowledged optimum technique for achieving the best predictive performance in every possible application. Amongst the most prediction-effective techniques, it is worth mentioning the following ones: Gaussian mixture models (GMMs; [13]), DBSCAN and DBSCAN++ [14, 15, 16], OPTICS [17], BIRCH [18], k-means [19], Mean shift [20] and spectral clustering [21, 22]. The main drawback of these techniques in terms of explainability is to rely on an opaque model.

In the following, further details for the traditional clustering techniques exploited within the ExACT algorithm are provided.

2.1.1. Gaussian Mixture Models

GMMs can be applied to perform (soft) clustering since they are probabilistic models assuming that all data points have been generated by a mixture of a finite number of Gaussian distributions having parameters to be determined. GMMs are more flexible for clustering than (for instance) k-means, since they can find clusters of data that are not only spherical. In addition, soft clustering is provided, since each GMM prediction is associated with a corresponding probability.

The performance of a GMM is strongly impacted by the number of Gaussian components to consider during the model training. The tuning of this parameter can be automated by using the Bayes information criterion (BIC). It is sufficient to train several instances of a GMM, with a different number of components, then calculate the BIC score for every instance and finally pick the one with the lowest associated BIC score.

2.1.2. DBSCAN

The DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise; [14, 15]) is an unsupervised clustering technique to find subsets of data having arbitrary shapes. It is based on a density criterion, so clusters are created by aggregating data samples w.r.t. a user-defined parameter, usually called ε, representing the maximum distance between two data points inside the same cluster. For this reason, the ε parameter is the most important to tune for DBSCAN. An automated procedure for this purpose has been proposed in [23]. The peculiarities of DBSCAN make it a good choice to perform outliers removal from clusters found by other clustering techniques.

2.2. Explainable Clustering

During the last decades, a number of researchers have focused their attention on the explanation of clusters, in particular in the field of medical applications, one of the most relevant critical areas. Several works in the literature are related to tree-based clustering, such as the IMM algorithm [24] and others [25, 26, 27]. For inducing a decision tree, there are two main approaches: top-down and bottom-up. The top-down is the most used and it will also be used in our proposal. This approach starts building a root node, which contains all objects of a training database. After that, the root node is split into partitions (usually named child nodes) and this is recursively repeated over the child nodes until a stopping criterion is met. It is worth noting that these approaches share a common trait, i.e., they partition the input feature space with cutting hyperplanes perpendicular to the most relevant features, taking into account one feature for each cut.

Cluster explanation via rectangular input space partitioning has been proposed in [28], whereas a less human-readable density-based clustering has been recently proposed in [29] for the CLASSIX algorithm. The former enables higher degrees of human interpretability since it describes clusters in terms of only 2 interval inclusion preconditions. As a drawback, it may combine input features in the output description to create new composite features, thus hindering interpretability from the human perspective.

Other examples of explainable clustering techniques, in particular applied to image data sets and medical time series, have been presented in [30].

W.r.t. the aforementioned techniques, ExACT is able to consider all the input features – similarly to tree-based clustering methods as IMM – to create a density-based partitioning—as DBSCAN and CLASSIX. The main difference with all the existing methods is the highly human-readable format of the identified clusters, that are approximated via hypercubic regions. Thus, ExACT offers a more compact and human-readable representation than existing techniques, even though the hypercubic approximation may hinder its performance when applied to overlapping clusters.

ExACT provides global explanations about the input feature space partitioning into disjoint clusters, that may be used to obtain local explanations about single cluster assignments.

In the following we provide an overview of two explainable clustering techniques used as benchmarks in the experiments presented here, namely CLASSIX and IMM.

2.2.1. CLASSIX

CLASSIX (contrived acronym defined by the authors as “CLustering by Aggregation with Sorting-based Indexing” and the letter “X” for “eXplainability”) has been recently proposed in [29] as an explainable clustering procedure based on two phases and denoted by small computational time requirements. During the first phase, a greedy aggregation is performed in order to create groups of training instances having small distances from each other—where “small” is defined via an input parameter. A preceding sorting step is required to complete the aggregation. The second phase consists of merging the groups into definitive clusters. The merging phase may be density- or distance-based and it is described in detail in [29].

Users adopting CLASSIX need to provide a pair of parameters to define the minimum size of the clusters, intended as the number of instances, and the maximum distance between training samples belonging to the same group (with reference to the aggregation phase).

CLASSIX is able to provide both local and global explanations. The global explanation is based on the coordinates of the initial points for each one of the groups created at the end of the first phase. Local explanations may describe the reason behind the cluster assignment for a single instance as well as why two instances are assigned to the same cluster or not. Local explanations are provided by listing the operations performed during CLASSIX’s merging phase.

2.2.2. IMM

In [24] the IMM (Iterative Mistake Minimization) clustering procedure is presented as an accurate, efficient, and interpretable method based on the induction of decision trees. Induced decision trees are binary and their internal nodes are associated with training data partitions. Node splits involve single features. The algorithm requires growing k leaves to identify as many clusters, trying to keep the tree size as small as possible. During the tree construction, the cluster’s fragmentation is minimised. Fragmentation is intended as spreading instances from a single cluster over multiple subtrees.

To provide explanations for a cluster assignment it is sufficient to describe the complete path from the tree root through the leaf associated with that assignment. As for the tree growth complexity, a clustering identifying k clusters may be described by a tree with depth equal to $k - 1$, in the worst case. This implies describing any clustering assignment with at most $k - 1$ constraints on the input features.

3. Explainable Clustering with E ACT

In this section we propose the design and implementation of a new explainable clustering technique. ExACT (EXplainable Automated Clustering Technique) is a hierarchical clustering algorithm based on the induction of binary trees where each node represents an input space region approximating a cluster and having the shape of a hypercube (or of a difference of two hypercubes). ExACT is a supervised technique since it distinguishes between input and output features. It is suitable to be applied to continuous input attributes and continuous, discrete, or categorical output variables.

