Learning contextual discounting and contextual reinforcement from labelled data

David Mercier, Frédéric Pichon, Éric Lefèvre, and François Delmotte

Univ. Lille Nord de France, F-59000 Lille, France
UArtois, LGI2A, F-62400, Béthune, France
`{firstname.lastname@univ-artois.fr}`

Abstract. This paper addresses the problems of learning from labelled data contextual discounting and contextual reinforcement, two correction schemes recently introduced in belief function theory. It shows that given a particular error criterion based on the plausibility function, for each of these two contextual correction schemes, there exists an optimal set of contexts that ensures the minimization of the criterion and that finding this minimum amounts to solving a constrained least-squares problem with as many unknowns as the domain size of the variable of interest.

Keywords: Belief functions, Learning, Contextual Discounting, Contextual Reinforcement, Source Biases, Expert Tuning.

1 Introduction

Classically, in belief function theory, the correction of the information provided by a source concerning a variable of interest $\mathbf{x}$ defined on a finite domain $\mathcal{X}$, is achieved using the discounting operation [9–11]. This operation admits one parameter, a real β belonging to $[0, 1]$, reflecting the degree of *reliability* of the source of information [10, Section 5.7] [6, Section 2.5].

Discounting operation has been extended by Mercier *et al.* in [6], where it is considered that one may have some knowledge about the reliability of a source, conditionally on different subsets (*contexts*) of $\mathcal{X}$, the set of contexts forming a partition of $\mathcal{X}$. This operation, called *contextual discounting based on a coarsening*, is controlled by a vector of parameters β_A, each β_A belonging to $[0, 1]$ and reflecting the degree of reliability of the source given context $A \subseteq \mathcal{X}$ (in other words, knowing that the true value of $\mathbf{x}$ lies in A). In this same article [6, Section 5], following preceding work from Elouedi *et al.* for the classical discounting [4], a computationally efficient method to automatically learn from labelled data the parameters β_A of a contextual discounting based on a coarsening, once a partition (a set of contexts) has been fixed, is also introduced; the idea is to find the parameters values which minimize a measure of discrepancy between the ground truth and the outputs of the source corrected according to the parameters values. This method is potentially useful to improve a source performance in, *e.g.*, a classification application, as well as to discover its contextual reliability. However, the problem of finding the optimal partition of $\mathcal{X}$ for a given source was left open.

In [7], Mercier *et al.* have extended this contextual discounting based on a coarsening to be applicable to any set of contexts (*i.e.*, the set of contexts no longer needs to form a partition of $\mathcal{X}$). This mechanism is therefore simply called *contextual discounting (CD)*. The *contextual reinforcement (CR)* of a source is also introduced as the dual of CD. This new correction operation is also controlled by a vector of parameters β_A in $[0, 1]$ associated with a set of subsets (contexts) A of $\mathcal{X}$. However, the interpretation of CR was not clear (it was only known that CR amounts to the negation [3] of the CD of the negation of the information provided by the source) and the problem of learning CD and CR from labelled data was not tackled. Recently, in [8], Pichon *et al.* gave an interpretation to CR: it amounts to assuming, for each context A, that the source is truthful with mass β_A, and that with mass $(1 - \beta_A)$ it lies only when it tells that the true value of $\mathbf{x}$ is in $\overline{A}$; but the question of learning CD and CR from data remained open.

In this paper, we address the problems of learning from labelled data contextual discounting and contextual reinforcement, the former problem being only partially addressed so far – its solution is restricted to the case where a set of contexts has been fixed beforehand and where this set must also form a partition of $\mathcal{X}$ – and the latter problem not being addressed at all. Especially, we show that given the discrepancy measure used in [6], there exists an optimal set of contexts for the most general form of CD (the CD proposed in [7] and that does not require the set of contexts to form a partition of $\mathcal{X}$) that ensures the minimization of the measure, and that finding this minimum amounts to a computationally simple optimization problem (a constrained least-squares problem with K unknowns, K being the size of $\mathcal{X}$). Furthermore, we show that a similar result holds for CR. In addition, an illustrative example of the proposed learning of CD and CR is given. This example is also useful to make insightful additional remarks on CD and CR, and in particular the potential superiority of the recently introduced CR mechanism over CD, to improve a source performance.

