A Multi-level Consensus Clustering Framework for Customer Choice Modelling in Travel Industry

Sujoy Chatterjee[1,2] and Nicolas Pasquier[1](✉)

[1] CNRS, I3S, Université Côte d'Azur, Sophia Antipolis, France
pasquier@i3s.unice.fr
[2] Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea
sujoy@unist.ac.kr

Abstract. In the travel industry context, customer segmentation, that is the clustering of travelers to distinguish segments of customers with similar needs and desires, is a major issue for improving the personalization of recommendations in flight search queries. Indeed, when booking travel itineraries, different customers purchase tickets according to different criteria, like price, duration of flight, layover time, etc. However, clustering algorithm application is a challenging task because of two central issues inherent to the unsupervised nature of the grouping of instances: The choice of the clustering algorithm and parameterization and the evaluation of the resulting clusters of instances. Essentially, each clustering algorithm and evaluation measure relies on an assumption of the distribution model of the instances in the data space. The relevance of the resulting clustering mainly depends to which extent they are adapted to the analyzed data space properties. We present a Multi-level Consensus Clustering framework that combines the results of several clustering algorithmic configurations to generate a multi-level consensus clustering solution in which each cluster represents an agreement between the different clustering results. Relevant agreements are identified using a closed sets-based approach and represented in a hierarchical representation providing the end-user a representation of the consensus cluster construction process and their inclusion relationships. We show how this framework developed for Customer Choice Modeling in travel context can provide a better segmentation and refine the customer segments by identifying relevant sub-segments represented as sub-clusters in the hierarchical representation, and we present the technical and scientific challenges posed by the approach.

Keywords: Consensus clustering · Ensemble clustering · Multi-level clustering · Closed sets · Travel search queries · Customer Choice Modelling

1 Introduction

In travel industry, the Customer Choice Modelling (CCM) application aims to model the decision process of a traveler, or a category of travelers, the analysis and the prediction

M. H. Miraz et al. (Eds.): iCETiC 2020, LNICST 332, pp. 142–157, 2020.
https://doi.org/10.1007/978-3-030-60036-5_10

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.

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.

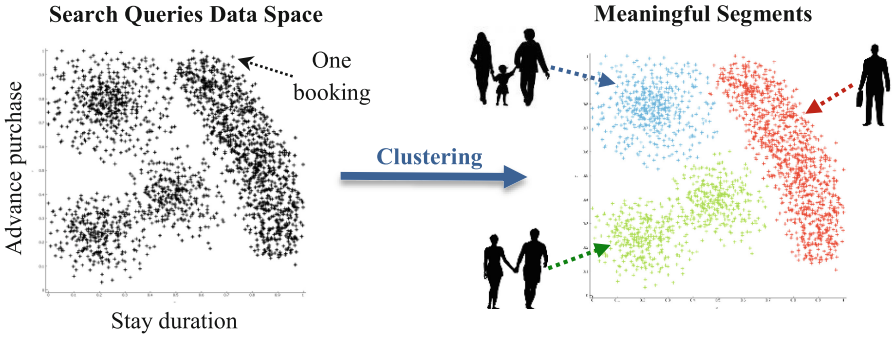

Fig. 1. Clustering of search queries for customer segment identification.

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.

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 [7, 16, 23] 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 analyzed [12, 24].

The resulting clusterings of a dataset are usually evaluated using unsupervised evaluation measures. These measures are called *internal validation measures* as they are

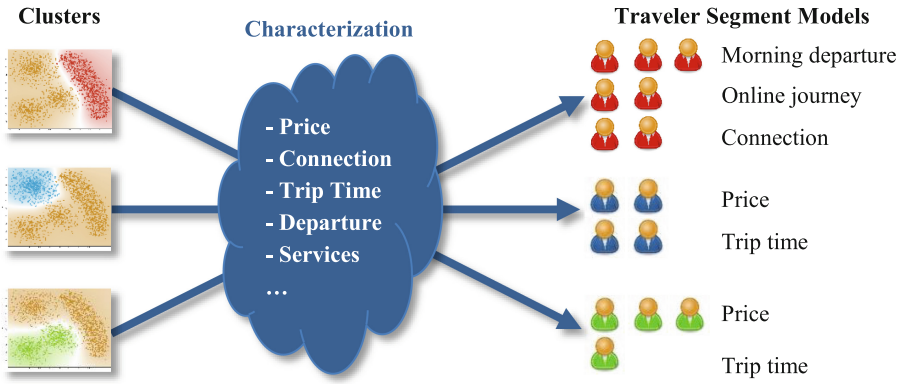

Fig. 2. Characterization of search query clusters for traveler segment modelling.