A key peculiarity of the algorithm is to keep the memory of the output associated with the clusters, other than the mere membership of instances to clusters. In the case of regression data sets, the clusters' outputs are real values or regression laws involving the input variables, instead of the discrete outputs adopted for classification and clustering data sets. More in detail, we stick to the generalisation proposed in [31, 32], associating: *(i)* the most common label to regions containing classification or clustering data points; *(ii)* the mean value calculated on the data points' outputs to regions containing regression data points if there is a high degree of similarity between these output values; *(iii)* a linear combination of the input variables to regions containing regression data points, otherwise. For this reason ExACT has predictive capabilities that go beyond the usual clustering assignments and its identified clusters may be evaluated with classical clustering scores but also with metrics borrowed from classification and regression tasks (i.e., predictive error can be evaluated through the F_1 or R^2 scores).

In the following we use the notion of *predictive error* for the regions detected at the end of the procedure. The predictive error is evaluated through the difference between the outputs associated with the ExACT's clusters and the corresponding expected predictions (i.e., the ground truth). The predictive error assessment differs based on the kind of output feature at hand. Indeed, it is defined as inversely proportional to the accuracy score for data sets having discrete outputs and as the mean absolute error for real-valued outputs.

3.1. Properties of E ACT

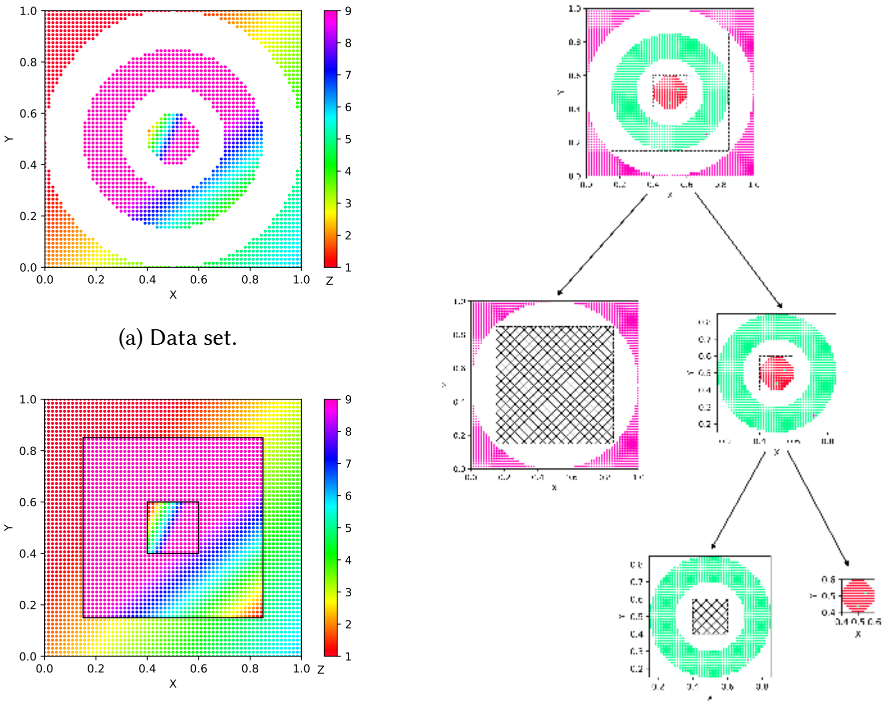
The algorithm relies on two well-known clustering techniques—namely, GMMs and DBSCAN. In a nutshell, GMMs are exploited for detecting the clusters inside an input space region, then DBSCAN is applied to the detected clusters for the deletion of possible outliers. Finally, hypercube approximation is performed to approximate each cluster to a hypercube that can "explain" the cluster in a human-interpretable format. The goal of ExACT is to approximate the set of clusters detected by GMMs (and cleaned by DBSCAN) with as many regions (cubes), with the following criteria:

exhaustivity of the approximations all the input feature space is covered by regions, i.e., every input instance belongs to at least one region;

disjointness of the regions each input instance inside the input feature space belongs to at most one region;

strict hierarchy of the regions each found hypercubic region lies within a wider region having the same shape.

The resulting decision tree provides an explanation of the clusters, explainability via interpretability [33]. The strict hierarchy of the regions trivially implies that instances belonging to an inner region also belong to the outer, enclosing ones, apparently violating the disjointness property. However, when calculating the membership of data points to regions, ExACT assigns every instance to the smallest region containing that point, so predictions are completely unambiguous and easily human-understandable, following the induced tree structure from the rightmost leaf through the root node, and disjointness is preserved.



(a) Data set.

(b) Approximation performed by ExACT.

(c) Binary tree induced by ExACT.

Figure 1: Example of ExACT partitioning performed on an artificial data set having concentric clusters.

ExACT is a recursive algorithm, starting from the surrounding cube – i.e., the minimal hypercube enclosing all the data set samples – and iteratively building smaller, inner hypercubes, inducing a binary tree structure. The rationale behind our method is to create at every iteration a difference cube enclosing data points belonging to a single cluster, with the goal of minimising the total amount of cubes and the cluster fragmentation. As detailed in the following, the difference cube is the one obtained by subtracting the best cube (right child) from the starting hypercube (parent node).

3.2. Algorithm and Parameters

The algorithm details are summarised in Algorithm 1 and an example of the performed partitioning on an artificial data set described by 3 concentric clusters is reported in Figure 1. In particular, Figure 1a depicts the input data set, having 2 continuous input features and 1 continuous output feature. The approximation provided by ExACT is reported in Figure 1b, whereas the binary tree induced by the algorithm during the recursive input space partitioning is reported in Figure 1c.