This paper is organized as follows. Required basic concepts on belief functions and contextual correction mechanisms are exposed in Section 2. Learning of CD and of CR is formally studied in Section 3 and illustrated in Section 4, where a comparison of CD and CR correction capacities is also presented. Finally, Section 5 concludes the paper.

2   Belief functions and contextual correction mechanisms: basic concepts and notations

2.1   Representation and combination of beliefs

The necessary background material on the representation and combination of beliefs is given here.

Representation of beliefs A *mass function (MF)* represents an agent's opinion regarding a variable of interest $\mathbf{x}$ taking values in a finite domain $\mathcal{X}$ [9, 11].

It is defined as a mapping $m : 2^{\mathcal{X}} \to [0,1]$ verifying $\sum_{A \subseteq \mathcal{X}} m(A) = 1$. The negation $\overline{m}$ of a mass function m is defined as $\overline{m}(A) = m(\overline{A})$, $\forall A \subseteq \mathcal{X}$.

A MF m defined by $m(\mathcal{X}) = w$ and $m(A) = 1 - w$, with $w \in [0,1]$ and $A \subset \mathcal{X}$, can be conveniently noted A^w. Likewise a MF m such that $m(\emptyset) = v$ and $m(A) = 1 - v$, with $v \in [0,1]$, $A \subseteq \mathcal{X}$, $A \neq \emptyset$, can be conveniently noted A_v.

A MF m is in one-to-one correspondence with a plausibility function pl, a commonality function q and an implicability function b, which are respectively defined by: $pl(A) = \sum_{B \cap A \neq \emptyset} m(B)$, $q(A) = \sum_{A \subseteq B} m(B)$ and $b(A) = \sum_{B \subseteq A} m(B)$, for all $A \subseteq \mathcal{X}$.

Combination Two MF m_1 and m_2 can be combined using the *conjunctive rule of combination* [11] denoted by $\owedge$ and defined by:

$$(m_1 \owedge m_2)(A) = \sum_{B \cap C = A} m_1(B) \cdot m_2(C), \quad \forall A \subseteq \mathcal{X}. \tag{1}$$

Numerous combination rules exist [12] to merge mass functions. The other combination of particular interest in this paper is the *disjunctive rule of combination* $\ovee$ [3, 10] defined by replacing the symbol $\cap$ in (1) by $\cup$.

Let us also recall that if $m = m_1 \owedge m_2$, the corresponding commonality functions verify $q = q_1 \cdot q_2$, and if $m = m_1 \ovee m_2$, the corresponding implicability functions verify $b = b_1 \cdot b_2$.

2.2 Contextual discounting and reinforcement of a belief function

Throughout this paper, m_S is a MF defined on $\mathcal{X}$, provided by a source S, and $\mathcal{A}$ is a set of subsets (contexts) of $\mathcal{X}$.

Contextual discounting (CD) The contextual discounting [6–8] of m_S is the MF m defined, with $\beta_A \in [0,1]$, $\forall A \in \mathcal{A}$, by:

$$m = m_S \ovee_{A \in \mathcal{A}} A_{\beta_A} . \tag{2}$$

The classical discounting [9, 10] is retrieved when $\mathcal{A}$ is composed of just one element which is the whole domain $\mathcal{X}$:

$$m = m_S \ovee \mathcal{X}_\beta = \beta \, m_S + (1 - \beta) m_{\mathcal{X}} , \tag{3}$$

with $m_{\mathcal{X}}$ defined by $m_{\mathcal{X}}(\mathcal{X}) = 1$.

In practice, β_A represents the proportion of $m_S(B)$ which remains on B, and $(1 - \beta_A)$ represents the part of $m_S(B)$ transferred to $B \cup A$, $\forall A \in \mathcal{A}$ and $\forall B \subseteq \mathcal{X}$.

