HAL
archives-ouvertes.fr

A clustering model for uncertain preferences based on belief functions

Yiru Zhang, Tassadit Bouadi, Arnaud Martin

▶ **To cite this version:**

Yiru Zhang, Tassadit Bouadi, Arnaud Martin. A clustering model for uncertain preferences based on belief functions. DaWaK: Data Warehousing and Knowledge Discovery, Sep 2018, Regensburg, Germany. 20th International Conference on Big Data Analytics and Knowledge Discovery, 2018, <10.1007/978-3-319-98539-8_9>. <hal-01880391>

HAL Id: hal-01880391
https://hal.archives-ouvertes.fr/hal-01880391

Submitted on 24 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A clustering model for uncertain preferences based on belief functions

Yiru Zhang, Tassadit Bouadi, and Arnaud Martin

Univ Rennes 1, CNRS, IRISA
firstname.lastname@irisa.fr
http://www-druid.irisa.fr/

Abstract. Community detection is a popular topic in network science field. In social network analysis, preference is often applied as an attribute for individuals' representation. In some cases, uncertain and imprecise preferences may appear in some cases. Moreover, conflicting preferences can arise from multiple sources. From a model for imperfect preferences we proposed earlier, we study the clustering quality in case of perfect preferences as well as imperfect ones based on weak orders (orders that are complete, reflexive and transitive). The model for uncertain preferences is based on the theory of belief functions with an appropriate dissimilarity measure when performing the clustering steps. To evaluate the quality of clustering results, we used Adjusted Rand Index (ARI) and silhouette score on synthetic data as well as on Sushi preference data set collected from real world. The results show that our model has an equivalent quality with traditional preference representations for certain cases while it has better quality confronting imperfect cases.

Keywords: Clustering for orders; Imperfect preference modeling; Theory of belief functions

1 Introduction

Community detection is a very popular topic in network science field, and has received a great deal of attention. It is a key task for identifying groups (*i.e. clusters*) of objects that share common properties and/or interact with each other. Many algorithms have been developed for efficient community detection. Depending on the information source used to perform the *clustering* task, these algorithms can be classified into tow main categories: graph structure based techniques [12], and node attribute based techniques [18]. The first one considers the *relationships* and the *connections* between the objects (*e.g.* friendships or professional relationships between social network agents[1], proteins interactions, etc.), while the second one analyses the similarity between objects based

[1] In decision making theory, different terms may refer to the same concepts. To avoid ambiguity, we unify the terminology concerning preferences. In this article, "agents" is used for individuals expressing their preferences, "alternatives" for items which are compared in preferences.

on their attribute and feature information (*e.g.* gender, location, personal interests, etc.). More recently, some works [14] discuss hybrid techniques using both node attributes and network topology for community detection. In this paper, we are particularly interested in attribute based techniques in the context of social networks. More precisely, we consider the case of interest-based social networks (*e.g.* Pinterest, Flickr, etc.) where agents sharing similar interests, opinions or viewpoints on some topics belong to the same community. In many real life applications, preferences (*i.e.* a preference describes how an agent orders any two alternatives) are considered to be very useful to efficiently express and model agent's interests, needs or wishes. The aim of this work is to propose a novel community detection method based on clustering agents according to their preferences.

Few work has been done on clustering agents based on their preferences. Kamishima *et al.* in [10] proposed the k'o-means clustering method, an adaptation of the k-means method, adjusted to support preference orders. In [17], the authors introduced a new community detection algorithm based on preference network. The communities are constructed according to the node preferences (*i.e.* each node gives information about its preferred nodes in order to be in the same group).

However, preferences are not always expressed firmly or consistently, sometimes a preference may be uncertain or imprecise facing an unknown situation, or conflicting when dealing with multiple sources. To the best of our knowledge, none of the work mentioned above has investigated preference-based clustering methods when preferences are imperfect (*i.e.* uncertain, imprecise or conflicting). To form meaningful groups of agents according to their preferences, a clustering algorithm need to capture the preference data structure and to cope with imperfect information.

In previous works [19, 11], a qualitative and expressive preference modeling strategy based on the theory of belief functions to model imperfect preferences was proposed. In this paper, starting from this model for agent's preference modeling, we develop a preference-based clustering approach in the space of theory of belief functions. We discuss the clustering quality of our method by considering the Adjusted Rand Index (ARI) and silhouette coefficient as evaluation criteria. To highlight the relevance of the proposed solution, we perform experiments on synthetic and real data to compare our method with different preference modelings, reference in the field, and found the advantage in the expressiveness of the uncertainty and the conflict of the preferences.

Outline of the paper is as follows. In section 2, we give background information related to preference orders, similarity measures over orders, and theory of belief functions. We then explain in section 3 our previous preference model based on theory of belief functions and in section 4 our clustering approach in detail. Experiments and their analysis are given in section 5. We finish with the conclusion and perspectives in section 6.

