Possibilistic Estimation of Distributions to Leverage Sparse Data in Machine Learning

Andrea G. B. Tettamanzi[1(✉)], David Emsellem[2], Célia da Costa Pereira[3], Alessandro Venerandi[4], and Giovanni Fusco[4]

[1] Université Côte d'Azur, CNRS, Inria, I3S, Sophia Antipolis, France
andrea.tettamanzi@univ-cotedazur.fr
[2] Kinaxia SA, Sophia Antipolis, France
david.emsellem@kcitylabs.fr
[3] Université Côte d'Azur, CNRS, I3S, Sophia Antipolis, France
celia.da-costa-pereira@univ-cotedazur.fr
[4] Université Côte d'Azur, CNRS, ESPACE, Nice, France
{alessandro.venerandi,giovanni.fusco}@univ-cotedazur.fr

Abstract. Prompted by an application in the area of human geography using machine learning to study housing market valuation based on the urban form, we propose a method based on possibility theory to deal with sparse data, which can be combined with any machine learning method to approach weakly supervised learning problems. More specifically, the solution we propose constructs a possibilistic loss function to account for an uncertain supervisory signal. Although the proposal is illustrated on a specific application, its basic principles are general. The proposed method is then empirically validated on real-world data.

Keywords: Possibility theory · Machine learning · Weakly supervised learning

1 Introduction

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs [18]. Each example consists of an input record, which collects the values of a number of input variables, and the associated value of the output variable (also called the supervisory signal). The learnt function can then be used to "predict" the value of the output variable for new unlabeled input records, whose output value is not known.

In many real-world problems, obtaining a fully labeled dataset is expensive, difficult, or outright impossible. An entire subfield of machine learning, called weakly (or semi-) supervised learning has thus emerged, which studies how datasets where the supervisory signal is not available or completely known

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1237, pp. 431–444, 2020.
https://doi.org/10.1007/978-3-030-50146-4_32

for all the records can be used for learning. According to a useful taxonomy of classification problems that can arise in that field [10], four broad classes of problems can be identified:

- single-instance, single-label (SISL), which corresponds to the standard setting where all the examples consist of a single instance, to which a single (i.e., certain) class label is assigned;
- single-instance, multiple-label (SIML), where some examples, consisting of single instances, are assigned a (disjunctive) set of possible class labels, including, to an extreme, the set of all the class labels, which corresponds to the case of a missing supervisory signal;
- multiple instance, single-label (MISL), when examples may consist of sets of instances, being assigned a single label as a whole;
- multiple instance, multiple-label (MIML), when, in addition, some examples are assigned a (disjunctive) set of possible values.

This taxonomy of course assumes that the output variable of the underlying objects, which one seeks to predict, can only have a single true value. It should be mentioned that other problems exist, called *multi-label* classification [19], where, for a given underlying object, described by an input record, multiple values can be active at the same time, which would then be described by a *conjunctive* set of output values. In that case, the learnt function is set-valued.

The above taxonomy can be extended to regression or "predictive modeling" problems, where the "label" is a number, ranging on a discrete or continuous interval.

In the framework of an interdisciplinary research project applying machine learning and urban morphology theory to the investigation of the influence of the urban environment on the value of residential real estate [21], we faced the problem of incorporating into a predictive model uncertain information associated with prices, addressing issues of data sparsity, a problem that falls within the SIML category of the above taxonomy. This prompted us to propose an original method to deal with output variable uncertainty in predictive modeling, based on possibility theory, which we present below.

As it has been argued for example by Bouveyron *et al.* [1], while the problem of noise in data has been widely studied in the literature on supervised learning, the problem of label noise remains an important and unsolved problem in supervised classification. Nigam *et al.* [15] proposed Robust Mixture Discriminant Analysis (RMDA), a supervised classification whose aim is to detect inconsistent labels by comparing the labels given for labeled data set with the ones obtained through an unsupervised modeling based on the Gaussian mixture model. To solve the problem of automatic building extraction from aerial or satellite images with noise labels, Zhang *et al.* [23] propose to capture the relationship between the true label and the noisy label, a general label noise-adaptive (NA) neural network framework consisting of a combination of a base network with an additional probability transition module (PTM) introduced to capture the relationship between the true label and the noisy label. Other researchers prefer to focus on constructing a loss function that is robust to noise [9].

