A mean distance between elements of same class for rich labels

Arthur Hoarau, Constance Thierry, Jean-Christophe Dubois, Yolande Le Gall

A mean distance between elements of same class for rich labels

Arthur Hoarau, Constance Thierry,
Jean-Christophe Dubois, and Yolande Le Gall

Univ Rennes, CNRS, IRISA, DRUID, France

Abstract. The prevalence of imperfections in data, characterized by uncertainty and imprecision, prompts the need for effective modeling techniques. The theory of belief functions offers a mathematical framework to address this challenge. In this paper, we tackle the problem of calculating the mean distance between elements of the same class, especially when class membership is uncertain and imprecise. Leveraging belief functions and a notion of similarity between elements, we propose a solution and validate its efficacy through experimental evaluations. The proposed method proves effective when labels exhibit low imprecision, whereas unsupervised methods may be more effective for labels closer to complete ignorance.

Keywords: Belief Functions · Rich Labels · Mean Distance

1 Introduction

The imperfection [9] in data is now prevalent in many application domains. It may be uncertainty (lack of knowledge, *e.g.* "Tomorrow it might be sunny") or imprecision (quantitative or completeness deficiency, *e.g.* "It will be sunny tomorrow or the day after"). The possibility of representing them provides a better way of taking them into account. The theory of belief functions [1, 8] allows for the mathematical modeling of this uncertainty and imprecision, and the notion of distance between multiple bodies of evidence has been extensively studied [5] in this context. In connection with the notion of distances, there are numerous applications in unsupervised machine learning, such as clustering [2], notably with the evidential c-means [7]. Additionally, a related problem, which approaches the issue we are addressing, is that of missing value imputation [6]. In all cases, the representation of imperfection is an expanding field, especially in machine learning, and several recent works have been conducted with the aim of collecting real uncertain and imprecise labels from actual users.

In this paper, we focus on the mean distance between elements of the same class, which is easily calculable when the classes are known but less straightforward when the membership of elements to a class is defined uncertainly and imprecisely. We propose to address this issue using the theory of belief functions and a notion of similarity between labels. A degenerate and informal application of this method has been practiced in an active learning setup [3] (reduced to

the similarity of the element's imperfect label with the labels of other elements and not that of the true classes). Since the notions of distances and belief functions are addressed, it may be interesting to note that an impossibility has been demonstrated [10] between the conjunctive combination of beliefs and the use of distances. However, this does not concern this case directly since the classical definition of distance is used, which separates two points, such as the Euclidean distance. The theory of belief functions intervenes here on the classes and not on the explanatory variables.

The document is structured as follows: Section 2 introduces the problem by recalling the calculation of the average distance between elements of the same class for hard labels. Section 3 presents the method by introducing the theory of belief functions. Section 4 proposes two experiments that help understand the behavior of the proposed method as well as its performance in a practical case. Finally, Section 5 concludes this article.

2 Mean distance between elements for hard classes

Let $\mathcal{X} = \{x^n = (x_1^n, \ldots, x_P^n) | n = 1, \ldots, N\}$ represent a P features collection of N samples, and $\Omega = \{q_1, \ldots, q_C\}$ a set of C classes. Let d be a distance over the features space, for the proofs and experiments in this paper, the Euclidean distance will be adopted; however, any other distance can be used. It is defined as follows:

$$\begin{aligned} d(x^i, x^j) &= \sqrt{\sum_{k=1}^{K} (x_k^i - x_k^j)^2}, \\ &= ||x^i - x^j||, \end{aligned} \tag{1}$$

with $||x||$ the Euclidean norm of x, for convenience we denote $d(x^i, x^j) = d^{i,j}$.

A simple way to compute the mean distance d between all elements is to sum all pairwise distances and divide by the total number of pairs (excluding the distance between an element and itself):

$$\mathsf{d} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} d^{i,j}}{N^2 - N}. \tag{2}$$

This equation can be simplified for complexity reasons by computing only half of the matrix (of dimension $N * N$), dividing the number of elements by 2.

The interest lies in accounting for the class of each observation. The mean distance between elements of the same class d_q can be calculated as follows:

$$\mathsf{d}_q = \frac{\sum_{i=1}^{N_q} \sum_{j=1}^{N_q} d^{i,j}}{N_q^2 - N_q}, \tag{3}$$

with q as the corresponding class, and N_q the number of elements of class q (for the sums over N_q, we simplify notation by implying that the summed distances $d^{i,j}$ refer to those between x^i and x^j belonging to class q).

