Unequal Singleton Pair Distance for Evidential Preference Clustering

Yiru Zhang, Arnaud Martin

Unequal Singleton Pair Distance for Evidential Preference Clustering

Yiru Zhang[1,2] and Arnaud Martin[3]

[1] St. Francis Xavier University, Antigonish, Canada
[2] Hainan University, Haikou, China
[3] University Rennes, CNRS, IRISA, DRUID, Lannion, France

Abstract. Evidential preference based on belief function theory has been proposed recently, simultaneously characterizing preference information with uncertainty and imprecision. However, traditional distances on belief functions do not adapt to some intrinsic properties of preference relations, especially when indifference relation is taken into comparison, therefore may cause inconsistent results in preference-based applications. In order to solve this issue, Unequal Singleton Pair (USP) distance has been proposed previously, with applications limited in preference aggregation. This paper explores forward the effectiveness of USP distance in preference clustering, especially confronting multiple conflicting sources. Moreover, a combination strategy for multiple conflicting sources of preference is proposed. The experiments on synthetic data show that USP distance can effectively improve the clustering results in Adjusted Rand Index (ARI).

Keywords: Belief function theory · Preference clustering · Distance.

1 Introduction

With the blossomy development of the digital world, there are various ways to describe one's preference information, such as binary choice (like, dislike), rated with ranks, scores, even colors. Indeed, it is challenging to accurately and effectively cluster preferences, and data quality is one problem. Low quality may be caused by uncertainty, conflicts, incompleteness, or other flaws. We refer to such preference data as "imperfect" in this paper.

Many works have been devoted to modeling imperfect preferences. For example, fuzzy preference [11], possibilistic model [1], probabilistic model [8] and Plackett-Luce model [9] have been proposed to deal with preference with uncertainty and have gained success in various scenarios of applications. However, these methods are usually limited to uncertainty information with uncertainty by proportional or probabilistic values by imposing distribution assumptions.

Dissimilarity measures play an important role in preference analysis, notably in preference aggregation [13, 5] and preference learning [7, 14] applications. The former application concerns combining multiple preferences into a consensus one,

while the latter one concentrates on machine learning over preference information, usually applied in ranking problems [4]. Preference clustering is a mission in preference learning, aiming at categorize the preference information based on their similarities, often applied in recommendation systems and community detection tasks [6, 16, 17]. Indeed, some preference aggregation strategies are intrinsically identical to the minimization of distance sums, as demonstrated in a work of Viappini [15].

Naturally, dissimilarity measure methods in BFT come into the focus for evidential preferences. Even though many dissimilarity measures have been proposed in BFT, they are proved not suitable for evidential preference in [19] because of the conflicts between inherent properties of preference relations. An important one is the equal dissimilarity value between singletons. Formally, in a framework of discernment (FoD) $\Omega = \{\omega_1, \omega_2, \ldots, \omega_H\}$, $\forall \omega_p, \omega_q \in \Omega$, to the limit of our knowledge, the dissimilarity function $d(\cdot)$ over two singletons $d(\omega_p, \omega_q)$ is a constant, usually normalized as 1. However, dissimilarity between singletons should be naturally discriminated in preference relation. For example, the dissimilarity d_Δ between three binary preference relations "strict prefer to" (denoted as $\succ$), "indifferent to" (denoted as $\approx$), and "inverse strict prefer to" (or "preferred by", denoted as $\prec$) is naturally $d_\Delta(\succ, \prec) > d_\Delta(\succ, \approx)$ while all dissimilarity measures in BFT output $d_\Delta(\succ, \prec) = d_\Delta(\succ, \approx)$. This valuation set ignores the intermediate role of "indifference" between the two directions of "strict preference", which is detrimental in distance based applications with weak preferences. Zhang *et. al.* [19] discussed negative consequences of such valuation in preference aggregation application and proposed Unequal Singleton Pair (USP) distance, solving the issue by discriminating the dissimilarity between different singleton pairs with other important properties in BFT still guaranteed.

The effectiveness of USP distance in evidential preference aggregation has already been demonstrated [19], while not applied in evidential preference clustering.