However, the uncertainty brought about by data sparsity in our problem is hard to characterize as a probability distribution, making the application of one of the probabilistic approaches to weak supervised learning proposed in the literature unattractive.

Vannoorenberghe *et al.*, instead, proposed a different approach [20] in which the induction of decision trees is based on the theory of belief functions. In their framework, it is supposed that the training examples have uncertain or imprecise labels. In the same spirit, Quost *et al.* [17] also proposed a belief-function-based framework to be used for supervised learning, in which the training data are associated with uncertain labels. They supposed that each example in the training data set is associated to a belief assignment that represents the actual knowledge of the actual class of the example and used a boosting method to solve the classification problem. Denœux *et al.* [4] introduce a category of learning problems in which the labels associated to the examples in the training data set are assessed by an expert and encoded in the form of a possibility distribution. Although this work is very relevant to what we are proposing here, the authors obtain their possibility distributions from human experts, which can be expensive and difficult, whereas the method we propose automatically computes those distributions from data.

Traditionally, housing market valuations are modeled, in a linear fashion, through a combination of intrinsic and extrinsic features evaluated for each dwelling. Although such linear models provide easy-to-read results, they are severely limited as they assume linearity and independence among variables. However, this might not be the case. For example, a specific variable might change its behaviour for different subsets of observations. More recently, researchers have applied Machine Learning (ML) techniques to study the same phenomenon [3,8,16]. However, their aim was mainly predictive. Thus, although linearity and independence among variables were tackled through the use of such algorithms, their results lacked interpretability. Finally, the intrinsic/extrinsic dichotomy does not hold when the goal of the analysis is the valuation of urban subspaces (like neighborhoods or street segments) instead of individual dwellings. To tackle these issues, we have devised an approach rooted in Urban Morphology to explain housing values at a fine level of spatial granularity, that of street segments and we designed a sequence of appropriate ML techniques that output interpretable results. To be more specific, the proposed approach, firstly, computes street-based measures of housing values, urban form, functions, and landscape and then models the relationship between them through an ensemble method comprising of Gradient Boosting (GB) [7], topological Moran's test [14], and SHAP [13], a recently developed technique to interpret outputs of ML algorithms. The approach has been used to explain the median valuation of street segments in the French Riviera, using housing transactions from the period 2008–2017, through more than 100 metrics of urban form, functions, and landscape.

One difficulty this approach runs into is that transaction data, which are the only source of observations of the output variables (the measures of housing values), are rather sparse at the scale of the street segment. One possible way to

overcome this difficulty would be to limit our study to those street segments for which enough observations are available, e.g., at least ten transactions within the observation period; however, this drastically reduces the number of street segments that can be studied and introduces a bias toward neighborhoods having a relatively high turnover. An alternative and more attractive solution, which is the subject of this paper, is to use all available observations, while taking into account the uncertainty brought about by data sparsity.

Essentially, the solution we propose is to model the uncertainty relevant to the output variable within the framework of possibility theory and modify the loss function of the ML technique, used to model the phenomenon, so that it can weight the error based on the uncertainty of the output variable. This approach is very much in the same spirit as the fuzzy loss function proposed by Hüllermeier [11,12]. An important advantage of such solution is that it is readily transferable to any supervised or semi-supervised ML technique using a loss function (which is the case for the vast majority of such techniques).

The rest of the paper is organized as follows: Sect. 2 provides some background on possibility theory, which is required in order to understand the proposed approach; Sect. 3 states the main question we address, as it emerges from the real-estate price study that motivated our proposal. The proposed solution itself is presented in Sect. 4, while Sect. 5 discusses its empirical validation. Section 6 draws some conclusions and proposes some ideas for further research.

2 Background on Possibility Theory

Fuzzy sets [22] are sets whose elements have degrees of membership in $[0, 1]$. Possibility theory [6] is a mathematical theory of uncertainty that relies upon fuzzy set theory, in that the (fuzzy) set of possible values for a variable of interest is used to describe the uncertainty as to its precise value. At the semantic level, the membership function of such set, π, is called a *possibility distribution* and its range is $[0, 1]$. A possibility distribution can represent the available knowledge of an agent. $\pi(\mathcal{I})$ represents the degree of compatibility of the interpretation $\mathcal{I}$ with the available knowledge about the real world if we are representing uncertain pieces of knowledge. By convention, $\pi(\mathcal{I}) = 1$ means that it is totally possible for $\mathcal{I}$ to be the real world, $1 > \pi(\mathcal{I}) > 0$ means that $\mathcal{I}$ is only somehow possible, while $\pi(\mathcal{I}) = 0$ means that $\mathcal{I}$ is certainly not the real world.