The tackled issue arises when the class assignment of an observation lacks certainty and precision. Thus, we introduce a mean distance between elements belonging to the same class, tailored for rich (uncertain and imprecise) label representations.

3 Mean distance between elements for rich labels

This section presents the proposed method of mean distance between elements of the same class when the class is known uncertainly and imprecisely. The theory of belief functions [1,8] is used to model these rich labels.

3.1 Mean distance for rich labels (MDRL)

The goal of the proposed method is to extend Equation (3) when the labels are uncertain and imprecise (when N_q is unknown). For this purpose, a similarity measure between the target class q and the rich label is used to weigh the contribution of each observation in the total calculation. In this paper, we arbitrarily choose the similarity measure $1 - d_J$, where d_J is the Jousselme distance [4] between two mass functions. The method is defined by the following equation:

$$\mathsf{MDRL}_q = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(1-d_J^{q,i})(1-d_J^{q,j})d^{i,j}}{[\sum_{i=1}^{N}(1-d_J^{q,i})]^2 - \sum_{i=1}^{N}(1-d_J^{q,i})^2}, \tag{4}$$

with $d_J^{q,i}$ the Jousselme distance between m_q (the categorical mass function on class q) and m_i the mass function defining the class of x^i, and with $d^{i,j}$ the Euclidean distance between x^i and x^j.

Proposition 1: This equation is equal to the classical mean distance between all observations (2) when considering complete ignorance.
Proposition 2: Equation (4) is equal to (3) for hard labels.
Proposition 3: This mean distance is null for identical objects, positive if an object is distinct from others, and symmetric under permutation of elements.

Propositions 1 and 2 are proven below. For Proof 1, all labels are completely ignorant ($m_i(\Omega) = 1, \forall i \in [0, N]$), therefore d_J becomes constant, let $(1-d_J^{q,i}) = \Delta^{\Omega}$. For Proof 2, and thus in the case of hard labels, the Jousselme distance between two elements of the same class becomes 0, and 1 otherwise for a different class.

Proof 1:

$$\begin{aligned}
\mathsf{MDRL}_q &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(1-d_J^{q,i})(1-d_J^{q,j})d^{i,j}}{[\sum_{i=1}^{N}(1-d_J^{q,i})]^2-\sum_{i=1}^{N}(1-d_J^{q,i})^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(\Delta^{\Omega})(\Delta^{\Omega})d^{i,j}}{[\sum_{i=1}^{N}(\Delta^{\Omega})]^2-\sum_{i=1}^{N}(\Delta^{\Omega})^2} \\
&= \frac{(\Delta^{\Omega})^2\sum_{i=1}^{N}\sum_{j=1}^{N}d^{i,j}}{(\Delta^{\Omega})^2[(\sum_{i=1}^{N}1)^2-\sum_{i=1}^{N}1]} \\
&= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}d^{i,j}}{N^2-N} \iff (2)
\end{aligned} \tag{5}$$

Proof 2:

$$\begin{aligned}
\mathsf{MDRL}_q &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(1-d_J^{q,i})(1-d_J^{q,j})d^{i,j}}{[\sum_{i=1}^{N}(1-d_J^{q,i})]^2-\sum_{i=1}^{N}(1-d_J^{q,i})^2} \\
&= \frac{\sum_{i=1}^{N_q}\sum_{j=1}^{N_q}(1)(1)d^{i,j}}{[\sum_{i=1}^{N_q}1]^2-\sum_{i=1}^{N_q}(1)^2} \\
&= \frac{\sum_{i=1}^{N_q}\sum_{j=1}^{N_q}d^{i,j}}{{N_q}^2-N_q} \iff (3)
\end{aligned} \tag{6}$$

Example: We consider students who have obtained grades in three subjects (they belong to class 1 or class 2: $\Omega = \{1, 2\}$). The goal is to determine the homogeneity[1] of the students' level in the two classes, the mean distance between the students (on the grades) according to their class is then calculated. This intra-class inertia allows us to compare the homogeneity level of each class. Students, grades, and their true class are described in left hand part of Table 1. A numerical conversion is made (from F to A[2]). The mean distance is calculated using Equation (3), and the obtained values are 11.2 for students in true class 1 and 5.5 for students in true class 2. Class 2 is thus much more homogeneous than class 1. Now, suppose that the students' class is partially known, this uncertainty and imprecision are described in the right hand part of Table 1. The formula used is no longer applicable[3]. With the proposed Equation (4), we obtain MDRL values of 10.0 for class 1 and 6.78 for class 2. This also indicates that class 2 is more homogeneous than class 1.