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 [6, 11, 21] 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 [4, 9, 20] 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 [14, 18, 19]. 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 in order 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, Sect. 3 describes the technical and the scientific challenges addressed, and Sect. 4 concludes the article.

2 Multiple Consensus Clustering Framework

The proposed framework was developed on the basis of the Multiple Consensus Clustering approach introduced with the *MultiCons* algorithm [2]. 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 [8].

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.

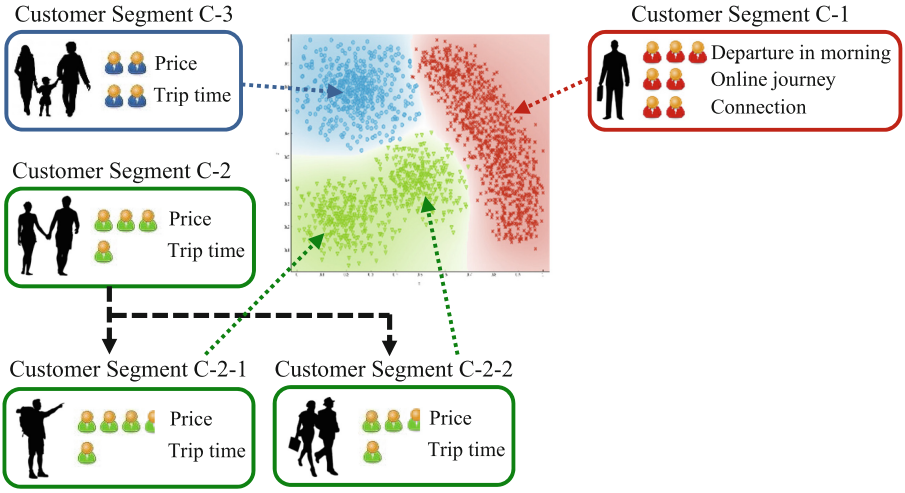

Fig. 3. Business segment specialization by multi-level consensus cluster analysis. (Color figure online)

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 neighborhood 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 [1, 10, 15] 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:

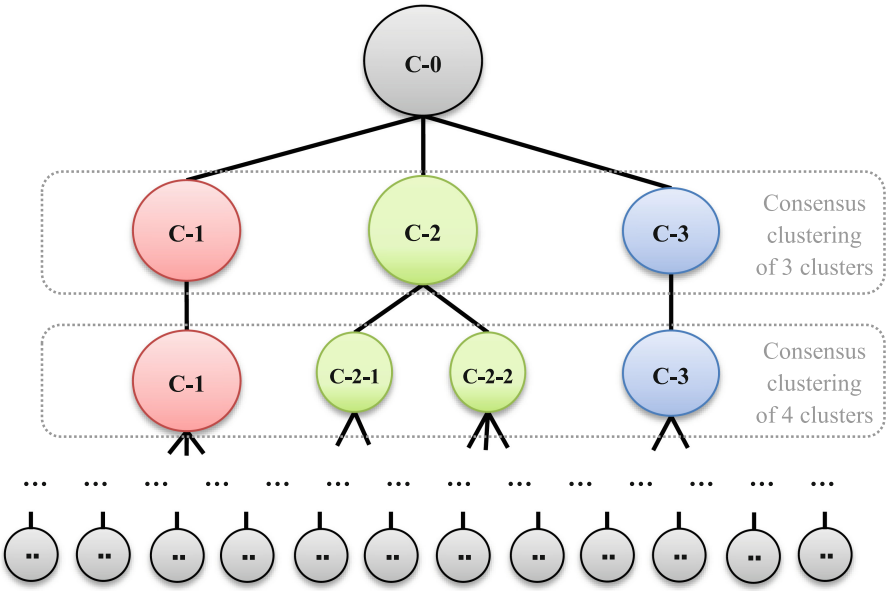

Fig. 4. Example representation of business segment specialization in the multiple consensus clustering tree-like hierarchy.