In this paper, we study USP distance in evidential preference clustering applications. The evidential preferences are obtained from conflicting preference sources over identical alternative pairs. In our method, the conflicts between multiple sources are interpreted as the ignorance of an agent. The experiments show that the clustering results are improved by applying USP distance in terms of Adjust Rand Index (ARI).

The paper is organized as follows: in Section 2, basic notions on belief functions as well as evidential preference model are introduced, followed by the calculation tutorial of USP distance and clustering model in Section 3. Afterward, the comparison experiments of clustering with other distances are depicted in Section 4. Conclusion and discussions are given finally in Section 5.

2 Preliminary

2.1 Belief functions

Let $\Omega = \{\omega_1, \ldots, \omega_H\}$ be a finite set representing all possible status of a categorical attribute, the uncertainty and imprecision of this attribute is expressed by Basic Belief Assignment (BBA).

Definition 1. *(Basic Belief Assignment (BBA))* A *Basic Belief Assignment* (BBA) on Ω is a function $m : 2^{\Omega} \rightarrow [0, 1]$ such that:

$$m(\emptyset) = 0 \text{ and } \sum_{X \subseteq \Omega} m(X) = 1. \tag{1}$$

The subsets X of Ω such that $m(X) > 0$ are called *focal elements*, while the finite set Ω is called *the framework of discernment (FoD)*. Ω is also considered as *total ignorance* since it represents all the possibilities. A BBA representing *total ignorance* ($m(\Omega) = 1$) is also called a *vacuous* BBA. A BBA is *simple supported* if a non-zero value is assigned only to one singleton and Ω. Besides, a BBA m is called *categorical* on element $X, X \in 2^{\Omega}$ if $m(X) = 1$, denoted as X^0. We refer to a categorical BBA on one singleton as *categorically simple supported*.

2.2 Evidential Preference Model

Preference modeling is usually based on order theory. In this paper, we use the widely accepted notions in studies of preferences from [12].

Definition 2. *(Preference relation)* Between any two alternatives a_i, a_j, only three exclusive relations possibly exist $\{\succ, \approx, \sim\}$, defined from binary relation R, with $\neg$ denoting logic negation, as:

$$\begin{aligned}
\text{Strict preference: } & a_i \succ a_j \text{ iff } a_i R a_j \text{ and } a_j \neg R a_i; \\
\text{Indifference: } & a_i \approx a_j \text{ iff } a_i R a_j \text{ and } a_j R a_i; \\
\text{Incomparability: } & a_i \sim a_j \text{ iff } a_i \neg \succ a_j \text{ and } a_i \neg \prec a_j \text{ and } a_i \neg \approx a_j.
\end{aligned}$$

Definition 3. *(Preference Structure)* A preference structure is a collection of binary relations defined on the set $\mathcal{A}$ and such that:

- for each couple (a_i, a_j), $a_i, a_j \in \mathcal{A}$, at least one relation is satisfied;
- for each couple (a_i, a_j), $a_i, a_j \in \mathcal{A}$, if one relation is satisfied, any other relation cannot be satisfied.

The evidential preference model is originally proposed by [10] on weak orders and extended to quasi orders with the consideration of *incomparability* by [18].

Definition 4. *(Evidential preference) For any alternative pair $a_i, a_j \in \mathcal{A}$, four relations are possible. Therefore, the preference FoD Ω_{ij}^{pref} is defined as:*

$$\Omega_{ij}^{pref} = \{\omega_{ij}^{R} | R \in \{\succ, \prec, \approx, \sim\}\}. \tag{2}$$

The degree of uncertainty on preference relation is represented by values on singletons. The imprecision is characterized by values on union sets.

With the combination rules in the framework of BFT, the evidential preference model is effective in group decision-making with imperfect preference information sources, as systematically discussed in [19].

3 Clustering model for evidential preferences with Unequal Singleton Pair (USP) distance

In this section, we introduce the clustering model over evidential preference with USP distance, followed by a brief tutorial for calculating USP distance.

3.1 Strategy of reasoning and clustering

The reasoning strategy is designed with the procedure depicted in Figure 1, where σ denotes a preference structure, u an agent, and D the matrix of pairwise distances.

