Annals of Emerging Technologies in Computing (AETiC)
Vol. 5, No. 2, 2021

Research Article

Customer Choice Modelling: A Multi-Level Consensus Clustering Approach

Nicolas Pasquier[1,*] and Sujoy Chatterjee[2]

[1]Université Côte d'Azur, CNRS, I3S, Sophia Antipolis, France
nicolas.pasquier@univ-cotedazur.fr
[2]University of Petroleum and Energy Studies, Dehradun, India
sujoy.chatterjee@ddn.upes.ac.in
*Correspondence: nicolas.pasquier@univ-cotedazur.fr

Received: 18th February 2021; Accepted: 17th March 2021; Published: 1st April 2020

Abstract: Customer Choice Modeling aims to model the decision-making process of customers, or segments of customers, through their choices and preferences identified by the analysis of their behaviors in one or more specific contexts. Clustering techniques are used in this context to identify patterns in their choices and preferences, to define segments of customers with similar behaviors, and to model how customers of different segments respond to competing products and offers. However, data clustering is an unsupervised learning task by nature, that is the grouping of customers with similar behaviors in clusters must be performed without prior knowledge about the nature and the number of intrinsic groups of data instances, i.e., customers, in the data space. Thus, the choice of both the clustering algorithm used and its parameterization, and of the evaluation method used to assess the relevance of the resulting clusters are central issues. Consensus clustering, or ensemble clustering, aims to solve these issues by combining the results of different clustering algorithms and parameterizations to generate a more robust and relevant final clustering result. We present a Multi-level Consensus Clustering approach combining the results of several clustering algorithmic configurations to generate a hierarchy of consensus clusters in which each cluster represents an agreement between different clustering results. A closed sets based approach is used to identified relevant agreements, and a graphical hierarchical representation of the consensus cluster construction process and their inclusion relationships is provided to the end-user. This approach was developed and experimented in travel industry context with Amadeus SAS. Experiments show how it can provide a better segmentation, and refine the customer segments by identifying relevant sub-segments represented as sub-clusters in the hierarchical representation, for Customer Choice Modeling. The clustering of travelers was able to distinguish relevant segments of customers with similar needs and desires (i.e., customers purchasing tickets according to different criteria, like price, duration of flight, lay-over time, etc.) and at different levels of precision, which is a major issue for improving the personalization of recommendations in flight search queries.

Keywords: ***Consensus Clustering; Ensemble Clustering; Multi-level Clustering; Closed Sets; Clusters Hierarchy; Customer Choice Modelling***

1. Introduction

This article is an extended version of article [1] published in the iCETiC'2020 international conference during which he received the best paper award. The Multi-level Consensus Clustering framework presented is extended here with the description of the algorithmic processes involved by the implementation of the framework in the general context of Customer Choice Modelling, considering both the context of unsupervised clustering, where no background information is used in the process, and semi-supervised clustering, where background knowledge can be introduced in the process to improve the relevance of the results. Technical and scientific challenges related to the different steps of the approach workflow in both unsupervised and semi-supervised application contexts are highlighted. The

Nicolas Pasquier and Sujoy Chatterjee. "Customer Choice Modelling: A Multi-Level Consensus Clustering Approach", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 103-120, Vol. 5, No. 2, 1st April 2021, Published by International Association of Educators and Researchers (IAER), DOI: 10.33166/AETiC.2021.02.009, Available: http://aetic.theiaer.org/archive/v5/v5n2/p9.html.

developments of the approach and it experimental evaluation conducted with Amadeus IT Group for the optimization of the flight search recommendation engine through Customer Choice Modelling are presented.

In travel industry, the Customer Choice Modelling application aims to model the decision process of a traveler, or a category of travelers, the analysis and the prediction of his preferences and the choices he makes in different contexts. Since the needs and wishes of travelers vary according to different features, like the number of children they have, the trip duration or the price of the tickets for example, a better understanding of travelers behaviors, through the segmentation of travelers according to their distinct characteristics, is necessary for improving travel search query recommendations.

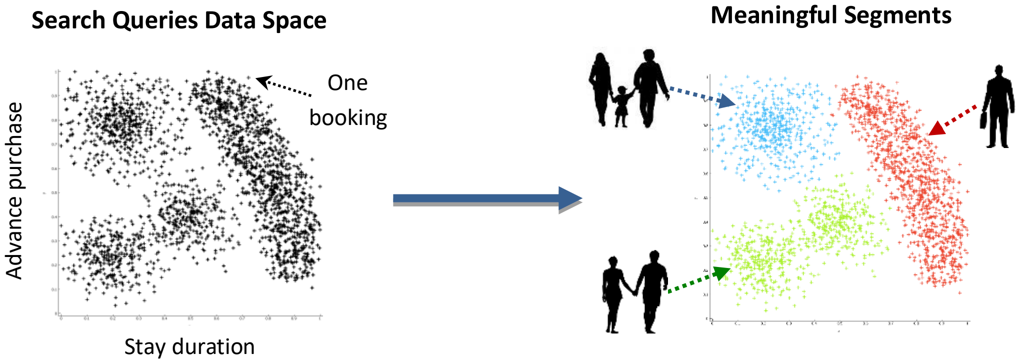


Figure 1. Clustering of Search Queries for Customer Segment Identification.

The use of clustering techniques in Customer Choice Modeling aims to discriminate the segments of customers, or business classes, according to their properties in the data space as outlined in Fig. 1. Customer segments are identified as clusters, i.e. groups with similar properties, of customers in the data space of travel search queries. This data space is defined by the traveler search query parameters and their results, such as the booking of a proposed travel or service.

The characterization of the resulting clusters aims to identify the different segments of customers, each segment corresponding to a category of travelers with different needs and requirements as outlined in Fig. 2. During this step, the specific features of each cluster and their weight in the booking result probabilities are extracted by a comparative analysis of the clusters. Finally, for each segment, personalized booking options can be defined according to this characterization of clusters. New search queries recommendations can then be adapted according to the segment they correspond to.

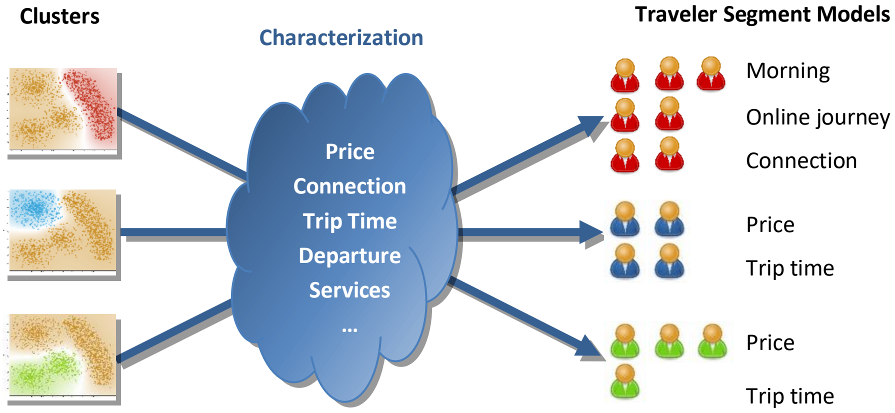


Figure 2. Characterization of Search Query Clusters for Customer Segment Modelling.

While many clustering algorithms have been proposed in the literature, it is widely agreed that none of them can generate a relevant clustering result in all contexts. Indeed, each clustering algorithm is based on a different assumption about the subjacent model of the distribution of instances in the data space, e.g., density-based or centroid-based. The parameterization of the algorithm defines a way to put this model into practice on the dataset. See [2-6] for comprehensive reviews about clustering algorithms. Choosing an adequate algorithmic configuration, that is choosing an algorithm and setting its parameters, for clustering a dataset is a challenging central issue since the relevance of the resulting clustering relies on how well it is suitable for the characteristics of the data space being analysed [7-9].

The resulting clusterings of a dataset are usually evaluated using unsupervised evaluation measures. These measures are called internal validation measures as they are based solely on the properties of clusters in the data space and do not use other information, making them unsupervised by nature. Each internal validation measure evaluates how much the clusters match a specific underlying model of the distribution of instances in the data space. Hence, different measures can provide different results for the same clustering and overrate clustering results from algorithms that are based on the same assumption about the data distribution as the measure. See [10-14] for extensive studies about clustering validation measures.

To overcome the issue of the algorithmic configuration choice, different algorithmic configurations providing different clustering solutions for the same dataset, consensus clustering approaches were proposed. These approaches combine clusters extracted by diverse clustering algorithmic configurations, called base clusterings, to generate consensus clusters corresponding to agreements between base clusters for improving clustering robustness. The set of base clusterings is also called the ensemble and the consensus clustering approach called ensemble clustering in the literature. See [15-18] for comprehensive reviews and studies on ensemble clustering algorithmic approaches. The evaluation of the relevance of a consensus clustering is performed by the analytical comparison between clusters in the clustering solution and clusters in the base clusterings. The most frequently used measures are the Adjusted Rand Index (ARI) and the Normalized Mutual Information (NMI) that evaluates the relevance of the consensus clustering as its average similarity with all base clusterings in the ensemble [19-22]. Such consensus clustering validation measures provide an efficient solution to identify and rank the best agreements among all the base clusterings regarding the possible different data distribution models, e.g., density-based or centroid-based, in sub-spaces of the data space corresponding to clusters.

