Confidence path regularization for handling label uncertainty in semi-supervised learning: use case in bipolar disorder monitoring

Kamil Kmita
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
kmita@ibspan.waw.pl
0000-0001-8829-2420

Gabriella Casalino
Computer Science Dept.
University of Bari Aldo Moro
Bari, Italy
gabriella.casalino@uniba.it
0000-0003-0713-2260

Giovanna Castellano
Computer Science Dept.
University of Bari Aldo Moro
Bari, Italy
giovanna.castellano@uniba.it
0000-0002-6489-8628

Olgierd Hryniewicz
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
hryniewi@ibspan.waw.pl
0000-0001-9877-508X

Katarzyna Kaczmarek-Majer
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
k.kaczmarek@ibspan.waw.pl
0000-0003-0422-9366

***Abstract*—Semi-supervised learning has gained great interest because of its ability to combine unlabeled data with – potentially few – labeled observations in a training process. However, in some application contexts, one can question whether all available labels are equally valid. For example, in the context of bipolar disorder (BD) remote monitoring, a common practice is to extrapolate the psychiatrist's assessment onto some fixed time window surrounding the visit, the so-called ground truth period. In consequence, all data from this period are labeled with the same category. Such an approach may potentially result in misguided supervision affecting the model's performance. In this paper, we consider the problem of label uncertainty, assuming that the labels are crisp, but they may be assigned to particular observations with varying confidence. We propose a novel method called Confidence Path Regularization (CPR) that incorporates this uncertainty into the fuzzy c-means semi-supervised learning. The proposed CPR approach is a novel method for automatic, data-driven handling of label uncertainty. We achieve it by estimating the confidence factor for each labeled observation. In addition, CPR allows for the exploration of potential class-specific patterns in the adjusted confidence. The proposed method is illustrated with experiments on partially labeled data about speech characteristics collected from smartphone application for BD monitoring. In this particular applied scenario, we also use additional contextual data to improve the construction of confidence paths. It is shown that the proposed CPR approach enables to reflect the varying confidence in labels as compared with the nominal approach which assigns the majority of observations to the same class associated with relevant ground truth period.**

***Index Terms*—semi-supervised learning, prediction, label uncertainty, weak learning, regularization, bipolar disorder, process monitoring, acoustic features, smartphones, intelligent data analysis**

I. Introduction

Semi-supervised learning has been increasingly attracting interest from researchers in recent years. It is a setting classified between two well-established fields of machine learning: 1) unsupervised learning, where one tries to extract information from unlabeled samples, and 2) supervised learning, where a goal is to infer a predictive model from labeled observations. Significant progress has been made in the development of learning schemes that leverage unlabeled samples by incorporating additional supervised data into the learning process. A common approach is to enhance clustering or fuzzy clustering with partial supervision [1], [2], leading to flexible semi-supervised approaches that have proven very useful in a wide range of practical tasks, e.g. in medical decision support [3], [4]. Adding even a relatively small amount of labeled data may improve the results of clustering.

In general, semi-supervised learning can be regarded as a particular case of a more general learning scheme known as weak-supervised learning [5], where supervision of a training instance is expressed as a vector which elements indicate the membership of the instance for each semantic category. This turns out to be useful when a crisp association of an instance to the class is difficult due to uncertainty and ambiguity.

Many motivating examples for this research come from the medical domain. In particular, in this work, we focus on the mental illness remote monitoring where it is assumed that the remotely collected behavioral data (e.g. voice characteristics from patient's phone calls) are assigned a label that is extrapolated from psychiatric assessment obtained during a medical appointment.

This label is extrapolated onto the so-called ground truth period, the fixed time window surrounding the visit (e.g. from 7 days before the visit up to 2 days after the visit). Nonetheless, even if we can assume that the majority of phone calls from the ground truth period indeed share some common characteristics of that disease phase, we do not have labels directly associated with these supervised calls. Uncertainty naturally arises about whether all calls should be equally treated as supervised to the same extent. In this paper, we focus on such uncertainty about the validity of the crisp labels.

To cope with label uncertainty, we propose a novel semi-supervised clustering method that takes into account uncertainty related to the labels or the labeling process itself. This method, called Confidence Path Regularization (CPR), extends the existing semi-supervised fuzzy c-means clustering algorithm by enabling a data-driven assessment of certainty associated with labels. We analyze a case study of bipolar disorder (BD) remote monitoring based on phone calls. Experimental results show that our CPR method enables differentiating between phone calls that are highly certain for the particular label and those that are not.

The structure of the paper is as follows. Section II describes the related work. Section III describes the main characteristics of the proposed Confidence Path Regularization approach. In Section IV, we discuss the obtained experimental results that illustrate the usefulness of the proposed approach in the BD application scenario. In Section V, the main conclusions are stated and future work is outlined.

II. Related work

Usually, data labeling relies on assigning each point unambiguously to a single class for the purpose of statistical learning. However, uncertainty in the data labeling regards different domains and affects automatic analyses. For example, a movie can be labeled to different genres, or a protein sequence could be assigned to several structural subcategories. Such uncertainty is common in many biological and medical applications [6]. Data annotation is often uncertain due to the intrinsic subjective nature of the labeling process [7]. Moreover, when different experts are involved to reduce process subjectivity, intra-observer and inter-observer variability issues are introduced in the analysis [8], leading to uncertainty.

Most methods, e.g. [9], [10], and literature surveys [8], [11], [12] interpret uncertain labels as labels corrupted by noise, stating that the problem is to reduce this noise. Several recent studies have shown the negative effects of training deep learning models with noisy annotations [13]. Existing methods propose to use conditional random fields [14] and neural networks [15] to achieve high-quality annotations by correcting the noise in the annotations. Other approaches [16], [17] apply resampling to the training samples and evaluate the importance of each sample during the training process by additional modules in order to obtain a more robust model.