A possibility distribution π is said to be normalized if there exists at least one interpretation $\mathcal{I}_0$ s.t. $\pi(\mathcal{I}_0) = 1$, i.e., there exists at least one possible situation which is consistent with the available knowledge.

Definition 1 *(Possibility and Necessity Measures).* *A possibility distribution π induces a* possibility measure *and its dual* necessity measure*, denoted by Π and N respectively. Both measures apply to a classical set $S \subseteq \Omega$ and are defined as follows:*

$$\Pi(S) = \max_{\mathcal{I} \in S} \pi(\mathcal{I}); \tag{1}$$

$$N(S) = 1 - \Pi(\bar{S}) = \min_{\mathcal{I} \in \bar{S}} \{1 - \pi(\mathcal{I})\}. \tag{2}$$

In words, $\Pi(S)$ expresses to what extent S is consistent with the available knowledge. Conversely, $N(S)$ expresses to what extent S is entailed by the available knowledge. It is equivalent to the impossibility of its complement $\bar{S}$—the more $\bar{S}$ is impossible, the more S is certain. A few properties of Π and N induced by a normalized possibility distribution on a finite universe of discourse Ω are the following. For all subsets $A, B \subseteq \Omega$:

1. $\Pi(A \cup B) = \max\{\Pi(A), \Pi(B)\}$;
2. $\Pi(A \cap B) \leq \min\{\Pi(A), \Pi(B)\}$;
3. $\Pi(\emptyset) = N(\emptyset) = 0; \quad \Pi(\Omega) = N(\Omega) = 1$;
4. $N(A \cap B) = \min\{N(A), N(B)\}$;
5. $N(A \cup B) \geq \max\{N(A), N(B)\}$;
6. $\Pi(A) = 1 - N(\bar{A})$ (duality);
7. $N(A) > 0 \Rightarrow \Pi(A) = 1; \quad \Pi(A) < 1 \Rightarrow N(A) = 0$;

A consequence of these properties is that $\max\{\Pi(A), \Pi(\bar{A})\} = 1$. In case of complete ignorance on A, $\Pi(A) = \Pi(\bar{A}) = 1$.

3 Problem Statement

To make this paper self-contained, we briefly recall here some elements, relevant to the solution we are going to describe in Sect. 4, of the problem that motivated our proposal. The interested reader can refer to [21] for a more detailed explanation.

3.1 Pre-processing

Housing transactions of different years and housing typologies cannot be directly compared due to yearly inflation, housing market cycles (e.g., economic recession, upturn), and specific market behaviours affecting different housing types. For example, bigger properties tend to be sold less frequently as they are more expensive and subject to long term investments, while smaller properties, due to their relative lower valuations, tend to be exchanged more easily and tend to be subject of shorter-term investments. The average price per square meter tends also to be structurally higher for small flats for technical reasons (even the smallest flat needs sanitary and cooking equipment, which proportionally weigh more on the average price per surface unit compared to a larger property). The very notion of average price per square meter can thus be challenged when applied to such diverse housing markets. To address these issues, instead of the conventional price per square meter, our method requires to separate the transactions by year and housing type and compute ventiles of prices for each subset year of transaction - housing type. We consider such statistics as appropriate normalized values, which account for different market segments and years, thus making transactions comparable among them.

3.2 Computation of the Median of Values

Having classified each transaction in a ventile of value, the next step requires to first assign to each data point the street segment to which they belong and, second, aggregate the information on value at the street level through the computation of a measure of central tendency (i.e., median). Such measure provides information on the central point of the distribution of the value of the housing market, for each street. We perform such computation for the ensemble of each street segment and its immediate neighbouring streets. This for two reasons: firstly, transactions located in streets directly connected to one another tend to have similar valuations (due to the influence of the same location factors, presence of properties at the intersection of several streets segments, etc.); secondly, data on house prices tend to be quite sparse, even for several years, and thus a local interpolation allows us to increase the data coverage. Nevertheless, most street segments end up having less than ten transactions per housing type, which introduces uncertainty into the computation of the median statistics.

