Cluster Decomposition of the Body of Evidence

Alexander Lepskiy

National Research University – Higher School of Economics,
Moscow, Russia

BELIEF 2022, October 26 - 28, 2022, Paris, France

Research Motivation

The body of evidence can be complex. For example, it may consist of many focal elements with a complex intersection structure.
We have the following problems:

- complex belief structures are difficult to interpret;
- high computational complexity of performing operations on complex belief structures.

Therefore, the following problems are relevant:

- analysis of the structure of the set of focal elements $\mathcal{A}$ of the body of evidence $F = (\mathcal{A}, m)$;
- finding an enlarged (simplified) structure of the set of focal elements $\widetilde{\mathcal{A}}$;
- redistribution of masses of focal elements of the set $\mathcal{A}$ to focal elements from $\widetilde{\mathcal{A}}$. As a result, we obtain a new mass function $\widetilde{m}$, etc.

Research Motivation

We proposes to solve these problems based on the **clustering** of a set of focal elements. We suggest that the complex structure of the body of evidence may be the result of aggregation of contradictory (conflict) information.

Therefore, the general approach to clustering the body of evidence can be as follows. **The inconsistency should be minimal within clusters and maximal between clusters in the resulting partition into clusters.**

This approach is similar to the **compactness principle** in cluster analysis. Distances should be minimal between elements of the same cluster and maximal between clusters.

Outline of Presentation

- Background of the Belief Function Theory;
- Restriction and Extension of the Mass Function;
- Statements of the Clustering Problem Based on Conflict Optimization;
- Clustering Based on the Conflict Density;
- The k-means Algorithm for the Body of Evidence;
- Evaluation of the Internal Conflict Based on Clustering;
- Conclusions.

Background of the Belief Function Theory

Let

- X be some set;
- $\mathcal{A} \subseteq 2^X$ be some finite subset of focal elements;
- $m : 2^X \to [0,1]$, $\sum_{A\in\mathcal{A}} m(A) = 1$ be some mass function, $m(A) > 0 \ \forall \ A \in \mathcal{A}$;
- the pair $F = (\mathcal{A}, m)$ is called a body of evidence;
- categorical evidence $F_A = (A, 1)$;
- if $F = (\mathcal{A}, m)$, then $F = \sum_{A\in\mathcal{A}} m(A)F_A$;
- in particular, simple evidence $F_A^\alpha = \alpha F_A + (1-\alpha)F_X$, $\alpha \in [0,1]$;
- the measure of external conflict of two bodies of evidence $F_1 = (\mathcal{A}_1, m_1)$ and $F_2 = (\mathcal{A}_2, m_2)$:

$$Con(F_1, F_2) = \sum_{A\cap B=\emptyset} m_1(A)m_2(B).$$

Restriction and Extension of the Mass Function

Let $F = (\mathcal{A}, m)$ be the body of evidence and $\mathcal{A}' \subseteq \mathcal{A}$.

The set function $m' : 2^X \to [0, 1]$, $m'(A) = m(A)\ \forall\ A \in \mathcal{A}'$ and $m'(A) = 0\ \forall\ A \notin \mathcal{A}'$ is called the **restriction of the mass function** m to the set $\mathcal{A}' \subseteq \mathcal{A}$.

In the general case, the mass function m' does not satisfy the condition $\sum_A m'(A) = 1$. It is necessary to extend m' to the mass function $\tilde{m}'$ so that this extension reflects the distribution of the m'.

Examples of some **extensions**:

- proportional extension: $\tilde{m}'(A) = m'(A) / \sum_{B \in \mathcal{A}'} m'(B)\ \forall\ A \in \mathcal{A}'$.
- vacuous extension: $\tilde{m}'(A) = m'(A)\ \forall A \neq X$, $\tilde{m}'(X) = m'(X) + 1 - \sum_{B \in \mathcal{A}'} m'(B)$.

If a certain rule for the extension of the mass function is fixed, then the new body of evidence $F' = (\mathcal{A}', \tilde{m}')$ will be uniquely determined by the original body of evidence $F = (\mathcal{A}, m)$ and the set $\mathcal{A}' \subseteq \mathcal{A}$. Therefore, such a body of evidence will be denoted as $F(\mathcal{A}') = (\mathcal{A}', \tilde{m}')$.

In particular, if the vacuous extension is used, then the body of evidence $F(\{A\}) = F_A^{m(A)} = m(A)F_A + (1 - m(A))F_X$ will be simple $\forall A \in \mathcal{A}$.