Binary trees built by ExACT are obtained by partitioning the input feature space into hypercubes. ExACT starts by taking into consideration all the input space, that represents the initial hypercube. To build the tree the following steps are recursively executed:

Algorithm 1 ExACT pseudocode

```
Require: maximum depth $\delta$, default value: 2
Require: predictive error threshold $\theta$, default value: 0.1
Require: maximum amount of clusters $\xi$, default value: 2
Provide: the root node of the induced tree

function ExACT($D$)
    $H_0 \leftarrow$ SurroundingCube($D$)
    $N_0 \leftarrow$ NewNode($H_0, D$)
    Split($N_0, 1$)
    return $N_0$

function SurroundingCube($D$)
    return the minimal cube enclosing all the points of cluster $D$

function Split($node$, $depth$)
    $clusters \leftarrow$ CreateClusters($node.data$)
    $eligible \leftarrow$ ClustersToNodes($clusters, node.cube$)
    if $eligible = \emptyset$ then return
    $best \leftarrow \arg\max_{n \in eligible} \{$ Volume($n.cube$) $\}$
    $node.right \leftarrow best$
    $node.left \leftarrow$ NewNode($node.cube, node.data \setminus best.data$)
    $error \leftarrow$ PredictiveError($node.right$)
    if $(error > \theta) \wedge (depth < \delta)$ then Split($node.right, depth + 1$)    ▷ Recursion

function NewNode($H$, $D$)
    $node \leftarrow$ new Node()
    $node.cube \leftarrow H$,    $node.data \leftarrow D$
    $node.right \leftarrow \emptyset$,    $node.left \leftarrow \emptyset$
    return $node$

function CreateClusters($D$)
    return at most $\xi$ clusters containing the data of $D$

function ClustersToNodes($C$, $H$)
    $nodes \leftarrow \emptyset$
    for all $cluster \in C$ do
        $data \leftarrow cluster \setminus \{ c \in cluster \mid c$ is an outlier $\}$
        $cube \leftarrow$ SurroundingCube($data$)
        if $cube \neq H$ then $nodes \leftarrow nodes \cup \{$ NewNode($cube, data$) $\}$
    return $nodes$

function Volume($H$)
    return the volume of hypercube $H$

function PredictiveError($node$)
    return average predictive error for $node$
```

1. given a hypercubic portion of the data set, associated with a tree node (a.k.a., the *current node*), apply GMMs to determine both the optimal number of clusters and the clusters themselves;

2. for each found cluster, apply DBSCAN to remove the outliers for that cluster (to avoid

creating dirty, too big hypercubes);

3. construct the minimal surrounding cube enclosing the data points of each cluster (i.e., approximate clusters to hypercubes): each cluster is associated with the smallest hypercube containing all the points within it;
4. select the *best* cube amongst all the created hypercubes, that is the one having the biggest volume, excluding hypercubes equivalent to the current node's cube (i.e., the one enclosing the whole data set when performing the first algorithm recursion);
5. assign the best cube, with all the contained data points, to the right child of the current node;
6. assign the difference cube to the left child of the current node—the difference cube is the one obtained by subtracting the best cube (right child) from the starting hypercube (parent node);
7. repeat all the steps above considering the right child as the current node if the predictive error measured for the right child's cube is greater than a user-defined threshold θ.

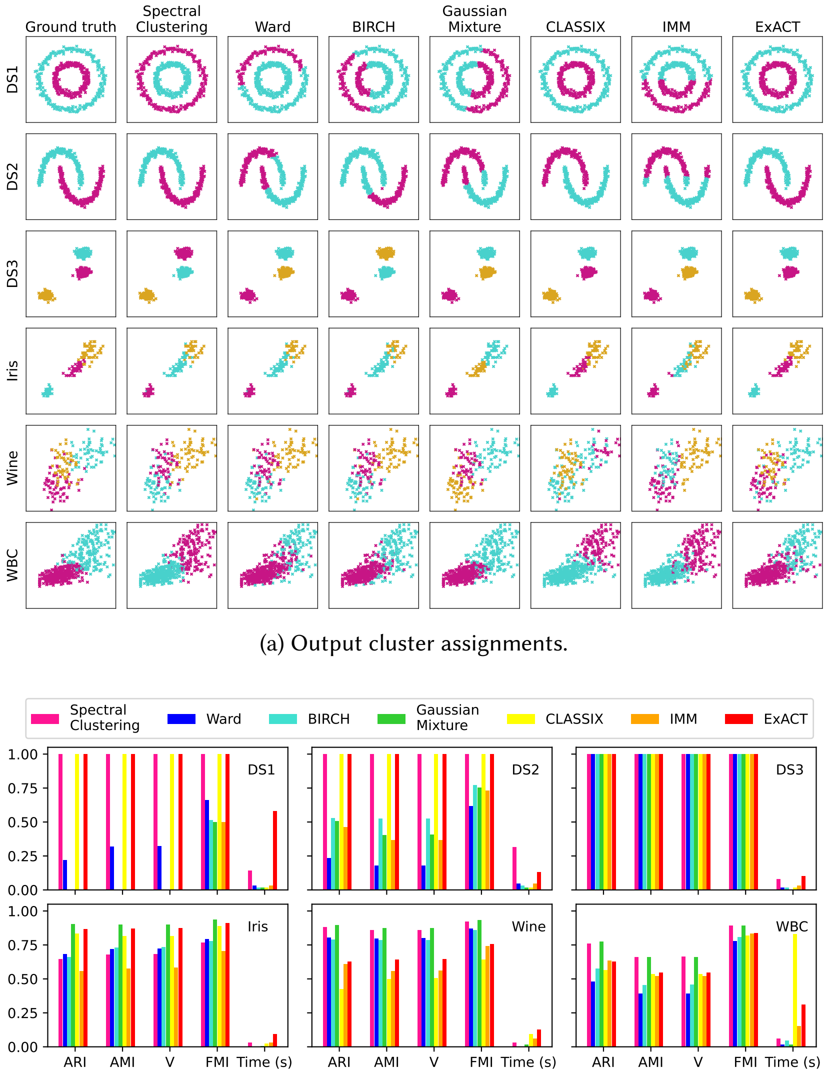
Note that the algorithm is not going to split the left child, since the difference hypercube is assumed to be sufficiently precise. The difference hypercube, in fact, typically approximates a single cluster, or a cluster portion if ExACT is not able to avoid fragmentation via hypercubic approximation.