Contextual reinforcement (CR) The contextual reinforcement [7, 8] of m_S is the MF m defined, with $\beta_A \in [0,1]$, $\forall A \in \mathcal{A}$, by:

$$m = m_S \owedge_{A \in \mathcal{A}} A^{\beta_A} . \tag{4}$$

In practice, β_A represents the fraction of $m_S(B)$ remaining on B, and $(1 - \beta_A)$ the part of $m_S(B)$ transferred to $B \cap A$, $\forall A \in \mathcal{A}$ and $\forall B \subseteq \mathcal{X}$.

3   Learning CD and CR from labelled data

3.1   Description of the learning process

In this section, we study how to automatically learn CD and CR from:

1. A training set describing the outputs of a source (expressed in the form of a MF) regarding the classes in $\mathcal{X} = \{x_1, \ldots, x_K\}$ of n objects o_i, $i \in \{1, \ldots, n\}$ (A small illustrative example is given in Section 4 in Table 2);
2. And a measure of discrepancy to be minimized between the corrections of the mass functions provided by the source and the reality.

In this paper, the following measure of discrepancy between the corrected information and the ground truth has been chosen:

$$E_{pl}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} (pl\{o_i\}(\{x_k\}) - \delta_{i,k})^2 \ , \tag{5}$$

where $\forall i \in \{1, \ldots, n\}$, $pl\{o_i\}$ is the plausibility function obtained from a contextual correction of the output m_S of the source with a vector of coefficients $\boldsymbol{\beta} \in [0,1]^{|\mathcal{A}|}$. The binary variable $\delta_{i,k}$ indicates the class of o_i as follows: $\forall i \in \{1, \ldots, n\}$, $\forall k \in \{1, \ldots, K\}$, $\delta_{i,k} = 1$ if object o_i belongs to the class x_k, and $\delta_{i,k} = 0$ otherwise.

Our choice to use measure E_{pl} (5) is mostly based on the fact that, as it will be seen in Propositions 2 and 4, its minimization has the advantage to yield constrained least-squares problems, which can be solved efficiently. Moreover, it was the one used in the approach proposed in [6], which we are clearly extending with this current work. At last, using the plausibility on singletons is in accordance with the Shafer [9] and Smets [11] *singular* [2] interpretation of belief functions, which is adopted in this paper. However, we may note that other measures of discrepancy could be used, *e.g.*, a measure based on the pignistic probability [11] or on a distance measure [5], but then it is not guaranteed that their minimization can be performed efficiently.

3.2   Learning CD

Plausibilities on the singletons after having applied CD on a MF m_S provided by a source are given by next proposition.

Proposition 1. *Let* $m = m_S \textcircled{\cup}_{A \in \mathcal{A}} A_{\beta_A}$, $\beta_A \in [0,1]$, $\forall A \in \mathcal{A}$, *be the CD of a MF* m_S. *The plausibility function associated with* m *is defined for all* $x \in \mathcal{X}$ *by:*

$$pl(\{x\}) = 1 - (1 - pl_S(\{x\})) \prod_{A \in \mathcal{A}, x \in A} \beta_A \ . \tag{6}$$

Proof. See Appendix A.1

Next proposition indicates that the minimization of E_{pl} when CD has been applied, is obtained using the vector $\boldsymbol{\beta}$ composed of the K parameters $\beta_{\{x_k\}}$, which means the parameters associated with the singletons of $\mathcal{X}$. Moreover the minimization of E_{pl} using this vector constitutes a constrained least-squares problem which can then be solved efficiently using standard algorithms.

Proposition 2. *The minimization of E_{pl} with CD is obtained using the vector $\boldsymbol{\beta} = (\beta_{\{x_k\}}, k \in \{1, \ldots, K\})$ and constitutes a constrained least-squares problem as (5) can then be rewritten as:*

$$E_{pl}(\boldsymbol{\beta}) = \|\boldsymbol{Q}\boldsymbol{\beta} - \boldsymbol{d}\|^2 \text{ with } \boldsymbol{Q} = \begin{bmatrix} diag(\boldsymbol{pl_1} - 1) \\ \vdots \\ diag(\boldsymbol{pl_n} - 1) \end{bmatrix} \text{ and } \boldsymbol{d} = \begin{bmatrix} \boldsymbol{\delta_1} - 1 \\ \vdots \\ \boldsymbol{\delta_n} - 1 \end{bmatrix}, \quad (7)$$