In order to characterize the behavior of customers, appropriate segmentation of customers is highly needed. On the other hand, most of the clustering algorithms assume some specific dataspace distribution over the dataset while producing the clusters. Therefore, the different clustering algorithms applied even on the same dataset may generate the different diverse clustering solutions. Moreover, from the perspective of customer search data in the travel context, it is very difficult to know the prior information regarding the number of clusters over the customers. There is limited research that address the issues of customers segmentation resulting from different clustering algorithms. Note that, each clustering algorithm seeks to provide the actual number of clusters when applied to the dataset. Therefore, motivated by these shortcomings, consensus clustering can act as a major role to find better clustering over the dataset. In this proposed conceptual model, the effort is made to find the better segmentation of customers without specifying the actual number of clustering from the individual base clustering having number of clusters in a certain range.

The article is organized as follows. Section 2 presents the proposed framework, Section 3 details the algorithmic process of the proposed framework, Section 4 describes the technical and scientific challenges addressed, and Section 5 concludes the article.

2. Multiple Consensus Clustering Framework

The proposed framework was developed based on the Multiple Consensus Clustering approach introduced with the MultiCons algorithm [23]. This approach is a multi-level clustering approach providing as a result a hierarchical decomposition of the consensus clusters generated. In this hierarchy, named *ConsTree* for tree of consensuses, the levels depict consensus clusterings of the dataset, each level corresponding to a different number of agreements between the base clusterings. In multi-level clustering, a cluster at a level in the produced hierarchy can be decomposed into several smaller clusters in the sub-levels of the hierarchy. This hierarchy can then be presented to the end-user as tree-like graphical representation where nodes are clusters and edges represent inclusion relationships between clusters of successive levels. The proposed framework can be adapted to other multi-level clustering approaches.

The benefit of multi-level clustering in Customer Choice Modelling is to provide a data representation context that can both discriminate the business classes, i.e., segments of customers, according to their properties in the data space and refine them by distinguishing different sub-classes of a class, representing sub-segments of customers, according to the different modeling properties of each sub-cluster in the data space [24].

2.1. Multiple Consensus Clustering Approach

Multi-level clustering provides a relevant framework for the simultaneous identification of business classes and sub-classes as illustrated in Fig. 3. In this example, we assume the original dimensions of the dataset representing travel characteristics are summarized through a two-dimensional reduction, such as obtained by a component reduction approach for example, and the generated clusters in this two-dimensional data space, representing customer segments, are characterized by their distinctive features regarding dimensions in the initial data space. In this schematic example, the customer segment C-2 is specialized into two customer sub-segments, namely C-2-1 and C-2-2, corresponding to two sub-clusters. These sub-clusters can be identified as two subspaces corresponding to significant variations in density in the data space of segment C-2 represented as a green area.

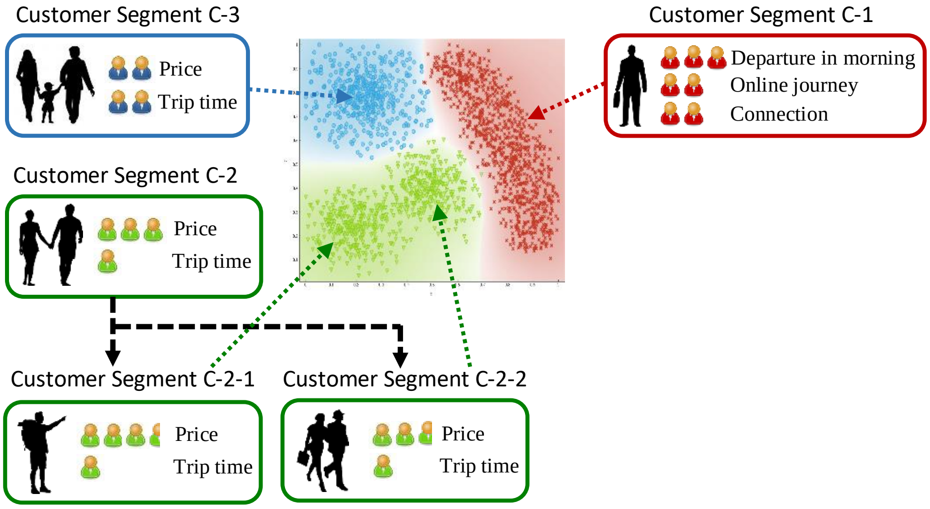


Figure 3. Business Segment Specialization by Multi-level Consensus Cluster Analysis.

The objective of multiple consensus clustering is to identify such a specialization of business classes in the generated hierarchy of consensus clusters. We can observe in the example two-dimensional data space in Fig. 3 that the variations in the density of data points in the sub-spaces corresponding to clusters C-1, C-2 and C-3 can enable their identification using a density-based clustering algorithm by choosing appropriate values for the size and density of neighbourhood algorithm parameters. Furthermore, the sub-spaces corresponding to clusters C-2-1 and C-2-2 can be distinguished in the sub-space of cluster C-2 by choosing different adequate values for these parameters. Then, in the resulting hierarchy of consensus clusters such as represented in the tree of consensuses shown in Fig. 4, a level of the hierarchy will correspond to clusters C-1, C-2 and C-3 and a lower level in the hierarchy will contain the four clusters C-1, C-2-1, C-2-2 and C-3. The second of the above-mentioned levels will be a sub-level of the first that corresponds to a higher rate of agreements among the base clusterings. Note that in the tree of consensuses representation, the size of nodes is proportional to the number of instances the corresponding cluster contains.

2.2. Traveler Choice Modelling Problem Decomposition

The proposed multiple consensus clustering framework can be viewed as a semi-supervised algorithmic process in the sense that it combines unsupervised internal validation of multi-level consensus clusters and supervised business metric based *external validation* of multi-level consensus clusters. Interested readers can refer to [25-27] for definitions and studies related to semi-supervised clustering concepts. It relies on the decomposition of the problem of traveler choice modelling into the three following tasks:

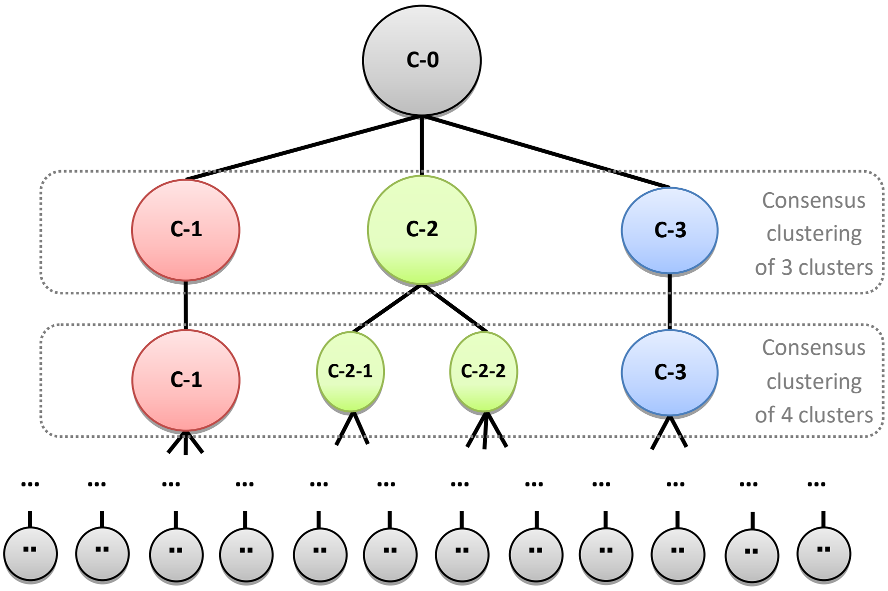

C-0
C-1
C-2
C-3
Consensus clustering of 3 clusters
C-1
C-2-1
C-2-2
C-3
Consensus clustering of 4 clusters

Figure 4. Example Representation of Business Segment Specialization in the Multiple Consensus Clustering Tree-like Hierarchy.

1. **Identify traveler segments: How can search queries be grouped by similarity?**
   The first task is to identify segments of travelers, each segment corresponding to a category of travelers with different needs and requirements. A segment can be refined and represented as several clusters in the data space corresponding to slightly different features, i.e., sub-segments.
2. **Understand traveler choice patterns: What is the likelihood of a search offer to be booked?**
   The second task consists to learn a predictive model for assessing the probability of a travel search query to lead to a booking or not through the analysis of the features of successful and unsuccessful search queries.
3. **Optimize bookings for each segment: What really matters and to which extent it does?**
   The third task is to connect clusters with traveler classes so that each cluster is representative of a segment, or a sub-segment, of travelers, and to identify discriminative feature of clusters, i.e. search queries feature values that distinguish the segments.

This decomposition of the problem of Customer Choice Modeling relies on the capability of multi-level consensus clustering to distinguish sub-segments of the predefined customer segments when each sub-segment corresponds to slightly different properties regarding its instance modeling in the data space compared to other sub-segments.