The algorithm terminates when the node assigned to the right child contains a cube whose predictive error is smaller than the θ threshold or, otherwise, after a number of recursive iterations equal to the maximum user-defined depth δ. The set of hyper-parameters to tune for ExACT is completed by ξ, which represents the maximum number of clusters that it is possible to find with the GMMs. We stress here the fact that ξ is only an upper bound since the optimal number of clusters to be identified is automatically assessed through the BIC score during the execution of ExACT. It is recommended to keep ξ larger than the actual cluster amount, if known, in order not to perform a wrong clustering. Very large values for ξ do not affect the performance of ExACT, since they do not imply selecting with the automated BIC-based procedure a large number of clusters to be detected with the GMMs. Analogously, the ε parameter of DBSCAN is automatically set as suggested in [23], so it is not a parameter to be chosen by users executing ExACT. The δ parameter should be set according to the consideration that a depth equal to n produces at most $n+1$ explainable clusters. It is important to notice that only one explainable cluster (the innermost in the hierarchy) is a hypercube, all the others are difference cubes. Finally, θ strongly depends on the task at hand. When dealing with categorical output features, as for ExACT applied on clustering or classification data sets, θ needs to be defined as a predictive accuracy threshold. In this case, a cube is further partitioned only if its predictive accuracy is lower than the given threshold. On the other hand, when performing clustering on regression data sets, the threshold represents the maximum mean absolute error allowed for individual cubes. Consequently, hypercubes whose predictive error exceeds the θ threshold are further partitioned.

4. Experiments

Experiments to assess the capabilities of ExACT applied to clustering, classification, and regression tasks in comparison with state-of-the-art clustering and other predictors are reported in the following. The adopted ExACT implementation is included in the PSyKE framework[1] [34, 35, 36, 37].

4.1. E ACT for Explainable Clustering



(a) Output cluster assignments.

(b) Clustering performance assessments.

Figure 2: Example of ExACT clustering compared with other state-of-the-art techniques.

[1] Code available at https://github.com/psykei/psyke-python

Listing 1 Clustering rules provided by ExACT for the Iris data set.

```
Cluster 1 if PetalWidth in [1.6, 2.5] and PetalLength in [4.8, 6.9] and
             SepalWidth in [2.5, 3.8] and SepalLength in [5.7, 7.9].
Cluster 2 if PetalWidth in [1.0, 2.5] and PetalLength in [3.0, 6.9] and
             SepalWidth in [2.2, 3.8] and SepalLength in [4.9, 7.9].
Cluster 3 otherwise.
```

The capabilities of ExACT in clustering labelled data have been assessed on six different data sets. Three of them are synthetic clustering data sets included in the Scikit-Learn library[2]. These data sets are described by 2 continuous input features and they have 2 or 3 clusters to be identified. The other three are real-world classification data sets:

Iris data set [38], composed of 4 continuous input features and a categorical output feature assuming 3 different values. Only the petal length and width are reported in the figures shown in this section;

Wine data set [39], having 13 real-valued features and 3 possible discrete output values representing as many classes. Only alcohol and proline input features are shown in the figures reported here;

Wisconsin breast cancer (WBC) data set [40], a binary classification task described by 30 continuous input features. Worst smoothness and worst symmetry are the input features reported in the figures.

Figure 2 depicts our experiments involving clustering. All the data sets are represented in the leftmost column of Figure 2a. The other columns report the results of traditional and explainable clustering techniques applied to the same data sets. We selected spectral clustering, Ward, BIRCH, and GMMs as traditional clustering benchmarks, whereas CLASSIX and IMM are the explainable alternatives. The clustering assignments performed by ExACT are reported in the rightmost column. In Figure 2b the performance assessments for all the aforementioned clustering techniques applied to all the selected data sets are summarised. The adopted metrics are the following: *(i)* adjusted rand index (ARI; [41]); *(ii)* adjusted mutual score (AMI; [42]); *(iii)* Fowlkes-Mallows index (FMI; [43]); *(iv)* V-measure (V; [44]); *(v)* computational time, reported in seconds and averaged over 100 runs. The other 4 indices are defined in the [0, 1] interval, with values close to 1 identifying good clustering assignments. Being metrics explicitly designed for clustering tasks, they are not susceptible to permutations and/or renaming of the clusters' labels. For this reason, they are not applicable to evaluate the accuracy score of clustering techniques applied to perform classification tasks.

Figure 2a enables a qualitative assessment of the performance achieved by the clustering techniques. ExACT, CLASSIX, and spectral clustering appear as the best procedures. A quantitative assessment can be carried out on the results of Figure 2b, highlighting the superiority of these 3 algorithms. Unfortunately, none of these techniques can achieve the best performance on all the data sets. A drawback of ExACT is that it may be slower than the others. However, it completes its task in less than 1 second in all the case studies.

[2] https://scikit-learn.org/stable/modules/clustering.html

Clustering rules extracted via ExACT for the Iris data set are exemplified in Listing 1.