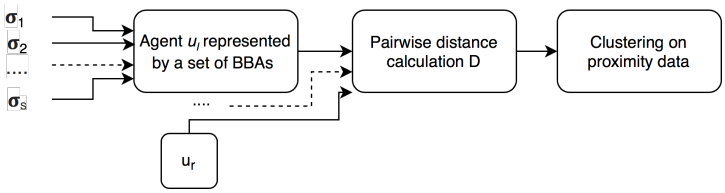


Fig. 1: Strategy of clustering

In a case that an identical agent u's ($u \in \mathcal{U}$) preference is expressed by multiple sources, agent u is therefore represented by a list of pairwise evidential preferences obtained by the combination of multiple sources. Afterward, pairwise distances between different agents are calculated for the clustering process. In this strategy, three main steps are included:

1. Combination of multiple conflicting preference sources for one agent;
2. Calculation of distances between different agents;
3. Clustering over agents based on the proximity distances.

In the following parts, we introduce the combination of multiple preferences and the calculation of distances, while the clustering method is out of the scope because any clustering method for proximity data is available.

3.2 Evidential preference reasoning and combination

Evidential preferences are reasoned from conventional crisp preference information, wildly conflicting preferences from multiple sources. We develop an evidential preference reasoning strategy for multiple (conflicting) sources.

Given multiple preference structures (from multiple sources) $S = \{\sigma_1, \sigma_2, \ldots\}$ for one agent u on identical alternative set $\mathcal{A}$, the average mutual conflicting (AMC) κ_{AMC} among S is defined as:

$$\kappa_{AMC} = \frac{1}{\binom{|S|}{2}} \sum_{\substack{\sigma_p, \sigma_q \in S \\ p<q}} d(\sigma_p, \sigma_q), \tag{3}$$

where d denotes a distance function for preference orders (rather than pair-wise preferences). In this paper, we apply Fagin's distance [3], which is an extended version of Kendall's distance.

The BBAs' values are obtained by mean rule combination with normalization of AMC. For agent u's opinion between a_i and a_j, denote the crisp preference from source s as $\sigma_s(a_i, a_j)$. The BBA m_{ij}^u representing agent u's evidential preference opinion between a_i, a_j is calculated by:

$$m_{ij}^u(X) = \begin{cases} \dfrac{1 - \kappa_{AMC}}{|S|} \displaystyle\sum_{s \in S} m_{ij}^s(X), & \text{if } X \neq \Omega; \\ \kappa_{AMC}, & X = \Omega, \end{cases}$$

where m_{ij}^s is categorical as it comes from a crisp preference source without uncertainty nor imprecision.

The distance between two agents u_r and u_l are calculated by the mean value of their pairwise preference distance, defined as:

$$d(u_r, u_l) = \frac{1}{\binom{|\mathcal{A}|}{2}} \sum_{\substack{a_i, a_j \in \mathcal{A} \\ i<j}} d_{BFT}(m_{ij}^{u_r}, m_{ij}^{u_l}), \tag{4}$$

where d_{BFT} denotes a distance function for BBAs in the theory of belief functions, $\binom{|\mathcal{A}|}{2}$ the combination number of 2 elements out of A. In our work, we apply USP distance as introduced below to avoid the flaw mentioned in the Section 1.

3.3 USP distance

Unequal Singleton Pair distance is originally proposed to solve a flaw existing in all dissimilarity measures in BFT. Before USP distance, all measures value the dissimilarity between singletons equally. In a FoD $\Omega = \{\omega_1, \omega_2, \ldots, \omega_H\}$, the dissimilarity between any two different singletons is a constant (normalized as 1), formally, $\forall \omega_m, \omega_n \in \Omega, \omega_m \neq \omega_n$:

$$d(\{\omega_m\}^0, \{\omega_n\}^0) \equiv 1. \tag{5}$$

USP distance, which is an extensive version of Jousselme distance, can solve this flaw. given for two BBAs m_1 and m_2 in Ω, USP distance is defined by:

$$d_{USP}(m_1, m_2) = \sqrt{(\boldsymbol{m_1} - \boldsymbol{m_2})^T \boldsymbol{\Sigma} (\boldsymbol{m_1} - \boldsymbol{m_2})}, \tag{6}$$

where Σ denotes the similarity matrix between elements in 2^{Ω}. In Jousselme distance, Σ is a Jaccard matrix defined on the structure of elements, while in USP distance, Σ is defined by resemblance $resemb$ and entirety $entire$ of the two elements. The value of resemblance and entirety are calculated by the difference in the similarity between singleton pairs.

