On learning evidential contextual corrections from soft labels using a measure of discrepancy between contour functions

Siti Mutmainah[1,2], Samir Hachour[1], Frédéric Pichon[1], and David Mercier[1]

[1] Univ. Artois, EA 3926 LGI2A, Béthune, F-62400, France
firstname.lastname@univ-artois.fr
[2] UIN Sunan Kalijaga, Yogyakarta, Indonesia
siti.mutmainah@uin-suka.ac.id

Abstract. In this paper, a proposition is made to learn the parameters of evidential contextual correction mechanisms from a learning set composed of soft labelled data, that is data where the true class of each object is only partially known. The method consists in optimizing a measure of discrepancy between the values of the corrected contour function and the ground truth also represented by a contour function. The advantages of this method are illustrated by tests on synthetic and real data.

Keywords: Belief functions · Contextual corrections · Learning · Soft labels.

1 Introduction

In Dempster-Shafer theory [15, 17], the correction of a source of information, a sensor for example, is classically done using the discounting operation introduced by Shafer [15], but also by so-called contextual correction mechanisms [10, 13] taking into account more refined knowledge about the quality of a source.

These mechanisms, called contextual discounting, negating and reinforcement [13], can be derived from the notions of *reliability (or relevance)*, which concerns the competence of a source to answer the question of interest, and *truthfulness* [12, 13] indicating the source's ability to say what it knows (it may also be linked with the notion of bias of a source). The contextual discounting is an extension of the discounting operation, which corresponds to a partially reliable and totally truthful source. The contextual negating is an extension of the negating operation [13, 12], which corresponds to the case of a totally reliable but partially truthful source, the extreme case being the negation of a source [5]. At last, the contextual reinforcement is an extension of the reinforcement, a dual operation of the discounting [11, 13].

In this paper, the problem of learning the parameters of these correction mechanisms from soft labels, meaning partially labelled data, is tackled. More specifically, in our case, soft labels indicate the true class of each object in an imprecise manner through a contour function.

A method for learning these corrections from labelled data (hard labels), where the truth is perfectly known for each element of the learning set, has already been introduced in [13]. It consists in minimizing a measure of discrepancy between the corrected

contour functions and the ground truths over elements of a learning set. In this paper, it is shown that this same measure can be used to learn from soft labels, and tests on synthetic and real data illustrate its advantages to 1) improve a classifier even if the data is only partially labelled; and 2) obtain better performances than learning these corrections from approximate hard labels approaching the only available soft labels.

This paper is organized as follows. In Section 2, the basic concepts and notations used in this paper are presented. Then, in Section 3, the three applied contextual corrections as well as their learning from hard labels are exposed. The proposition to extend this method to soft labels is introduced. Tests of this method on synthetic and real data are presented in Section 4. At last, a discussion and future works are given in Section 5.

2 Belief functions: basic concepts used

Only the basic concepts used are presented in this section (See for example [15, 17, 3] for further details on the belief function framework).

From a frame of discernment $\Omega = \{\omega_1, ..., \omega_K\}$, a *mass function (MF)*, noted m^Ω or m if no ambiguity, is defined from 2^Ω to $[0, 1]$, and verify $\sum_{A \subseteq \Omega} m^\Omega(A) = 1$.

The focal elements of a MF m are the subsets A of Ω such that $m(A) > 0$.

A MF m is in one-to-one correspondence with a *plausibility function* Pl defined for all $A \subseteq \Omega$ by

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{1}$$

The *contour function* pl of a MF m is defined for all $\omega \in \Omega$ by

$$\begin{array}{rl} pl : \Omega \rightarrow & [0, 1] \\ \omega \mapsto & pl(\omega) = Pl(\{\omega\}) \ . \end{array} \tag{2}$$

It is the restriction of the plausibility function to all the singletons of Ω.

The knowledge of the reliability of a source is classically taken into account by the operation called *discounting* [15, 16]. Let us suppose a source S provides a piece of information represented by a MF m_S. With $\beta \in [0, 1]$ the degree of belief of the reliability of the source, the discounting of m_S is defined by the MF m s.t.

$$m(A) = \beta \, m_S(A) + (1 - \beta) m_\Omega(A) \ , \tag{3}$$

for all $A \subseteq \Omega$, where m_Ω represents the total ignorance, i.e. the MF defined by $m_\Omega(\Omega) = 1$.