Amongst the 3 explainable clustering techniques, IMM is the one having the lowest associated scores. On the other hand, CLASSIX seems to be equivalent or only slightly worse than ExACT from the clustering performance standpoint. The interpretability of these techniques cannot be easily assessed, since they do not provide the same representation. IMM and ExACT are comparable since they follow a tree structure and therefore their clusters can be described by reading the tree paths from the root to the leaves. CLASSIX, conversely, provides a different sort of explainability. For instance, when queried about an individual assignment, CLASSIX provides the numeric code associated with the corresponding output cluster and the one of the group identified during its first grouping phase. No further information about the clusters or groups is provided unless to query CLASSIX for a global explanation. In this case, the centroid coordinates of each group are listed, together with the indication of the final cluster where the groups have flowed into. It is clear how this kind of explanation is not straightforward to be understood by humans, since it involves a chain of concepts encoded with numbers and coordinates, such as the radius of the CLASSIX groups. Furthermore, if CLASSIX is executed on normalised data, also its response will contain normalised coordinates. Conversely, ExACT may be provided with the normalisation schema in order to obtain cluster explanations that do not require further manipulation to enable human analysis. Finally, ExACT is more general-purpose than the other techniques, as discussed in the following.

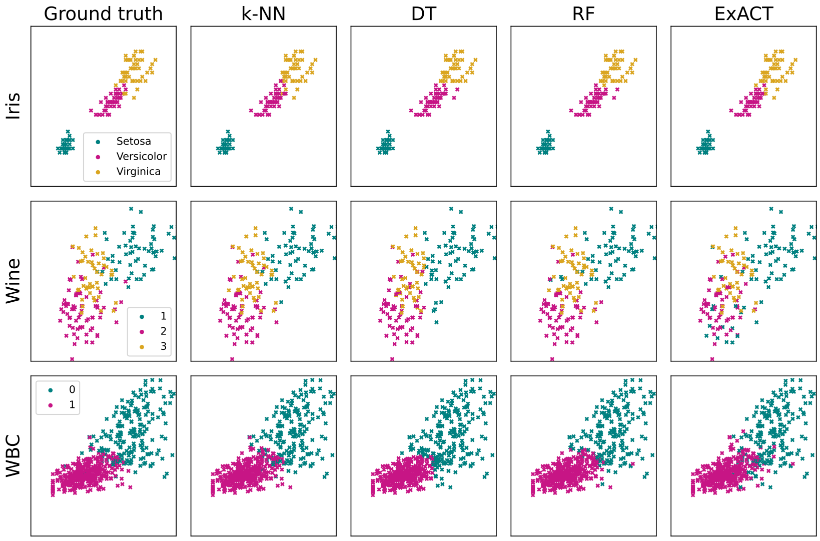
4.2. E ACT for Explainable Classification

Given its ability to output (explainable) predictions when queried with samples to be classified, ExACT is also suitable to carry out classification tasks. Figure 3 depicts the results of ExACT applied to perform classification on the Iris, Wine, and WBC data sets. Its predictions are compared with those of state-of-the-art machine learning predictors, namely: a k-nearest neighbours (k-NN), a decision tree (DT), and a random forest (RF) model. The performance of these models is assessed and compared for each data set through the classification accuracy score, representing the rate of correct predictions over all the predictions. Predictions and measured accuracy scores are shown in Figures 3a and 3b, respectively.

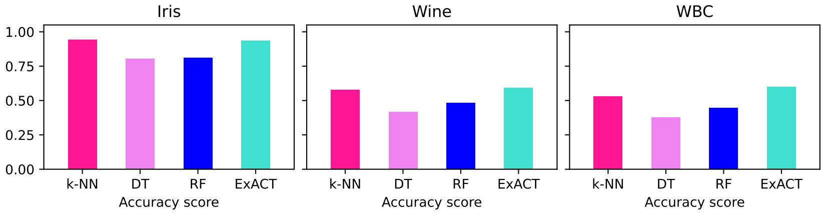
ExACT achieves a comparable or better predictive performance w.r.t. the other models. In addition, its predictions are more valuable, since they are human-interpretable. We exemplify in Table 1 the clusters obtained for the Iris data set and in Figure 4 the corresponding explainable tree. The ExACT instance to obtain these results has been parametrised with a maximum depth $\delta = 2$, an error threshold $\theta = 0.1$ and a maximum amount of clusters $\xi = 3$. The selected δ value enables the creation of up to 4 clusters. Being a classification data set, $\theta = 0.1$ implies that hypercubic regions having an accuracy score smaller than $1 - \theta = 0.9$ are further split during the recursive iterations of ExACT. The provided clustering is human-interpretable since for each possible Iris output class an associated hypercubic input space region is provided and such regions are described in terms of interval inclusion constraints over input variables.

4.3. E ACT for Explainable Regression

ExACT has been applied to several regression data sets, namely:



(a) Output predictions for classification tasks.



(b) Classification performance assessments.

Figure 3: Example of ExACT compared with other state-of-the-art classifiers.

Input feature	Petal width	Petal length	Sepal width	Sepal length	Iris class
Cluster 1	1.6 – 2.5	4.8 – 6.9	2.5 – 3.8	5.7 – 7.9	Virginica
Cluster 2	1.0 – 2.5	3.0 – 6.9	2.2 – 3.8	4.9 – 7.9	Versicolor
Cluster 3	0.1 – 2.5	1.2 – 6.9	2.2 – 4.1	4.4 – 7.9	Setosa

Table 1
Example of ExACT clustering applied for classification on the Iris data set.