4 Experiments

In this section, we propose two experiments to demonstrate the usefulness of the proposed method on several datasets presented in Table 2. These datasets contain quantitative variables that have been processed to remove the mean and scale to unit variance. Each draw is performed 100 times (one draw corresponds to the selection of noised observations). Firstly, a preliminary experiment describes the behavior of the method with respect to the quality of the labels and

[1] Homogeneity is represented by the mean distance between students of the same class.
[2] Grades are: $A, A^-, B^+, B, B^-, C^+, C, C^-, D^+, D, D^-, F$.
[3] For the class that maximizes the pignistic probability, the mean distances are 9.3 for class 1 and 7.8 for class 2.

Table 1: Students' grades for each course (on the left) with true class (in the middle) and rich labels indicating class membership (on the right).

Student	Course 1	Course 2	Course 3	True Class	Class 1	Class 2	Ω
Alice	11	8	11	1	1	0	0
Bob	6	0	11	2	0	1	0
Carol	4	2	0	1	0.8	0	0.2
Dave	1	11	5	1	0	0.1	0.9
Eve	8	4	9	2	0	0.8	0.2
Mallory	10	8	7	2	0.1	0	0.9
Oscar	8	0	3	1	1	0	0
Trudy	7	6	10	2	0	1	0

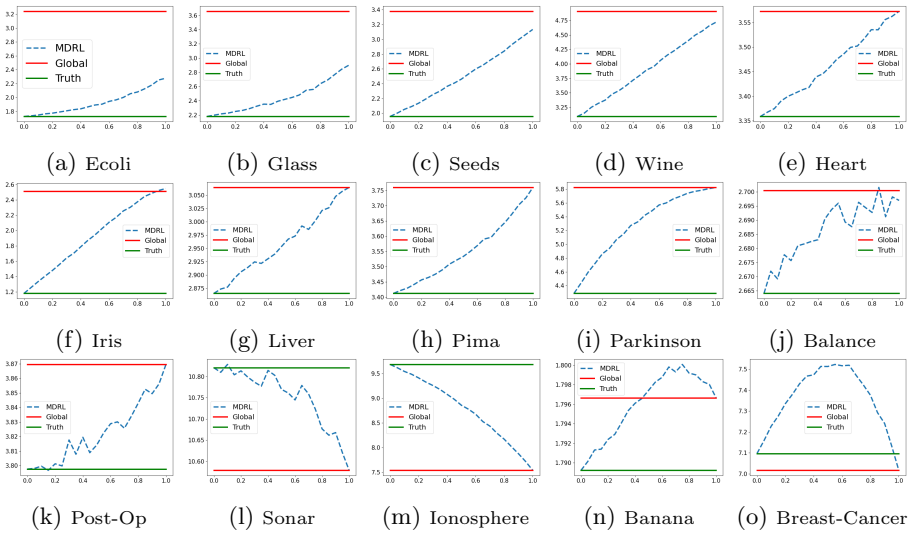
to its theoretical limit between the true mean distance based on classes and a naive mean distance over the entire dataset. The second experiment compares the performance of the proposed method with other methods, both supervised and unsupervised. For both experiments, the mean distances are calculated based on the noise level as follows.

Imprecision noise: An observation is randomly chosen and the corresponding label loses one degree of precision, with another class chosen at random in Ω (*e.g.* If an observation is labeled *Virginica* on Iris dataset, the noisy label becomes either *Virginica* $\cup$ *Setosa* or *Virginica* $\cup$ *Versicolor*). A 50% noisy dataset would mean that half of the labels have lost a degree of precision.

Table 2: Datasets description, with class distribution entropy.