Another approach to overcome the issues caused by noisy crisp labels is to consider fuzzy or uncertain labels so that samples are assigned to each label with some membership degree [18]. Probability and possibility theories are also used to handle uncertainty from another point of view. Examples include Bayesian Neural Networks for leveraging uncertain labels in Chest X-rays in [19] or a multivariate multinomial mixture model for DNA barcoding [20].

All the above methods assume a fully supervised scenario, while few papers discuss semi-supervised learning with label uncertainty. In [6] and [21] overclustering is used to detect sub-structures of uncertain labels in order to improve classification through deep learning. Soft labels are used in [22] to improve two semi-supervised multiple classifier frameworks. Uncertainty-aware pseudo-labels are proposed in [23], whilst a semi-supervised support vector regression based on self-training with label uncertainty is proposed in [24]. Data uncertainty and semi-supervision adjustments in clustering are also discussed in [25], where multiple fuzzification coefficients are applied to implement the semi-supervision component.

In this paper, we go beyond the state-of-the-art and enable to model the varying confidence in labels within the semi-supervised scenario of fuzzy c-means.

III. The proposed confidence path regularization in semi-supervised learning

The proposed method builds on the Semi-Supervised Fuzzy C-Means (SSFCM) algorithm proposed in [2]. Of many existing approaches to including partial supervision, this method allows for an intuitive and interpretable way to introduce label uncertainty. It shall be noted that authors of the original algorithm briefly discuss augmentation of their original method that we further develop into a wider framework.

The core SSFCM being an extension to the Fuzzy C-Means (FCM) algorithm aims at grouping observations $\mathbf{x}_j \in \mathbb{R}^p, j = 1, \ldots, N$ into K clusters $c_1, \ldots, c_K$. Contrary to hard clustering algorithms, each observation $\mathbf{x}_j$ may be allocated to more than one cluster. This is expressed by membership values $u_{jk} \in [0, 1]$: the greater the value of u_{jk}, the more observation $\mathbf{x}_j$ belongs to k-th cluster.

Partition matrix $U = [u_{jk}]$ must satisfy two conditions:

$$\sum_{k=1}^{K} u_{jk} = 1, \forall j, \tag{i}$$

$$0 < \sum_{j=1}^{N} u_{jk} < N, \forall k. \tag{ii}$$

The optimal allocation of observations is achieved by iteratively minimizing objective function J that quantifies distances between observations and prototypes of the clusters $\mathbf{v}_k \in \mathbb{R}^p$:

$$J = \sum_{k=1}^{K} \sum_{j=1}^{N} u_{jk}^m d_{jk}^2 + \alpha \sum_{k=1}^{K} \sum_{j=1}^{N} (u_{jk} - b_j f_{jk})^m d_{jk}^2. \tag{1}$$

Here, d_{jk} is the Euclidean distance between an observation and a cluster prototype, $F = [f_{jk}]$ is a matrix introducing partial supervision that contains assumed membership values, $b_j \in \{0, 1\}$ is an indicator variable equal to 1 iff $\mathbf{x}_j$ is labeled, m is a fuzzification coefficient, and $\alpha \geq 0$ is a scalar weighting the proportional contribution of partial supervision. As the parameter of the method, α must be provided a priori. Following [2], we assume:

$$\alpha = \frac{N}{\sum_{j=1}^{N} b_j}. \tag{2}$$

We also restrict $f_{jk} \in \{0, 1\}$ and set $m = 2$. The latter is a frequent assumption leading to convenient analytical properties (see [2]).

Label uncertainty can be incorporated into SSFCM by means of a confidence factor $\text{conf}_j \in [0, 1]$ defined as a level of confidence assigned to the actual membership grades. Then, the modified objective function J_c takes the following form:

$$J_c = \sum_{k=1}^{K} \sum_{j=1}^{N} u_{jk}^{m} d_{jk}^{2} + \alpha \sum_{k=1}^{K} \sum_{j=1}^{N} (u_{jk} - b_j f_{jk})^m \cdot \text{conf}_j \cdot d_{jk}^{2}. \tag{3}$$

Confidence factor is not a weight itself, but rather an observation-wise adjustment of α that results in modified weights $\alpha_j = \alpha \cdot \text{conf}_j$. Making note of that, we further use only conf_j notation to reason in terms of label uncertainty. Note that because of the assumption $\text{conf}_j \in [0, 1]$ the confidence factor can only decrease the impact of a supervised observation in the objective function J_c (3). We also restrict that $(b_j = 0) \implies (\text{conf}_j = 0)$ to include uncertainty only for the supervised observations. Let us define a new index $i = 1, \ldots, M$ indexing M supervised observations out of all N observations. This convention should simplify calculations that include only the supervised observations. Consequently, we will denote confidence factors for supervised data instances as conf_i.

In practice, one can rarely evaluate uncertainty upfront and arbitrary assumptions about exact values must be made a priori. To handle this problem, we introduce the Confidence Path Regularization (CPR) method to adjust the default confidence values in an automatic, data-driven way. This novel approach is parameterized with: a number of regularization passes R, a strength of each regularization $\text{reg}_r \in [0, 1]$, and a weight of each regularization pass' outcome $w_r \in \mathbb{N}$.

The proposed CPR method builds on a regularization assumption that highly certain supervised observations should be consistently assigned high degrees of membership to the supervised class by the SSFCM method across varying values of confidence factor.