Several justifications for this mechanism can be found in [16, 10, 13].

The contour function of the MF m resulting from the discounting (3) is defined for all $\omega \in \Omega$ by (see for example [13, Prop. 11])

$$pl(\omega) = 1 - (1 - pl_S(\omega))\beta \ , \tag{4}$$

with pl_S the contour function of m_S.

3 Contextual corrections and learning from labelled data

In this Section, the contextual corrections we used are first exposed, then their learning from hard labels. The proposition to extend this method to soft labels is then introduced.

3.1 Contextual corrections of a mass function

For the sake of simplicity, we only recall here the contour functions expressions resulting from the applications of contextual discounting, reinforcement and negating mechanisms in the case of K contexts where K is the number of elements in Ω.

It is shown in [13] that these expressions are rich enough to minimize the discrepancy measure used to learn the parameters of these corrections, this measure being presented in Section 3.2.

Let us suppose a source S providing a piece of information m_S.

The contour function resulting from the *contextual discounting (CD)* of m_S and a set of contexts composed of the singletons of Ω is given by

$$pl(\omega) = 1 - (1 - pl_S(\omega))\beta_{\{\omega\}} \ , \tag{5}$$

for all $\omega \in \Omega$, with the K parameters $\beta_{\{\omega\}}$ which may vary in $[0, 1]$.

For the *contextual reinforcement (CR)* and the *contextual negating (CN)*, the contour functions are respectively given, from a set of contexts composed of the complementary of each singleton of Ω, by

$$pl(\omega) = pl_S(\omega)\beta_{\overline{\{\omega\}}} \ , \tag{6}$$

and

$$pl(\omega) = 0.5 + (pl_S(\omega) - 0.5)(2\beta_{\overline{\{\omega\}}} - 1) \ , \tag{7}$$

for all $\omega \in \Omega$, with the K parameters $\beta_{\overline{\{\omega\}}}$ able to vary in $[0, 1]$.

3.2 Learning from hard labels

Let us suppose a source of information providing a MF m_S concerning the true class of an object among a set of possible classes Ω.

If we have a learning set composed of n instances (or objects) the true values of which are known, we can learn the parameters of a correction by minimizing a discrepancy measure between the output of the classifier which is corrected (a correction is applied to m_S) and the ground truth [7, 10, 13].

Introduced in [10], the following measure E_{pl} yields a simple optimization problem (a linear least-squares optimization problem, see [13, Prop. 12, 14 et 16]) to learn the vectors $\boldsymbol{\beta}_{CD}$, $\boldsymbol{\beta}_{CR}$ and $\boldsymbol{\beta}_{CN}$ composed of the K parameters of corrections CD, CR and CN:

$$E_{pl}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} (pl_i(\omega_k) - \delta_{i,k})^2 \ , \tag{8}$$

where pl_i is the contour function regarding the class of the instance i resulting from a contextual correction (CD, CR or CN) of the MF provided by the source for this instance, and $\delta_{i,k}$ is the indicator function of the truth of all the instances $i \in \{1, \ldots, n\}$, i.e. $\delta_{i,k} = 1$ if the class of the instance i is ω_k, otherwise $\delta_{i,k} = 0$.

3.3 Learning from soft labels

In this paper, we consider the case where the truth is no longer given precisely by the values $\delta_{i,k}$, but only in an imprecise manner by a contour function $\tilde{\delta}_i$ s.t.

$$\begin{array}{rcl} \tilde{\delta}_i : \Omega & \rightarrow & [0,1] \\ \omega_k & \mapsto & \tilde{\delta}_i(\omega_k) = \tilde{\delta}_{i,k} \ . \end{array} \tag{9}$$

The contour function $\tilde{\delta}_i$ gives information about the true class in Ω of the instance i.