Combined Cycle Power Plant (CCPP) data set [45], having 4 input continuous features. Only ambient temperature and exhaust vacuum input features are reported in the following figures;

Istanbul Stock Exchange (ISE) data set [46], described by 7 continuous input features. Only the stock market return index of UK and the MSCI European index are shown in the figures;

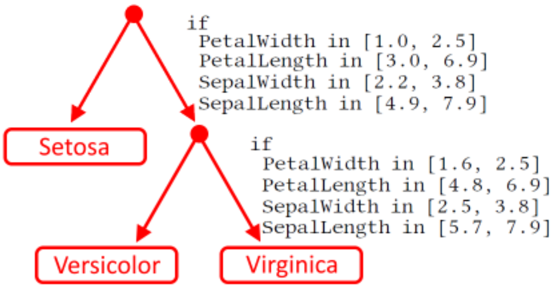
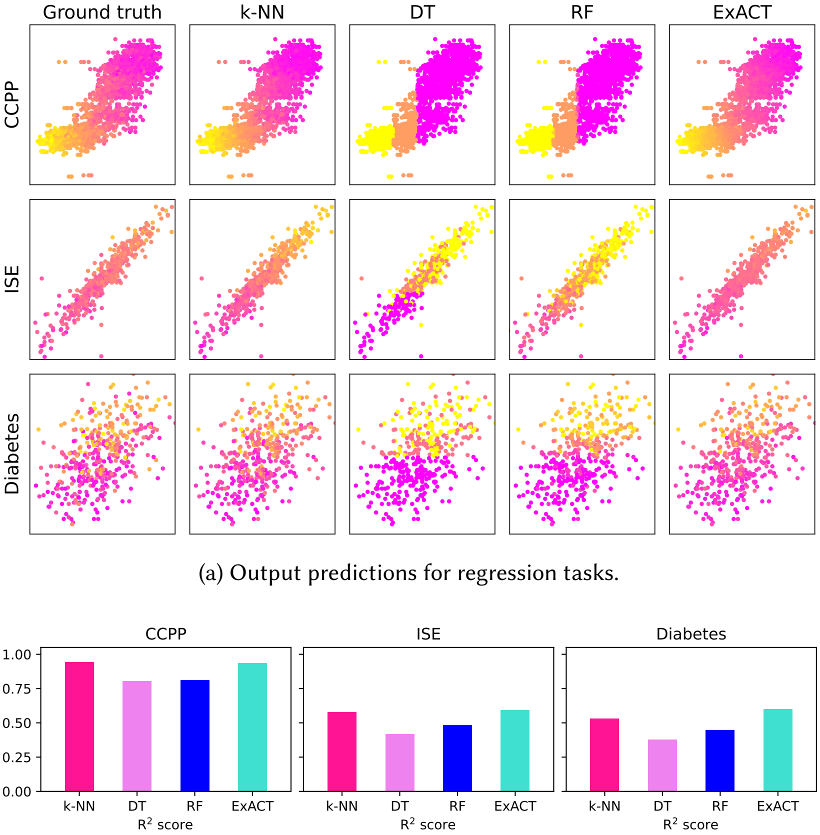


Figure 4: Decision tree provided by ExACT for explainable classification on the Iris data set.



(a) Output predictions for regression tasks.

(b) Regression performance assessments.

Figure 5: Example of ExACT compared with other state-of-the-art regressors.

Diabetes data set [47], containing 10 input variables. In the figures the S1 and S5 features are reported.

We applied to these data sets ExACT and a pool of ML regressors (a k-NN, a DT, and a RF) as for the classification case study. Ground truth and predictions are reported in Figure 5a. To assess the predictive performance of the algorithms the R^2 value has been adopted. Corresponding

	Ambient temp. (AT)	Exhaust vacuum (EV)	Ambient pressure (AP)	Rel. humidity (RH)	Net hourly electrical energy output
Cluster 1	6.2 – 32.5	35.4 – 50.2	998.1 – 1026.4	35.6 – 100.1	499.9 - 2.2 AP - 0.3 AT - 0.1 EV
Cluster 2	6.2 – 32.5	34.0 – 50.2	997.9 – 1026.4	35.6 – 100.1	697.9 - 1.8 AP - 2.0 AT + 0.6 EV
Cluster 3	6.2 – 35.8	25.4 – 81.6	997.8 – 1026.5	25.6 – 100.1	234.7 - 1.4 AP - 0.3 AT + 0.3 RH
Cluster 4	2.3 – 35.8	25.4 – 81.6	992.9 – 1033.3	25.6 – 100.2	628.2 - 2.2 AP - 0.5 AT - 0.2 EV

Table 2
Example of ExACT clustering for the CCPP data set.

measurements are shown in Figure 5b. Once again, ExACT has a predictive performance comparable or even superior to that of ML models, and its predictions are human-interpretable due to its hypercubic approximation strategy.

An example of ExACT's explainable clustering applied to a regression task is reported in Table 2 for the CCPP data set. The corresponding algorithm parameters are $\delta = 3$, $\theta = 0.02$ and $\xi = 2$. We stress here that when ExACT is applied for regression the recursive refinement of hypercubic approximations is performed only for cubes having mean absolute predictive error greater than the θ threshold. It is worthwhile to notice that the outputs shown in Table 2 are linear combinations of the input variables. It is as well possible to obtain constant outputs, to the benefit of human interpretability but at the expense of predictive performance.

5. Conclusions

In this paper we present an explainable clustering technique named ExACT, applicable to any kind of task described by a data set having continuous input features. No constraints are assumed on the output feature. This algorithm takes advantage of GMMs and DBSCAN to detect clusters and approximate them with human-interpretable hypercubic regions described in terms of interval inclusion constraints on the input features. Our experiments prove the effectiveness of ExACT performing explainable clustering, classification and regression in comparison to other state-of-the-art traditional and explainable clustering techniques, but also w.r.t. ML classifiers and regressors.

Our future works will be focused on enhancing the rationale behind the ExACT's region approximation and possibly on the adoption of deep clustering techniques instead of GMMs and DBSCAN, as in the current version. Furthermore, ExACT may benefit from an automatic technique enabling parameter auto-tuning and in particular we plan to implement a procedure aimed at highlighting the best values for the maximum depth and the predictive error threshold parameters.

Acknowledgments

This work has been supported by European Union's Horizon Europe AEQUITAS research and innovation programme under grant number 101070363.

References

[1] European Commission, C. Directorate-General for Communications Networks, Technology, Ethics guidelines for trustworthy AI, Publications Office, 2019. doi:doi/10.2759/346720.