2 Basic notions

2.1 Preference Order

Definition 1 *(Binary Relation) Let* $A = \{a_1, a_2, \ldots, a_{|A|}\}$ *be a finite set of alternatives, a* ***binary relation*** O *on the set* A *is a subset of the Cartesian product* $A \times A$*, that is, a set of ordered pairs* (a_i, a_j) *such that* a_i *and* a_j *are in* $A : O \subseteq A \times A$ *[13].*

A binary relation satisfies any of the following properties: *reflexive, irreflexive, symmetric, antisymmetric, asymmetric, complete, strongly complete, transitive, negatively transitive, semitransitive, and Ferrers relation* [13]. The detailed definitions of the properties are not in the scope of this article.

Based on definition 1, we denote the binary relation "prefer" by $\succeq$. The relation[2] $a_i \succeq a_j$ means "a_i is at least as good as a_j". Inspired by a four-valued logic introduced in [1, 11, 4], we introduce four relations between alternatives. Given the alternative set A and a preference order $\succeq$ defined on A, we have $\forall a_i, a_j \in A$, the four possible relations defined by:

- **Strict preference** denoted by P: $a_i \succ a_j$
  (a_i is strictly preferred to a_j) $\Leftrightarrow a_i \succeq a_j \wedge \neg(a_j \succeq a_i)$
- **Inverse strict preference** denoted by $\neg P$: $a_j \succ a_i$
  (a_i is inversely strictly preferred to a_j) $\Leftrightarrow \neg(a_j \succeq a_i) \wedge a_j \succeq a_i$
- **Indifference** denoted by I: $a_i \approx a_j$
  (a_i is indifferent, or equally preferred, to a_j) $\Leftrightarrow a_i \succeq a_j \wedge a_j \succeq a_i$
- **Incomparability** denoted by J: $a_i \sim a_j$
  (a_i is incomparable to a_j)$\Leftrightarrow \neg(a_i \succeq a_j) \wedge \neg(a_j \succeq a_i)$

A preferences structure $\langle P, I, J\rangle$ on multiple alternatives can therefore be presented by a binary relation [13].

Definition 2 *(Preference Structure) A preference structure is a collection of binary relations defined on the set* A *such that for each pair* a_i, a_j *in* A*:*

- *at least one relation is satisfied*
- *if one relation is satisfied, another one cannot be satisfied.*

The model for uncertain preferences detailed in [19] is compatible with quasi-orders while our dissimilarity measure is suitable for weak orders, which is a subset of quasi-orders, defined as [13]:

Definition 3 *(Weak Order) Let* O *be a binary relation* ($O = P \cup I$) *on the set* A*,* O *being a characteristic relation of* $\langle P, I\rangle$*, the following three definitions are equivalent:*

1. O *is a weak order.*
2. O *is reflexive, strongly complete and transitive.*
3. $\begin{cases} I \text{ is transitive} \\ P \text{ is transitive} \\ P \cup I \text{ is reflexive and complete.} \end{cases}$

[2] As $a_i \succeq a_j$ is equivalent to $a_j \preceq a_i$, to avoid repetitive comparisons between two alternatives, we assume $i > j$ in this article.

2.2 Dissimilarity between orders

To measure the dissimilarity between two preferences represented by total orders, metrics such as Euclidean distance and Kendall distance are often adopted. Given two preference orders O_1 and O_2 on the same alternatives, we give some basic concepts on such metrics.

Euclidean distance The rank function $r(O, a)$ denotes the position of the alternative a according to the order O. For example, for the order $O = a_1 \succ a_3 \succ a_2$, $r(O, a_1) = 1$ and $r(O, a_2) = 3$. Thus, for two orders O_1 and O_2 on the same alternative set A, Euclidean distance ($l^2 - norm$) between two orders is defined by:

$$d_{l^2}(O_1, O_2) = \sqrt{\sum_{a \in A} (r(O_1, a) - r(O_2, a))^2} \tag{1}$$

Kendall's τ distance and Fagin distance Kendall τ distance measures the dissimilarity with "penalty". Fagin proposed a more general metric in [6] adapting for orders with indifference based on Kendall distance, we name it Fagin distance in this article. In Fagin distance, for alternatives a_i, a_j, the penalty between two orders O_1 and O_2 on a_i and a_j, denoted as $\bar{K}^{(p)}_{i,j}(O_1, O_2)$, is defined as follows:

- **Case 1**: a_i and a_j are in both O_1 and O_2. If a_i and a_j are ordered in the same way (such as $a_i \succ a_j$ in both O_1, O_2), $\bar{K}^{(p)}_{i,j}(O_1, O_2) = 0$, this corresponds to "no penalty" for a_i and a_j. If a_i, a_j are ordered reversely (such as $a_i \succ a_j$ in O_1 while $a_i \prec a_j$ in O_2), the penalty of a_i, a_j $\bar{K}^{(p)}_{i,j}(O_1, O_2) = 1$.
- **Case 2**: a_i and a_j are tied in both O_1 and O_2, $\bar{K}^{(p)}_{i,j}(O_1, O_2) = 0$. Intuitively, both partial orders agree that a_i and a_j are tied.
- **Case 3**: a_i and a_j are indifference in one of the partial order (say O_1) and of different rank in the other order (therefore O_2), we give a penalty parameter $\bar{K}^{(p)}_{i,j}(O_1, O_2) = p$ [3].