We formulate this assumption by calculating $\text{conf}_i^{\star}$, the adjusted confidence factor calculated from a sequence of R SSFCM models fitted to the data with decreasing values of default conf_i, see (4). In each pass $r = 1, \ldots, R$, confidence factor for every observation conf_i is multiplied by the regularization factor $\text{reg}_r \in [0, 1]$ before fitting the model. Membership values of the supervised class obtained from R models are then weighted by scalars w_r. The smaller the reg_r, the greater the corresponding w_r should be to take into account the decreased strength of supervision. These are $\{\text{reg}_r\}_{r=1,\ldots,R}$ that form confidence regularization path, and together with weights w_r they become parameters of the method we propose.

The adjusted confidence factor $\text{conf}_i^{\star}$ is a result of following normalization:

$$\text{conf}_i^{\star} = \frac{1}{\sum_{r=1}^{R} w_r} \cdot \sum_{r=1}^{R} u_{i,s(i)}^{r} w_r, \tag{4}$$

where $s(i)$ is an index of column in F matrix for $i-$th observation that contains the ground truth-based label (i.e. $f_{i,s(i)} = 1$).

The proposed **Semi-supervised Fuzzy C-means with Confidence Path Regularization** algorithm is defined as follows:

```
input: data $X, F, \alpha, \{\text{conf}_i\}, \{\text{reg}_r\}, \{w_r\}$
output: $\{\text{conf}_i^{\star}\}$
for $r \in \{1, \ldots, R\}$ do
    for $i \in \{1, \ldots, M\}$ do
        $\text{conf}_i^{\text{reg}} = \text{conf}_i \cdot \text{reg}_r$
    end for
    $\text{model}_r = \text{SSFCM}(X, F, \alpha, \{\text{conf}_i^{\text{reg}}\})$
    persist $\{u_{i,s(i)}^{r}\}_{i=1,\ldots,M}$ membership values
        from $\text{model}_r$
end for
for $i \in \{1, \ldots, M\}$ do
    derive $\text{conf}_i^{\star}$ according to (4)
end for
return $\{\text{conf}_i^{\star}\}$
```

Details of the iterative algorithm optimizing the objective function J_c in SSFCM model fitting can be found in [2]. We only provide below the formula for updating the membership values in each iteration of SSFCM fitting:

$$u_{jk} = \frac{1}{1 + \alpha \cdot \text{conf}_j} \cdot \left(\frac{1 + \alpha \cdot \text{conf}_j \cdot \left(1 - b_j \cdot \sum_{s=1}^{K} f_{js}\right)}{\sum_{s=1}^{K} \left(d_{jk}^{2} / d_{js}^{2}\right)} \right) + \frac{1}{1 + \alpha \cdot \text{conf}_j} \cdot \left(\alpha \cdot \text{conf}_j \cdot f_{jk} \cdot b_j\right). \tag{5}$$

It is clear from (5) that the final value of u_{jk} is an interplay between the evidence coming from the observed data and the strength of supervision. This supports our weighting regularization assumption: even if we decrease the strength of supervision, observations that are highly representative of the supervised class should be still close to the corresponding cluster centers and achieve high membership values.

The proposed CPR approach can be applied either when (i) one does not assume anything about uncertainty, or (ii) contextual knowledge is used to assign different default conf_i values to different observations. In (i), $\forall i \; \text{conf}_i = 1$ is simply assumed, and CPR will yield modified confidence factors. The example of (ii) is psychiatric disease remote monitoring. Partial supervision is frequently obtained by extrapolating the label provided by the psychiatrist during a stationary visit onto the data collected during a period surrounding the visit. In such a case, one could assign gradually decreasing confidence values for the data collected further from the visit. CPR would then operate on such values and adjust them, potentially increasing $\mathrm{conf}_i^\star$ closer to 1 for these distant data that are estimated to be highly representative of the supervised class.

IV. Results

A. About dataset and labeling

We illustrate the performance of the proposed CPR method for real-life sensors and psychiatric data about bipolar disorder patients participating in a prospective observational study. The data were collected from a dedicated mobile application installed on patients' smartphones. The application recorded objective data, such as statistics of calls and text messages, and acoustic features of patients' speech. The latter were extracted from the signal with the OpenSmile [26] software installed on smartphones. In this work, we selected five voice characteristics that describe different types of jitter, shimmer, and spectrum of the signal (spectral flux and spectral centroid). Observations $\mathbf{x}_j \in \mathbb{R}^5$ summarise specific calls. In particular, each voice characteristic measure throughout the call was summarized by the mean value. We will further use interchangeably *calls* and *observations* to refer to $\mathbf{x}_j$ in our experiments.

Labels for partial supervision were obtained from psychiatric assessments performed during patient visits. In-depth diagnosis consisted, inter alia, of *CGI-BD* categorization of disease phase. Psychiatrists assigned one of $\{$depression, mixed, euthymia, dysfunction$\}$ labels at each visit that we treat as the supervised class. In addition, doctors expressed their opinion on how long the given phase has been present: $\{$days, weeks$\}$.

For the purpose of the experiments, we chose data from a single patient that had been assigned each of four *CGI-BD* disease phase categories throughout the study. In total, there were 1295 calls available. Extrapolation results for each visit are summarised in Table I. Note that the number of calls treated as supervised depends on the extrapolation strategy that we describe in detail below.