2.3. Multi-Level Consensus Clustering Framework for Customer Choice Modelling

The proposed framework relies on a sequential process that integrates successively unsupervised, semi-supervised and supervised techniques to identify customer segments and sub-segments, according to the similarity of their searching and booking activities, that are as significant as possible from a business process viewpoint.

An overview of the framework process is shown in Fig. 5. This process first builds multi-level consensus clusters, evaluates these clusters, and selects the most relevant ones considering both internal and external validations. Then, an interactive analysis of the hierarchical relationships between clusters depicted in the tree-like representation provides the end-user with a visual illustration for exploring and identifying the most relevant clusters and the business segments they correspond to. The most important criteria (ranges of values for variables price, trip duration, connections, etc.) for delimitating each customer segment are then identified according to prior expertise and the automatic characterization of the clusters they correspond to. This distinctive characterization of segments will then allow to predict the segment of a new customer by assigning him/her to the segment represented by the cluster which characterization vector is the most similar to the customer, that is the closest cluster in the data space.

This interactive process starts with the preprocessing of the dataset according to end-users choices, arising from dataset exploration, in order to ensure the applicability of clustering algorithmic configurations used to generate the base clusterings. These algorithmic configurations are defined to ensure that two central properties of the clustering ensemble are satisfied. The first is the required diversity of the search space for consensus clusters, that is the ensemble of base clusterings should cover a sufficiently wide range of clustering approaches and parameterizations. The second is to ensure the robustness of the final solution by centering this search space on the number of clusters corresponding to optimal internal and external validation measures according to the number of base clustering connected components. Then, the clustering ensemble is represented as a refined membership matrix depicting assignments to base clusters for each instance. Galois closed patterns are extracted from the matrix to identify all existing agreements to cluster instances together between the base clusterings. These closed patterns correspond each to a maximal, regarding inclusion relation, set of instances clustered together and its associated maximal, regarding inclusion relation, set of base clusters containing these instances. They are then iteratively processed in increasing order of their number of base clusters for generating clustering patterns, each one representing an agreements for clustering a (maximal) set of instances. A consensus function is then applied to the clustering patterns as a merge/split process, considering their properties regarding the number of agreements and disagreements between base clusterings on grouping the sets of instances they correspond to, for generating consensus clusters. This closed patterns-based process can treat datasets with very large number of instances N since, contrarily to most other consensus clustering approaches, it does not require the processing of a co-association matrix of size N^2 but only of a membership matrix which size is N.M, where M is the number of base clusters, with M << N, and regarding the demonstrated scalability properties of Galois closed sets extraction algorithms [28-30].

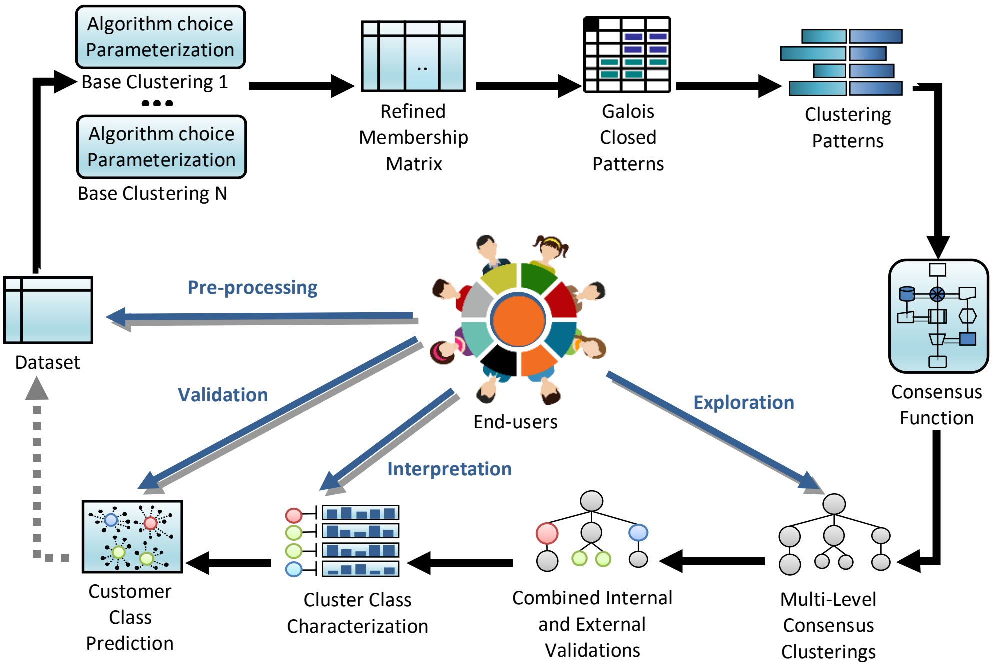


Figure 5. Multi-Level Consensus Clustering Framework.

Generated consensus clusters and their hierarchical relationships, regarding inclusion relation, are graphically represented in the tree of consensuses. Each level of this graphical representation depicts a consensus clustering, i.e., a partitioning of all instances in the dataset, and each node of a level represents a consensus cluster, that is a maximal grouping of instances agreed among base clusterings. The edges between nodes of two successive levels represent cluster regroupings leading to a new consensus cluster of instances. Depicting the consensuses creation process, this visualization allows the end-users to choose the most relevant result among the different consensus clustering solutions, i.e., between different levels of agreements among the base clusterings. The clustering solution having the best overall similarity with the clustering ensemble is recommended in the graphical representation as the final consensus clustering solution. This MultiCons approach visualization is extended in this framework to facilitate and precise the interpretation of the consensus cluster creation process and their properties, and to allow the end-users to choose the most relevant consensus multi-level clusters that can originate from different consensus clusterings, i.e., different levels of the hierarchy. Algorithmic and statistical methods developed for this extension consider the properties of the structure of the hierarchy, e.g., the stability of consensus clusters and not only the stability of consensus clusterings, and the relationships between clusters in the data space, e.g., overlapping sets between sets of instances and sets of base clusters that define the clustering patterns and weighting of base clusterings according to their number of clusters. The stability of a consensus cluster refers to the individual recurrence of a group of instances among successive levels of the hierarchy while the stability of a consensus clustering refers to the recurrence of a partitioning of all instances, i.e., a set of clusters, among successive levels of the hierarchy.

This automatic, or semi-automatic depending on end-user preferences, processing of the hierarchical tree of consensuses structure allows to generate new internal and external validation measurements for each cluster, based on closed pattern properties in the data space, that are significant to characterize each selected consensus cluster and distinguish it from others selected consensus clusters. From these characterizations, a vector that is representative and distinctive of each cluster is generated. Then, the business segment of new instances, regarding business metrics, is predicted using a mapping function that assigns new instances to their closest cluster in the data space identified as the most similar cluster characterization vector. Preliminary experimental results on the comparison of this closed patterns-based multiple consensus clustering approach and other state-of-the-art consensus clustering approaches were conducted in collaboration with Amadeus IT Group. They showed the relevance of the resulting consensus clusters regarding Amadeus business metrics used for flight search recommendations.

The most relevant and significant results of the validation by the end-users of the predictions of the assigned segment to instances can be integrated in subsequent iterations of the process. These results can be represented as cannot-link and must-link constraints in order to use semi-supervised clustering algorithms among the base clusterings for example.

3. Multiple Consensus Clustering Process

This section presents the process of the proposed ensemble clustering based framework as a flowchart with identified scientific and technological challenges for each step. This process can be decomposed in the three following phases:

1. Data Exploration and Preprocessing.
2. Multi-Level Ensemble Consensus Clustering.
3. Clusters to Classes Learning.

These three phases, with their different steps and associated challenges, are detailed in the three following subsections. For each phase, a flowchart depicting its workflow, with its successive steps and the related scientific (theoretical) and technological (technical) challenges, is given.

3.1. Data Exploration and Preprocessing

The workflow of this phase is depicted in Fig. 6. It aims at generating the dataset that will be processed by clustering algorithms from internal data source, and, depending on the application objectives, from potential external data sources, such as changes in currency conversion rates for instance.

The *background knowledge integration* step deals with all questions relating to the use of external knowledge in the data preprocessing phase of the process. This step is optional in the process depending on the application addressed. In the context of Customer Choice Modelling for Amadeus flight search recommendation engine, no external data source was involved. However, in different contexts, such as financial or accounting applications for example, external information such as historical data about conversion rates of currencies can be involved in the data integration process.

The *data preprocessing step* aims to generate a data representation as much as possible adequate for an efficient processing by clustering algorithms, regarding both the computation cost and the relevance of the results. This step involves the use of classical tasks for the integration of heterogenous data from different sources, the processing of data noise and the optimization of data representation and storage considering the constraints of the application and the clustering algorithms used for the generation of base clusterings.

