Application of the Choquet Integral in Clustering Method: K-MeansChoqu

ABBAS Moncef *[1] and KERBOUI Roumeissa †[2]

[1,2]USTHB, Faculty of Mathematics, Labo. AMCD & RO, algiers

Abstract

Preference models often represent a global degree of utility of an alternative in terms of an aggregation of several local utility degrees, each of which pertains to a specific criterion. Methods for learning preference models from observed preference data, have mainly focused on fitting the aggregation function while assuming the local utility functions to be given. This paper advocates the so-called Choquet integral as a mathematical tool for learning for classification. While being widely used as a flexible aggregation function in fields such as multiple criteria decision making, the Choquet integral is much less known in machine learning so far. Apart from combining monotonicity and flexibility in a mathematically sound and elegant manner, the Choquet integral has additional features making it attractive from a machine learning point of view. For example, it offers measures for quantifying the importance of individual predictor variables and the interaction between groups of variables, thereby supporting the interpretability of a model. Within the same concept, we get to interpret Choquet integral in a multicriteria distance to the most delegated, partition based clustering algorithms namely k-means. Where this aggregation function takes into consideration the interaction between criteria without losing criteria information. Moreover, the proposed approach is rather modest in multicriteria clustering, compared with the classical Euclidean distance which makes the combination with k-means clustering and the famous aggregation function Choquet Integral. A generalized data was illustrate in the problem which is compared with the usual k-means algorithm. By comparing the result of this practical approach, it was found that the results obtained are more accurate, easy to understand and above all need less time to process.

Mots-clef: Decision Theory and Analysis, Fuzzy Sets and Systems, Machine Learning, Multi-Criteria Decision Analysis.

The contact person (who attend the conference): KERBOUI Roumeissa

1 Introduction

The field of Multi-Criteria Decision Aid (MCDA), looks closer at the way in which we view objects on which we can express a preference and studies the way in which we reach certain decisions on them. In comparison to Data Analysis, the information available on these objects is richer. The objects become, in this context, decision alternatives from the perspective of one or several persons called decision makers (DMs).

We may be able to find certain similarities between the problems of classification and clustering from Data analysis and those of sorting and ranking from MCDA. The problem of clustering, however, has not received

*moncef_abbas@yahoo.com
†romahisam@gmail.com

a very large interest in the field of MCDA. Several methods that have been developed use concepts native to the field of Data Analysis, and thus not use the richer information that is available in MCDA, namely the decision-makers's preferences on the decision alternatives, while others try to construct distance measures to characterize globally how similarly two alternatives compare to the rest. One of the most popular techniques of clustering ease implementation and speed execution is the k-means algorithm [MM14]. The algorithm is based on the initial choice of cluster centers and there is an extensive research on initialization methods.

Clustering such complex data is particularly advantageous for exploratory data analysis. Researchers have generally shown that, clustering by using well-known conventional algorithms generate clusters with acceptable structural quality and consistency, and are partially efficient in terms of execution time. However, classic machine learning and data mining algorithms do not work well for time series due to their unique structure. The high dimensionality, very high feature correlation, and the large amount of noise that characterize time series data present difficult challenges for clustering. Accordingly, massive research efforts have been made to present an efficient approach for time series clustering.

In this study, we introduce a relatively new fuzzy decision support technique based on an aggregation function named the Choquet Integral (CI) [Cho54] into the MCDM clustering problem. The generalization of the weighted arithmetic mean, is the Discrete Choquet Integral, while being widely used as a flexible aggregation function in fields such as multiple criteria decision making was proposed by many authors [DM04], [LLVR13]. This integral, which is constructed from the concept of fuzzy measure were used in the sense of interaction between criteria.

The aim is to modifier the objective function of k-means, that is to use another principal to choose the closest one. To use the application of Choquet Integral in the k-Means algorithm is to calculate the aggregation of all Actions including the centroids of the groups, Then detect each action nearest the centroid of each group. Using the Choquet Integral technique, the Clustering process have shown an efficient result compared to the main algorithm of clustering K-Means.

2 Clustering Application

In the MCDA methodology, clustering alternatives into homogeneous groups consists in assigning a set of n actions $X = \{x_1, x_2, ..., x_n\}$ evaluated on m criteria $\{f_1, f_2, ..., f_m\}$ to one of the groups while examining their intrinsic value. From the comparison of the evaluation of alternative on all criteria with the centers groups, result get the assignment of an alternative to specific group.