Knowing then the truth only partially, we propose to learn the corrections parameters using the following discrepancy measure $\tilde{E}_{pl}$, extending directly (8):

$$\tilde{E}_{pl}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} (pl_i(\omega_k) - \tilde{\delta}_{i,k})^2 \ . \tag{10}$$

The discrepancy measure $\tilde{E}_{pl}$ also yields, for each correction (CD, CR et CN), a linear least-squares optimization problem. For example, for CD, $\tilde{E}_{pl}$ can be written by

$$\tilde{E}_{pl}(\boldsymbol{\beta}) = \|\boldsymbol{Q\beta} - \tilde{\boldsymbol{d}}\|^2 \tag{11}$$

with

$$\boldsymbol{Q} = \begin{bmatrix} diag(\boldsymbol{pl_1} - 1) \\ \vdots \\ diag(\boldsymbol{pl_n} - 1) \end{bmatrix}, \ \tilde{\boldsymbol{d}} = \begin{bmatrix} \tilde{\boldsymbol{\delta}}_{\mathbf{1}} - 1 \\ \vdots \\ \tilde{\boldsymbol{\delta}}_{\boldsymbol{n}} - 1 \end{bmatrix} \tag{12}$$

where $diag(\boldsymbol{v})$ is a square diagonal matrix whose diagonal is composed of the elements of the vector $\boldsymbol{v}$, and where for all $i \in \{1, \ldots, n\}$, $\tilde{\boldsymbol{\delta}}_{\boldsymbol{i}}$ is the column vector composed of the values of the contour function $\tilde{\delta}_i$, meaning $\tilde{\boldsymbol{\delta}}_{\boldsymbol{i}} = (\tilde{\delta}_{i,1}, \ldots, \tilde{\delta}_{i,K})^T$.

In the following, this learning proposition is tested with generated and real data.

4 Tests on generated and real data

We first expose how soft labels can be generated from hard labels to make the tests exposed afterwards on synthetic and real data.

4.1 Generating soft labels from hard labels

It is not easy to find partially labelled data in the literature. Thus, as in [1, 14, 8, 9], we have built our partially labelled data sets (soft labels) from perfect truths (hard labels) using the procedure described in Algorithm 1 (where $B\hat{e}ta, \mathcal{B}$, and $\mathcal{U}$ means respectively Bêta, Bernoulli and uniform distributions).

Algorithm 1 Soft labels generation

Input: hard labels δ_i with $i \in \{1, \ldots, n\}$, where for each i, the integer $k \in \{1, \ldots, K\}$ s.t. $\delta_{i,k} = 1$ is denoted by k_i.
Output: soft labels $\tilde{\delta}_i$ with $i \in \{1, \ldots, n\}$.

procedure HARDTOSOFTLABELS
  for each instance i **do**
    Draw $p_i \sim B\hat{e}ta(\mu = .5, v = .04)$
    Draw $b_i \sim \mathcal{B}(p_i)$
    if $b_i = 1$ **then**
      Draw $k_i \sim \mathcal{U}_{\{1,\ldots,K\}}$
    $\tilde{\delta}_{i,k_i} \leftarrow 1$
    $\tilde{\delta}_{i,k} \leftarrow p_i$ for all $k \neq k_i$

Algorithm 1 allows one to obtain soft labels that are all the more imprecise as the most plausible class is false.

4.2 Tests performed

The chosen evidential classifier used as a source of information is the evidential k-nearest neighbor classifier (EkNN) introduced by Denœux in [2] with $k = 3$. We could have chosen another one with other settings, it can be seen as a black box.

The first test set we consider is composed of synthetic data composed of 3 classes built from 3 bivariate normal distributions with respective means $\mu_{\omega_1} = (1, 2)$, $\mu_{\omega_2} = (2, 1)$ and $\mu_{\omega_3} = (0, 0)$, and a common covariance matrix Σ s.t.

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}. \tag{13}$$

For each class, 100 instances have been generated. They are illustrated in Figure 1.

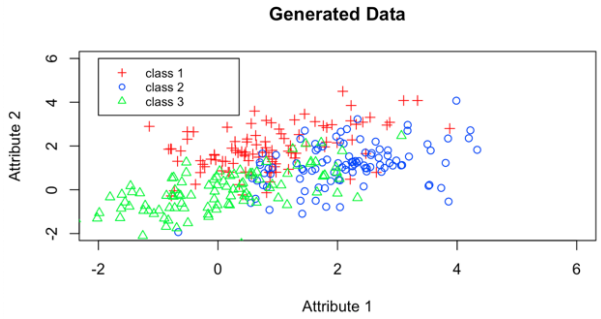

Fig. 1. Illustration of the generated dataset (3 classes, 2 attributes).