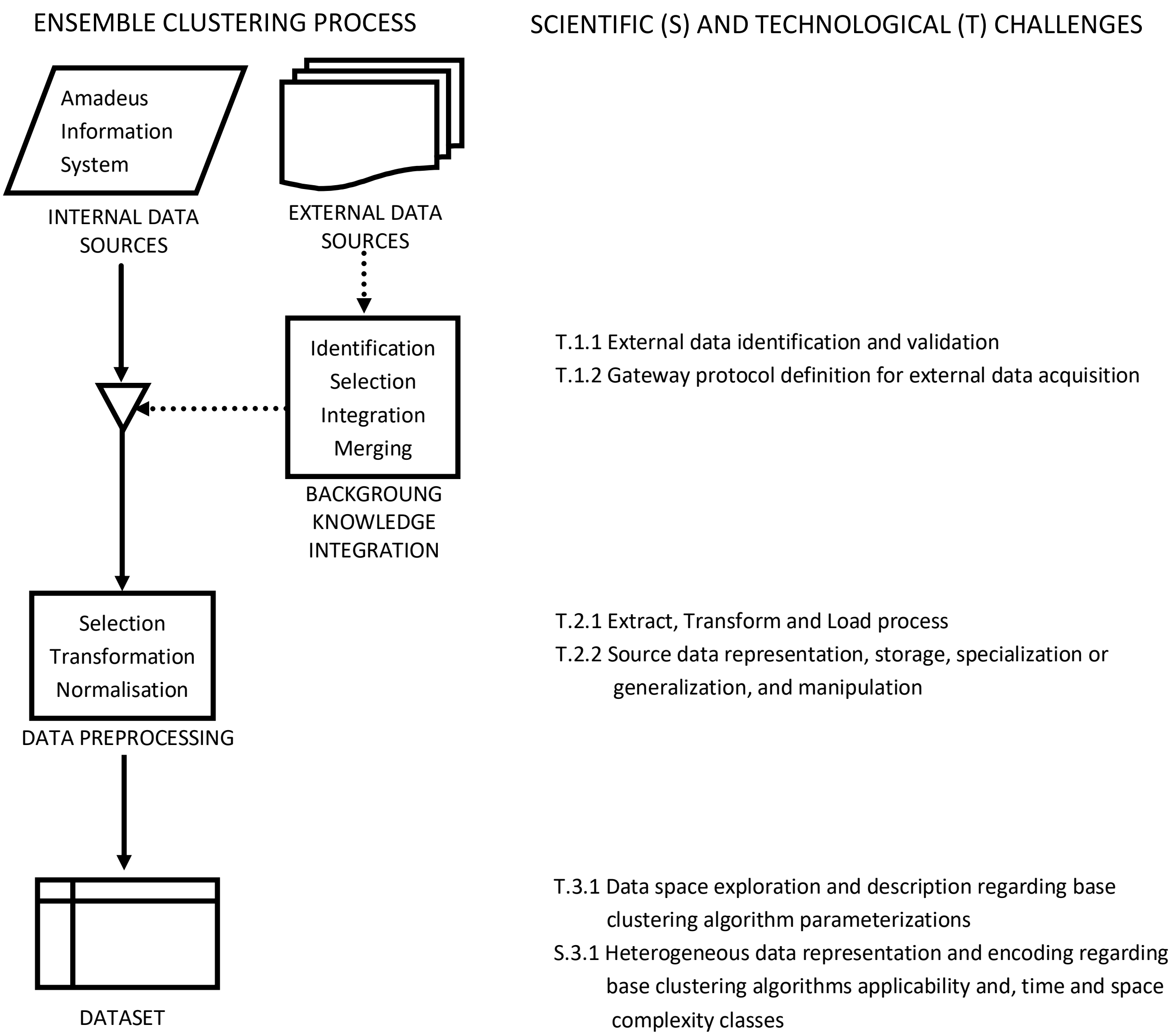

ENSEMBLE CLUSTERING PROCESS
SCIENTIFIC (S) AND TECHNOLOGICAL (T) CHALLENGES
Amadeus Information System
INTERNAL DATA SOURCES
EXTERNAL DATA SOURCES
Identification Selection Integration Merging
BACKGROUNG KNOWLEDGE INTEGRATION
T.1.1 External data identification and validation
T.1.2 Gateway protocol definition for external data acquisition
Selection Transformation Normalisation
DATA PREPROCESSING
T.2.1 Extract, Transform and Load process
T.2.2 Source data representation, storage, specialization or generalization, and manipulation
DATASET
T.3.1 Data space exploration and description regarding base clustering algorithm parameterizations
S.3.1 Heterogeneous data representation and encoding regarding base clustering algorithms applicability and, time and space complexity classes

Figure 6. Workflow of the Data Exploration and Preprocessing Phase.

The *data space exploration step* aims to better understand the data space (e.g., types and number of variables, number of objects, presence of noise or correlated variables, etc.) using statistical and algorithmic tools. The objective is to automatize as far as possible the definition of algorithmic configurations used for the generation of base clusterings (e.g., appropriate range of values for extracted number of clusters) during the next phase. This step is detailed in section 4.1 regarding the challenges it involves.

3.2. Multi-level Ensemble Consensus Clustering

The workflow of this phase is depicted in Fig. 7. It aims to create a hierarchy of consensus clusters from the dataset and generate a tree-like graphical representation of this hierarchy depicting the creation process of consensus clusters.

During the first step, different clustering algorithmic configurations are applied to generate the base clusterings represented in a unified format in the clustering ensemble. The main challenge of this step is the definition of the algorithmic configurations used to generate the set of base clusterings, i.e., the clustering ensemble, that will define the search space for the consensus generation. The estimation of the most probable number of clusters inherent to the data space structure is a crucial step to define an interval of values for the number of clusters generated in base clusterings. This interval must also ensure a sufficient diversity in the base clustering solutions and the levels of resolution (e.g., size) of their clusters. This step is detailed in section 4.2 regarding the challenges it involves, and the solution based on connected components used in the proposed approach.