Statements of the Clustering Problem Based on Conflict Optimization

Suppose we have a body of evidence $F = (\mathcal{A}, m)$. It is required to find a partition of the set $\mathcal{A}$ into subsets $\{\mathcal{A}_1, \ldots, \mathcal{A}_l\}$ such that:

- maximize external conflict between bodies of evidence (clusters):
$$Con(F(\mathcal{A}_1), \ldots, F(\mathcal{A}_l)) \to \max;$$
- minimize total internal conflict within clusters
$$\sum_{i=1}^{l} Con_{in}(F(\mathcal{A}_i)) \to \min;$$
- minimize the overall conflict between the centers of clusters C_i and the bodies of evidence formed by the focal elements of these clusters
$$\sum_{i=1}^{l} \sum_{B \in \mathcal{A}_i} Con\,(F(\{B\}), C_i) \to \min\,.$$

In a more general setting, it is required to find a covering of the set of focal elements $\mathcal{A}$ instead of a partition.

Clustering Based on the Conflict Density

A mapping $\psi_F : 2^X \to [0,1]$ is called a **conflict density function** of the body of evidence $F = (\mathcal{A}, m)$ if it satisfies the following conditions:

1. $\psi_F(A) = 0$, if $B \cap A \neq \emptyset \ \forall B \in \mathcal{A}$;
2. $\psi_F(A) = 1$, if $B \cap A = \emptyset \ \forall B \in \mathcal{A}$;
3. $\psi_{\alpha F_1 + \beta F_2} = \alpha \psi_{F_1} + \beta \psi_{F_2} \ \forall F_1, F_2 \in \mathcal{F}(X)$, where $\alpha + \beta = 1$, $\alpha \geq 0$, $\beta \geq 0$.

It is easy to show that a set function satisfying conditions 1)-3) is equal to $\psi_F(A) = \sum_{B: A \cap B = \emptyset} m(B) = 1 - Pl(A)$.

The main idea of the clustering algorithm based on the conflict density function is that the 'centers' of the clusters should have a large value of the conflict density function.

We will use the function $\varphi_F(A) = m(A)\psi_F(A)$, $A \in \mathcal{A}$, which will take on large values for those focal elements that have not only a high density, but also a large mass.

Algorithm for Clustering Based on Conflict Density

1. Let's calculate the values of the set function $\varphi_F(A)$, $A \in \mathcal{A}$. If we have $\varphi_F(A) = 0$ for all $A \in \mathcal{A}$, then we stop the algorithm.

2. If there are $A \in \mathcal{A}$ for which $\varphi_F(A) > 0$, then we arrange such focal elements in descending order of function values φ_F:
$\varphi_F(A_1) \geq \varphi_F(A_2) \geq \ldots$. We choose the number of clusters l by analyzing the rate of decrease of the sequence $\{\varphi_F(A_i)\}$. Selected focal elements will be initial clusters: $\mathcal{A}_i^{(0)} = \{A_i\}$, $i = 1, \ldots, l$.

3. The remaining focal elements are redistributed among clusters $\mathcal{A}_1^{(0)}, \ldots, \mathcal{A}_l^{(0)}$ according to the **principle of maximizing the conflict between evidence clusters**. We will assign a focal element $B \in \mathcal{A} \backslash \left\{ \mathcal{A}_1^{(0)}, \ldots, \mathcal{A}_l^{(0)} \right\}$ to the cluster $\mathcal{A}_i^{(0)}$ for which the maximum conflict measure is reached:

$$\mathcal{A}_i^{(0)} = \underset{j: B \in \mathcal{A}_j^{(0)}}{\arg\max}\, Con\left(F\left(\mathcal{A}_1^{(0)}\right), \ldots, F\left(\mathcal{A}_j^{(0)} \cup \{B\}\right), \ldots, F\left(\mathcal{A}_l^{(0)}\right)\right).$$

If equal maximum values of the conflict are obtained by assigning B to several clusters $\mathcal{A}_j^{(0)}$, $j \in J$, then we include B in all these clusters, and the mass value $m(B)$ is evenly distributed among the updated clusters. In this case, B will be included in each cluster with weight $m(B)/|J|$. As a result, we obtain a coverage $\{\mathcal{A}_1, \ldots, \mathcal{A}_l\}$ of the set of all focal elements $\mathcal{A}$. The values of the mass function m_i on the $\mathcal{A}_i$, $i = 1, \ldots, l$ are calculated using the given restriction and extension procedures.