3.3 Street-Based Metrics of the Urban Environment and Landscape

To characterize the context of each street segment in the most comprehensive way possible, we compute a set of descriptors that quantify aspects of the urban fabric, street-network configuration, functions, housing stock, and landscape. Their definition is out of the scope of this paper.

4 Proposed Solution

In abstract terms, we can describe the problem as follows. We are given a sample of observed variates $x_1, x_2, \ldots, x_n$ and a number of probability distributions (the hypothesis space). We want to assess the possibility degree of these probability distributions based on the given sample, i.e., the degree to which they are compatible with the observations.

4.1 Possibility of a Price Distribution

We can limit ourselves to parametric families of distributions. In this case, this problem can be described as a sort of possibilistic parameter estimation. In the Incertimmo project, we consider distributions described by three deciles: $(d_1, d_5 = \text{median}, d_9)$. We recall that the ith decile d_i is the smallest number x satisfying $\Pr[X \leq x] \geq \frac{i}{10}$. By definition, $d_1 \leq d_5 \leq d_9$. Furthermore, for a probability distribution defined in the $[a, b]$ interval, we know that 10% of the probability mass is in the $[a, d_1]$ interval, 40% in the $(d_1, d_5]$ interval, 40% in the $(d_5, d_9]$ interval, and 10% in the $(d_9, b]$ interval. We make the additional simplifying assumption that the probability mass is uniformly distributed within each of the above intervals. While this might look like a very restrictive assumption, on the one hand it is motivated by the type of qualitative descriptions of price

distribution that are of interest to the human geographers (i.e., by the application at hand) and, on the other hand, could easily be relaxed by selecting other parametric families of distributions without serious consequences on the proposed approach.

This yields a parametric family of probability distributions on the $[a, b]$ interval whose density (in the continuous case) or probability (in the discrete case) function is

$$f(x; d_1, d_5, d_9) = \begin{cases} \frac{1}{10(d_1 - a)}, & a \leq x \leq d_1; \\ \frac{2}{5(d_5 - d_1)}, & d_1 < x \leq d_5; \\ \frac{2}{5(d_9 - d_5)}, & d_5 < x \leq d_9; \\ \frac{1}{10(b - d_9)}, & d_9 < x \leq b. \end{cases} \tag{3}$$

The degenerate cases that arise when $d_1 = d_5$, $d_5 = d_9$, or $d_9 = b$ are treated by adding to the result the contributions of the lines whose condition becomes empty.

In the specific application described in Sect. 3, as we have seen, the observed variates are ventiles of the general distribution of housing prices, taking up values in the discrete set $\{1, 2, \ldots, 20\}$.

Given the sample $x_1, x_2, \ldots, x_n$, it is easy to compute the probability that it is produced by a given distribution of the parametric family, having parameters (d_1, d_5, d_9). This is

$$\Pr[x_1, x_2, \ldots, x_n \mid (d_1, d_5, d_9)] = \prod_{i=1}^{n} f(x_i; d_1, d_5, d_9). \tag{4}$$

This probability is in fact a likelihood function over the distribution of price ventiles, which we will denote by $\mathcal{L}(d_1, d_5, d_9)$.

The link between likelihoods and possibility theory has been explored in [5]. The main result of that study was that possibility measures can be interpreted as the supremum of a family of likelihood functions. It should be stressed that this is an *exact* interpretation, not just an approximation. Based on this result, we transform the likelihood function $\mathcal{L}$ into a possibility distribution over the set of parametric probability distributions of the form (d_1, d_5, d_9).

Notice that the parameters d_1, d_5, d_9 of the probability distributions are the elementary events of this possibility space. A possibility distribution over that space is obtained by letting

$$\pi(d_1, d_5, d_9) = \mathcal{L}(d_1, d_5, d_9)/\mathcal{L}_{\max}, \tag{5}$$

where

$$\mathcal{L}_{\max} = \max_{1 \leq x \leq y \leq z \leq 20} \mathcal{L}(x, y, z),$$

with $x, y, z \in \{0, 1, \ldots, 20\}$, so that all the maximum-likelihood probability distributions have a possibility degree of 1, thus yielding a normalized possibility distribution. Alternatively, the logarithm of the likelihood could be used instead. We restrict the values of the three parameters to the set $\{0, 1, \ldots, 20\}$, because

prices are relative and expressed in ventiles in the application at hand. Of course, for other applications, different ranges of values should be considered.