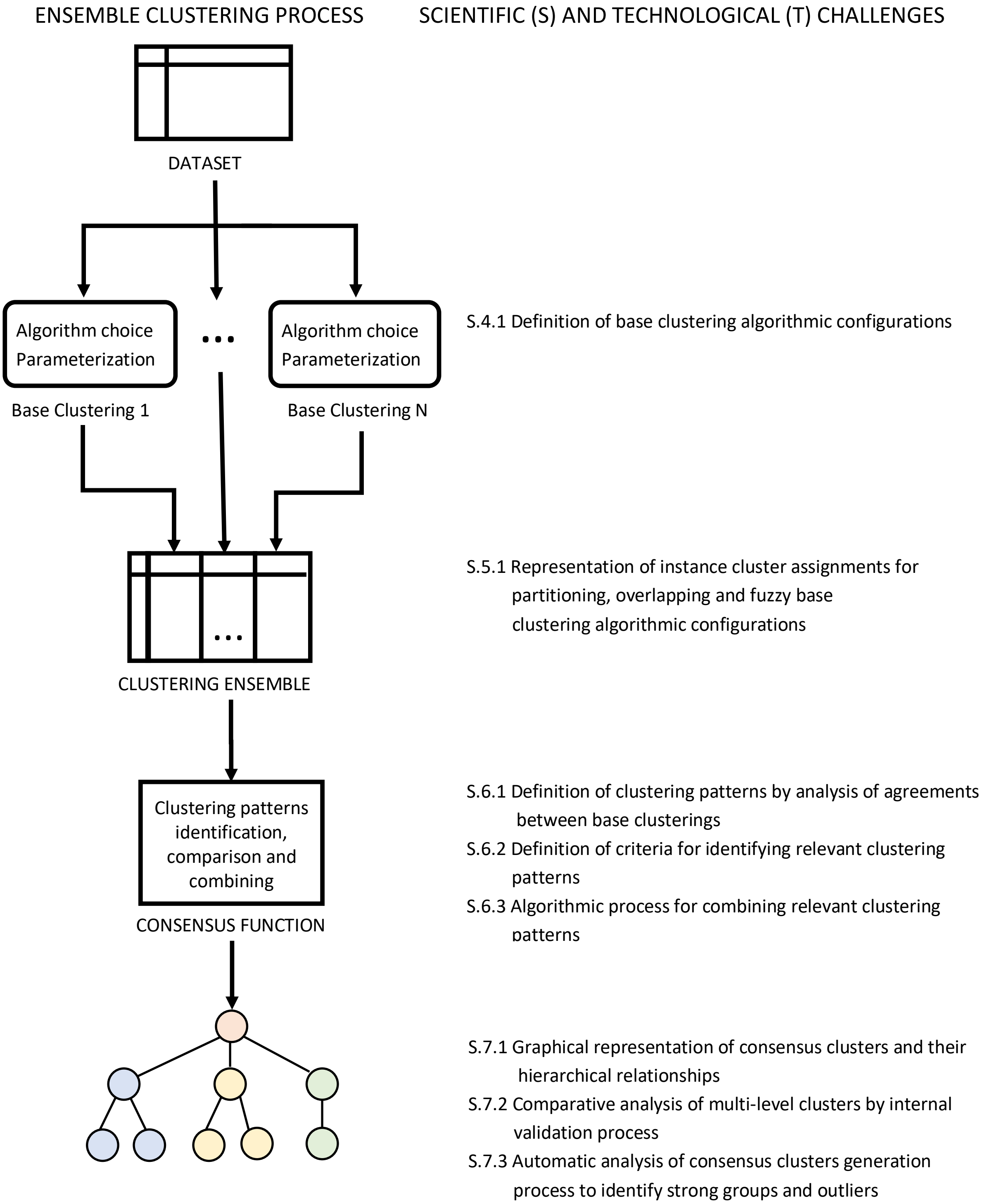

ENSEMBLE CLUSTERING PROCESS
SCIENTIFIC (S) AND TECHNOLOGICAL (T) CHALLENGES
DATASET
Algorithm choice
Parameterization
Base Clustering 1
...
Algorithm choice
Parameterization
Base Clustering N
S.4.1 Definition of base clustering algorithmic configurations
...
CLUSTERING ENSEMBLE
S.5.1 Representation of instance cluster assignments for partitioning, overlapping and fuzzy base clustering algorithmic configurations
Clustering patterns identification, comparison and combining
CONSENSUS FUNCTION
S.6.1 Definition of clustering patterns by analysis of agreements between base clusterings
S.6.2 Definition of criteria for identifying relevant clustering patterns
S.6.3 Algorithmic process for combining relevant clustering patterns
S.7.1 Graphical representation of consensus clusters and their hierarchical relationships
S.7.2 Comparative analysis of multi-level clusters by internal validation process
S.7.3 Automatic analysis of consensus clusters generation process to identify strong groups and outliers

Figure 7. Workflow of the Multi-level Ensemble Consensus Clustering Phase.

A *consensus function* is then applied to the clustering ensemble to identify agreements between base clusterings, deduce clustering patterns from the relevant agreements identified, and combine the clustering patterns in a relevant manner. The closed set based method identifies the clustering patterns that correspond to different level of agreement, i.e., number of agreements, among base clusterings. These patterns identify the most frequent groupings of instances at different levels of precision, enabling to select finally the clusters corresponding to the strongest bonds between instances among base clusters. This step and the challenges it involves are detailed in section 4.3, as well as the method developed to address them in the proposed approach.

The *hierarchical decomposition of consensus clusters* represents in a tree-like format the relationships between the clusters and the successive steps (groupings) of their generation starting from the minimal groups of objects (maximal agreements between base clusterings). These relationships and other properties of clusters are then used for internal validation and ranking of consensus clusters. In the context of semi-supervised learning, partial prior knowledge such as cannot-link and must-link constraints defined from instances of known classes can also be used for this validation and ranking step.

3.3. Clusters to Classes Learning

The workflow of this phase is depicted in Fig. 8. It aims to generate the business class prediction model from the ranked hierarchy of consensus clusters by integrating external validation through business metric application.

The first step consists to identify and select the most relevant consensus clusters regarding both internal and external, i.e. Amadeus business metric based, validations. Selected clusters can belong to different levels of the hierarchical structure, in order that each cluster is as far as possible representative of a class of business objects and discriminative of other business classes. A business class can then be represented by one or several clusters depending on validation results. This step, based on combined internal and external validation processes, is detailed in section 4.4.

The selected multi-level consensus clusters constitute, together with instances identified as outliers, a clustering of the dataset, that is a partitioning of all instances in the dataset. The automated analysis of the generation process of consensus clusters to identify potential strong clusters, i.e., groups of instances representing maximal agreements between base clusters, and outlier instances is detailed in section 4.5. The visualization and analytical exploration of selected consensus clusters in the tree-like graphical representation of the hierarchy objective is to help the user understanding the inherent structures in the data space and validate the selected clusters from a business application perspective.

The cluster class characterization step aims to identify the business metric related criteria that discriminate the business class of each cluster from other business classes, i.e., the segments of customers. The discriminative criteria of cluster business classes are validated through the comparative analysis of internal and external validation results regarding the business application objectives. The validated discriminative criteria then provide the information required to define a classifier, i.e., a prediction model of the class of instances, deployed in the operational business application during the last step. This predictive model is based on the analysis of the similarity between the features of the instance to classify and the discriminative criteria of clusters. The definition of the class prediction model and the results obtained with the different algorithmic solutions tested are presented in section 4.6.

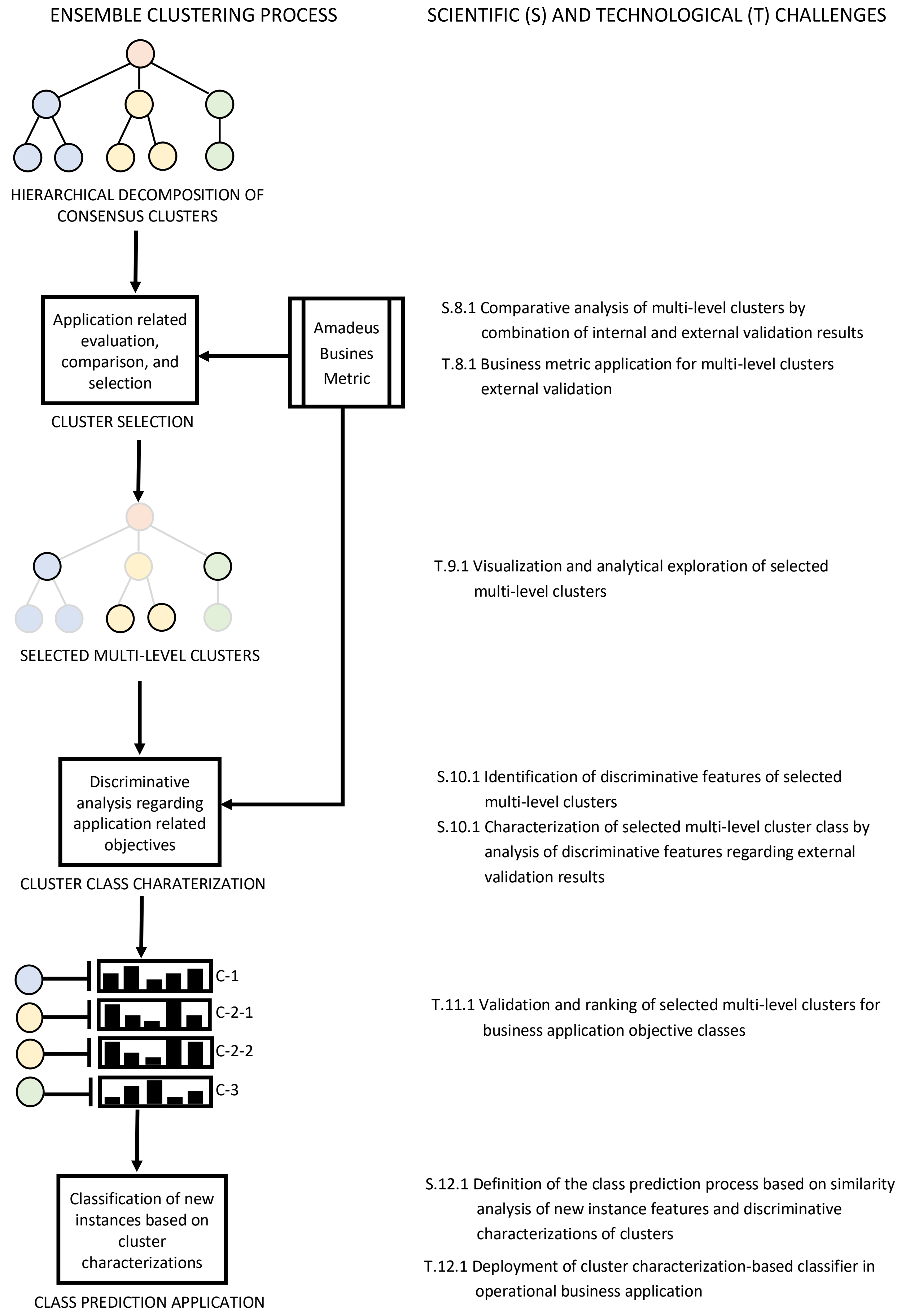
4. Technical and Scientific Challenges

This section details the central scientific and technological challenges addressed during the development, implementation, and experimental application of the proposed framework in the context of the Amadeus flight search recommendation engine, with central results and findings, and future extensions of the realizations.

4.1. Data Space Exploration and Description Regarding Base Clustering Algorithm Parameterizations

To conduct experimental and comparative studies an initial dataset was constructed by extracting search queries of flight bookings for flights departing from the U.S.A. during one week of January 2018. This dataset contains the 9 most relevant variables identified according to Amadeus business expertise and metrics: Distance between the airports, geography, number of passengers, number of children, advance purchase, stay duration, day of the week of the departure, day of the week of the return, and day of the week of search. The Geography variable values are encoded as categorical ordinal values: 0 for domestic flights with departure and arrival airports in the same country, 1 for continental flights with departure and arrival airports on the same continent and 2 for intercontinental flights. This dataset contains a very large number of instances representing customers, in the order of millions.

The exploratory analysis of the dataset space showed that an important proportion of the instances have very similar variable values, and the populations are divided into several strata based on similar characteristics.