Based on these cases, the *Kendall distance with penalty parameter p* (*i.e.* Fagin distance) is defined as follow:

$$K^{(p)}(O_1, O_2) = \sum_{i,j \in [1,|A|]} \bar{K}^{(p)}_{i,j}(O_1, O_2) \tag{2}$$

2.3 Belief functions

The theory of belief functions (also referred to as Dempster-Shafer or Evidence Theory) was firstly introduced by Dempster [2] then developed by Shafer [16]

[3] In our work, we take $p = 0.5$

as a general model of uncertainties. It is applied widely in information fusion and decision making. By extending probabilistic and set-valued representations, it allows to represent degrees of belief and incomplete information in an unified framework. Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be a finite set. A *(normalized) mass function* on Ω is a function $m : 2^\Omega \to [0, 1]$ such that:

$$m(\emptyset) = 0 \tag{3}$$

$$\sum_{X \subseteq \Omega} m(X) = 1 \tag{4}$$

The subsets X of Ω such that $m(X) > 0$ are called *focal elements* of m, while the finite set Ω is called *framework of discernment*. A mass function is called *simple support* if it has only two focal elements: $X \subseteq \Omega$ and Ω. A mass function having only one focal element $A \in \Omega$ is called a *categorical mass function*.

For example, if we consider the simple support mass $m(\omega_1 \cup \omega_2) = 0.8$, $m(\Omega) = 0.2$, this mass function represents an uncertainty with the degree 0.8 on the imprecise element $\{\omega_1$ or $\omega_2\}$ and a partial ignorance with the degree 0.2 on Ω.

2.4 Distance on belief functions

Several distances can be used on belief functions and Jousselme distance is considered as a reliable similarity measure between different mass functions [5]. It considers coefficients on the elements composed by singletons. Jousselme distance is defined as follows, denoted by d_J:

Definition 4 *Let m_1 and m_2 be two mass functions on the same frame of discernment Ω, containing $|\Omega| = n$ mutually exclusive and exhaustive hypotheses. The distance between m_1 and m_2 is:*

$$d_J(m_1, m_2) = \sqrt{\frac{1}{2}(m_1 - m_2)^T \underline{\underline{D}}(m_1 - m_2)} \tag{5}$$

where $\underline{\underline{D}}$ is the $2^n \times 2^n$ Jaccard matrix given by:

$$D(A, B) = \frac{|A \cap B|}{|A \cup B|}; A, B \subseteq \Omega. \tag{6}$$

3 Preference model under uncertainty

In this section, we detail the preference model proposed in [19] as well as different dissimilarity measures for orders.

3.1 Problem setting

We denote by U a set of agents, $U = \{u_1, u_2, \ldots, u_{|U|}\}$, and A a set of mono-criterion alternatives, $A = \{a_1, a_2, \ldots, a_{|A|}\}$. Every agent $u_i \in U$ expresses his/her preferences over A by a quasi order in the space of $A \times A$, denoted by O_i. Expert's preferences order O_i may come from two different sources $S = \{s_1, s_2\}$, denoted by O_1 and O_2.

3.2 Preference model on belief functions

The objective is to cluster the experts represented by their preferences under uncertainty. We consider the model from [19] to represent this uncertain preference by the theory of belief functions. The framework of discernment is defined on possible relations:

$$\Omega_{ij} = \{\omega_{ij}^1, \omega_{ij}^2, \omega_{ij}^3, \omega_{ij}^4\} \tag{7}$$

where ω_{ij}^1, ω_{ij}^2, ω_{ij}^3 and ω_{ij}^4, represent respectively $a_i \succ a_j$, $a_i \prec a_j$, $a_i \approx a_j$ and $a_i \sim a_j$. This procedure consists in two steps:

1. Initialization of mass functions
2. Clustering on quasi orders represented by mass functions.

This model is used in our contribution. Its utilization is described in detail in the following sections.