We considered two strategies when extrapolating labels from the day of psychiatric assessment to the calls surrounding the visit: baseline (BL) and extended (EXT). In the BL approach, all calls in an interval spanning from 7 days before a visit up to 2 days after the visit were assigned the label with the confidence factor equal to 1. In the EXT approach, we made use of the contextual information about the duration of the phase provided by psychiatrists. Owing to that, we considered different confidence values gradually decreasing over the days before the visit. A summary of confidence factors assigned in BL and EXT approaches for the experiments is presented in Table II. Each strategy (a, b, c, d, e) describes the rule for assigning conf_i values for a call falling into the respective time window in each of the approaches: BL, EXT (duration: days), and EXT (duration: weeks).

BL extrapolation technique is frequently assumed in the literature as the way of deriving ground truth period, see e.g. [27]. Let us note the binary character of this method. If a call is recorded 7 days before the visit, it is treated as a completely confident supervised observation ($\mathrm{conf}_i = 1$), whereas for a call recorded 8 days before the visit no supervision is applied at all.

Table I
Total number of supervised calls labeled in BL and EXT approaches.

Visit	Label	Duration[1]	BL # data	EXT # data
1	depression	days	58	76
2	mixed	days	55	68
3	euthymia	weeks	85	182
4	dysfunction	days	63	98
	total		261	424

[1] Duration of this mental state was assessed by the psychiatrist. Options available: days, weeks. Providing this information was not compulsory.

Table II
Summary of confidence factor values assigned in BL and EXT extrapolation procedures.

time window identifier	time window[1]		conf_i for BL	conf_i for EXT[2]	
	start	end		duration: days	duration: weeks
a	-3	2	1	1	1
b	-7	-4	1	0.5	0.5
c	-10	-8	0	0.5	0.5
d	-17	-11	0	0	0.5
e	-21	-18	0	0	0.25

[1] Time window is defined by start and end days relative to the visit day spanning the interval [start, end] for a given labeling strategy.
[2] EXT extrapolation approach differs for duration of the phase assessed by psychiatrist.

Contrary to such a procedure, the EXT approach we propose in this article allows for flexibility in terms of the initial confidence factor values. Calls that had been recorded a long time before the visit took place were getting lower conf_i values to quantify uncertainty intrinsically related to this extrapolation technique.

The strength of supervision for such calls was thus decreased as $\mathrm{conf}_i < 1$. However, this impact can be increased (or decreased) in a data-driven way as a result of the CPR procedure. It shows the potential of combining the EXT approach with the CPR procedure in handling label uncertainty in semi-supervised scenarios.

B. Confidence path regularization

We considered six experimental scenarios to assess CPR performance compared with non-adjusted semi-supervision. The scenarios differed in terms of:

(i) label extrapolation strategy: BL or EXT,
(ii) confidence factor: **nominal** conf_i values or CPR-driven **adjusted** $\mathrm{conf}_i^\star$ values,
(iii) parameter α: **default** α from (2) relevant for given extrapolation strategy (note that α depends on the number of calls treated as supervised), or **sensitivity** α assessing impact of changing α only. We perform sensitivity analyses only for BL extrapolation, setting α to the value

Table III
Illustrative example of confidence and assignment to clusters given $K = 4$ for two exemplary calls following BL nominal, BL adjusted, EXT nominal and EXT adjusted scenarios.

id	confidence	u_1	u_2	u_3	u_4	$s(i)$[1]
scenario 1) BL nominal (default $\alpha = 4.96$)						
359	$\text{conf}_{359} = 1$	0.08	0.87	0.03	0.02	2
1016	$\text{conf}_{1016} = 1$	0.99	0.01	0.0	0.0	1
scenario 2) BL adjusted (default $\alpha = 4.96$)						
359	$\text{conf}^{\star}_{359} = 0.8$	0.09	0.85	0.04	0.02	2
1016	$\text{conf}^{\star}_{1016} = 0.52$	0.95	0.04	0.01	0.0	1
scenario 3) EXT nominal (default $\alpha = 3.05$)						
359	$\text{conf}_{359} = 0.5$	0.18	0.7	0.07	0.05	2
1016	$\text{conf}_{1016} = 1$	0.98	0.02	0.0	0.0	1
scenario 4) EXT adjusted (default $\alpha = 3.05$)						
359	$\text{conf}^{\star}_{359} = 0.77$	0.06	0.82	0.09	0.04	2
1016	$\text{conf}^{\star}_{1016} = 0.42$	0.61	0.13	0.24	0.02	1

[1] $s(i)$ is an index pointing to the column from F matrix containing supervised class for a given call. For each entry in the table, $u_{i,s(i)}$ value was highlighted for greater readability.

obtained in EXT labeling approach (as more calls are considered supervised in EXT strategy).

While the type of confidence factor (ii) and the value of α (iii) are parameters of the SSFCM algorithm, the label extrapolation strategy (i) is quite arbitrary and depends on clinical assumptions made.

Below we describe six scenarios in more detail:

1) BL nominal (default α): BL scenario with nominal confidence factor values, $\alpha = 4.96$ set according to (2),
2) BL adjusted (default α): BL scenario with adjusted confidence factors using CPR method, $\alpha = 4.96$,
3) EXT nominal (default α): EXT scenario with nominal confidence factor values, $\alpha = 3.05$ set according to (2),
4) EXT adjusted (default α): EXT scenario with adjusted confidence factor values, $\alpha = 3.05$,
5) BL nominal (sensitivity α): similar to scenario 1, but with $\alpha = 3.05$ set to the corresponding EXT extrapolation approach. This scenario serves as a sensitivity assessment.
6) BL adjusted (sensitivity α): similar to scenario 2, but with $\alpha = 3.05$. This scenario serves as a sensitivity assessment.