with $diag(\boldsymbol{v})$ a square diagonal matrix with the elements of vector $\boldsymbol{v}$ on the main diagonal, and with $\boldsymbol{pl_i} = (pl_S\{o_i\}(\{x_1\}), \ldots, pl_S\{o_i\}(\{x_K\}))^T$, and $\boldsymbol{\delta_i} = (\delta_{i,1}, \ldots, \delta_{i,K})^T$ the column vector of 0-1 class indicator variables for object o_i.

Proof. See Appendix A.2

This answers a prospect given in [6] concerning the study of the set of contexts which yields the best possible value for the measure of discrepancy E_{pl}. The answer given here is that there will be no smaller value reachable for E_{pl} than the one obtained with the set of the singletons of $\mathcal{X}$ with associated coefficients $\boldsymbol{\beta} = (\beta_{\{x_k\}}, k \in \{1, \ldots, K\})$.

3.3   Learning CR

Plausibilities on the singletons after having applied CR are given in next proposition.

Proposition 3. *Let $m = m_S \bigcirc_{A \in \mathcal{A}} A^{\beta_A}$, $\beta_A \in [0, 1]$, $\forall A \in \mathcal{A}$, be the CR of a MF m_S. The plausibility function associated with m is defined for all $x \in \mathcal{X}$ by:*

$$pl(\{x\}) = pl_S(\{x\}) \prod_{A \in \mathcal{A}, x \notin A} \beta_A . \qquad (8)$$

Proof. See Appendix A.3

Proposition 4. *The minimization of E_{pl} with CR is obtained using the vector $\boldsymbol{\beta} = (\beta_{\overline{\{x_k\}}}, k \in \{1, \ldots, K\})$ and constitutes a constrained least-squares problem as (5) can then be written as:*

$$E_{pl}(\boldsymbol{\beta}) = \|\boldsymbol{P}\boldsymbol{\beta} - \boldsymbol{\delta}\|^2, \text{ with } \boldsymbol{P} = \begin{bmatrix} diag(\boldsymbol{pl_1}) \\ \vdots \\ diag(\boldsymbol{pl_n}) \end{bmatrix} \text{ and } \boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\delta_1} \\ \vdots \\ \boldsymbol{\delta_n} \end{bmatrix}, \qquad (9)$$

with the same notations as in Proposition 2.

Proof. See Appendix A.4

4 CD and CR learnings: comments and illustration

4.1 CD and CR respective correction capacities

The differences between CD and CR concerning their respective plausibilities ranges on singletons after having been applied are briefly discussed here.

With CD, as $pl(\{x_k\}) = 1 - (1 - pl_S(\{x_k\}))\beta_{\{x_k\}}$ for each $k \in \mathcal{X}$, $k \in \{1,\ldots,K\}$, with $\beta_{\{x_k\}}$ varying in $[0,1]$, $pl(\{x_k\})$ can take any values in the interval $[pl_S(\{x\}), 1]$. It means that with CD the value on each singleton $pl_S(\{x\})$ can be shifted as close to 1 as required, in other words weakened as required.

In contrast, with CR, as $pl(\{x_k\}) = pl_S(\{x_k\})\beta_{\overline{\{x_k\}}}$ for each $x_k \in \mathcal{X}$, $k \in \{1,\ldots,K\}$, $pl(\{x_k\})$ can take any values in $[0, pl_S(\{x_k\})]$ with $\beta_{\overline{\{x_k\}}}$ varying from 0 to 1. With CR, the value of the plausibility on each singleton can then be carried as close to 0 as necessary. In other words, CR strengthens the information provided by the source by decreasing the plausibilities on certain singletons.

The following example illustrates these different capacities of adjustment to the reality on simple scenarios for CD and CR.

Example 1. Let us suppose that $\mathcal{X} = \{a, b, c\}$ and, without lack of generality, that the ground truth is a.