4 Contribution: agent clustering based on their preferences

In this section, we explain how the agents are represented and clustered from two sources of preferences. The clustering procedure is straightforward, concisely illustrated in figure 1. The first block concerns the representation of agents

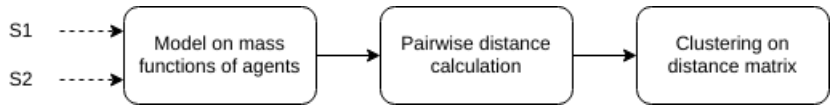

Fig. 1. Flowchart of preference based clustering

and modeling of mass functions from two preference sources S_1, S_2 (sections 4.1 and 4.2). The second block concerns the measure of dissimilarity between agents (section 4.3). The third block concerns clustering algorithm, we use EkNNclus algorithm in our work (section 4.4).

4.1 Representation of agents

We consider the case that a group of $|U|$ agents expressing their preferences between each pair of alternatives from the set A. Therefore, the preferences of an agent u, denoted by O_u, is represented by a mass function on all possible alternative pairs:

$$O_u := \left[m_{1,2}; m_{1,3}; \ldots; m_{1,K}; m_{2,3}; \ldots; m_{|A|-1,|A|}\right] \tag{8}$$

Hence, for $|A|$ alternatives, the representation of an agent is made up by $\frac{(|A|-1)(|A|-2)}{2}$ mass functions.

4.2 Modeling of mass functions

In our model, we take advantage of the possibility of expressing on ignorance in the framework given by equation (7). Given two preference order sources O_1, O_2 from one agent, we interpret Fagin distance $d_F(O_1, O_2)$ as ignorance degree when conflict encountered. That is to say: for a_i, a_j $(i < j)$ in both O_1, O_2 of agent u, with their ranking denoted by $r(O, a)$, the mass function value is given according to following conditions:

1. **Case 1**: a_i, and a_j are in the same relation in both O_1, O_2 (say $a_i \succ a_j$), $m_{i,j}$ is a categorical mass function ($m_{i,j}(\omega^1) = 1$).
2. **Case 2**: a_i and a_j are in the conflicting relations in O_1 and O_2, respectively denoted as $\omega_{o_1}, \omega_{o_1} \in \Omega_{ij}$, $\omega_{o_1} \neq \omega_{o_2}$ (say $a_i \succ a_j$ in O_1 while $a_i \approx a_j$ in O_2, $\omega_{o_1} = \omega^1, \omega_{o_2} = \omega^3$), the mass function values are given by:

$$\begin{aligned} &m_{i,j}(\Omega) = d_F(O_1, O_2) \\ &m_{i,j}(\omega_{o_1}) = m_{i,j}(\omega_{o_2}) = (1 - d_F(O_1, O_2))/2 \end{aligned} \tag{9}$$

4.3 Dissimilarity between different agents

The dissimilarity measure is based on Jousselme distance [8] for mass functions. Given two mass functions modeling preference relations between alternatives i and j from agents u_1 and u_2 expressing preference orders O_1, O_2. We denote Jousselme distance as $d_J(m_{ijO_1}, m_{ijO_2})$. The dissimilarity between two agents' preferences is denoted via Jousselme distance as:

$$d(O_1, O_2) = \sum_{j=1}^{k} \sum_{i=1, i<j}^{k} d_J(m_{ijO_1}, m_{ijO_2}) \tag{10}$$

Where m_{ijO} denotes the mass function of alternative pair (a_i, a_j) according to the order O. Therefore, a normalized distance is given by

$$d_{Normalize}(O_1, O_2) = \frac{1}{Nb_{total}} d(O_1, O_2) \tag{11}$$

where $Nb_{total} = \frac{(|A|-1)(|A|-2)}{2}$, is the amount of all alternative pairs.

To simplify the expression, we use BF model to refer to our model and the corresponding dissimilarity function.

4.4 Unsupervised classifier–Ek-NN [3]

For dissimilarity spaces in which only pairwise distances are given (such as Kendall distance), the centroid of several agents is a metric k-center problem and is proved to be NP-hard. Therefore, we avoid using clustering methods requiring the calculation of centroid, such as k-means.

We applied Ek-NNclus method [3] as classifier. Ek-NNclus is a clustering algorithm based on the evidential k-nearest-neighbor rule, thus it requires only

the pairwise metric for k-nearest-neighbor searching. Another advantage of EkNNclus is that the number of clusters does not need to be determined in advance, only the neighborhood size k should be set. A determination of k is proposed in [20].

5 Experiments

Although the model was originally designed for preferences under uncertainty, we still wonder its quality for clustering on certain preferences. Thus the clustering quality of our model can be divided into two aspects: on certain preferences and on uncertain preferences.

5.1 Evaluation criteria

With the similar aforementioned reasons, it's NP-hard to calculate centroids. Thus, we choose two evaluation criteria that do not require a cluster centroid calculation: Adjusted Rand Index (ARI) [7] for data with ground truth and silhouette coefficient [15] for any dataset.

We tested different metrics on synthetic certain and uncertain preferences. We also compared different metrics on a real world certain preferences from SUSHI data set [9].