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

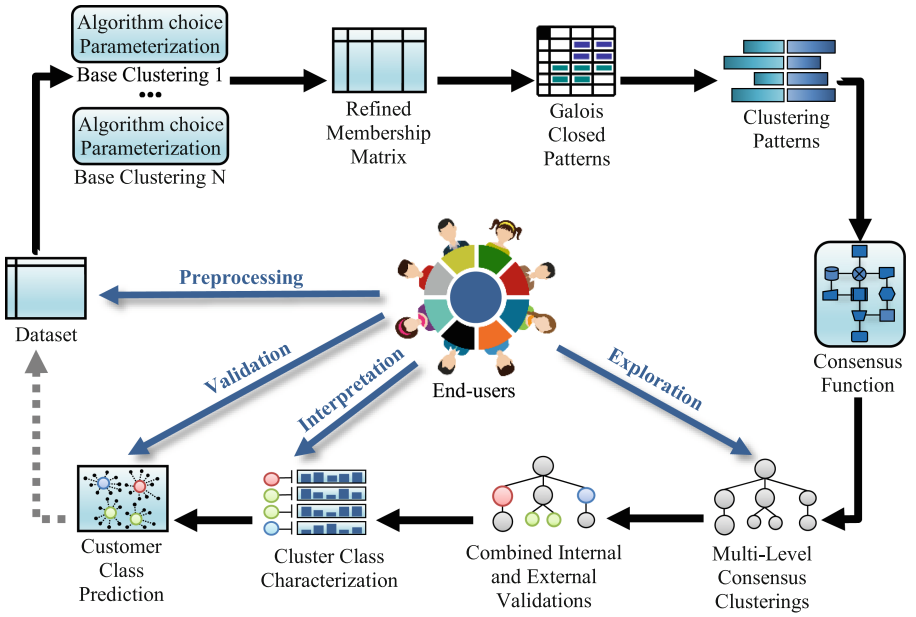

Fig. 5. Multi-level consensus clustering framework.

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 $M \ll N$, and regarding the demonstrated scalability properties of Galois closed sets extraction algorithms [3, 17, 25].

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 Technical and Scientific Challenges

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

3.1 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. 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.

3.2 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.

3.3 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.

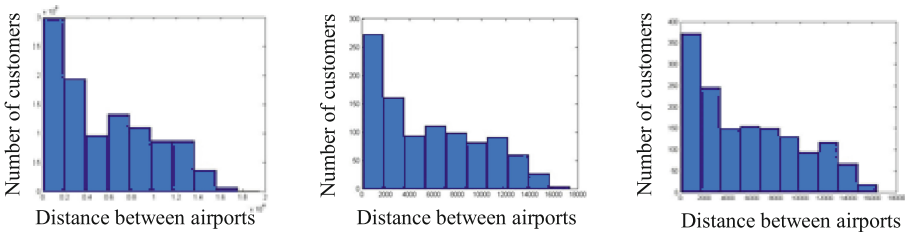

Fig. 6. Distribution of values for the 'distance between airports' variable in the original dataset (left), the stratified sample of size 1000 (middle) and the stratified sample of size 1500 (right).

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.

3.4 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.

3.5 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.

3.6 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.

4 Conclusion

During the development of the proposed multi-level consensus clustering framework, several consensus clustering algorithms, internal and external clustering validation measures and integrations of supervised, semi-supervised and unsupervised techniques were studied, with the objective 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 [12].

We have described the technical and scientific challenges encountered during the development and implementation of the proposed framework in collaboration with Amadeus IT Group. 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 [5]. 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.