Let us suppose that a source n°1 outputs a mass $m_S(\{b,c\}) = 1$ which means that $pl_S(\{a\}) = 0$ and $pl_S(\{b\}) = pl_S(\{c\}) = 1$. To bring closer source n°1 output and the reality: CD can increase $pl_S(\{a\})$ to 1; CR can decrease $pl_S(\{b\})$ to 0 and $pl_S(\{c\})$ to 0.

This example is taken again in Table 1, and two more situations are considered: a source n°2 giving $m_S(\{c\}) = 1$, that is $pl_S(\{a\}) = pl_S(\{b\}) = 0$ and $pl_S(\{c\}) = 1$ and a source n°3 giving $m_S(\{a,b\}) = 1$, that means $pl_S(\{a\}) = pl_S(\{b\}) = 1$ and $pl_S(\{c\}) = 0$.

Table 1. Attainable plausibilities with CD and CR for three sources outputs.

	Ground truth	Source n°1	CD	CR	Source n°2	CD	CR	Source n°3	CD	CR
pl({a})	1	0	1	0	0	1	0	1	1	1
pl({b})	0	1	1	0	0	0	0	1	1	0
pl({c})	0	1	1	0	1	1	0	0	0	0
		CD: $E_{pl} = 2$			CD: $E_{pl} = 1$			CD: $E_{pl} = 1$		
		CR: $E_{pl} = 1$			CR: $E_{pl} = 1$			CR: $E_{pl} = 0$		

As it can be observed in Table 1, CD can improve only one value of plausibility: the plausibility on the ground truth by increasing it as close as possible to 1, whereas CR can improve all the other plausibility values (all except the one associated with the ground truth) by decreasing them as near as possible to 0. CR has then more degrees of flexibility to improve the plausibility output of the source. Situations where CD can be of more help than CR, are in particular

those where all the plausibilities on singletons which are not the ground truth are equal to zero, for example: $pl_S(\{b\}) = pl_S(\{c\}) = 0$ and $pl_S(\{a\}) = 0.5$, which means $m_S(\{a\}) = 0.5$ and $m_S(\emptyset) = 0.5$.

4.2   An illustrative example

Inspired from [4] and [6, Section 5], we consider the following small example of target recognition illustrated in Table 2.

Example 2. Two sensors are in charge of recognizing flying objects which can be airplanes (a), helicopters (h) or rockets (r). Data are composed of 4 known objects on which two sensors have expressed their outputs as MF on $\mathcal{X} = \{a, h, r\}$.

Table 2. Ouputs of two sensors regarding the classes of 4 objects which can be airplanes (a), helicopters (h) or rockets (r). Data come from [4, Table 1].

		$\{a\}$	$\{h\}$	$\{r\}$	$\{a,h\}$	$\{a,r\}$	$\{h,r\}$	$\mathcal{X}$	Ground truth
Sensor 1	$m_{S_1}\{o_1\}$	0	0	0.5	0	0	0.3	0.2	a
	$m_{S_1}\{o_2\}$	0	0.5	0.2	0	0	0	0.3	h
	$m_{S_1}\{o_3\}$	0	0.4	0	0	0.6	0	0	a
	$m_{S_1}\{o_4\}$	0	0	0	0	0.6	0.4	0	r
Sensor 2	$m_{S_2}\{o_1\}$	0	0	0	0.7	0	0	0.3	a
	$m_{S_2}\{o_2\}$	0.3	0	0	0.4	0	0	0.3	h
	$m_{S_2}\{o_3\}$	0.2	0	0	0	0	0.6	0.2	a
	$m_{S_2}\{o_4\}$	0	0	0	0	0	1	0	r

Results of the minimization of E_{pl} for CD and CR are summarized in Table 3 for both sensors 1 and 2. Let us recall that $\boldsymbol{\beta} = (\beta_{\{a\}}, \beta_{\{h\}}, \beta_{\{r\}})$ for CD, and $\boldsymbol{\beta} = (\beta_{\overline{\{a\}}}, \beta_{\overline{\{b\}}}, \beta_{\overline{\{c\}}})$ for CR with different meanings for each correction mechanism.