[2] European Commission, AI Act – Proposal for a regulation of the european parliament and the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, 2021.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (2018) 1–42. doi:10.1145/3236009.

[4] S. Ayache, R. Eyraud, N. Goudian, Explaining black boxes on sequential data using weighted automata, in: International Conference on Grammatical Inference, PMLR, 2019, pp. 81–103.

[5] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.

[6] E. M. Kenny, C. Ford, M. Quinn, M. T. Keane, Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies, Artificial Intelligence 294 (2021) 103459. doi:10.1016/j.artint.2021.103459.

[7] M. Moshkovitz, S. Dasgupta, C. Rashtchian, N. Frost, Explainable k-means and k-medians clustering, in: International conference on machine learning, PMLR, 2020, pp. 7055–7065.

[8] F. Sabbatini, R. Calegari, Explainable clustering with CREAM, in: P. Marquis, C. S. Tran, G. Kern-Isberner (Eds.), 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023), IJCAI Organization, Rhodes, Greece, 2023, pp. 593–603. doi:10.24963/kr.2023/58.

[9] R. Calegari, G. Ciatto, A. Omicini, On the integration of symbolic and sub-symbolic techniques for XAI: A survey, Intelligenza Artificiale 14 (2020) 7–32. doi:10.3233/IA-190036.

[10] J. Huysmans, B. Baesens, J. Vanthienen, ITER: An algorithm for predictive regression rule extraction, in: Data Warehousing and Knowledge Discovery (DaWaK 2006), Springer, 2006, pp. 270–279. doi:10.1007/11823728_26.

[11] F. Sabbatini, G. Ciatto, A. Omicini, GridEx: An algorithm for knowledge extraction from black-box regressors, in: D. Calvaresi, A. Najjar, M. Winikoff, K. Främling (Eds.), Explainable and Transparent AI and Multi-Agent Systems. Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers, volume 12688 of *LNCS*, Springer Nature, Basel, Switzerland, 2021, pp. 18–38. doi:10.1007/978-3-030-82017-6_2.

[12] F. Sabbatini, R. Calegari, Symbolic knowledge extraction from opaque machine learning predictors: GridREx & PEDRO, in: G. Kern-Isberner, G. Lakemeyer, T. Meyer (Eds.), Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, KR 2022, Haifa, Israel. July 31 - August 5, 2022, 2022. URL: https://proceedings.kr.org/2022/57/. doi:10.24963/kr.2022/57.

[13] K. P. Murphy, Machine learning – A probabilistic perspective, Adaptive computation and machine learning series, MIT Press, 2012.

[14] R. F. Ling, On the theory and construction of k-clusters, The Computer Journal 15 (1972) 326–332. URL: https://doi.org/10.1093/comjnl/15.4.326. doi:10.1093/comjnl/15.4.326. arXiv:https://academic.oup.com/comjnl/article-pdf/15/4/326/1005965/150326.pdf.

[15] M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: E. Simoudis, J. Han, U. M. Fayyad (Eds.), Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA, AAAI Press, 1996, pp. 226–231. URL: http://www.aaai.org/Library/KDD/1996/kdd96-037.php.

[16] J. Jang, H. Jiang, DBSCAN++: towards fast and scalable density clustering, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 3019–3029. URL: http://proceedings.mlr.press/v97/jang19a.html.

[17] M. Ankerst, M. M. Breunig, H. Kriegel, J. Sander, OPTICS: ordering points to identify the clustering structure, in: A. Delis, C. Faloutsos, S. Ghandeharizadeh (Eds.), SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA, ACM Press, 1999, pp. 49–60. URL: https://doi.org/10.1145/304182.304187. doi:10.1145/304182.304187.

[18] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: an efficient data clustering method for very large databases, in: H. V. Jagadish, I. S. Mumick (Eds.), Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996, ACM Press, 1996, pp. 103–114. URL: https://doi.org/10.1145/233269.233324. doi:10.1145/233269.233324.

[19] S. P. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (1982) 129–136. URL: https://doi.org/10.1109/TIT.1982.1056489. doi:10.1109/TIT.1982.1056489.

[20] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (1995) 790–799. URL: https://doi.org/10.1109/34.400568. doi:10.1109/34.400568.

[21] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 888–905. URL: https://doi.org/10.1109/34.868688. doi:10.1109/34.868688.

[22] S. X. Yu, J. Shi, Multiclass spectral clustering, in: 9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France, IEEE Computer Society, 2003, pp. 313–319. URL: https://doi.org/10.1109/ICCV.2003.1238361. doi:10.1109/ICCV.2003.1238361.

[23] N. Rahmah, I. S. Sitanggang, Determination of optimal epsilon (eps) value on DBSCAN algorithm to clustering data on peatland hotspots in Sumatra, in: IOP conference series: earth and environmental science, volume 31, IOP Publishing, 2016, p. 012012.

[24] S. Dasgupta, N. Frost, M. Moshkovitz, C. Rashtchian, Explainable k-means and k-medians clustering, CoRR abs/2002.12538 (2020). URL: https://arxiv.org/abs/2002.12538. arXiv:2002.12538.

[25] J. Basak, R. Krishnapuram, Interpretable hierarchical clustering by constructing an unsupervised decision tree, IEEE Trans. Knowl. Data Eng. 17 (2005) 121–132. URL: https://doi.org/10.1109/TKDE.2005.11. doi:10.1109/TKDE.2005.11.

[26] R. Fraiman, B. Ghattas, M. Svarc, Interpretable clustering using unsupervised binary trees, Adv. Data Anal. Classif. 7 (2013) 125–145. URL: https://doi.org/10.1007/s11634-013-0129-3. doi:10.1007/s11634-013-0129-3.

[27] D. Bertsimas, A. Orfanoudaki, H. M. Wiberg, Interpretable clustering via optimal trees, CoRR abs/1812.00539 (2018). URL: http://arxiv.org/abs/1812.00539. arXiv:1812.00539.