Acknowledgments. 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. Agovic, A., Banerjee, A.: Semi-Supervised Clustering. Data Clustering: Algorithms and Applications, Chapter 20. Chapman & Hall, pp. 505–534 (2013)
2. Al-Najdi, A., Pasquier, N., Precioso, F.: Using frequent closed itemsets to solve the consensus clustering problem. Int. J. Softw. Eng. Knowl. Eng. **26**(10), 1379–1397 (2016)
3. Bertet, K., Demko, C., Viaud, J.F., Guérin, C.: Lattices, closures systems and implication bases: a survey of structural aspects and algorithms. Theor. Comput. Sci. **743**, 93–109 (2018)
4. Boongoen, T., Iam-On, N.: Cluster ensembles: a survey of approaches with recent extensions and applications. Comput. Sci. Rev. **28**, 1–25 (2018)
5. Chatterjee, S., Pasquier, N., Nanty, S., Zuluaga, M.A.: Multi-objective consensus clustering framework for flight search recommendation, 17 p. Cornell University (2020). https://arxiv.org/abs/2002.10241
6. Dalton, L., Ballarin, V., Brun, M.: Clustering algorithms: on learning, validation, performance, and applications to genomics. Curr. Genomics **10**(6), 430–445 (2009)
7. Fahad, A., et al.: A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Trans. Emerg. Top. Comput. **2**(3), 267–279 (2014)
8. Färber, I., et al.: On using class-labels in evaluation of clusterings. In: KDD MultiClust International Workshop on Discovering, Summarizing and Using Multiple Clusterings. ACM, Washington DC (2010)
9. Ghosh, J., Acharya, A.: A Survey of Consensus Clustering. Handbook of Cluster Analysis, Chapter 22. Chapman & Hall, pp. 497–518 (2016)
10. Grira, I., Crucianu, M., Boujemaa, N.: Unsupervised and semi-supervised clustering: a brief survey. Rev. Mach. Learn. Tech. Process. Multimed. Content **1**, 9–16 (2005)
11. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. J. Intell. Inf. Syst. **17**, 107–145 (2001)
12. Hamid, O.H., Braun, J.: Reinforcement learning and attractor neural network models of associative learning. In: Sabourin, C., Merelo, J.J., Madani, K., Warwick, K. (eds.) IJCCI 2017. SCI, vol. 829, pp. 327–349. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-16469-0_17
13. Hennig, C.: Clustering Strategy and Method Selection. Handbook of Cluster Analysis, Chapter 31. Chapman & Hall, pp. 703–730 (2016)
14. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)
15. Jain, A., Jin, R., Chitta, R.: Semi-Supervised Clustering. Handbook of Cluster Analysis, Chapter 20. Chapman & Hall, pp. 443–468 (2016)
16. Kriegel, H.-P., Kröger, P., Zimek, A.: Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. Knowl. Discov. Data **3**(1), 1–58 (2009). Article 1
17. Mondal, K.C., Pasquier, N., Mukhopadhyay, A., Maulik, U., Bandhopadyay, S.: A new approach for association rule mining and bi-clustering using formal concept analysis. In: Perner, P. (ed.) MLDM 2012. LNCS (LNAI), vol. 7376, pp. 86–101. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31537-4_8
18. Van der Hoef, H., Warrens, M.J.: Understanding information theoretic measures for comparing clusterings. Behaviormetrika **46**, 353–370 (2019)

19. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)
20. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. Int. J. Pattern Recognit Artif Intell. **25**(3), 337–372 (2011)
21. Rendón, E., Abundez, I., Arizmendi, A., Quiroz, E.M.: Internal versus external cluster validation indexes. Int. J. Comput. Commun. **5**(1), 27–34 (2011)
22. Xiong, H., Li, Z.: Clustering Validation Measures. Data Clustering Algorithms and Applications, Chapter 23. CRC Press, pp. 571–605 (2014)
23. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. Ann. Data Sci. **2**(2), 165–193 (2015)
24. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)
25. Yahia, S.B., Hamrouni, T., Nguifo, E.M.: Frequent closed itemset based algorithms: a thorough structural and analytical survey. ACM SIGKDD Explor. Newsl. **8**(1), 93–104 (2006)