In the following parts, we introduce the method of generating synthetic preferences and compare the clustering quality of different metrics. To simplify the experiments, all preferences are expressed in a space of 10 alternatives.

5.2 Certain preferences

On synthetic data Certain preferences are those who are from non-conflicting sources. In this case, we only consider and generate one source of preferences. To study the clustering quality, we firstly generate preferences with different ranges to their centroids. The data is generated in following steps in algorithm 1.

By increasing the number of switching operations T, we obtain clusters with different densities.

To avoid random errors, we generate different preference sets 10 times and take the average value of ARI and silhouette score. Besides, the optimal parameter K in EkNN-clus algorithm varies with the size of data and distribution of the samples. The selection of K is not in the scope of this article. We test on various K and choose the one that returns the largest ARI and average silhouette coefficient[4] as our result.

In figures 2, 3, 4, ARI and silhouette coefficient performed on generated data with neighbors in different ranges (switch time from 1 to 3) and different sizes (neighbor size[5] varies from 10 to 100) are illustrated.

[4] Without special remark, we use term "silhouette coefficient" for "average" value on set of samples by default.

[5] By saying neighbor size, we mean the number of samples in each cluster.

Algorithm 1 Generate preferences in $|C|$ clusters

Input: Cluster number $|C|$
Switch time T
neighbour size NS
Alternative size in each order $|A|$
Output: $|C|$ clusters of preferences
// Centroids initialization
randomly generate centroid c_1 of $|A|$ elements.
for i_c in $2 : |C|$ **do**
  $dist\_max = 0$
  for s in $1 : 5000$ **do**
    randomly generate preference o_s of $|A|$ elements
    $dist\_sum = \sum_{i=1}^{i_c-1} d_{Kendall}(o_s, c_i)$
    if $dist\_sum > dist\_max$ **then**
      $dist\_max = \sum_{i=1}^{i_c-1} d_{Kendall}(o_s, c_i)$
      centroid $c_{c_i} = o_s$
    end if
  end for
end for
Generate neighbors
for each centroid o_c **do**
  for ns in $1 : NS$ **do**
    for t in $1 : T$ **do**
      randomly generate index i, j
      exchange ranking order of a_i, a_j in o_c, making a new order
    end for
  end for
end for

According to these results, one can conclude that the BF model and Kendall distance have equivalent good quality both in terms of ARI and silhouette score, while Euclidean distance always has a poor quality. A high value in ARI usually corresponds to a high silhouette score, signifying a good clustering result.

On real data SUSHI preference dataset [9] is collected from a survey on Japanese consumer preferences over different sushis. It has a data set containing 5000 complete strict rank orders (*i.e.* total orders) of 10 different kinds of sushi.

We applied these three metrics in clustering on real data of Sushi Preference Data Set. Figure 5 illustrates silhouette plots of clusters with different metrics.[6] Kendall distance and BF model have similar quality. Euclidean distance has a relatively poor quality. This result is consistent with the synthetic data in figures 2, 3 and 4.

Among the three metrics, none of them has an absolutely high silhouette score (larger than 0.5). This is due to the quality of the data. SUSHI dataset does not guarantee the existence of obvious communities among the agents.

From both synthetic data and real world data without uncertainty, we observe that BF model and Kendall distance have very similar clustering. In fact, in case of certain problems, where all mass functions are categorical, for total orders, the normalized distance in BF model is degraded to Kendall distances. This can be easily proved by their definitions.

[6] As different K in EKNN-clus algorithm returns different clustering results, we compare clustering result who returns relatively high silhouette coefficient.