Now, specific values or intervals of one parameter will constitute complex events, i.e., sets of elementary events, and their possibility and necessity measures can be computed as usual, as the maximum of the possibilities of the distributions that fit the specification and $1 - \max$ of all the others, respectively. For instance, the possibility measure over the median price ventile (d_5) of a given street segment will be given, for all $d_5 \in \{1, \ldots, 20\}$, by

$$\Pi(d_5) = \max_{1 \le x \le d_5 \le y \le 20} \pi(x, d_5, y), \tag{6}$$

where $\pi(\cdot, \cdot, \cdot)$ is defined as in Eq. 5.

4.2 Loss Function Under Possibilistic Uncertainty

The error made by the model predicting that the median price for a street segment is in ventile $\hat{y}$ when all we know is the possibility distribution π over the probability distributions of the prices for that segment, can be defined as

$$L(\hat{y}, \pi) = \int_0^1 \min_{\Pi(y) \ge \alpha} (\hat{y} - y)^2 d\alpha, \tag{7}$$

where $\Pi(y)$ is the possibility measure of the distributions having y as their median. Equation 7 is based on an underlying square error function $e(y) = (\hat{y} - y)^2$, but it could be easily generalized to use an arbitrary error function.

In practice, if $\Lambda = (0 = \lambda_1, \lambda_2, \ldots, 1)$ is the list of possibility levels of π, such that $\forall i > 1, \exists (z, y, z) : \pi(z, y, z) = \lambda_i$, Eq. 7 can be rewritten as

$$L(\hat{y}, \pi) = \sum_{i=2}^{\|\Lambda\|} (\lambda_i - \lambda_{i-1}) \min_{\Pi(y) \ge \lambda_i} (\hat{y} - y)^2. \tag{8}$$

This loss function has been coded in Python in such a way that it could be provided as a custom evaluation function to an arbitrary machine learning method offering this possibility. Most of the loss computation requires iterating through all price distributions of the parametric family (with the three parameters $0 \le d_1 \le d_5 \le d_9 \le 20$, there are 1,540 of them). To optimize performance, we pre-computed the loss function into a lookup table. Since we are using Gradient Boosting Gradient descent (Newton version in XGBoost), we have provided also functions that return its gradient and Hessian.

5 Experiments and Results

In this section we report the experiments we carried out to validate our method. We use real-world data consisting of all the housing transactions made on the

French Riviera over the period 2008–2017.[1] Each record contains detailed intrinsic features of the dwelling that was sold/bought, including its address and the price paid. From these data, we compiled a dataset whose records correspond to street segments, described by more than 100 metrics of urban form, functions, and landscape, and a distribution of price ventiles for each type of dwelling.

Our goal is to compare the performance of a predictive model trained on this dataset, where the labels are uncertain, due to sparseness of transaction data, to the performance of a predictive model trained on a dataset where the labels are certain (in our case, estimated based on a sufficient number of transactions). If the model trained under uncertainty is able to obtain results similar to those of the model trained without uncertainty, this can be taken as evidence that our method is successful at compensating for the loss of transaction data.

5.1 Experimental Protocol

We proceeded as follows. From our dataset, we extracted the set of street segments having at least 10 recorded transactions. These are the street segments for which we consider that the distribution of prices can be estimated in a reliable way. Let us call this dataset D.

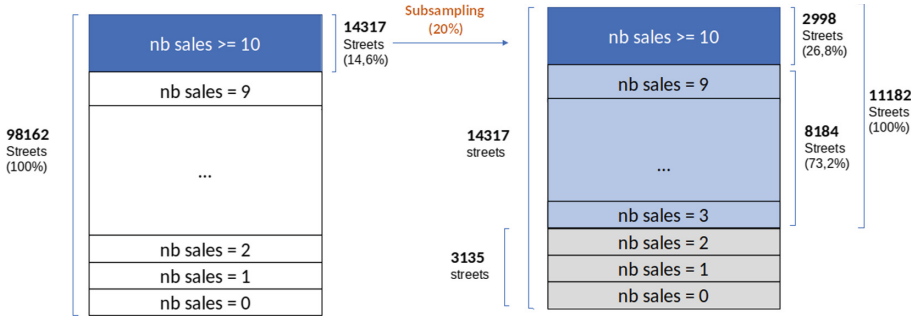

Fig. 1. A graphical illustration of the sampling mechanism used to create subsamples of the original dataset.