Example

Let we have $X = \{1, 2, 3\}$ and the body of evidence

$$F = 0.3F_{\{1\}} + 0.2F_{\{2\}} + 0.3F_{\{1,3\}} + 0.2F_{\{2,3\}}$$

is given on X, i. e. $\mathcal{A} = \{\{1\}, \{2\}, \{1, 3\}, \{2, 3\}\}$.

Step 1. Find the values of the function φ_F:
$\varphi_F(\{1\}) = \varphi_F(\{2\}) = 0.12$, $\varphi_F(\{1, 3\}, \{2, 3\}) = 0.06$.

Step 2. Let us assign the number of clusters $l = 2$ and $\mathcal{A}_1^{(0)} = \{\{1\}\}$, $\mathcal{A}_2^{(0)} = \{\{2\}\}$.

Step 3. Let's distribute the remaining two focal elements among clusters. As a result, we get $\{1, 3\} \in \mathcal{A}_1$, $\{2, 3\} \in \mathcal{A}_2$. We get the final clustering $\mathcal{A}_1 = \{\{1\}, \{1, 3\}\}$, $\mathcal{A}_2 = \{\{2\}, \{2, 3\}\}$.

The Evidence k-means Algorithm

It is required to find such a coverage of the set of all focal elements $\mathcal{A}$ by subsets (clusters) $\mathcal{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_l\}$ that would **minimize intracluster conflict**.

We will use the concept of center of a set (cluster) of focal elements by analogy with the classical k-means algorithm.

By the center of the i-th cluster $\mathcal{A}_i$, we mean some body of evidence C_i constructed from the pair $(\mathcal{A}_i, m_i)$, where m_i is the restriction of the mass function to $\mathcal{A}_i \subseteq \mathcal{A}$, $i = 1, \ldots, l$.

We will consider the total conflict between the centers of clusters and the bodies of evidence generated by the focal elements of these clusters as a minimized functional by analogy with the k-means algorithm:

$$\Phi = \sum_{i=1}^{l} \sum_{B \in \mathcal{A}_i} Con\,(F(\{B\}), C_i), \tag{1}$$

where $F\left(\{B\}\right) = m(B)F_B + (1 - m(B))F_X$.

Let us assume that the covering $\mathcal{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_l\}$ is fixed and the center of the i-th cluster has the form

$$C_i = \sum_{A \in \mathcal{A}_i} \alpha_i(A) F_A, \tag{2}$$

where $\alpha_i = (\alpha_i(A))_{A \in \mathcal{A}_i} \in S_{|\mathcal{A}_i|}$,
$S_k = \left\{ (t_1, \ldots, t_k) : t_i \geq 0, i = 1, \ldots, k, \sum_{i=1}^k t_i = 1 \right\}$ is the k-dimensional simplex. Then we have for the vacuous extension

$$\Phi = \sum_{i=1}^{l} \sum_{B \in \mathcal{A}_i} Con\left(F(\{B\}), C_i\right) = k_{\mathcal{C}} - \sum_{i=1}^{l} Q_i(\alpha_i),$$

where $k_{\mathcal{C}} = \sum_{i=1}^l \sum_{B \in \mathcal{A}_i} m(B) \geq 1$, $Q_i(\alpha_i) = \sum_{A \in \mathcal{A}_i} \alpha_i(A) Pl_{\mathcal{A}_i}(A)$ and $Pl_{\mathcal{A}_i}(A) = \sum_{\substack{B \in \mathcal{A}_i: \\ A \cap B \neq \emptyset}} m(B)$ is the restriction of the plausibility function to the set $\mathcal{A}_i$.

The minimum of the functional Φ for a fixed coverage $\mathcal{C} = \{\mathcal{A}_1, \dots, \mathcal{A}_l\}$ will be equal

$$\min \Phi = k_{\mathcal{C}} - \sum_{i=1}^{l} \max_{A \in \mathcal{A}_i} Pl_{\mathcal{A}_i}(A) \tag{3}$$

and it achieved for cluster centers

$$C_i = \sum_{A \in \overline{\mathcal{A}_i}} \alpha_i(A) F_A, \quad \alpha_i = (\alpha_i(A))_{A \in \overline{\mathcal{A}}_i} \in S_{|\overline{\mathcal{A}_i}|}, \quad i = 1, \dots, l, \tag{4}$$

where $\overline{\mathcal{A}_i} = \left\{ A \in \mathcal{A}_i : A = \arg\max_{A \in \mathcal{A}_i} Pl_{\mathcal{A}_i}(A) \right\}$.
This minimum does not depend on the choice of parameters
$\alpha_i = (\alpha_i(A))_{A \in \overline{\mathcal{A}}_i} \in S_{|\overline{\mathcal{A}_i}|}$, $i = 1, \dots, l$.