Confidence regularization path consisted of $R = 3$ models with $\{\text{reg}_1 = 0.1, \text{reg}_2 = 0.5, \text{reg}_3 = 1\}$ and $\{w_1 = 10, w_2 = 2, w_3 = 1\}$. This choice is based on simple reasoning: if the confidence factor is decreased 10 times when multiplied by $\text{reg}_1 = 0.1$, then the corresponding weight w_1 should be 10. If conf_i is decreased twice by multiplying it by $\text{reg}_2 = 0.5$, then $w_2 = 2$. Finally, when no regularization happens ($\text{reg}_3 = 1$), then weight $w_3 = 1$.

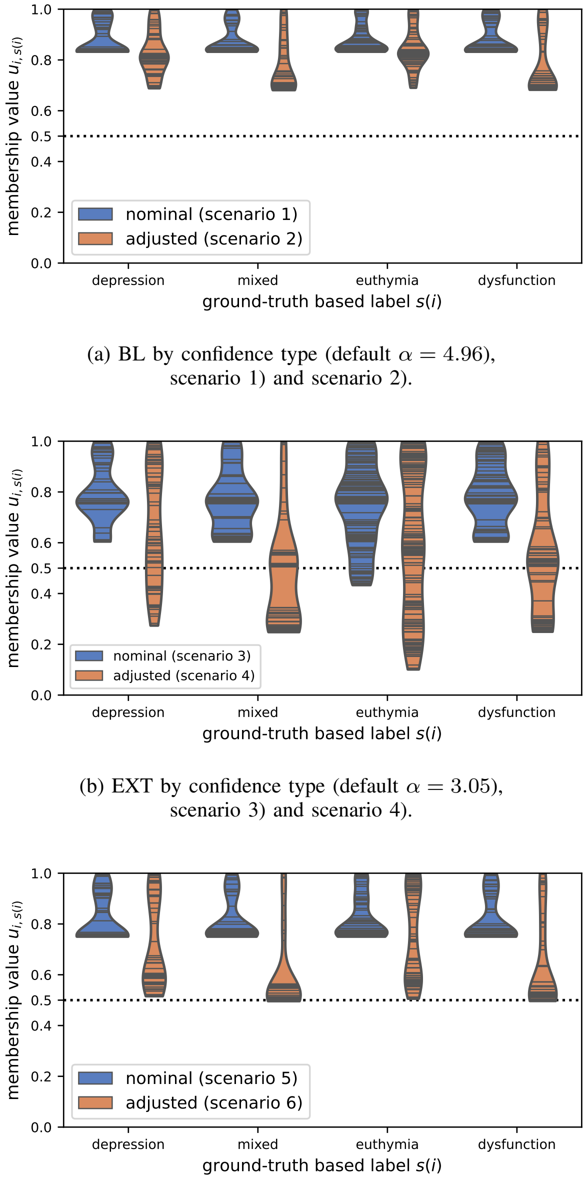
We provide an illustrative example of the experimental results for scenarios 1-4 (default α scenarios) in Table III. The table contains final membership values for two calls described by unique identifiers: 359 and 1016. In the EXT labeling approach, call 1016 was assigned $\text{conf}_{1016} = 1$, and call 359 was assigned $\text{conf}_{359} = 0.5$ since EXT approach treats calls further from the visit day with decreased conf_i. Call 359 was labeled as **mixed**, and call 1016 was labeled as **depression**. Relevant rows from estimated partition matrix U are displayed in Table III. In addition, the information about the confidence factor used in the corresponding SSFCM model fitted, the value of α, and the ground truth-based label are included in the table. Note that in this setting, entries f_{ik} in matrix F have values only in $\{0, 1\}$. Thus, the last column in Table III points to the relevant index $s(i)$ for $i-$ th observation such that $f_{i,s(i)} = 1$. Specifically, for call 359 $s(i = 359) = 1$, and for call 1016 $s(i = 1016) = 2$. The membership values of the supervised class $u_{i,s(i)}$ are highlighted in blue for greater readability.

In our experiments, we aim at discovering if scenarios that are not adjusted for label uncertainty are consistently leading to higher, less variable membership values than corresponding CPR-adjusted counterparts. In Table III, one can see that for call 1016, all scenarios apart from EXT adjusted (default α) scenario resulted in very high degrees of membership. It was only this scenario 4 that significantly decreased the membership grade – down to $u_1 = 0.61$ – providing insight that label certainty of this call may be questioned.

On the other hand, for call 359, all scenarios led to high membership values. It is worth noting the impact of CPR in case of EXT scenarios for this call: the default $\text{conf}_{359} = 0.5$ in scenario 3 was increased up to $\text{conf}^{\star}_{359} = 0.77$ in scenario 4, leading to higher degree of membership as well.

While our method allows for detailed comparisons on the level of individual calls, we focus on high-level evaluations in this paper. Building on the illustrative example, we would like to assess whether the CPR approach resulted generally in more variable degrees of membership related to the supervised classes. Fig. 1 presents overall distributions of such $u_{i,s(i)}$ membership values for M supervised calls by each scenario. We obtained them using `Python seaborn` package, specifically `seaborn.violinplot` function. It plots a combination of boxplot and kernel density estimate (KDE), allowing for effective comparison of multiple scenarios. In the experiments, we used `cut=0` option to prevent extending density past the observed values. We also plotted black sticks to present actual data points for greater clarity.

In all scenarios, CPR adjustment led consistently to higher variance and lower measures of central tendency. It shows that the CPR-adjusted confidence approach could differentiate between highly certain and uncertain supervised calls, while nominal procedure (without CPR-adjusting) was prone to taking the supervised labels for granted when estimating degrees of membership.



(a) BL by confidence type (default $\alpha = 4.96$), scenario 1) and scenario 2).