ENSEMBLE CLUSTERING PROCESS
SCIENTIFIC (S) AND TECHNOLOGICAL (T) CHALLENGES
HIERARCHICAL DECOMPOSITION OF CONSENSUS CLUSTERS
Application related evaluation, comparison, and selection
Amadeus Busines Metric
CLUSTER SELECTION
SELECTED MULTI-LEVEL CLUSTERS
Discriminative analysis regarding application related objectives
CLUSTER CLASS CHARATERIZATION
C-1
C-2-1
C-2-2
C-3
Classification of new instances based on cluster characterizations
CLASS PREDICTION APPLICATION

S.8.1 Comparative analysis of multi-level clusters by combination of internal and external validation results

T.8.1 Business metric application for multi-level clusters external validation

T.9.1 Visualization and analytical exploration of selected multi-level clusters

S.10.1 Identification of discriminative features of selected multi-level clusters

S.10.1 Characterization of selected multi-level cluster class by analysis of discriminative features regarding external validation results

T.11.1 Validation and ranking of selected multi-level clusters for business application objective classes

S.12.1 Definition of the class prediction process based on similarity analysis of new instance features and discriminative characterizations of clusters

T.12.1 Deployment of cluster characterization-based classifier in operational business application

Figure 8. Workflow of the Clusters to Classes Learning Phase.

For the purpose of rapid prototyping and testing of the developed and compared algorithmic approaches, and to enable the application of algorithms that have limitations regarding the number of instances processed, a sampling was performed on the sub-populations to generate a stratified sampling of the whole dataset while preserving the distribution properties of the original dataset. For experimental evaluations, three stratified samples containing respectively 500, 1000 and 1500 instances were created. The effect of the stratified sampling for the 'distance between airports' variable can be observed in Fig. 6 showing the histograms of the distribution of the variable values in the original dataset and in the two largest stratified samples created.

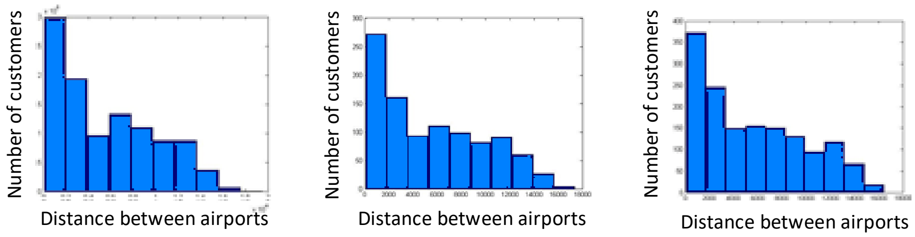


Figure 6. Distribution of values for the 'Distance between airports' variable in the original dataset (left), the stratified sample of size 1000 (middle) and the stratified sample of size 1500 (right).

4.2. Definition of Base Clustering Algorithmic Configurations

Consensus clustering results depend to a significant extent on the relevance of the set of base clusterings used to generate the clustering ensemble, which constitutes the search space for the consensus function. A major concern for generating a relevant set of base clusterings is to define an interval of values for the number of clusters generated by base clusterings that ensures diversity in both the solutions and the levels of resolution of clusters. This parameter, usually denoted as the K parameter, is required by most classical clustering algorithms.

This work showed the important impact of the clustering ensemble properties regarding both a sufficient diversity in the search space, i.e., the potential consensus clusters explored, and a centering of this search space on the most stable number of connected components, for defining an interval of K values for K-parameter based algorithms. Ensuring these properties are satisfied through the generation of an enhanced search space, in the refined clustering ensemble and membership matrix, is a major step for obtaining relevant consensus clusters.

In presence of base clustering solutions with diverse number of clusters, assessing the most common number of clusters present among the base clustering solutions is essential. In this context, we use an iterative way to estimate the number of clusters by forming a graph from the co-association matrix between the objects. The association among any pairwise objects can be quantified by the number of clustering solutions in which they are in the same cluster. Thus, the co-association of any two objects can be found by forming a co-association matrix, and this co-association can be thought of as the bonding among any two objects. This co-association matrix can be transformed into a weighted graph where the vertices are basically the objects and the edge weight define their bonding based on their concurrent occurrence in a same cluster over different base clustering solutions. In this situation, if a small amount of edge weight is reduced step-by-step, we obtain a graph where the loosely bonded vertices become disconnected. In this context, the connected components basically represent the clusters as the bonding among the objects inside the same component is higher than the bonding of two objects lying in different components. If for deletion of a very small amount of weight from each edge, the number of connected component changes rapidly then this means that the bonding of the objects in some components is not strong. However, in contrary, it is seen that even if the deletion of some weights (i.e., edges) keeps the number of connected components constant, this means that the bonding is very high. In this way, the edge weight is reduced step-by-step iteratively, and the number of connected components is observed. Therefore, in each step, the number of connected components corresponds to the number of clusters that can be present in the base clustering solutions, and a stable number of connected components (i.e., number of times it remains constant over the iterations) denotes the most likely number of clusters in the corresponding dataset.

In the experiments, initially the number of components is sorted based on the frequency of staying same even after the edge deletion. Then, to ensure the quality, the final number is selected from the base clustering solutions having the highest similarity with respect to the external cluster validity index (ARI measure). We may think of about the utility of using this connected component finding process while the ultimate choosing criteria is using the similarity metric, i.e., ARI. To explain this, suppose a clustering solution (e.g., with K=3) has a highest similarity value with respect to the other base clustering solutions. But while performing the step-by-step edge-weight deletion process, there is no occurrence of a number of three connected components. This means that it is less likely that the objects have strong relationships with this number of clusters. So, even though it has a highest similarity among the base clustering solutions, it cannot be chosen as the most likely number of clusters that can be present in the clustering solution. However, if the number of three connected components occurs at least once while having the highest similarity, with respect to the base clusterings in terms of ARI, then it can be selected as the most probable number of clusters that can be present in the clustering solutions.

4.3. Definition of Clustering Patterns by Analysis of Agreements Between Base Clusterings

Closed patterns extracted from the refined membership matrix consist each of a set of instances and a set of base clusters that agreed to cluster together these instances. They constitute the initial clustering patterns of the algorithmic process that generates new clustering patterns by combination of existing ones in an incremental manner. This process was enhanced during this work to extend the comparative analysis of the final clustering patterns and thus optimize the generation of consensus clusters.

A new measure for evaluating the relevance of each clustering pattern, that is a set of instances and the corresponding set of base clusters, was developed to compare, select, and combine them using the maximum information at our disposal. This measure considers at the same time:

- The number of agreements and disagreements between base clusterings on grouping the set of instances of the clustering pattern.
- The inclusion relationships between sets of instances and sets of base clusters of compared clustering patterns.
- The sizes of the sets of instances of the closed patterns extracted from base clusterings.
- The number of clusters in the base clusterings that affects the probability of co-occurrence of instances in a cluster.

This new measure was shown to be able, contrarily to the initial measure, to provide distinct values for clustering patterns with different properties regarding the base clusterings they correspond to.

4.4. Comparative Analysis of Multi-level Clusters by Internal and External Validation Criteria

The problem of the evaluation of the quality of both consensus clusterings and consensus clusters is a central issue to generate a relevant solution. The state-of-the-art and comparative study of validation measures of clusterings and clustering ensembles shows that, basically, two types of performance evaluation are used:

- Internal validation in which the evaluation is done with the dataset itself only. This evaluation is based on the analysis of relationships between instances in clusters regarding their distribution in the data space and their common properties. For this, many indices are defined in literature, like Silhouette index, Entropy, R-Squared (RS), Root-Mean-Square Standard Deviation (RMSSTD), Semi-Partial R-squared (SPR), Distance between two clusters (CD), Partition Coefficient (PC), Classification Entropy (CE), Partition Index (PC), Separation Index (S), Xie and Beni's index (XB), Inter-Cluster Density (ID), Davies-Bouldin (DB) index, Dunn's Index (DI), Alternative Dunn Index (ADI), etc.
- External validation in which existing prior knowledge about the dataset is involved. This prior knowledge is represented either as class labels for the dataset instances, when each instance can be assigned a business segment, or as another clustering result in which assigned clusters are considered as instance segment labels and the evaluated clustering is then compared to this existing clustering result. The most commonly used indices for this are the Average Rand Index (ARI) and the Normalized Mutual Information (NMI), although several other indices were proposed in the literature such as Accuracy, Cohesion, Entropy, F-measure, Purity, etc.

The new measures developed for internal and external validation aim to extend the information classically used for internal and external validations, that is the list of co-occurrences of pairs of instances in the clusters, by integrating in the calculation the information provided by the clustering patterns, e.g., the new clustering pattern relevance measure developed, and their hierarchical relationships such as depicted in the tree of consensuses.

The new measures developed are based on the closed sets-based framework of Formal Concept Analysis. The main motivation relies on the fact that the ARI and NMI popular metrics basically compare the similarities among pairs of clustering solutions (external evaluation concept). However, in a specific clustering solution the quality of individual clusters (internal evaluation concept) is not considered and all clusterings are treated the same way which is not realistic in the considered type of scenarios. Frequent closed sets-based measures become an interesting solution in this context being more effective when little or no information is available regarding the number of actual clusters in the dataset, as well as when only base clustering solutions are available instead of the initial dataset.