Here we give a tutorial for USP distance calculation. Define a set of elements in 2^{Ω}, $W = \{X_1, X_2, \ldots, X_M\}$, therefore $W \subseteq 2^{\Omega}$. Denote $resemb(W)$ for $resemb(X_1, X_2, \ldots, X_M)$ and $entire(W)$ for $entire(X_1, X_2, \ldots, X_M)$ to simplify the expression. The size of W is defined by the number of elements $X \in 2^{\Omega}$, denoted by $|W|$. Singletons in W is defined by the union of all elements in W, formally:

$$\cup W = \bigcup_{X_i \in W} X_i. \tag{7}$$

To guarantee the uniqueness of the solution, the entirety value of a singleton is set as 1. Denote the subset of W by W_{sub}, $entire(W)$ is defined as a generalized version of cardinal function on the union sets:

$$entire(W) = \sum_{\omega \in \cup W} entire(\omega) + \sum_{t=1}^{|2^{\Omega}|} \sum_{\substack{W_{sub} \subseteq W \\ |W_{sub}|=t}} resemb(W_{sub}) \times (-1)^t. \tag{8}$$

To simplify the calculation, we assume that the resemblance values are non-zero only between two singletons and the entirety of a singleton is 1, formally:

$$resemb(W) = 0, \quad \forall W \subseteq 2^{\Omega}, |W| \geq 3, \tag{9}$$

$$entire(\omega) = 1, \quad \forall \omega \in \Omega. \tag{10}$$

Inserting above equations into Equation (8), we have:

$$entire(X, Y) = \sum_{\omega \in X \cup Y} entire(\omega) - \sum_{\substack{\omega_m \in X \\ \omega_n \in Y \\ m \neq n}} resemb(\omega_m, \omega_n). \tag{11}$$

Hence, the similarity between two elements X and Y is calculated by:

$$sim(X_1, X_2) = \frac{\sum\limits_{\substack{\omega_m \in X_1 \\ \omega_n \in X_2 \\ m \neq n}} resemb(\omega_m, \omega_n)}{\sum\limits_{\omega \in X_1 \cup X_2} entire(\omega) - \sum\limits_{\substack{\omega_m \in X_1 \\ \omega_n \in X_2 \\ m \neq n}} resemb(\omega_m, \omega_n)}. \tag{12}$$

To guarantee Equation (9), the following constraint can be deduced:

$$\sum_{\substack{\omega_m, \omega_n \in \Omega \\ \omega_m \neq \omega_n}} sim(\omega_m, \omega_n) \leq 1. \tag{13}$$

3.4 Value setting of USP distance for evidence preference

Assume that similarities between categorical BBA representing preferences are:

$$\begin{aligned} d_{\Delta}(\omega^{\succ}, \omega^{\approx}) = d_{\Delta}(\omega^{\prec}, \omega^{\approx}) \quad &= x; \\ d_{\Delta}(\omega^{\succ}, \omega^{\prec}) &= 1. \end{aligned} \tag{14}$$

Assume $resemb(\omega^{\succ}, \omega^{\approx}) = p$, from Equation (12), we get:

$$p = \frac{2x}{1+x} \tag{15}$$

In this work, we take the extreme value as in [19], shown in Table 1, with which the similarity matrix Sim over 2^{Ω} can be obtained by Equation (12).

Table 1: Similarity between singletons

sim	$\omega^{\succ}$	$\omega^{\prec}$	$\omega^{\approx}$
$\omega^{\succ}$	1	0	1/3
$\omega^{\prec}$	0	1	1/3
$\omega^{\approx}$	1/3	1/3	1

For preference structures, by applying Equation (4), the USP distance degrades to Fagin's distance. Due to the space limitations, the proof will be provided in an extended version.