(b) EXT by confidence type (default $\alpha = 3.05$), scenario 3) and scenario 4).

(c) BL by confidence type (sensitivity $\alpha = 3.05$) scenario 5) and scenario 6).

Figure 1. Comparison of $u_{i,s(i)}$ membership values for 4 ground truth-based labels in each of scenarios 1-6. KDE distributions are presented by type of the confidence approach: nominal (conf_i) or CPR-adjusted ($\text{conf}_i^\star$).

Fig. 1a and Fig. 1c assess the impact of α on differences between nominal conf_i and adjusted $\text{conf}_i^\star$ in case of BL scenarios with different α: the default one, and the sensitivity one. Changing the default $\alpha = 4.96$ to sensitivity $\alpha = 3.05$ decreased the overall strength of supervision. In turn, higher variability was observed in both nominal and adjusted approaches. The adjusted approach responded stronger to this sensitivity analysis: the estimated distributions for each class in scenario 6 cover approximately two times wider range of membership values than in scenario 2, whereas for the nominal approach the increase in variability in scenario 5 compared with scenario 1 was relatively weaker.

Finally, Fig. 1b presents the true power of CPR. Let us recall that scenarios 3 and 4 had already some variability introduced into nominal conf_i values by the extrapolation strategy. Confidence factors of $0.25, 0.5$, or 1 were assigned based on the contextual knowledge. Also, there were more calls treated as supervised because of the mechanism of the EXT approach.

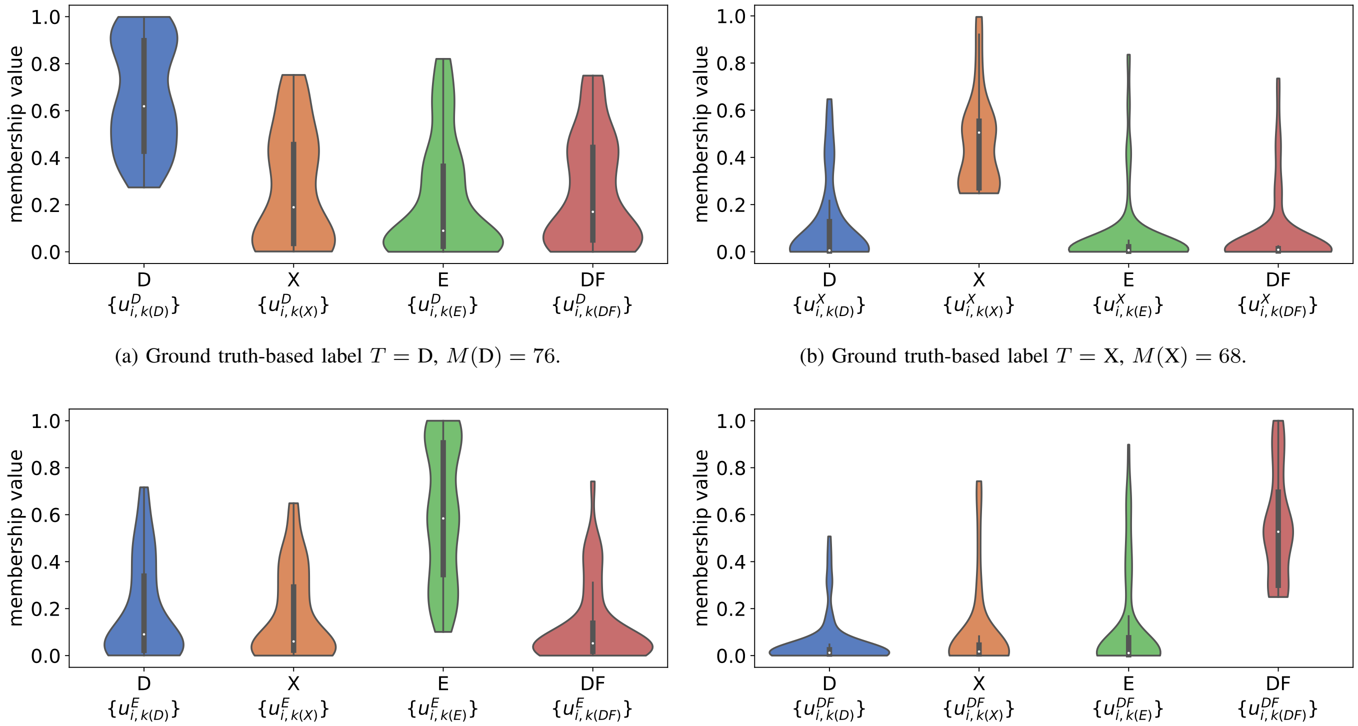
Considering EXT nominal approach (scenario 3), estimated distributions covered a wider range of degrees of membership compared with nominal BL scenarios. Note that only some calls labeled as *euthymia* were assigned membership values < 0.5. This is an important experimental result that justifies the need for confidence path regularization adjustment. Without it, barely any supervised calls were assigned a degree of membership < 0.5.

EXT adjusted (nominal α) scenario 4 generated membership degrees spanning wider range of values. Fig. 1b shows that there were many calls assigned values < 0.5 for every class. Our method enabled differentiation between supervised calls of high certainty and supervised calls of questionable certainty. An interesting observation is that for *mixed* and *dysfunction* categories the estimated degrees of membership tended to cluster together, with fewer observations getting very high membership values, while for *depression* and *euthymia* categories the estimated values were uniformly distributed across the whole spectrum of values.