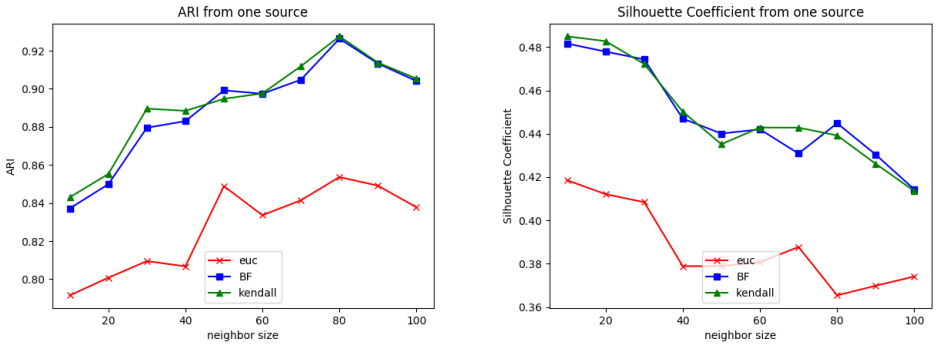

Fig. 2. ARI and silhouette coefficient, switch = 1

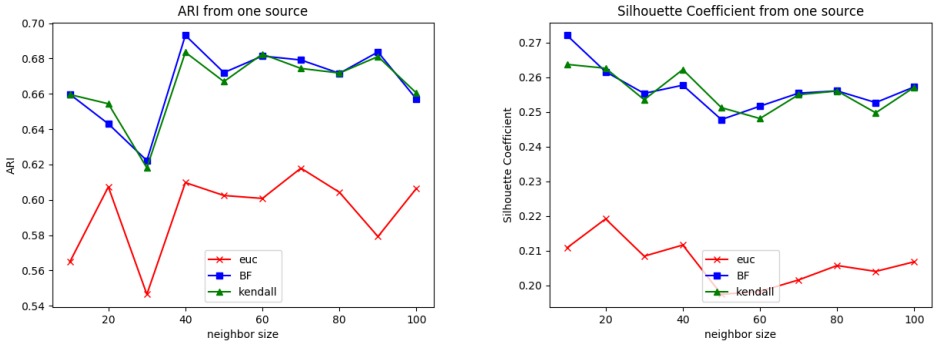

Fig. 3. ARI and silhouette coefficient, switch = 2

5.3 Uncertain preferences

In this part, we suppose a case that two preferences are given with different representations: ranking and score. The ranking preferences are generated in the same way as in subsection 5.2. Scores are generated by the following steps: scores range from 1 to 5 are generated respecting to a given rank preference. In this way, indifference relations are introduced, causing conflicts between two preference sources. Given a ranking preference O_r of 10 alternatives a_1 to a_{10}, the scores are generated by the following rules:

- For least preferred two alternatives (2 alternatives at the end of the O_r, *i.e.* ranking no. 9 and 10), we give score 1.
- For alternatives sorted at the positions 7 and 8, we give score 2.
- With the similar rule, for most preferred two sushis (ranking no. 1 and 2), we give score 5.

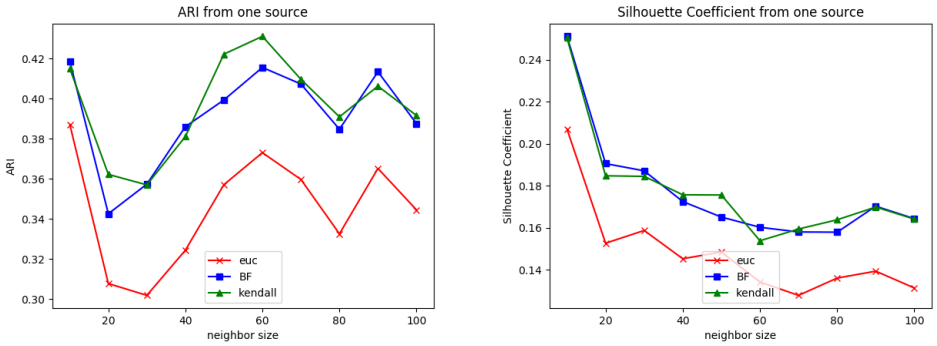

Fig. 4. ARI and silhouette coefficient, switch = 3

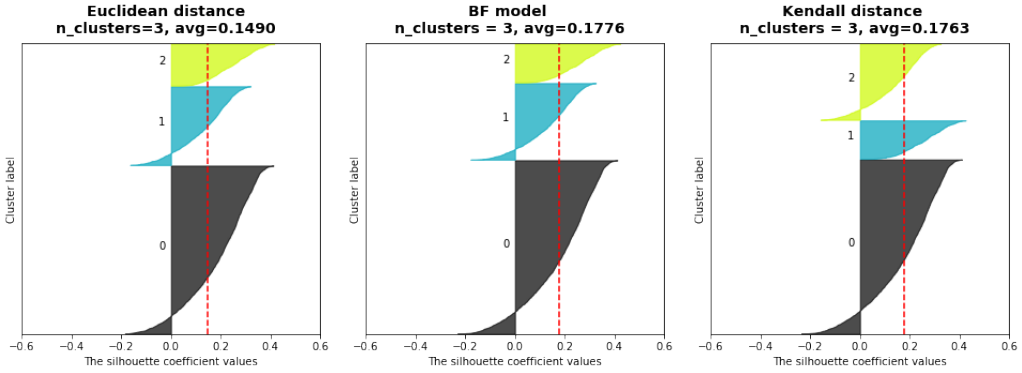

Fig. 5. Silhouette plots of different metrics on SUSHI

For example with:

$$a_1 \succ a_2 \succ a_3 \succ a_4 \succ a_5 \succ a_6 \succ a_7 \succ a_8 \succ a_9 \succ a_{10}$$

the scores are: $a_1 : 5, a_2 : 5, a_3 : 4, a_4 : 4, a_5 : 3, a_6 : 3, a_7 : 2, a_8 : 2, a_9 : 1, a_{10} : 1$.