Table 3. Results for the minimization of E_{pl} with the data in Table 2 for each contextual correction mechanism for both sensors 1 and 2.

Contextual correction	Sensor 1	Sensor 2
CD	$\boldsymbol{\beta} = (0.76, 1.00, 1.00)$ $E_{pl}(\boldsymbol{\beta}) = 3.39$	$\boldsymbol{\beta} = (0.74, 1.00, 1.00)$ $E_{pl}(\boldsymbol{\beta}) = 4.81$
CR	$\boldsymbol{\beta} = (0.94, 0.66, 0.38)$ $E_{pl}(\boldsymbol{\beta}) = 2.33$	$\boldsymbol{\beta} = (0.65, 0.22, 0.55)$ $E_{pl}(\boldsymbol{\beta}) = 2.39$

For CD it can be observed that $\beta_{\{h\}} = \beta_{\{r\}} = 1$ for both sensors, which means that both sensors are reliable to detect objects h and r. There is no need to transfer a portion of mass $m_S(B)$ to $B \cup \{h\}$ or $B \cup \{r\}$ with $B \subseteq \mathcal{X}$

$(1 - \beta_{\{h\}} = 1 - \beta_{\{r\}} = 0)$. It is not the case for objects of type a which cause problems for both sensors, sensor 1 being slightly more reliable.

Minimizing E_{pl} with CR confirms that both sensors are more truthful to recognize objects h and r as $\beta_{\overline{\{a\}}} = 0.94 > \beta_{\overline{\{h\}}} > \beta_{\overline{\{r\}}}$ for sensor 1 and $\beta_{\overline{\{a\}}} = 0.65$ for sensor 2. In terms of mass transfers, there is less need to transfer a portion of mass $m_S(B)$ to $B \cap \overline{\{a\}} = B \cap \{h, r\}$ $(1 - 0.94 = 0.06$ for sensor 1, and $1 - 0.65 = 0.35$ for sensor 2) than to $B \cap \overline{\{h\}} = B \cap \{a, r\}$ (0.34 for sensor 1, and 0.78 for sensor 2) or $B \cap \overline{\{r\}} = B \cap \{a, h\}$ (0.62 for sensor 1, and 0.45 for sensor 2). With these data, CR also permits to obtain lower values for E_{pl} than those reached with CD, which confirms the advantages of CR over CD exposed in Section 4.1 and in Example 1 concerning the minimization of E_{pl}.

4.3   On the absence of link between learning CR and CD

Even if CR and CD are related (CR amounts to the negation of the CD of the negation of the information provided by the source [7]), CR and CD parameters minimizing E_{pl} (5) cannot be deduced analytically from each other.

Let us consider next example which is a slight modification of example 2.

Example 3. By modifying in Table 2, MF $m_{S_1}\{o_1\}$ by $m_{S_1}\{o_1\}(\{r\}) = 0.5282$, $m_{S_1}\{o_1\}(\{h, r\}) = 0.3$ and $m_{S_1}\{o_1\}(\mathcal{X}) = 0.1718$ (information coming from Sensor 1 is slightly deteriorated, the truth being a), learning of CD parameters for sensors 1 and 2 yields the same vector $\boldsymbol{\beta} = (0.74, 1.00, 1.00)$, while the learning of CR parameters yields $\boldsymbol{\beta} = (0.92, 0.68, 0.38)$ for sensor 1 and $\boldsymbol{\beta} = (0.65, 0.22, 0.55)$ for sensor 2.

Example 3 shows that knowing the vector $\boldsymbol{\beta}$ minimizing E_{pl} for CD does not imply knowing the vector $\boldsymbol{\beta}$ minimizing E_{pl} for CR.

5   Conclusion

In this paper, we have studied the learning of CD and of CR from labelled data, given a measure of discrepancy based on the plausibility function. We have shown that for each of these two contextual correction schemes, there exists an optimal set of contexts that ensures the minimization of the measure and that finding this minimum amounts to solving a constrained least-squares problem with K unknowns. These results can find applications in at least two domains: learning the biases of a source of information (what are the characteristics of a source?) and in source tuning (how to tune a source to obtain the best performances?). Future work will consist in exploiting these mechanisms in more complex applications and investigating the tuning of the combination of several sources.