This type of insight can lead to further exploration of inter-class specific patterns that we explore below. We focus on scenario 4 EXT adjusted (default α). We would like to compare distributions of membership values of the supervised class (i.e. $u_{i,k=s(i)}$) with membership degrees of the rest of the classes ($u_{i,k\neq s(i)}$) for each label category separately. Let $t \in \{$ depression (D), mixed (X), euthymia (E), dysfunction (DF) $\}$ denote a given label, and $k(t)$ an index of the column in F corresponding to the given label; $k(t) = 1, 2, 3, 4$ in our experiments. Let also T denote a ground truth-based label provided by partial supervision. Using this notation, $\{u^T_{i,k(t)}\}_{i=1,\dots,M(T)}$ defines a set of membership values of class $k(t)$ for $M(T)$ supervised calls that were labeled with category T. For example, $\{u^{T=\text{D}}_{i,k(t=\text{D})}\}$ denotes membership values of *depression* class for $M(\text{D})$ calls labeled as *depression* (D), and $\{u^{T=\text{D}}_{i,k(t=\text{X})}\}$ denotes membership values of *mixed* class for the same $M(\text{D})$ calls labeled as *depression*.

Fig. 1 compared membership values of the supervised class $u_{i,s(i)}$ across scenarios, while Fig. 2 compares membership values of the supervised class $\{u^T_{i,k(t=T)}\}$ with membership values of the non-supervised classes $\{u^T_{i,k(t\neq T)}\}$ across different label categories T.

KDE distributions presented in Fig. 2 explain how clearly membership values of the supervised class were separated from other classes. For example, Fig. 2b presents a clean separation for *mixed* label: these were mainly outliers from classes other than $t = X$ that achieved degrees of membership greater than approximately 0.25. A similar situation was



(a) Ground truth-based label $T = \text{D}$, $M(\text{D}) = 76$.

(b) Ground truth-based label $T = \text{X}$, $M(\text{X}) = 68$.

(c) Ground truth-based label $T = \text{E}$, $M(\text{E}) = 182$.

(d) Ground truth-based label $T = \text{DF}$, $M(\text{DF}) = 98$.

Figure 2. Each subfigure presents KDE plots of membership values $\{u^{T}_{i,k(t)}\}, i = 1, \ldots, M(T)$ for the supervised calls labeled as T. T denotes the ground-truth based label, $M(\text{T})$ the number of calls labeled as T, $k(t)$ denotes the index from matrix F corresponding to the given label t.

observed in Fig. 2d.

Contrary to that, Fig. 2a provides insight that supervised observations from *depression* ground truth period were harder to distinguish from other classes. Even though the distribution of $\{u^{D}_{i,k(D)}\}$ was characterized by higher measures of central tendency, the tails of distributions for other classes were heavy. In consequence, many observations were getting high degrees of membership to the classes other than *depression*. A similar pattern follows for Fig. 2c.

One could try to form a hypothesis based on the results presented in Fig. 2 that in this experimental setting, the ground truth-based labels of *depression* and *euthymia* were less separable in terms of label certainty, whereas *mixed* and *dysfunction* membership values clearly separated calls between highly representative of the relevant class and less so.

V. Conclusion and Further work

Uncertainty is intrinsically related to many labeling procedures, e.g. manual annotation of the data, or automatic extrapolation. For this reason, crisp labels may not always be suitable for statistical learning. Therefore, uncertain or soft labels have been proposed in the literature, and semi-supervised learning algorithms are gaining attention. Nonetheless, they often require a priori assumptions about the exact level of uncertainty. Therefore, we introduced the Confidence Path Regularization (CPR) method for automatic, data-driven handling of label uncertainty in the semi-supervised scenario. Specifically, the Semi-Supervised Fuzzy C-Means (SSFCM) clustering algorithm was extended into a wider CPR framework.

We presented a use case in bipolar disorder (BD) remote monitoring. In this medical context, data about voice characteristics collected from phone calls were gathered continually, allowing for patient monitoring. Partial supervision was implemented by extrapolating the psychiatrist's assessment obtained during the medical appointment onto calls from the surrounding period. SSFCM algorithm has already proven effective in detecting the onset of the new disease phase. However, it also introduced the inherent uncertainty and questions about viable limits of the ground truth period commonly applied in the literature.

In order to increase the number of labeled data while controlling for the uncertainty, we introduced a concept of *extended* extrapolation. Contrary to state-of-the-art, it assigns lower confidence to the calls that were recorded further from the visit. This way, more calls are treated as supervised with varying confidence about the adequateness of the label.

The main result of this paper is the Confidence Path Regularization algorithm which adjusts default confidence factors by fitting a series of models with varying regularization of the default confidence. The resulting membership values from all models are then weighted appropriately to reflect the strength of regularization in a given model, and normalized. Adjusted confidence factors obtained this way are used in the final model to estimate membership values automatically corrected for existing label uncertainty.

Results of the experiments on BD data showed that CPR-adjustment was able to differentiate between highly certain and uncertain supervised calls, while nominal procedures (without CPR-adjusting) were prone to taking the default ground truth-based labels for granted.

Future directions to improve the proposed approach include different applications and research ideas. The first of them is to extend the proposed approach to the online learning variants of semi-supervised learning, such as the incremental semi-supervised fuzzy clustering [28]. Another idea is to examine the differences between certain and uncertain supervised calls in the feature space to discover if any specific patterns exist. In practice, to establish monitoring of the phone calls for early detection of phase change, statistical process control approaches could use only highly certain calls when modeling control limits to be used. Finally, the concept of CPR seems flexible enough to be introduced in other learning frameworks than SSFCM.