Dataset	Observations	Classes	Features	Entropy
Ecoli	336	8	7	0.73
Glass	214	6	9	0.83
Seeds	210	3	7	1.00
Wine	178	3	13	0.99
Heart	303	2	7	1.00
Iris	150	3	4	1.00
Liver	345	2	6	0.98
Pima	768	2	8	0.93
Parkinson	195	2	22	0.81
Balance	625	3	4	0.83
Post-Operative	86	2	8	0.85
Sonar	208	2	60	1.00
Ionosphere	351	2	34	0.94
Banana	5300	2	2	0.99
Breast Cancer	569	2	30	0.95

4.1 Experiment 1: Average behavior



(a) Ecoli (b) Glass (c) Seeds (d) Wine (e) Heart

(f) Iris (g) Liver (h) Pima (i) Parkinson (j) Balance

(k) Post-Op (l) Sonar (m) Ionosphere (n) Banana (o) Breast-Cancer

Fig. 1: Mean Distance for Rich Labels Vs. Noise. (Class 0)

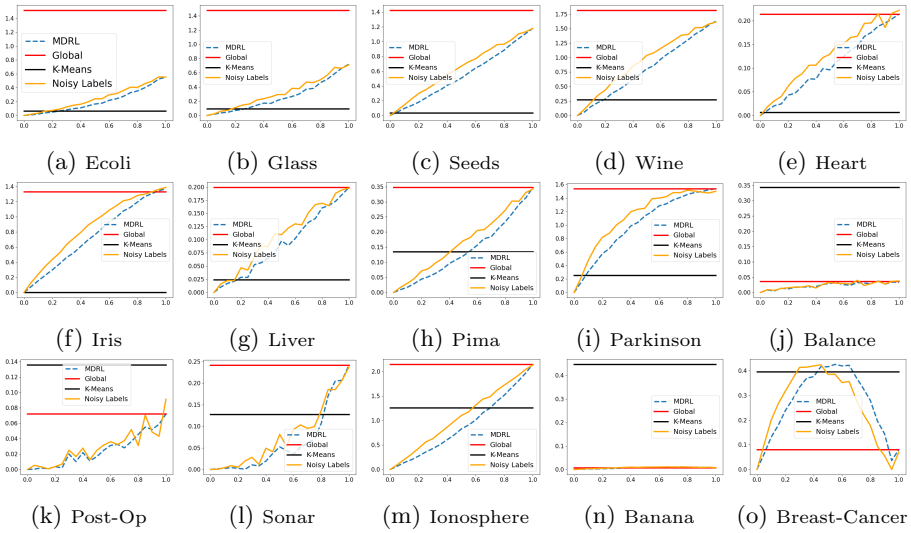
This first experiment focuses on the evolution of the proposed Mean Distance for Rich Labels (MDRL) across multiple datasets, limited by the mean distances between elements of the same true class (3) and the global naive distance between all elements (2), these values are thus invariant to noise. Figure 1 illustrates the behavior of the proposed method with regard to the noise. It varies from 0 (unnoisy dataset) to 1 (fully noised). One class[4] is depicted for each dataset, and the *Ground Truth* line represents the true mean distance between elements of this class. The *Global* line represents the mean distance between all elements of the dataset.

For datasets with two classes (Heart, Liver, Pima, Parkinson, Post-Operative, Sonar, Ionosphere, Banana, and Breast Cancer), the proposed method starts, as theoretically expected, exactly at the true mean distance and converges to the global mean distance when the noise level reaches 100%. Indeed, the noise used translates to total ignorance for datasets limited to two classes. For datasets with a large number of classes (Ecoli and Glass), the proposed method remains closer to the true value. If the noise added total ignorance instead of a degree of imprecision, the curve would also converge to the global mean distance when the dataset is fully noisy. Only the Breast Cancer dataset makes the task very challenging for estimating the mean distance between elements of the same class with respect to noise, due to the particular distribution of observations in the

[4] The first class present in each dataset is always depicted.

variables space for this dataset. The method is therefore largely capable of representing a mean distance that varies between the truth and the least informative value (without using any labels at all). The second experiment then aims to determine whether this method is relevant in terms of performance.

4.2 Experiment 2: Performance of the method



(a) Ecoli (b) Glass (c) Seeds (d) Wine (e) Heart

(f) Iris (g) Liver (h) Pima (i) Parkinson (j) Balance

(k) Post-Op (l) Sonar (m) Ionosphere (n) Banana (o) Breast-Cancer

Fig. 2: Error across different methods Vs. Noise. (Class 0)

In this experiment, the proposed Mean Distance for Rich Labels (MDRL) method is compared with three other methods for estimating the mean distance between elements of the same class on the datasets presented earlier. Once again, only one class per dataset is being studied.