References

1. Denœux, T.: Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. Artif. Intell. 172, 234-264 (2008)

2. Destercke, S., Burger, T.: Toward an axiomatic definition of conflict between belief functions. IEEE Trans. Cybern. 43(2), 585–596 (2012)
3. Dubois, D., Prade, H.: A Set-Theoretic View of Belief Functions: logical operations and approximations by fuzzy sets. Int. J. Gen. Syst. 12, 193–226 (1986)
4. Elouedi, Z., Mellouli, K., Smets, P.: Assessing Sensor Reliability for Multisensor Data Fusion Within the Transferable Belief Model. IEEE Trans. Syst. Man Cybern. Part B. 34(1), 782-787 (2004)
5. Jousselme, A.-L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations, Int. J. Approx. Reason. 53(2), 118–145 (2012)
6. Mercier, D., Quost, B., Denœux, T.: Refined modeling of sensor reliability in the belief function framework using contextual discounting. Inf. Fusion. 9(2), 246-258 (2008)
7. Mercier, D., Lefèvre, É., Delmotte, F.: Belief functions contextual discounting and canonical decompositions. Int. J. Approx. Reason. 53(2), 146-158 (2012)
8. Pichon, F., Mercier, D., Lefèvre, É., Delmotte, F.: Truthfulness in contextual information correction. In: Cuzzolin, F. (Ed.) Belief Functions: Theory and Applications. Third International Conference, BELIEF 2014, Oxford, UK, September 26-28, 2014. Proceedings. LNCS, vol. 8764, pp. 11-20. Springer, Heidelberg (2014)
9. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton, N.J., (1976)
10. Smets, P.: Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. Int. J. Approx. Reason. 9, 1-35 (1993)
11. Smets, P., Kennes, R.: The Transferable Belief Model. Artif. Intell. 66, 191-243 (1994)
12. Smets, P.: Analyzing the combination of conflicting belief functions. Inf. Fusion 8, 387–412 (2007)

A   Appendices

A.1   Proof of Proposition 1

As $m = m_S \bigcup_{A \in \mathcal{A}} A_{\beta_A}$, CD is given in terms of implicability functions by: $b = b_S \prod_{A \in \mathcal{A}} b_{\beta_A}$ with $b_{\beta_A}(B) = 1$ if $A \subseteq B$, β_A otherwise, for all $B \subseteq \mathcal{X}$. Thus, for all $B \subseteq \mathcal{X}$: $b(B) = b_S(B) \prod_{A \in \mathcal{A}, A \not\subseteq B} \beta_A$, and consequently, for all $x \in \mathcal{X}$:

$$pl(\{x\}) = 1 - b(\overline{\{x\}}) = 1 - b_S(\overline{\{x\}}) \prod_{A \in \mathcal{A}, A \not\subseteq \overline{\{x\}}} \beta_A$$

$$= 1 - b_S(\overline{\{x\}}) \prod_{A \in \mathcal{A}, x \in A} \beta_A = 1 - (1 - pl_S(\{x\})) \prod_{A \in \mathcal{A}, x \in A} \beta_A \ .$$

A.2   Proof of Proposition 2

From Proposition 1, after having applied CD on m_S, the discrepancy measure E_{pl} (5) can be written: $E_{pl}(\boldsymbol{\beta}) = \sum_{k=1}^{K} E_{pl}(\boldsymbol{\beta}, x_k)$, with for all $k \in \{1, \dots, K\}$:

$$E_{pl}(\boldsymbol{\beta}, x_k) := \sum_{i=1}^{n} \left(\left(1 - (1 - pl_S\{o_i\}(\{x_k\})) \prod_{A \in \mathcal{A}, x_k \in A} \beta_A \right) - \delta_{i,k} \right)^2 . \quad (10)$$

As $E_{pl}(\boldsymbol{\beta}, x_k) \geq 0$ for all $k \in \{1, \ldots, K\}$, the minimum value of $E_{pl}(\boldsymbol{\beta})$ is obtained when each $E_{pl}(\boldsymbol{\beta}, x_k)$ reaches its minimum.