4.5. Automatic Analysis of Consensus Cluster Generation Process for identifying Strong Clusters and Outlier Instances

Using the proposed new measures for comparing clusters in the tree of consensuses, based on clustering patterns and an analysis of the hierarchical relationships in the tree, both outlier instances and multi-level strong groups of instances can be identified if present. Outlier instances are identified through their unstable behavior from the viewpoint of the clustering process: They are successively associated and separated with the same instances in different levels of the tree. Strong groups are identified through their stability over different successive levels of the tree of consensuses, such as the C-1 and C-3 cluster in Fig. 4, that thus represent strong clusters, with maximal agreement, regarding the base clusterings.

Results of initial experimentations of the proposed approach were able to identify such a structure of clusters, where a significant cluster from the viewpoint of the customer segment representation is divided into three sub-segments with significant distinctive features regarding the new measures results. These initial results were evaluated using Amadeus specific business metrics that validated the relevance of the three sub-segments identified regarding the prediction of query search result booking.

4.6. Definition of the Class Prediction Process Based on Similarity Analysis of New Instance Features and Discriminative Characterizations of Clusters

Once the selected multi-level consensus clusters have been validated regarding both internal and external validations, and business metric, each cluster is associated to the business segment or sub-segment of customers it corresponds to. The clusters are then characterized in the data space to identify the criteria that discriminate them, that is the features that distinguish the instances in a cluster from the instances in other clusters. These criteria are combined to generate a classifier, that is an algorithmic process for predicting the class of new instances, i.e., the business segment or sub-segment of each new customer.

Different approaches for defining the class prediction model were tested, considering both the relevance of the generated predictions and the computational efficiency and scalability of the process. These approaches consist to determine which cluster is the nearest to the new instance in the data space considering the assessed distance (minimal, maximal, average, etc.) between the new instance and each cluster. The best results were obtained when a representative vector consisting of variable value domains is computed for each cluster and the distance is evaluated between the new instance and each representative vector.

Once the new instance class prediction process is validated, the next step consists to evaluate the capability of the approach to efficiently distinguish and predict significant business segments and sub-segments according to business objective classes defined by the Customer Choice Modelling application context.

5. Conclusion

During the development of the proposed multi-level consensus clustering approach, several consensus clustering algorithms, internal and external clustering validation measures, and integrations of supervised, semi-supervised and unsupervised techniques were studied. The objective was to obtain a better aggregation of individual clustering solutions. From the results, a conceptual framework for implementing an improved customer segmentation and choice modelling solution in travel context was designed.

The techniques developed during this project first aim to solve central issues for the Customer Choice Modeling data clustering steps by providing a multi-level consensus clustering based solution that:

- Does not require the user to define the number of clusters to generate as a parameter of the clustering solution, but automatically determine the number of clusters according to base clustering properties.
- Generates multiples consensus clustering solutions corresponding to different levels of agreements between the base clusterings. This property allows to choose the most relevant consensus solution considering both internal and external validation criteria.
- Generates a robust clustering solution that does not rely solely on a particular modeling assumption of clusters, i.e., a unique category of algorithms and a unique parameterization.
- Provides a hierarchy of consensus clusters, allowing the end-users to select clusters at different levels of precision regarding the business segments. In this hierarchy, a segment can be refined as several sub-segments, each corresponding to the same business class of instances but with slight variations regarding their distinctive features or the business objectives.
- Automatically identifies strong clusters, i.e., groups of instances agreed by a maximal number of base clusterings, and outlier instances, i.e., instances with features that do not hold the general properties of similar instances or the instances in the same clusters. This identification relies on the analytical comparison of consensus clusters and their hierarchical relationships.
- Generates a graphical representation of hierarchical relationships of consensus clusters, depicting their generation process, to help the end-users in the interpretation of the resulting consensus clusters.
- Can automatically identify the best multi-level consensus clusters obtained according to internal validation criteria and their ranking based on their structural properties and hierarchical relationships.

The second category of techniques developed aim to connect, from a business viewpoint, the unsupervised results of clustering and the classes of instances, that is the customer segments and sub-segments. These techniques aim to:

- Combine the results of internal and external validations for identifying the most relevant multi-level consensus clusters from a business objective perspective. These clusters should represent significant groups of instances from both the viewpoints of their distinct features in the data space and the business class each one corresponds to.
- Provide a statistical and analytical exploration solution for the business-related evaluation of the generated multi-level consensus clusters regarding internal (data space based) and external (business metric based) cluster validations, and of the obtained consensus clustering solution.
- Identify the discriminative features of clusters, that are required to distinguish instances assigned to different clusters, regarding distribution model properties of the cluster data sub-spaces.
- Generate an instance class prediction model by the comparative analysis of discriminative features of the selected clusters.
- Provide support to the end-users for the semi-automatic tasks of the process, such as the evaluation and validation of classes of clusters regarding business related objectives, predefined business classes and external metrics.

The techniques developed meet the central needs identified for Customer Choice Modelling in travel industry. The first is the capability to identify relevant business segments and sub-segments by the grouping of search queries according to their similarity. The second is the understanding of customer choice patterns, in order to predict the likelihood of a search query recommendation to be booked. The third is the optimization of the rate of bookings of search query recommendations for each business segment by the identification of search query features that really matters and the quantification of how much they matter for each segment. Importantly, since the proposed framework relies, among other things, on semi-supervised techniques, it has the capacity to be adapted to situations in which preferences of customers can switch in response to contextual changes as might happen in situations where travel business might be influenced by unusual circumstances such as a pandemic like the Coronavirus pandemic [7].

We have described the algorithmic process that was designed to implement the proposed multi-level consensus clustering framework. This process consists the three central phases, namely the data exploration and preprocessing, the multi-level ensemble consensus clustering, and the clusters to classes learning. The workflows depicting these three phases also show the technical and theoretical challenges that arise during the implementation of each of them. We also have described the technical and scientific challenges encountered during the development and implementation of the proposed framework in collaboration with Amadeus IT Group for the improvement of the flight search recommendation engine. The experimental evaluations carried out on Amadeus data about search queries of flight bookings have shown the feasibility and relevance of the proposed approach for Customer Choice Modelling in travel industry [31]. The tests conducted have shown a significant increase in the probabilities of flight search queries booking using the recommendations generated from the prediction of the segments and sub-segments of travelers extracted by the multi-level consensus clustering process.

Acknowledgment

This work has been supported by the French government, through the UCA^{JEDI} Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01.

References

[1] Sujoy Chatterjee and Nicolas Pasquier, "A Multi-Level Consensus Clustering Framework for Customer Choice Modelling in Travel Industry", In *Proceedings of the iCETiC International Conference on Emerging Technologies in Computing*, LNICST, Vol. 332, pp. 142-157, Published by Springer International Publishing, 2020, iCETiC'2020 Best Paper Award, DOI: 10.1007/978-3-030-60036-5_10, Available: https://link.springer.com/chapter/10.1007%2F978-3-030-60036-5_10.

[2] Abla C. Benabdellah, Asmaa Benghabrit and Imane Bouhaddou, "A Survey of Clustering Algorithms for an Industrial Context", In *Procedia Computer Science*, Vol. 148, pp. 291–302, Published by Elsevier, 2019, DOI: 10.1016/J.PROCS.2019.01.022, Available: https://www.sciencedirect.com/science/article/pii/S1877050919300225.

[3] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil *et al.*, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", In *IEEE Transactions on Emerging Topics in Computing*, Vol. 2, No. 3, pp. 267–279, Published by IEEE, 2014, DOI: 10.1109/TETC.2014.2330519, Available: https://ieeexplore.ieee.org/document/6832486.

[4] Emrah Hancer, Bing Xue and Mengjie Zhang, "A Survey on Feature Selection Approaches for Clustering", In *Artificial Intelligence Review*, Vol. 53, pp. 4519–4545, 2020, DOI: 10.1007/s10462-019-09800-w, Available: https://link.springer.com/article/10.1007%2Fs10462-019-09800-w.

[5] Hans-Peter Kriegel, Peer Kröger and Arthur Zimek, "Clustering High-dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering", In *ACM Transactions on Knowledge Discovery from Data*, Vol. 3, No. 1, Article 1, 2009, DOI: 10.1145/1497577.1497578, Available: https://dl.acm.org/doi/10.1145/1497577.1497578.

[6] Dongkuan Xu and Yingjie Tian, "A Comprehensive Survey of Clustering Algorithms", In *Annals of Data Science*, Vol. 2, No. 2, pp. 165–193, Published by Springer, 2015, DOI: 10.1007/s40745-015-0040-1, Available: https://link.springer.com/article/10.1007/s40745-015-0040-1.

[7] Oussama H. Hamid and Jochen Braun, "Reinforcement Learning and Attractor Neural Network Models of Associative Learning", In *Studies in Computational Intelligence*, Vol. 829, pp. 327-349, Published by Springer, 2019, DOI: 10.1007/978-3-030-16469-0_17, Available: https://link.springer.com/chapter/10.1007/978-3-030-16469-0_17.