We have then considered several real data sets from the UCI database [6] composed of numerical attributes as the EkNN classifier is used. Theses data sets are described in Table 1.

Table 1. Characteristics of the UCI dataset used (number of instances without missing data, number of classes, number of numerical attributes used)

Data	# Instances	# Classes	# Attributes
Ionosphere	350	2	34
Iris	150	3	4
Sonar	208	2	60
Vowel	990	11	9
Wine	178	3	13

For each dataset, a 10-repeated 10-fold cross validation has been undertaken as follows:

- the group containing one tenth of the data is considered as the test set (the instances labels being made imprecise using Algorithm 1),
- the other 9 groups form the learning set, which is randomly divided into two groups of equal size:
  - one group to learn the EkNN classifier (learnt from hard truths),
  - one group to learn the parameters of the correction mechanisms from soft labels (the labels of the dataset are made imprecise using Algorithm 1).

For learning the parameters of contextual corrections, two strategies are compared.

1. In the first strategy, we use the optimization of Equation (8) from the closest hard truths from the soft truths (the most plausible class is chosen). Corrections with this strategy are denoted by CD, CR and CN.
2. In the second strategy, Equation (10) is directly optimized from soft labels (cf Section 3.3). The resulting corrections using this second strategy are denoted by CDsl, CRsl and CNsl.

The performances of the systems (the classifier alone and the corrections - CD, CR or CN - of this classifier according to the two strategies described above) are measured using $\tilde{E}_{pl}$ (10), where $\tilde{\delta}$ represents the partially known truth. This measure corresponds to the sum over the test instances of the differences, in the least squares sense, between the truths being sought and the system outputs.

The performances $\tilde{E}_{pl}$ (10) obtained from UCI and generated data for the classifier and its corrections are summed up in Table 2 for each type of correction. Standard deviations are indicated in brackets.

From the results presented in Table 2, we can remark that, for CD, the second strategy (CDsl) consisting in learning directly from the soft labels, allows one to obtain lower differences $\tilde{E}_{pl}$ from the truth on the test set than the first strategy (CD) where the correction parameters are learnt from approximate hard labels. We can also remark

Table 2. Performances $\tilde{E}_{pl}$ obtained for the classifier alone and the classifier corrected with CD, CR and CN using both strategies. Standard deviations are indicated in brackets.

Data	EkNN	CD	CDsl	CR	CRsl	CN	CNsl
Generated Data	23.8 (3.8)	16.6 (2.8)	7.9 (1.5)	26.8 (3.0)	23.5 (3.7)	11.5 (1.6)	9.8 (0.6)
Ionosphere	16.2 (2.5)	9.6 (2.2)	5.3 (1.0)	17.2 (1.9)	15.9 (2.3)	9.3 (1.3)	8.4 (0.9)
Iris	12.5 (2.4)	8.4 (2.1)	3.3 (0.9)	13.1 (2.0)	12.3 (2.2)	6.7 (1.5)	4.8 (0.5)
Sonar	7.8 (2.0)	6.3 (1.9)	3.5 (0.9)	9.0 (1.6)	7.7 (1.9)	5.1 (0.8)	5.0 (0.9)
Vowel	279 (24)	278 (23)	62 (5)	310 (21)	279 (24)	240 (21)	65 (5)
Wine	13.3 (2.6)	10.4 (2.3)	4.3 (1.0)	15.0 (2.1)	13.3 (2.5)	7.2 (1.6)	5.7 (0.6)

that this strategy yields lower differences $\tilde{E}_{pl}$ than the classifier alone, illustrating, in these experiments, the usefulness of soft labels even if hard labels are not available, which can be interesting in some applications.

The same conclusions can be drawn for CN.

For CR, the second strategy is also better than the first one but we can note that unlike the other corrections, there is no improvement for the first strategy in comparison to the classifier alone (the second strategy having also some close performances to the classifier alone).