Still, ARI and silhouette scores are applied as evaluation criteria. We compared our model with an average-based-euclidean metric calculated as follows:

Confronting a case of two preferences: ranking O_r and score O_s, the mean rank of alternative a_i is calculated:

$$\bar{r}(a_i) = \frac{1}{2}(r(O_r, a_i) + r(O_s, a_i))$$

Thus, agent u's average preference order is represented by:

$$\bar{O}_u := \left[\bar{r}(a_1), \bar{r}(a_2), \ldots, \bar{r}(a_{|A|})\right] \tag{12}$$

Therefore, the example above has a such vector:

$[1, 1.5, 3, 3.5, 5, 5.5, 7, 7.5, 9, 9.5]$.

For Kendall distance, we calculate the distance matrix from rankings and scores, then take the average value as the combined distance. As indifference relations exist in O_s, we apply Fagin distance for O_s. Given ranking preferences O_{r1}, O_{r2} and score preferences O_{s1}, O_{s2}, denoting the preference from agent u_1 and u_2, the average distance is thus given by:

$$\bar{d}_{kendall}(u_1, u_2) = \frac{1}{2}(d_{kendall}(O_{r1}, O_{r2}) + d_{Fagin}(O_{s1}, O_{s2})) \tag{13}$$

We compared the model based on Euclidean distance equation (12), Kendall distance (13) and BF model given in equation (8).

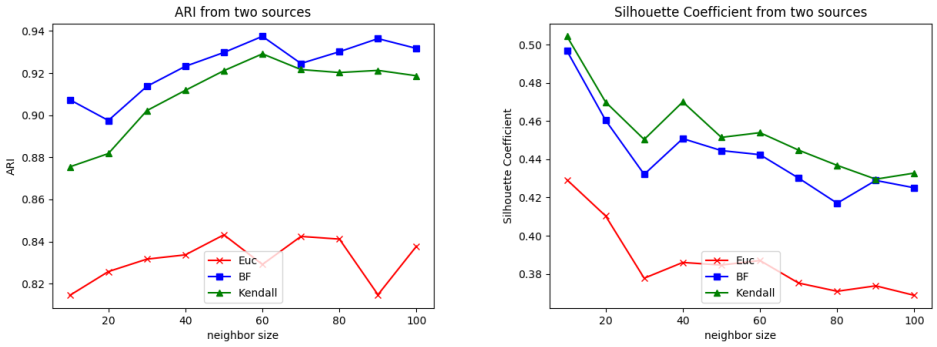

Fig. 6. ARI and silhouette coefficient on uncertain preferences, switch = 1

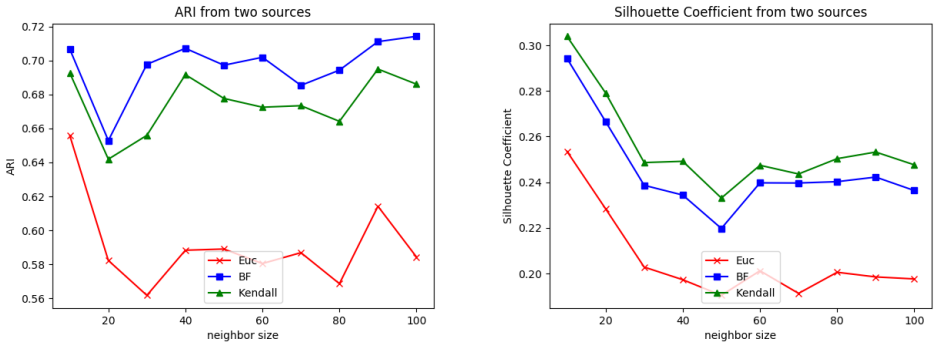

Fig. 7. ARI and silhouette coefficient on uncertain preferences, switch = 2

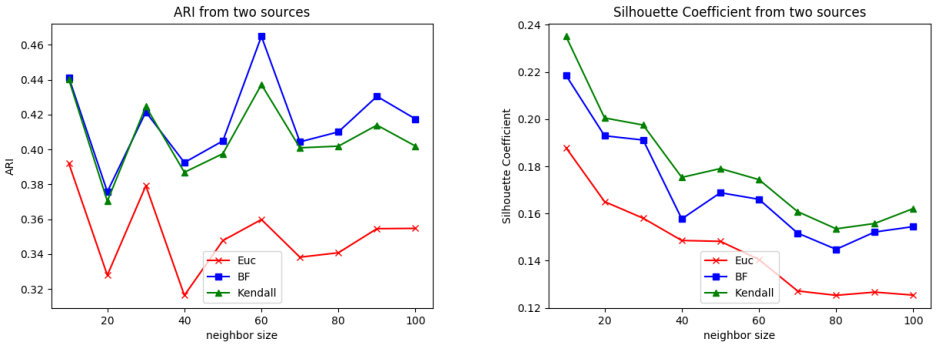

Fig. 8. ARI and silhouette coefficient on uncertain preferences, switch = 3

The results illustrated by these figures show the advantage of BF model over Euclidean distance and Kendall (Fagin) distance when dealing with two sources. Comparing Figure 6,7,8 from conflicting sources with Figure 2,3,4, we observe that both averaged Euclidean distance and Kendall distance are deteriorated more than BF model. The results prove the advantage of BF model on preferences under uncertainty. This advantage comes from the fact that in BF model, conflicts are partly interpreted as ignorance and have less impact in dissimilarity measuring. However, this compromise also causes a loss in criterion of silhouette coefficient.