We then constructed a second dataset D' by randomly subsampling the transactions from the street segments of D, in such a way as to obtain a similar distribution of the number of transactions per street segment as in the full dataset (i.e., the dataset including also segments having fewer than 10 recorded transactions). For example, if in the full dataset 15% of the street segments has more than 10 recorded transactions (by the way, this subset of the full dataset is what we have called D), then also 15% of the street segments of D' has more than 10 transactions. In general, if the percentage of street segments in the full dataset that have n transactions is x, then D is samples so that the percentage of street

[1] Extracted from the PERVAL database, https://www.perval.fr/.

Table 1. Distribution of available transactions per street segment for different subsampling rates. Column "≥ 10" gives the number of street segments with at least 10 transactions, "$[3, 9]$" the number of street segments with 3 to 9 transactions, and so on.

Rate	≥ 10	$[3, 9]$	$[1, 2]$	$= 0$
100%	14,317	–	–	–
90%	13,404	913	–	–
80%	12,366	1,951	–	–
70%	11,125	3,190	2	–
60%	9,939	4,364	14	–
50%	8,527	5,695	95	–
40%	6,917	6,989	406	6
30%	5,055	8,050	1,149	63
20%	2,998	8,184	2,791	344
10%	814	6,013	5,610	1,880

segments in D' that have n transactions be as close as possible to x. This way, we can say that D' is to D as D is to the set of all street segments in the study area and, as a consequence, any observation about the predictive power on D of models trained on D' can provide, by proportional analogy, an indication of the predictive power on the full dataset of models trained on D. Figure 1 graphically illustrates this sampling mechanism and Table 1 provides some statistics about the dataset D' that we obtain depending on the chosen sampling rate.

Notice that the two datasets D and D' have the same size (in our case, 14,317) and consist of exactly the same street segments; what differs between them is the number of transactions available to estimate the distribution of prices for each street segment. In dataset D, the distribution is known precisely, and it has the form (d_1, d_5, d_9). In dataset D', instead, all is known is a possibility distribution π, as explained in Sect. 4.

5.2 Validation of the Possibility Distribution

To show that the possibility distribution π defined as per Eq. 5, as well as its associated possibility and necessity measures, does indeed qualitatively describe the actual price distribution, we studied the possibility (computed according to Eq. 6) of the observed median price ventile m of street segments. Ideally, $\Pi(m)$ should be 1 for every street segment, if π perfectly described how the prices are distributed.

Figure 2 shows the probability distribution of $\Pi(m)$ for different sampling rates of the set of transactions. We can observe that when the possibility distributions π of transaction prices for each street segment is constructed using all the recorded transactions (which are at least 10 for any one of the 14,317 street segments considered for this study), the median is assigned a possibility

of 1 for most street segments, with some exceptions, which, upon inspection, turned out to be street segments whose price distribution is not unimodal. Since the parametric family of distribution used to fit the actual price distributions is unimodal, the most likely values for their parameters are those that make d_5 correspond to one of the modes. This is an intrinsic limitation introduced by the particular choice of a unimodal family of distribution, which was made to simplify the geographical interpretation of the result; however, despite this limitation, the results of the study seem to confirm the validity of the method used to construct the possibility distributions.

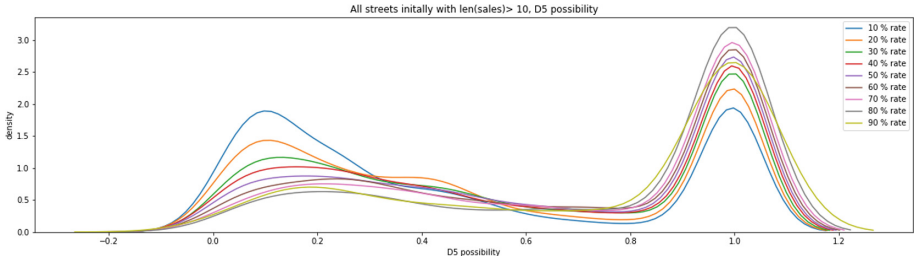

Fig. 2. Distribution of $\Pi(m)$ for different subsampling rates.