[28] J. Chen, Y. Chang, B. Hobbs, P. J. Castaldi, M. H. Cho, E. K. Silverman, J. G. Dy, Interpretable clustering via discriminative rectangle mixture model, in: F. Bonchi, J. Domingo-Ferrer, R. Baeza-Yates, Z. Zhou, X. Wu (Eds.), IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain, IEEE Computer Society, 2016, pp. 823–828. URL: https://doi.org/10.1109/ICDM.2016.0097. doi:10.1109/ICDM.2016.0097.

[29] X. Chen, S. Güttel, Fast and explainable clustering based on sorting, CoRR abs/2202.01456 (2022). URL: https://arxiv.org/abs/2202.01456. arXiv:2202.01456.

[30] L. Manduchi, M. Hüser, M. Faltys, J. E. Vogt, G. Rätsch, V. Fortuin, T-DPSOM: an interpretable clustering method for unsupervised learning of patient health states, in: M. Ghassemi, T. Naumann, E. Pierson (Eds.), ACM CHIL '21: ACM Conference on Health, Inference, and Learning, Virtual Event, USA, April 8-9, 2021, ACM, 2021, pp. 236–245. URL: https://doi.org/10.1145/3450439.3451872. doi:10.1145/3450439.3451872.

[31] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Hypercube-based methods for symbolic knowledge extraction: Towards a unified model, in: A. Ferrando, V. Mascardi (Eds.), WOA 2022 – 23rd Workshop "From Objects to Agents", volume 3261 of *CEUR Workshop Proceedings*, Sun SITE Central Europe, RWTH Aachen University, 2022, pp. 48–60. URL: http://ceur-ws.org/Vol-3261/paper4.pdf.

[32] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Towards a unified model for symbolic knowledge extraction with hypercube-based methods, Intelligenza Artificiale 17 (2023) 63–75. URL: https://doi.org/10.3233/IA-230001. doi:10.3233/IA-230001.

[33] A. Blanco-Justicia, J. Domingo-Ferrer, Machine learning explainability through comprehensible decision trees, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2019, pp. 15–26.

[34] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, On the design of PSyKE: A platform for symbolic knowledge extraction, in: R. Calegari, G. Ciatto, E. Denti, A. Omicini, G. Sartor (Eds.), WOA 2021 – 22nd Workshop "From Objects to Agents", volume 2963 of *CEUR Workshop Proceedings*, Sun SITE Central Europe, RWTH Aachen University, 2021, pp. 29–48. 22nd Workshop "From Objects to Agents" (WOA 2021), Bologna, Italy, 1–3 September 2021. Proceedings.

[35] F. Sabbatini, G. Ciatto, R. Calegari, A. Omicini, Symbolic knowledge extraction from opaque ML predictors in PSyKE: Platform design & experiments, Intelligenza Artificiale 16 (2022) 27–48. URL: https://doi.org/10.3233/IA-210120. doi:10.3233/IA-210120.

[36] F. Sabbatini, G. Ciatto, A. Omicini, Semantic Web-based interoperability for intelligent agents with PSyKE, in: D. Calvaresi, A. Najjar, M. Winikoff, K. Främling (Eds.), Explainable and Transparent AI and Multi-Agent Systems, volume 13283 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 124–142. URL: http://link.springer.com/10.1007/978-3-031-15565-9_8. doi:10.1007/978-3-031-15565-9_8.

[37] R. Calegari, F. Sabbatini, The PSyKE technology for trustworthy artificial intelligence 13796 (2023) 3–16. URL: https://doi.org/10.1007/978-3-031-27181-6_1. doi:10.1007/

978-3-031-27181-6_1, xXI International Conference of the Italian Association for Artificial Intelligence, AIxIA 2022, Udine, Italy, November 28 – December 2, 2022, Proceedings.

[38] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 179–188. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x. doi:https://doi.org/10.1111/j.1469-1809.1936.tb02137.x. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x.

[39] M. Forina, R. Leardi, C. Armanino, S. Lanteri, P. Conti, P. Princi, Parvus: An extendable package of programs for data exploration, classification and correlation, Journal of Chemometrics 4 (1988) 191–193.

[40] W. N. Street, W. H. Wolberg, O. L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis, in: R. S. Acharya, D. B. Goldgof (Eds.), Biomedical Image Processing and Biomedical Visualization, volume 1905, International Society for Optics and Photonics, SPIE, 1993, pp. 861 – 870. URL: https://doi.org/10.1117/12.148698. doi:10.1117/12.148698.

[41] L. Hubert, P. Arabie, Comparing partitions, Journal of classification 2 (1985) 193–218.

[42] X. V. Nguyen, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, Journal of Machine Learning Research 11 (2010) 2837–2854. URL: https://dl.acm.org/doi/10.5555/1756006.1953024. doi:10.5555/1756006.1953024.

[43] E. B. Fowlkes, C. L. Mallows, A method for comparing two hierarchical clusterings, Journal of the American statistical association 78 (1983) 553–569.

[44] A. Rosenberg, J. Hirschberg, V-Measure: A conditional entropy-based external cluster evaluation measure, in: J. Eisner (Ed.), EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic, ACL, 2007, pp. 410–420. URL: https://aclanthology.org/D07-1043/.

[45] P. Tüfekci, Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods, International Journal of Electrical Power & Energy Systems 60 (2014) 126–140. URL: https://www.sciencedirect.com/science/article/pii/S0142061514000908. doi:https://doi.org/10.1016/j.ijepes.2014.02.027.

[46] O. Akbilgic, H. Bozdogan, M. E. Balaban, A novel hybrid rbf neural networks model as a forecaster, Statistics and Computing 24 (2014) 365–375.

[47] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, The Annals of Statistics 32 (2004) 407 – 499. URL: https://doi.org/10.1214/009053604000000067. doi:10.1214/009053604000000067.