5 Discussion and future works

We have shown that contextual corrections may lead to improved performances in the sense of measure $\tilde{E}_{pl}$, which relies on the plausibility values returned by the systems for each class for each instance. We also note that by using the same experiments as those in Section 4.2 but evaluating the performances using a simple 0-1 error criterion, where for each instance the most plausible class is compared to the true class, the performances remain globally identical for the classifier alone as well as all the corrections (the most plausible class being often the same for the classifier and each correction).

For future works, we are considering the use of other performance measures, which would also take fully into account the uncertainty and the imprecision of the outputs. For example, we would like to study those introduced by Zaffalon et al. [18].

It would also be possible to test other classifiers than the EkNN. We could also test the advantage of these correction mechanisms in classifiers fusion problems.

At last, we also intend to investigate the learning from soft labels using another measure than $\tilde{E}_{pl}$ and in particular the evidential likelihood introduced by Denœux [4] and already used to develop a CD-based EkNN [9].

Acknowledgement

The authors would like to thank the anonymous reviewers for their helpful and constructive comments, which have helped them to improve the quality of the paper and to consider new paths for future research.

Mrs. Mutmainah's research is supported by the overseas 5000 Doctors program of Indonesian Religious Affairs Ministry (MORA French Scholarship)

References

1. E. Côme, L. Oukhellou, T. Denœux, P. Aknin. Learning from partially supervised data using mixture models and belief functions. Pattern Recognition, 42(3):334348, 2009.
2. T. Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man, and Cybernetics, 25(5):804-813, 1995.
3. T. Denœux. Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. Artificial Intelligence, 172:234-264, 2008.
4. T. Denœux. Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Transactions on Knowledge and Data Engineering, 25(1):119130, 2013.
5. D. Dubois, H. Prade. A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. International Journal of General Systems, 12(3):193-226, 1986.
6. D. Dua, C. Graff. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.
7. Z. Elouedi, K. Mellouli, P. Smets. The Evaluation of Sensors Reliability and Their Tuning for Multisensor Data Fusion within the Transferable Belief Model. Proceedings of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU'2001, pp. 350-361, Toulouse, 2001.
8. O. Kanjanatarakul, S. Kuson, T. Denœux. An Evidential K-Nearest Neighbor Classifier Based on Contextual Discounting and Likelihood Maximization. Proceedings of the 5th International Conference on Belief Functions, BELIEF'2018, pp. 155-162, Compigne, 17-21 September, 2018.
9. O. Kanjanatarakul, S. Kuson, T. Denœux. A New Evidential K-Nearest Neighbor Rule based on Contextual Discounting with Partially Supervised learning. International Journal of Approximate Reasoning, 113:287-302, 2019.
10. D. Mercier, B. Quost, T. Denœux. Refined Modeling of Sensor Reliability in the Belief Function Framework Using Contextual Discounting. Information Fusion, 9(2):246-258, 2008.
11. D. Mercier, E. Lefèvre, F. Delmotte. Belief functions contextual discounting and canonical decompositions. International Journal of Approximate Reasoning, 53(2):146-158, 2012.
12. F. Pichon, D. Dubois, T. Denoeux. Relevance and truthfulness in information correction and fusion. International Journal of Approximate Reasoning, 53(2):159-175, 2012.
13. F. Pichon, D. Mercier, E. Lefèvre, F. Delmotte. Proposition and learning of some belief function contextual correction mechanisms. International Journal of Approximate Reasoning, 72:4-42, 2016.
14. B. Quost, T. Denoeux, S. Li. Parametric classification with soft labels using the evidential EM algorithm: linear discriminant analysis versus logistic regression. Adv. Data Analysis and Classification, 11(4):659-690, 2017.
15. G. Shafer. A mathematical theory of evidence. Princeton University Press, Princeton, N.J, 1976.
16. P. Smets, Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem, International Journal of Approximate Reasoning, 9(1) : 1-35, 1993.
17. P. Smets, R. Kennes. The Transferable Belief Model. Artificial Intelligence, 66(2):191-234, 1994.
18. M. Zafallon, G. Corani, D.-D. Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. International Journal of Approximate Reasoning, 53(8):1282-1301, 2012.