Unsurprisingly, as the sampling rate decreases, the number of street segments for which the median is assigned a high possibility decreases.

5.3 Empirical Test of the Method

To conduct our tests, we selected, among all possible predictive modeling methods, XGBoost [2], which is the one that gave the best results when applied to dataset D based on a critical comparison and benchmarking of the most popular methods available. The rationale of this choice is that we wanted our solution to "prove its mettle" on a very challenging task, namely to prevent the degradation of the accuracy of the strongest available method when the supervisory signal becomes uncertain.

We trained a model to predict the median of ventile prices (i.e., d_5) on dataset D by using XGBoost regression with the standard loss function and we trained another model on dataset D' by using XGBoost regression instrumented with the proposed possibilistic loss function. We compare the results given by these two models when applied to a test set consisting of street segments not used to train the models. We treat the model trained on dataset D as the ground truth and we measure the deviation of the model trained on D' from such a target.

As a measure of prediction error, we compute the RMSD of the median (d_5) predicted by the model trained on D' for each segment, with respect to the median of that segment in D. We used a sampling rate of 20% to generate D' from D, i.e., D' contains only 20% of the transactions available in D. We split D' into

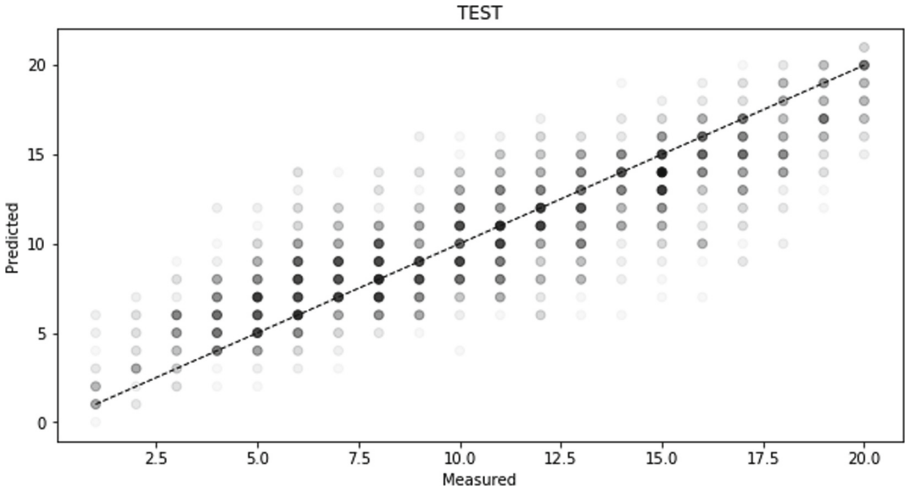

Fig. 3. Results obtained on the test set by applying XGBoost regression to D' with the possibilistic loss function.

a training set containing 80% of the street segments and a test set containing the remaining 20% of the street segments. After training the model on the training set, we obtain an RMSD of 2.778556 on the test set. Figure 3 shows a plot of predicted vs. actual price ventiles for the test set. For comparison, XGBoost using the standard loss function based on MSE trained on 80% of D (therefore, with full information), gives an RMSD of 2.283798 when tested on the remaining 20%. In other words, the possibilistic loss function allows the prediction error to increase by less than 22%, even though 80% of the transactions were removed from the dataset!

6 Conclusion

We have proposed a method based on possibility theory to leverage sparse data, which can be combined with any machine learning method to approach weakly supervised learning problems. The solution we propose constructs a possibilistic loss function, which can then be plugged into a machine learning method of choice, to account for an uncertain supervisory signal.

Our solution is much in the same spirit as the fuzzification of learning algorithms based on the generalization of loss function proposed in [11], in that it pursues model identification at the same time as data disambiguation, except that in our case ambiguity (i.e., uncertainty) affects the output variable only, which is only partially observable in the available data, while all the input variables are perfectly known and, thus, non-ambiguous. Furthermore, as in [11], the predictive model is evaluated by looking at how well its prediction fits the *most favorable* instantiation of the uncertain labels of the training data (this is the sense of the minimum operator in the possibilistic loss function definition).