6 Conclusion and perspectives

In this paper, we investigate the problem of clustering individuals according to their preferences, when dealing with multiple and conflicting sources (two in our case study). To cope with this issue, we apply the theory of belief functions (BF model) to express and interpret the contradictions and conflicts from different sources as uncertainty and ignorance. We introduce a new approach that captures the preference data structure and deal with uncertain information.

To highlight the relevance of the proposed solution, we perform experiments on synthetic and real data to compare our method with other preference models, and found the advantage in the expressiveness of the uncertainty and the incomparability of the preference orders. Indeed, we compare BF model on synthetic data between Euclidean distance and Kendall distance both in certain and uncertain cases, using Ek-NNclus algorithm for clustering. In certain cases, BF model has equivalent clustering-quality with Kendall distance and outperforms Euclidean distance. In uncertain cases, BF model has better clustering-quality over the other distances. We also applied this model on SUSHI preference data set and found that BF model has one of the most satisfying clustering-quality.

We applied the BF model on complete preference orders (*i.e. weak orders*) from only two sources. In the future, we will work on an ameliorated BF model

dealing with several conflicting preference sources. In fact, the combination of preferences from multiple sources is a social choice problem, and different combination rules can be applied, corresponding to different complexity. Moreover, a more general dissimilarity measure method for incomplete orders (*i.e. quasi-orders*) is also in the scope of our future work.

References

1. Belnap, N.D.: A Useful Four-Valued Logic, pp. 5–37. Springer Netherlands, Dordrecht (1977)
2. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Statist. 38(2), 325–339 (04 1967)
3. Denoeux, T., Kanjanatarakul, O., Sriboonchitta, S.: Ek-nnclus. Know.-Based Syst. 88(C), 57–69 (Nov 2015)
4. Elarbi, F., Bouadi, T., Martin, A., Ben Yaghlane, B.: Preference fusion for community detection in social networks. In: 24ème Conférence sur la Logique Floue et ses Applications. Poitiers, France (Nov 2015)
5. Essaid, A., Martin, A., Smits, G., Ben Yaghlane, B.: A Distance-Based Decision in the Credal Level. In: International Conference on Artificial Intelligence and Symbolic Computation (AISC 2014). pp. 147 – 156. Sevilla, Spain (Dec 2014)
6. Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E.: Comparing and aggregating rankings with ties. In: Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. pp. 47–58. PODS '04, ACM, New York, USA (2004)
7. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification 2(1), 193–218 (Dec 1985)
8. Jousselme, A.L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning 53(2), 118 – 145 (2012)
9. Kamishima, T.: Nantonac collaborative filtering: Recommendation based on order responses. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 583–588. KDD '03, ACM, New York, NY, USA (2003)
10. Kamishima, T., Akaho, S.: Efficient Clustering for Orders, pp. 261–279. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)
11. Masson, M.H., Destercke, S., Denoeux, T.: Modelling and predicting partial orders from pairwise belief functions. Soft Computing 20(3), 939–950 (2016)
12. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006)
13. Öztürké, M., Tsoukiàs, A., Vincke, P.: Preference Modelling, pp. 27–59. Springer New York, New York, NY (2005)
14. Qin, M., Jin, D., He, D., Gabrys, B., Musial, K.: Adaptive community detection incorporating topology and content in social networks. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 675–682. ACM (2017)
15. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53 – 65 (1987)

16. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
17. Tasgin, M., Bingol, H.O.: Community detection using preference networks. Physica A: Statistical Mechanics and its Applications 495, 126 – 136 (2018)
18. Yang, J., McAuley, J., Leskovec, J.: Community detection in networks with node attributes. In: 13th international conference on Data Mining (ICDM 2013). pp. 1151–1156. IEEE (2013)
19. Zhang, Y., Bouadi, T., Martin, A.: Preference fusion and Condorcet's paradox under uncertainty. In: 20th International Conference on Information Fusion, FUSION 2017. pp. 1–8. Xi'an, China (2017)
20. Zhang, Y., Bouadi, T., Martin, A.: An empirical study to determine the optimal k in ek-nnclus method. In: 5th International Conference on Belief Functions, BELIEF 2018. Compiègne, France (2018)