Besides, as all coefficients β_A belong to $[0, 1]$, for each x_k, $k \in \{1, \ldots, K\}$, the product $\prod_{A \in \mathcal{A}, x_k \in A} \beta_A$ of coefficients β_A in $E_{pl}(\boldsymbol{\beta}, x_k)$ (10) also belongs to $[0, 1]$ and can be denoted by a variable $\beta_k \in [0, 1]$. Hence, for each $k \in \{1, \ldots, K\}$, the minimum of $E_{pl}(\boldsymbol{\beta}, x_k)$ is reached for a particular value of β_k.

Now, we can remark that each coefficient $\beta_{\{x_k\}} \in [0, 1]$, $k \in \{1, \ldots, K\}$, only appears in the expression of $E_{pl}(\boldsymbol{\beta}, x_k)$ (10), $k \in \{1, \ldots, K\}$. Hence, choosing $\beta_k = \beta_{\{x_k\}}$ for all k (which means choosing $\mathcal{A}$ composed of the set of singletons of $\mathcal{X}$) constitutes then a solution, *i.e.*, a set of contexts for which the minimum value of $E_{pl}(\boldsymbol{\beta})$ is reached.

Each value of E_{pl} is then reachable using the vector $\boldsymbol{\beta}$ of coefficients $\beta_k := \beta_{\{x_k\}}$, $k \in \{1, \ldots, K\}$, and as already mentioned in [6, Section 5.1], the computation of the coefficient $\boldsymbol{\beta}$ with CD based on the singletons is a constrained least-squares problem. Indeed, for all $k \in \{1, \ldots, K\}$, and for all $i \in \{1, \ldots, n\}$:

$$pl\{o_i\}(\{x_k\}) - \delta_{i,k} = 1 - (1 - pl_S\{o_i\}(\{x_k\}))\beta_k - \delta_{i,k} \qquad (11)$$
$$= (pl_S\{o_i\}(\{x_k\}) - 1)\beta_k - (\delta_{i,k} - 1) . \qquad (12)$$

Then (5) can be rewritten as (7).

A.3   Proof of Proposition 3

As $m = m_S \bigcirc_{A \in \mathcal{A}} A^{\beta_A}$, the CR is determined in terms of commonality functions by $q = q_S \prod_{A \in \mathcal{A}} q^{\beta_A}$ with $q^{\beta_A}(B) = 1$ if $B \subseteq A$, β_A otherwise, for all $B \subseteq \mathcal{X}$. Then, for all $B \subseteq \mathcal{X}$: $q(B) = q_S(B) \prod_{A \in \mathcal{A}, B \nsubseteq A} \beta_A$, which means that after having applied CR, plausibilities on singletons are defined, for all $x \in \mathcal{X}$, by:

$$pl(\{x\}) = q(\{x\}) = q_S(\{x\}) \prod_{A \in \mathcal{A}, x \notin A} \beta_A = pl_S(\{x\}) \prod_{A \in \mathcal{A}, x \notin A} \beta_A$$

A.4   Proof of Proposition 4

From Proposition 3, for each $k \in \{1, \ldots, K\}$, coefficient $\beta_{\overline{\{x_k\}}}$ takes its values in $[0, 1]$ and only appears in $pl(x_k)$ when a CR has been applied. Then, with the same reasoning as for the CD case, the minimum value of E_{pl} with CR can be reached using the set of contexts $\{\overline{x_k} = \mathcal{X} \setminus \{x_k\}, k \in \{1, \ldots, K\}\}$.

The minimization of E_{pl} with CR based on the vector $\boldsymbol{\beta} = (\beta_k := \beta_{\overline{\{x_k\}}}, k \in \{1, \ldots, K\})$ is also a constrained least-squares problem as (5) can be written as (9) (as $\forall k \in \{1, \ldots, K\}$ and $\forall i \in \{1, \ldots, n\}$, $pl\{o_i\}(\{x_k\}) - \delta_{i,k} = pl_S\{o_i\}(\{x_k\})\beta_k - \delta_{i,k}$).