The development of the method we presented has been motivated and driven by a very specific application, namely by the need to leverage sparse data in a human geography setting. However, its working principle is quite general and could be extended to suit other scenarios. Indeed, distilling a completely general method is the main direction for future work.

Acknowledgments. This research was funded by the IDEX "UCA JEDI", within the AAP Partenariat 2016 (action 6.5).

Andrea Tettamanzi has been supported by the French government, through the 3IA Côte d'Azur "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

The authors would like to thank Denis Overal, director of the R&D at Kinaxia, for his support and insightful suggestions.

References

1. Bouveyron, C., Girard, S.: Robust supervised classification with mixture models: learning from data with uncertain labels. Pattern Recogn. **42**(11), 2649–2658 (2009)
2. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 785–794. ACM, New York (2016)
3. Chiarazzo, V., Caggiani, L., Marinelli, M., Ottomanelli, M.: A neural network based model for real estate price estimation considering environmental quality of property location. Transp. Res. Procedia **3**, 810–817 (2014)
4. Denœux, T., Zouhal, L.M.: Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy Sets Syst. **122**(3), 409–424 (2001)
5. Dubois, D., Moral, S., Prade, H.: A semantics for possibility theory based on likelihoods. J. Math. Anal. Appl. **205**, 359–380 (1997)
6. Dubois, D., Prade, H.: Possibility Theory–An Approach to Computerized Processing of Uncertainty. Plenum Press, New York (1988)
7. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
8. Gerek, I.H.: House selling price assessment using two different adaptive neuro-fuzzy techniques. Autom. Constr. **41**, 33–39 (2014)
9. Ghosh, A., Manwani, N., Sastry, P.S.: Making risk minimization tolerant to label noise. Neurocomputing **160**, 93–107 (2015)
10. Hernández-González, J., Inza, I., Lozano, J.A.: Weak supervision and other non-standard classification problems: a taxonomy. Pattern Recogn. Lett. **69**, 49–55 (2016)
11. Hüllermeier, E.: Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization. Int. J. Approx. Reason. **55**(7), 1519–1534 (2014)
12. Hüllermeier, E., Destercke, S., Couso, I.: Learning from imprecise data: adjustments of optimistic and pessimistic variants. In: Ben Amor, N., Quost, B., Theobald, M. (eds.) SUM 2019. LNCS (LNAI), vol. 11940, pp. 266–279. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35514-2_20
13. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp. 4765–4774 (2017)

14. Moran, P.A.: Notes on continuous stochastic phenomena. Biometrika **37**(1/2), 17–23 (1950)
15. Nigam, N., Dutta, T., Gupta, H.P.: Impact of noisy labels in learning techniques: a survey. In: Kolhe, M.L., Tiwari, S., Trivedi, M.C., Mishra, K.K. (eds.) Advances in Data and Information Sciences. LNNS, vol. 94, pp. 403–411. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-0694-9_38
16. Park, B., Bae, J.K.: Using machine learning algorithms for housing price prediction: the case of Fairfax County, Virginia housing data. Expert Syst. Appl. **42**(6), 2928–2934 (2015)
17. Quost, B., Denœux, T.: Learning from data with uncertain labels by boosting credal classifiers. In: KDD Workshop on Knowledge Discovery from Uncertain Data, pp. 38–47. ACM (2009)
18. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River (2010)
19. Tsoumakas, G., Katakis, I.: Multi-label classification: an overview. Int. J. Data Warehouse. Min. **3**(3), 1–13 (2007)
20. Vannoorenberghe, P., Denœux, T.: Handling uncertain labels in multiclass problems using belief decision trees. In: Proceedings of the 9th International Conference on Information Processing and Management, pp. 1919–1926 (2002)
21. Venerandi, A., Fusco, G., Tettamanzi, A., Emsellem, D.: A machine learning approach to study the relationship between features of the urban environment and street value. Urban Sci. **3**(3), 25 (2019)
22. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
23. Zhang, Z., Guo, W., Li, M., Yu, W.: GIS-supervised building extraction with label noise-adaptive fully convolutional neural network. IEEE Geosci. Remote Sens. Lett. 1–5 (2020). https://doi.org/10.1109/LGRS.2019.2963065