The Algorithm

1. Let's choose the number of clusters l. Let's assign some evidence bodies as initial cluster centers $C_i^{(0)}$, $i = 1, \ldots, l$. We fix the threshold of maximum conflict within clusters $Con_{\max} \in [0, 1]$. Put $s = 0$.

2. We redistribute focal elements among clusters according to the principle of minimizing the conflict between evidence clusters and cluster centers. The focal element $B \in \mathcal{A}$ is assigned to the cluster

$$\mathcal{A}_i^{(s)} = \arg\min_j Con\left(F(\{B\}), C_j^{(s)}\right)$$

and $\min_i Con\Big(F(\{B\}), C_i^{(s)}\Big) \leq Con_{\max}$. If $\min_i Con\Big(F(\{B\}), C_i^{(s)}\Big) > Con_{\max}$, then the focal element B is assigned as the center of the new cluster. We get clusters $\mathcal{A}_i^{(s)}$, $i = 1, \ldots, l$.

3. Let us calculate new cluster centers using the formula (4). We increase the counter $s \leftarrow s + 1$.

4. Steps 2 and 3 are repeated until the clusters (or their centers) stabilize.

Proposition.

Algorithm converges in a finite number of steps.

Cluster centers may depend on parameters $\alpha = (\alpha(A))_{A\in\overline{\mathcal{A}_i}} \in S_{|\overline{\mathcal{A}_i}|}$. In this case, it is necessary to use additional procedures for choosing parameters. The selection criteria can be considered, for example:

1. coverage minimization, i. e., we choose the parameters so that the coverage $\mathcal{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_l\}$ is 'closer' to the partition. For example, $\sum_{i=1}^{l} |\mathcal{A}_i| \to \min$.
2. minimizing the uncertainty of evidence-centers of clusters C_i, $i = 1, \ldots, l$. For example, this can be done using the generalized Hartley measure $H(C_i) = \sum_{A\in\overline{\mathcal{A}_i}} \alpha_i(A) \ln |A| \to \min$.
3. minimizing the distance between cluster centers and the original evidence body: $d(C_i, F) \to \min$, $i = 1, \ldots, l$;
4. maximizing distance between cluster centers $d(C_i, C_j) \to \max$ or maximizing conflict $Con(C_i, C_j) \to \max$, $i, j = 1, \ldots, l$ $(i \neq j)$ etc.

Example

Let's apply this algorithm for clustering into two clusters the body of evidence $F = 0.3F_{\{1\}} + 0.2F_{\{2\}} + 0.3F_{\{1,3\}} + 0.2F_{\{2,3\}}$ on $X = \{1, 2, 3\}$.

Step 1. Let $l = 2$, the initial centers of the clusters be equal to $C_1^{(0)} = F_{\{1\}}$, $C_2^{(0)} = F_{\{2\}}$; $s = 0$.

Step 2. We have the following conflict values

$Con\left(F(\{B\}), C_i^{(0)}\right)$	$\{1\}$	$\{2\}$	$\{1,3\}$	$\{2,3\}$
$C_1^{(0)}$	0	0.2	0	0.2
$C_2^{(0)}$	0.3	0	0.3	0

According to the principle of minimizing the conflict between evidence clusters and cluster centers, the initial clustering will have the form $\mathcal{A}_1^{(0)} = \{\{1\}, \{1,3\}\}$, $\mathcal{A}_2^{(0)} = \{\{2\}, \{2,3\}\}$.