In this paper we are interested on this problematic: clustering alternatives into groups which will remain as nearest as possible to a group. We refer the interested reader to this study. The k-means algorithm is one of the most usually unsupervised technique. This method allows to group the alternatives into clusters in which an alternative is assigned to class the nearest distance with the centers of different classes. while the distance between centers of different classes are largest. Following the multicriteria approach, we are interested in defining the nearest one to the best aggregation of each alternative using the aggregation function in MCDA called "Choquet Integral" that takes into account the multicriteria nature of the problem.

In this paper, we present an extension of the k-means algorithm to the multicriteria framework. The main idea of the method is the following: firstly aggregate all the alternatives with Choquet integral, take into consideration the interaction 2-additive between criteria and then assign the alternative to the right group which has the closest value to center of a group.

2.1 K-means algorithm

The k-Means is one of the useful and simplest learning algorithms that can solve the well-known clustering problem. The procedure classifies a data set through a priori fixed k groups following a simple and easy way. In this paper, we present an extension of the k-means algorithm to the multicriteria framework. The main idea of the method is the following: firstly aggregate all the alternatives with Choquet integral, take into consideration the interaction 2-additive between criteria and second assign alternative to the right group which has the closest value to the center of a group.

2.2 Choquet Integral

Although the Choquet integral has been widely applied as an aggregation operator in multiple criteria decision making. The problem of extracting a Choquet integral in a data-driven way has been addressed in the literature. Thus, we recall the basic definition of the (discrete) Choquet integral and related notions. Besides, the authors define the Choquet integral based on a so-called fuzzy measure, it was later on introduced by Choquet .

3 Main work

3.1 K-Means algorithm modify

The Algorithm partitions a set of N data $x_1, x_2, ..., x_N$ in to c $(1 < c < N)$ clusters and The result is a set of c centers $v_1, v_2, ..., v_c$. This structure arises though the minimization of the following objective function:

$$K = \sum_{k=1}^{c} \sum_{i=1}^{N} \mid C_\mu(v_k) - C_\mu(x_i) \mid$$

where $C_\mu(*)$ is the choquet integral w.r.t. a capacity μ given by:

For an alternative $x := (x_1, ..., x_n) \in X$, the expression of the choquet integral of a function $f : X \to \mathbb{R}_+$ with respect to capacity μ is given by:

$$C_\mu(f) := \sum_{i=1}^{n} (f(x_{(i)}) - f(x_{(i-1)})).\mu(E_{(i)}), \tag{1}$$

The algorithm is composed of the following steps:

1. Generate k group centers represented by the objects that are being clustered.

2. Aggregate all the alternatives set with the Choquet integral, in addition to the centers of each group.

3. Affect each alternative to the cluster that has the nearest center.

4. Recalculate k centers for all clusters, when all alternatives have been affected.

5. Repeat Steps 2 and 3 until no change produce for each group centers.

4 Perspective and remark

K-means algorithm use the Euclidean distance, which is based on aggregated data and thus may lose some information. Furthermore, we have proposed a modified k-means algorithm using Choquet integral. More specifically, we have used the Choquet integral as an aggregation function under the concept of fuzzy measure.

Without loss of genarality and in the final clustering of the method k-means, we can apply Choquet integral in running of the procedure, where we find many recent studies based in the use of this concept in Multicriteria decision problem. The algorithm has been tested on generalized data sets and shows encouraging results. By comparing the result of this practical approach, it was found that the results obtained are more accurate, easy to understand and above all the time to process is shorter. Our developed procedure addresses the above-mentioned challenges and shows promising results. Needless to say, this study is only a first step and should be complemented by more extensive experiments including diverse types of data sets in several field. In fact, due to the large number of constraints that have to be satisfied, this problem may become computationally complex. Dedicated techniques for solving it in a more efficient way are therefore desirable.

Future studies will be based on the application of the multicriteria clustering concept introduced in the framework to other clustering or classification algorithms with real-world data sets (country risk problem and diagnosis of firms).

References

[Cho54] Gustave Choquet. Theory of capacities. *Annales de l'institut Fourier*, 5:131–295, 1954.

[DM04] Y. De Smet and L. Montano Guzmán. Towards multicriteria clustering: An extension of the *k*-means algorithm. *European Journal of Operational Research*, 158(2):390–398, 2004.

[LLVR13] Gang Li, Rob Law, Huy Quan Vu, and Jia Rong. Discovering the hotel selection preferences of hong kong inboubd travelers using the choquet integral. *in Data Mining Application With R. Elsevier*, 2013.

[MM14] Igor Melnykov and Volodymyr Melnykov. On k-means algorithm with the use of mahalanobis distances. *Statistics and Probability Letters*, 84(C):88–95, 2014.