Acknowledgment

Kamil Kmita and Katarzyna Kaczmarek-Majer received funding from Small Grants Scheme (NOR/SGS/BIPOLAR/0239/2020-00) within the research project: "Bipolar disorder prediction with sensor-based semi-supervised Learning (BIPOLAR)". Datasets considered in this paper were collected in the CHAD project "Smartphone-based diagnostics of phase changes in the course of bipolar disorder" (RPMA.01.02.00-14-5706/16-00) that was financed from EU funds (Regional Operational Program for Mazovia) in 2017-2018. The authors thank the researchers Karol Opara and Weronika Radziszewska from Systems Research Institute, Polish Academy of Sciences for their support in data preparation. Gabriella Casalino acknowledges funding from the Italian Ministry of Education, University and Research through the European PON project AIM (Attraction and International Mobility), nr. 1852414, activity 2, line 1.

References

[1] D. Mavroeidis, "Accelerating spectral clustering with partial supervision," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 241–258, 2010.

[2] W. Pedrycz and J. Waletzky, "Fuzzy clustering with partial supervision." *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics*, vol. 27, no. 5, pp. 787–95, 1997.

[3] K. Kaczmarek-Majer, G. Casalino, G. Castellano, O. Hryniewicz, and M. Dominiak, "Explaining smartphone-based acoustic data in bipolar disorder: Semi-supervised fuzzy clustering and relative linguistic summaries," *Information Sciences*, vol. 588, pp. 174–195, 2022.

[4] G. Casalino, G. Castellano, K. Kaczmarek-Majer, and O. Hryniewicz, "Intelligent analysis of data streams about phone calls for bipolar disorder monitoring," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2021, pp. 1–6.

[5] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[6] L. Schmarje, J. Brünger, M. Santarossa, S.-M. Schröder, R. Kiko, and R. Koch, "Fuzzy overclustering: Semi-supervised classification of fuzzy labels with overclustering and inverse cross-entropy," *Sensors*, vol. 21, no. 19, 2021.

[7] P. F. Culverhouse, R. Williams, B. Reguera, V. Herry, and S. González-Gil, "Do experts make mistakes? a comparison of human and machine indentification of dinoflagellates," *Marine ecology progress series*, vol. 247, pp. 17–25, 2003.

[8] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, 2020.

[9] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox, "Self: Learning to filter noisy labels with self-ensembling," *arXiv preprint arXiv:1910.01842*, 2019.

[10] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," *arXiv preprint arXiv:2002.07394*, 2020.

[11] G. Algan and I. Ulusoy, "Image classification with deep learning in the presence of noisy labels: A survey," *Knowledge-Based Systems*, vol. 215, p. 106771, 2021.

[12] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *arXiv preprint arXiv:2007.08199*, 2020.

[13] K. Yi and J. Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7017–7025.

[14] A. Vahdat, "Toward robustness against label noise in training deep discriminative neural networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] K.-H. Lee, X. He, L. Zhang, and L. Yang, "Cleannet: Transfer learning for scalable image classifier training with label noise," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5447–5456.

[16] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness, "Unsupervised label noise modeling and loss correction," in *International conference on machine learning*. PMLR, 2019, pp. 312–321.

[17] P. Chen, B. B. Liao, G. Chen, and S. Zhang, "Understanding and utilizing deep neural networks trained with noisy labels," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1062–1070.

[18] A. M. Aung and J. Whitehill, "Harnessing label uncertainty to improve modeling: An application to student engagement recognition." in *FG*, 2018, pp. 166–170.

[19] H.-Y. Yang, J. Yang, Y. Pan, K. Cao, Q. Song, F. Gao, and Y. Yin, "Learn to be uncertain: Leveraging uncertain labels in chest x-rays with bayesian neural networks." in *CVPR Workshops*, 2019, pp. 5–8.

[20] C. Bouveyron, S. Girard, and M. Olteanu, "Supervised classification of categorical data with uncertain labels for dna barcoding." in *ESANN*. Citeseer, 2009.

[21] L. Schmarje, J. Brünger, M. Santarossa, S. Schröder, R. Kiko, and R. Koch, "Beyond cats and dogs: Semi-supervised classification of fuzzy labels with overclustering," *CoRR*, vol. abs/2012.01768, 2020.

[22] M. M. El-Zahhar and N. F. El-Gayar, "A semi-supervised learning approach for soft labeled data," in *2010 10th International Conference on Intelligent Systems Design and Applications*. IEEE, 2010, pp. 1136–1141.

[23] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *arXiv preprint arXiv:2101.06329*, 2021.

[24] P. Kang, D. Kim, and S. Cho, "Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing," *Expert Systems with Applications*, vol. 51, pp. 85–106, 2016.

[25] T. D. Khang, M.-K. Tran, and M. Fowler, "A novel semi-supervised fuzzy c-means clustering algorithm using multiple fuzzification coefficients," *Algorithms*, vol. 14, no. 9, 2021.

[26] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the 21st ACM Int. Conf. on Multimedia*, 2013, pp. 835–838.

[27] M. Dominiak, K.Kaczmarek-Majer, A. Z. Antosik-Wojcinska, K. R. Opara, M. Wojnar, A. Olwert, W. Radziszewska, O. Hryniewicz, L. Swiecicki, and P. Mierzejewski, "Behavioural data collected from smartphones in the assessment of depressive and manic symptoms for bipolar disorder patients: Prospective observational study," *Journal of Medical Internet Research*, 2021.

[28] G. Casalino, G. Castellano, and C. Mencar, "Data stream classification by dynamic incremental semi-supervised fuzzy clustering," *International Journal on Artificial Intelligence Tools*, vol. 28, no. 08, 2019.