4 Experiments

In this paper, we show our first experiments on synthetic data generated by Algorithm 1. The implementation is realised by Python 3.7, based on iBelief package[4]. After calculation of pairwise distance over agents, a proximity measure applicable clustering method is used. In this paper, EkNNclus [2] is chosen as the clustering learner. Parameter selection in EkNNclus is not in the scope of this paper. In this paper, we directly set the number of clusters as in the data generation process.

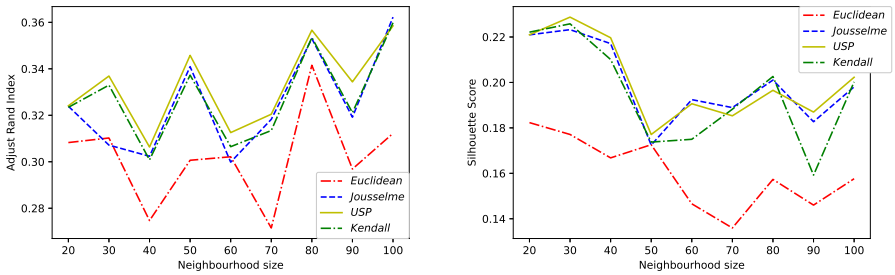
Algorithm 1 Generate conflicting preference sources in $|C|$ clusters

Require: Cluster number $|C|$
Switch time T
neighbour size N
Alternative size in each order $|A|$
Ensure: $|C|$ clusters of preferences
1: Initialise $|C|$ preference structures as centroids
2: **for** each centroid σ_c **do**
3: **for** n in $1:N$ **do**
4: **for** t in $1:T$ **do**
5: randomly generate index i, j;
6: exchange ranking order of a_i, a_j in σ_c to making a new order;
7: **end for**
8: **end for**
9: **end for**

[4] https://github.com/jusdesoja/iBelief_python

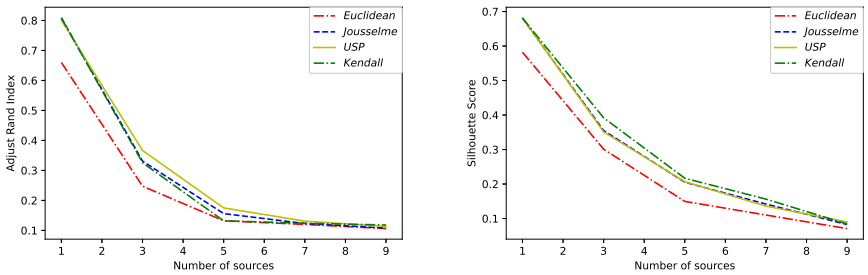
Confronting multiple preference sources, several methods are respectively compared with the average of Euclidean distance and Fagin (Kendall) distance. Clustering results are evaluated by Adjusted Rand Index and Silhouette score, depicted in Figure 2. To avoid random errors, the average value of 20 times experiments is calculated.

Two sets of experiments are conducted to demonstrate the effectiveness of USP distance in preference clustering. The first one is done with two conflicting sources, while the neighborhood size of preferences over 10 items increment, depicted in Figure 2. The second one is done with 8 clusters of preferences, with number of sources varying from 1 to 10 with step 2, depicted in Figure 3.



(a) Adjusted Rand Index. (b) Silhouette Score.

Fig. 2: Clustering results with different neighbourhood size



(a) Adjusted Rand Index. (b) Silhouette Score.

Fig. 3: Clustering results with different number of sources

It can be easily observed from Figure 2 that USP distance outperforms other distances, especially in terms of ARI. The advantages of USP distance are obtained by moderating the dissimilarity between $\prec, \succ$ and $\prec, \approx$, which respects better the natural definition of the preference relations. From Figure 3, the result is consistent with Experiment 1, that USP distance out performs other in ARI while worse in sihouette score. Moreover, with one source, experiments with

Jousslem distance, USP distance and Kendall distance return identical clustering results. This proves the assertation that Jousselme distance and USP distance degrade to Kendall distance confronting conventional preferences in total orders. We also observe that both ARI and silhouette score dramatically decrease with the number of conflicting sources augmenting. This is due to the fact that the alternative space is small (with only 10 alternatives), therefore one pair of conflicting preference already takes a big portion in all preference structure. In deed, with 10 sources of conflicting sources, two agents often become identical after the combination step. The results with 10 sources are similar in all distances, because the data is barely separable at this stage.