Step 3. It follows from the values of the restrictions of the plausibility function to clusters $Pl_{\mathcal{A}_1^{(0)}}(\{1\}) = \mathbf{0.6}$, $Pl_{\mathcal{A}_1^{(0)}}(\{1,3\}) = \mathbf{0.6}$, $Pl_{\mathcal{A}_2^{(0)}}(\{2\}) = 0.4$, $Pl_{\mathcal{A}_2^{(0)}}(\{2,3\}) = 0.4$ that the new clustering centers are equal to

$$C_1^{(1)} = \alpha F_{\{1\}} + (1-\alpha) F_{\{1,3\}}, \quad C_2^{(1)} = \beta F_{\{2\}} + (1-\beta) F_{\{2,3\}}, \ \alpha, \beta \in [0,1].$$

Step 4. We redistribute focal elements according to the criterion of least conflict with new centers:

$Con\Big(F(\{B\}), C_i^{(1)}\Big)$	$\{1\}$	$\{2\}$	$\{1,3\}$	$\{2,3\}$
$C_1^{(1)}$	0	0.2	0	0.2α
$C_2^{(1)}$	0.3	0	0.3β	0

We will get clusters $\mathcal{A}_1^{(1)} = \{\{1\},\{1,3\}\}$, $\mathcal{A}_2^{(1)} = \{\{2\},\{2,3\}\}$ after applying the coverage minimization rule. Stop of the algorithm.

As a result, we get a new body of evidence defined on the base set $\mathcal{A}$, the found coverage sets (clusters) $\mathcal{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_l\}$ will be focal elements, the mass function will be equal to

$$m(\mathcal{A}_i) = \sum_{B \in \mathcal{A}_i} m(B)/n(B),$$

where $n(B) = |\{\mathcal{A}_i : B \in \mathcal{A}_i\}|$ (the number of clusters containing the set B).
Such a body of evidence can be considered **second-order evidence**, which reflects the enlarged structure of the original evidence.

Evaluation of the Internal Conflict Based on Clustering

Let us assume that a cluster coverage $\mathcal{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_l\}$ of the body of evidence $F = (\mathcal{A}, m)$ is obtained. Then we can offer the following ways to evaluate the internal conflict of this body of evidence using some measure of external conflict Con:

1. $Con_1(F) = Con(F(\mathcal{A}_1), \ldots, F(\mathcal{A}_l))$;
2. $Con_2(F) = Con(C_1, \ldots, C_l)$, where $C_1, \ldots, C_l$ are centers of clusters $\mathcal{A}_1, \ldots, \mathcal{A}_l$ respectively.

For example, we will obtain the following estimates of the internal conflict for the body of evidence considered in Example. We have $\mathcal{A}_1 = \{\{1\}, \{1,3\}\}$, $\mathcal{A}_2 = \{\{2\}, \{2,3\}\}$ and
$Con_1(F) = Con(F(\mathcal{A}_1), F(\mathcal{A}_2)) = 0.18$,
$Con_2(F) = Con(C_1, C_2) = \alpha + (1-\alpha)\beta$, $\alpha, \beta \in [0,1]$.

Summary and Conclusion

- Two methods of evidence body clustering are discussed. Each of these methods assumes that weakly conflicting focal elements should belong to one cluster, and strongly conflicting focal elements should belong to different clusters;
- The first algorithm is based on the use of the distribution density function of conflicting focal elements;
- The second algorithm implements the idea of the k-means method. In this case, the cluster centers are formed in some optimal way. Further, focal elements are redistributed according to the principle of minimizing conflict with cluster centers;
- It shows how clustering can be used to evaluate the internal conflict of a body of evidence.

References

- Bronevich, A., Lepskiy, A.: Measures of conflict, basic axioms and their application to the clusterization of a body of evidence. Fuzzy Sets and Systems 446, 812–832 (2022)
- Denœux, T.: Inner and outer approximation of belief structures using a hierarchical clustering approach. Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems 9(4), 437–460 (2001)
- Harmanec, D.: Faithful approximations of belief functions. In: Laskey, K. B. and Prade, H., (eds) Uncertainty in Artificial Intelligence 15 (UAI99), Stockholm, Sweden, (1999)
- Lepskiy, A.: Analysis of Information Inconsistency in Belief Function Theory. Part I: External Conflict. Control Sciences 5, 2–16 (2021)
- Lepskiy, A.: Analysis of Information Inconsistency in Belief Function Theory. Part II: Internal Conflict. Control Sciences 6, 2–12 (2021)
- Petit-Renaud, S., Denœux, T.: Handling different forms of uncertainty in regression analysis: a fuzzy belief structure approach. In: Hunter, A. and Pearsons, S., (eds) Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU'99), pp. 340–351. Springer Verlag (1999)

Thanks for you attention

alex.lepskiy@gmail.com

alepskiy@hse.ru

https://www.hse.ru/en/org/persons/10586209