The naive *Global* method calculates the mean distance over all observations without considering the class, as mentioned earlier. Another method, *Noisy labels*, involves selecting the class that maximizes the pignistic probability for an observation and calculating the mean distance between elements of that class. The last compared method utilizes the unsupervised clustering algorithm of *K-means* to create clusters that maximize inter-class distance and minimize intra-class distance. The mean distance between the elements of this cluster is then calculated. A significant advantage given to this method is that the true number of classes is provided to the K-means algorithm to form its clusters. Moreover, the closest mean distance to the truth, among all created clusters, is chosen for comparison with the studied true class.

Figure 2 presents the difference between the estimated value and the true mean distance between elements of the studied class for each level of noise. Since

the Global and K-means methods are unsupervised, it is expected for their curves to be constant, as they do not depend on labels and therefore not on noise. The lower the curve, the closer the value is to the true mean distance, indicating better performance. The least performing method is naturally the Global mean, which does not take into account the labels of the observations. The only dataset where this method is particularly effective is the Breast-Cancer dataset. The proposed MDRL method is better performing than the hard *Noisy Labels* but follows its trend. This phenomenon is theoretically expected since the proposed method aims to be an improvement over it. Finally, the K-means method is much more competitive, often close to 0. However, since the proposed method equals the true mean value when there is no imprecision, it is always better performing than K-means at least with little noise. Then the performance degrades with the addition of noise. The relevance of using the proposed method therefore depends on the noise (and more generally on the uncertainty and imprecision of the sources).

5 Conclusion

In this paper, we propose a mean distance between elements of the same class when classes are not known with certainty and precision but represented by a belief function. This measure is shown to be limited by the true value of the mean distance between elements of the same class when labels are known with certainty and precision and the naive measure of the mean distance between all elements in the case of complete ignorance. Two proofs and experiments are also conducted to theoretically support these properties.

A distance and a dissimilarity measure are necessary. Therefore, the Euclidean distance and the Jousselme distance have been arbitrarily chosen here, but other distances and dissimilarity measures can be used. It has been observed during our experiments that this method can be useful under moderate noise (or imprecision), but it could be more appropriate to use unsupervised methods (such as K-means) when noise is significant. Many issues can be addressed with such a measure, and its practical use is already ongoing in machine learning problems, specifically in active learning.

Special thanks to Vincent Lemaire for conducting a pre-peer review.

References

1. Dempster, A.P.: Upper and Lower Probabilities Induced by a Multivalued Mapping. The Annals of Mathematical Statistics **38**(2), 325 – 339 (1967)
2. Denœux, T., Kanjanatarakul, O.: Evidential clustering: A review. In: Huynh, V.N., Inuiguchi, M., Le, B., Le, B.N., Denoeux, T. (eds.) Integrated Uncertainty in Knowledge Modelling and Decision Making. pp. 24–35 (2016)
3. Hoarau, A., Martin, A., Dubois, J.C., Le Gall, Y.: Imperfect labels with belief functions for active learning. In: Belief Functions: Theory and Applications. Springer International Publishing (2022)

4. Jousselme, A.L., Grenier, D., Éloi Bossé: A new distance between two bodies of evidence. Information Fusion **2**(2), 91–101 (2001)
5. Jousselme, A.L., Maupin, P.: Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning **53**(2) (2012)
6. Liu, Z., Pan, Q., Dezert, J., Martin, A.: Adaptive imputation of missing values for incomplete pattern classification. Pattern Recognition **52**, 85–95 (2016)
7. Masson, M.H., Denœux, T.: Ecm: An evidential version of the fuzzy c-means algorithm. Pattern Recognition **41**(4), 1384–1397 (2008)
8. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
9. Smets, P.: Imperfect Information: Imprecision and Uncertainty, pp. 225–254. Springer US, Boston, MA (1997)
10. Zhang, Y., Destercke, S., Zhang, Z., Bouadi, T., Martin, A.: On computing evidential centroid through conjunctive combination: An impossibility theorem. IEEE Transactions on Artificial Intelligence **PP**, 1–10 (06 2022)