5 Discussion and conclusion

This paper explores the usage of a previously proposed distance, Unequal Singleton Pair (USP) distance, into clustering applications over evidential preferences. A combination rule for multiple preference sources is also proposed by interpreting the conflicts as imprecision. By applying USP distance over evidential preferences, clustering results are improved in terms of ARI.

Compared with the simple average strategy, evidential reasoning with USP distance can moderate the conflict between different information sources. Unfortunately, this also causes some side effects on the clustering mission: The clustering results are improved while the clustering quality is jeopardized in terms of silhouette scores.

Despite that USP distance is empirically proven useful, its effectiveness over incomplete preference structure remains suspicious. In the evidential preference model, missing information is usually modeled by total ignorance, which is equivalent to complete imprecision. However, pieces of missing information are measured as identical by USP distance, making them easily clustered into one identical group. Such a phenomenon is ridiculously against logical facts. To correctly clustering incomplete data within BFT is in the scope of our future work.

References

1. Benferhat, S., Dubois, D., Prade, H.: Towards a possibilistic logic handling of preferences. Applied Intelligence **14**(3), 303–317 (2001)
2. Denœux, T., Kanjanatarakul, O., Sriboonchitta, S.: Ek-nnclus: a clustering procedure based on the evidential k-nearest neighbor rule. Knowledge-Based Systems **88**, 57–69 (2015)
3. 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 (2004)
4. Fürnkranz, J., Hüllermeier, E.: Preference learning and ranking by pairwise comparison. In: Preference learning, pp. 65–82. Springer (2010)
5. Jabeur, K., Martel, J.M., Khélifa, S.B.: A distance-based collective preorder integrating the relative importance of the group's members. Group Decision and Negotiation **13**(4), 327–349 (2004)

6. Kamis, N.H., Chiclana, F., Levesley, J.: Preference similarity network structural equivalence clustering based consensus group decision making model. Applied Soft Computing **67**, 706–720 (2018)
7. Kamishima, T., Akaho, S.: Efficient clustering for orders. In: Mining complex data, pp. 261–279. Springer (2009)
8. Lu, T., Boutilier, C.: Vote elicitation with probabilistic preference models: Empirical estimation and cost tradeoffs. In: International Conference on Algorithmic DecisionTheory. pp. 135–149. Springer (2011)
9. Luce, R.D.: Individual choice behavior: A theoretical analysis. Courier Corporation (2012)
10. Masson, M.H., Destercke, S., Denœux, T.: Modelling and predicting partial orders from pairwise belief functions. Soft Computing **20**(3), 939–950 (2016)
11. Orlovsky, S.: Decision-making with a fuzzy preference relation. Fuzzy sets and systems **1**(3), 155–167 (1978)
12. Öztürke, M., Tsoukiàs, A., Vincke, P.: Preference Modelling, pp. 27–59. Springer New York, New York, NY (2005)
13. Roy, B., Slowinski, R.: Criterion of distance between technical programming and socio-economic priority. RAIRO-Operations Research **27**(1), 45–60 (1993)
14. Tasgin, M., Bingol, H.O.: Community detection using preference networks. Physica A: Statistical Mechanics and Its Applications **495**, 126–136 (2018)
15. Viappiani, P.: Characterization of scoring rules with distances: application to the clustering of rankings. In: The Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015). pp. 104–110 (2015)
16. Wang, Y., Zhou, J.t., Li, X., Song, X.: Effective user preference clustering in web service applications. The Computer Journal **63**(11), 1633–1643 (2020)
17. Yang, Y., Hooshyar, D., Jo, J., Lim, H.: A group preference-based item similarity model: comparison of clustering techniques in ambient and context-aware recommender systems. Journal of Ambient Intelligence and Humanized Computing **11**(4), 1441–1449 (2020)
18. Zhang, Y., Bouadi, T., Martin, A.: Preference fusion and Condorcet's paradox under uncertainty. In: 20th International Conference on Information Fusion (2017)
19. Zhang, Y., Bouadi, T., Wang, Y., Martin, A.: A distance for evidential preferences with application to group decision making. Information Sciences **568**, 113–132 (2021)