[8] Christian Hennig, "Clustering Strategy and Method Selection", In *Handbook of Cluster Analysis*, Chapter 31, pp. 703–730, Published by Chapman & Hall/CRC, 2016, ISBN: 9780367570408, DOI: 10.1201/b19706-40, Available: https://www.routledgehandbooks.com/doi/10.1201/b19706-38.

[9] Rui Xu and Donald C. Wunsch "Survey of Clustering Algorithms", In *IEEE Transactions on Neural Networks*, Vol. 16, No. 3, pp. 645–678, Published by IEEE, 2005, DOI: 10.1109/TNN.2005.845141, Available: https://ieeexplore.ieee.org/document/1427769.

[10] Lori Dalton, Virginia Ballarin and Marcel Brun, "Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics", In *Current Genomics*, Vol. 10, No. 6, pp. 430–445, Published by Bentham Science, 2009, DOI: 10.2174/138920209789177601, Available: https://www.eurekaselect.com/69906/article.

[11] Maria Halkidi, Yannis Batistakis and Michalis Vazirgiannis, "On Clustering Validation Techniques", In *Journal of Intelligent Information Systems*, Vol. 17, pp. 107–145, Published by Springer, 2001, DOI: 10.1023/A:1012801612483, Available: https://link.springer.com/article/10.1023/A:1012801612483.

[12] Yang Lei, James C. Bezdek, Simone Romano, Nguyen X. Vinh, Jeffrey Chan *et al*., "Ground Truth Bias in External Cluster Validity Indices", In *Pattern Recognition*, Vol. 65, pp. 58–70, Published by Elsevier, 2017, DOI: 10.1016/j.patcog.2016.12.003, Available: https://www.sciencedirect.com/science/article/abs/pii/S0031320316303910.

[13] Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi and Elvia M. Quiroz, "Internal versus External Cluster Validation Indexes", In *International Journal of Computers and Communication*, Vol. 5, Issue 1, pp. 27–34, 2011, E-ISSN: 2074-1294, Available: http://www.universitypress.org.uk/journals/cc/20-463.pdf.

[14] Hui Xiong and Zhongmou Li, "Clustering Validation Measures", In *Data Clustering Algorithms and Applications*, Chapter 23, pp. 571–605, Published by Chapman & Hall/CRC Press, 2014, eBook ISBN: 9781315373515, DOI: 10.1201/9781315373515-23, Available: https://www.taylorfrancis.com/chapters/clustering-validation-measures-hui-xiong-zhongmou-li/e/10.1201/9781315373515-23.

[15] Tossapon Boongoen and Natthakan Iam-On, "Cluster Ensembles: A Survey of Approaches with Recent Extensions and Applications", In *Computer Science Review*, Vol. 28, pp. 1–25, Published by Elsevier, 2018, DOI: 10.1016/J.COSREV.2018.01.003.

[16] Joydeep Ghosh and Ayan Acharya, "A Survey of Consensus Clustering", in *Handbook of Cluster Analysis*, Chapter 22, pp. 497–518, Published by Chapman & Hall/CRC, 2016, ISBN: 9780367570408, DOI: 10.1201/b19706-28, Available: https://www.routledgehandbooks.com/doi/10.1201/b19706-28.

[17] Sandro Vega-Pons and José Ruiz-Shulcloper, "A Survey of Clustering Ensemble Algorithms", In *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 25, No. 3, pp. 337–372, Published by World Scientific Publishing, 2011, DOI: 10.1142/S0218001411008683.

[18] Xiuge Wu, Tinghuai Ma, Jie Cao, Yuan Tiand and Alia Alabdulkarim, "A Comparative Study of Clustering Ensemble Algorithms", In *Computers & Electrical Engineering*, Vol. 68, pp. 603–615, Published by Elsevier, 2018, DOI: 10.1016/j.compeleceng.2018.05.005.

[19] Lawrence Hubert and Phipps Arabie, "Comparing Partitions", In *Journal of Classification*, Vol. 2, No. 1, pp. 193–218, 1985, DOI: 10.1007/BF01908075, Available: https://link.springer.com/article/10.1007/BF01908075.

[20] Mayra Z. Rodriguez, Cesar H. Comin , Dalcimar Casanova, Odemir M. Bruno, Diego R. Amancio *et al*., "Clustering Algorithms: A Comparative Approach", In *PLoS ONE*, Vol. 14, No. 1, e0210236, 2019, DOI: 10.1371/journal.pone.0210236.

[21] Hanneke van der Hoef and Matthijs J. Warrens, "Understanding Information Theoretic Measures for Comparing Clusterings", In *Behaviormetrika*, Vol. 46, pp. 353–370, Published by Springer, 2019, DOI: 10.1007/s41237-018-0075-7, Available: https://link.springer.com/article/10.1007/s41237-018-0075-7.

[22] Nguyen X. Vinh, Julien R. Epps and James Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance", in *Journal of Machine Learning Research*, Vol. 11, pp. 2837–2854, 2010, Series Online ISSN: 1532-4435, Available: https://jmlr.org/papers/v11/vinh10a.html.

[23] Atheer Al-Najdi, Nicolas Pasquier and Frédéric Precioso, "Using Frequent Closed Itemsets to Solve the Consensus Clustering Problem", In *International Journal of Software Engineering and Knowledge Engineering*, Vol. 26, No. 10, pp. 1379–1397, Published by World Scientific Publishing, 2016, DOI: 10.1142/S021819401640009X, Available: https://www.worldscientific.com/doi/abs/10.1142/S021819401640009X.

[24] Ines Färber, Stephan Günnemann, Hans-Peter Kriegel, Peer Kröger, Emmanuel Müller *et al*., "On Using Class-Labels in Evaluation of Clusterings", In *MultiClust International Workshop on Discovering, Summarizing and Using Multiple Clusterings* held in conjunction with the *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2010)*, July 25-28, Washington, DC, United States, Published by ACM, 2010, Available: https://www.researchgate.net/publication/228374158_On_using_class-labels_in_evaluation_of_clusterings.

[25] Amrudin Agovic and Arindam Banerjee, "Semi-supervised Clustering", In *Data Clustering: Algorithms and Applications*, Chapter 20, pp. 505–534, Published by Chapman & Hall/CRC, 2013, eBook ISBN: 9781466558212, DOI: 10.1201/9781315373515-20, Available: https://www.taylorfrancis.com/chapters/semisupervised-clustering-amrudin-agovic-arindam-banerjee/e/10.1201/9781315373515-20.

[26] Nizar Grira, Michel Crucianu and Nozha Boujemaa, "Unsupervised and Semi-supervised Clustering: A Brief Survey", In *A Review of Machine Learning Techniques for Processing Multimedia Content*, pp. 9–16, 2005, Available: https://www.researchgate.net/publication/228704486_Unsupervised_and_Semi-supervised_Clustering_a_brief_survey.

[27] Anil Jain, Rong Jin and Radha Chitta, "Semi-supervised Clustering", In *Handbook of Cluster Analysis*, Chapter 20, pp. 443–468, Published by Chapman & Hall/CRC, 2016, ISBN: 9780367570408, DOI: 10.1201/b19706-26, Available: https://www.routledgehandbooks.com/doi/10.1201/b19706-26.

[28] Karell Bertet, Christophe Demko, Jean-François Viaud and Clément Guérin, "Lattices, Closures Systems and Implication Bases: A Survey of Structural Aspects and Algorithms", In *Theoretical Computer Science*, Vol. 743, pp. 93–109, Published by Elsevier, 2018, DOI: 10.1016/J.TCS.2016.11.021, Available: https://www.sciencedirect.com/science/article/abs/pii/S0304397516306806.

[29] Kartick C. Mondal, Nicolas Pasquier, Anirban Mukhopadhyay, Ujjwal Maulik and Sanghamitra Bandhopadyay, "A New Approach for Association Rule Mining and Bi-clustering using Formal Concept Analysis", In *Proceedings of the MLDM International Conference on Machine Learning and Data Mining in Pattern Recognition*, Lecture Notes in Artificial Intelligence (LNAI), Vol. 7376, pp. 86–101, Published by Springer, Heidelberg, 2012, DOI: 10.1007/978-3-642-31537-4_8, Available: https://link.springer.com/chapter/10.1007/978-3-642-31537-4_8.

[30] Sadok B. Yahia, Tarek Hamrouni and Engelbert M. Nguifo, "Frequent Closed Itemset based Algorithms: A Thorough Structural and Analytical Survey", In *ACM SIGKDD Explorations Newsletter*, Vol. 8, No. 1, pp. 93–104, Published by ACM, 2006, DOI: 10.1145/1147234.1147248.

[31] Sujoy Chatterjee, Nicolas Pasquier, Simon Nanty and Maria A. Zuluaga, "Multi-objective Consensus Clustering Framework for Flight Search Recommendation", In *Arxiv*, Article: arXiv:2002.10241, 17 pages, Published by Cornell University, 2020, Available: https://arxiv.org/abs/2002.10241.