Combination of uncertain ordinal expert statements: The combination rule EIDMR and its application to low-voltage grid classification with SVM

Sebastian Breker
EnergieNetz Mitte GmbH, Kassel, Germany
Email: sebastian.breker@energienetz-mitte.de

Bernhard Sick
Intelligent Embedded Systems, University of Kassel, Germany
Email: bsick@uni-kassel.de

Abstract—The use of expert knowledge is always more or less afflicted with uncertainties for many reasons: Expert knowledge may be imprecise, imperfect, or erroneous, for instance. If we ask several experts to label data (e.g., to assign class labels to given data objects, i.e. samples), we often state that these experts make different, sometimes conflicting statements. The problem of labeling data for classification tasks is a serious one in many technical applications where it is rather easy to gather unlabeled data, but the task of labeling requires substantial effort regarding time and, consequently, money. In this article, we address the problem of combining several, potentially wrong class labels. We assume that we have an ordinal class structure (i.e., three or more classes are arranged such as "light", "medium-weight", and "heavy") and only a few expert statements are available. We propose a novel combination rule, the Extended Imprecise Dirichlet Model Rule (EIDMR) which is based on a k-nearest-neighbor approach and Dirichlet distributions, i.e., second-order distributions for multinomial distributions. In addition, experts may assess the difficulty of the labeling task, which may optionally be considered in the combination. The combination rule EIDMR is compared to others such as a standard Imprecise Dirichlet Model Rule, the Dempster-Shafer Rule, and Murphy's Rule. In our evaluation of EIDMR we first use artificial data where we know the data characteristics and true class labels. Then, we present results of a case study where we classify low-voltage grids with Support Vector Machines (SVM). Here, the task is to assess the expandability of these grids with additional photovoltaic generators (or other distributed generators) by assigning these grids to one of five ordinal classes. It can be shown that the use of our new EIDMR leads to better classifiers in cases where ordinal class labels are used and only a few, uncertain expert statements are available.

I. INTRODUCTION

In general, human judgment can be gathered either in a quantitative or in a qualitative way. Often, humans feel not comfortable with the task to express their opinion or belief quantitatively, because they worry that a concrete numeric value could give the (false) impression that there is more confidence in a judgment than they really have. Even for application experts it is difficult to cope with high-dimensional and complex tasks (see, e.g., [1]). But, using words instead of concrete values is also ambiguous and imprecise [2]. Thus, making a statement based on predefined ordinal classes (e.g., "low", "medium", and "high") can be seen as a reasonable trade-off between eliciting a quantitative and a qualitative judgment, for instance.

In the following, we assume that we want to gather knowledge of experts in an application task, and only refer to "experts" as human knowledge source. These domain experts assess data objects (samples), e.g., they label data using predefined ordinal classes to solve machine learning problems with supervised learning techniques. In addition, the experts may assess the difficulty of providing that label. This optional value can be seen as a kind of uncertainty of a single expert regarding that statement.

Basically, the statement of experts using an ordinal scale is influenced by many subjective parameters, e.g.,

1) experts have individual experience levels,
2) their forms of the day may vary,
3) they have different notions of "strictness", and
4) experts have an individual tendency not to opt for extremes.

On an abstract level, experience level and form of the day can be regarded as affecting the probability that a correct statement is provided by an expert. The individual use of strictness can be seen as a probability that the expert tends to rate a sample with a higher (or lower) class label compared to the true class. Furthermore, the individual tendency not to opt for extremes leads to a probability that an expert tends to assess a sample whose true class is near a "boundary" class (e.g., the "lowest" or "highest" class) with a label which is nearer to a "middle" class. Depending on this tendency, the bandwidth of the actually exploited ordinal classes may be low.

Assessing the difficulty of the labeling tasks, the fraction of statements rated as either being "easy" or "difficult" by an expert will be larger or smaller for an expert with a high experience level compared to a low experienced expert. Likewise, these fractions will differ for a single expert with different forms of the day. Because of all these reasons, a decision regarding an ordinal classification made by an expert is more or less uncertain. Hence, a final classification decision in a concrete application should not only be made using statements from a single expert. As experts might disagree, uncertain statements have to be combined in an appropriate way. The aim of all *combination rules* is to generate fused class labels which predict the true class with a higher accuracy than with the more uncertain individual class labels.

In this article, we present a novel combination rule for

class labels which is shown to be superior to some existing ones if (1) we have to cope with ordinal classes as discussed above and (2) the number of labels (expert statements) for single samples is quite low. This rule, the *Extended Imprecise Dirichlet Model Rule (EIDMR)*, is based on a k-nearest-neighbor ($k$nn) approach and Dirichlet distributions, i.e., second-order distributions for multinomial distributions. Basically, it extends a standard Imprecise Dirichlet Model Rule (IDMR) by considering (1) the class order and (2) additional, also uncertain labels for similar samples (measured in an appropriate feature space). The uncertainty accompanied with the combination can be determined with the help of two probability boundaries. The usage of the resulting uncertainty values for the combined statement is optional. They can be used, e.g., for a comparison of the uncertainty of different data objects or for a comparison to another expert group in a second classification approach. In addition, experts may assess the difficulty of the labeling task, which may optionally be considered by EIDMR.

The accuracy of different rules (including EIDMR, IDMR, the Dempster-Shafer rule, and Murphy's rule) in predicting a class label is first investigated using three artificial data sets for which the true classes are known. Then, we compare the combination rules using real data from the field of low-voltage grid classification. With combined class labels, we train support vector machine (SVM) classifiers. The empirical data contain 300 samples of rural and suburban low-voltage grids with ten features. They were gathered in an extensive grid survey and labeled according to the grids' hosting capacity for distributed generators (e.g., photovoltaic generators) by five experts in distribution grid planning.

The remainder of the article is organized as follows: In Section II, we first discuss related work and IDMR. Then, we present EIDMR and illustrate its properties with an example. In Section III, the accuracy of these combination rules is investigated in more detail using three artificial data sets and compared to two other well-known rules derived from Dempster-Shafer theory. Section IV presents the results of our study on the classification of low-voltage grids. Section V summarizes the key findings and sketches our future research.

II. Combination of Class Labels

In this Section, we propose the new EIDMR, which is based on Imprecise Dirichlet Models (IDM) and can be used to fuse uncertain ordinal class labels. We start with a presentation of the current state of the art (Section II-A). After that, the methodical foundations of IDM and IDMR are given in Section II-B. Next, we propose our new EIDMR in Section II-C. The application of IDMR and EIDMR is illustrated with a simple example in Section II-D.

A. Combination Rules – The State of the Art

Before we present the two combination rules which are based on IDM in a formal way, some important, general properties of combination rules are identified. Four of these requirements on combination rules are claimed in [3], [4] and

extended by another fifth requirement in [5]. In summary, these requirements are [5]:

1) irrelevance of order of statements in knowledge fusion (commutativity and associativity),
2) decrease of ignorance with an increasing number of statements,
3) concordance of knowledge increases the belief in a statement,
4) conflicting knowledge decreases the belief in a statement, and
5) persistent conflict is reflected.

Most well-known combination rules are based on Dempster-Shafer theory (DST) [6], e.g. the Dempster-Shafer Rule (DSR) [6] or Murphy's Rule (MR) [3]. A survey on combination rules based on DST has been presented in [7]. Most of the alternative DST based rules consider conflict redistribution. The counter-intuitive behavior concerning paradox problems [8] caused by compatible evidence is rarely considered [9]. Beside the DST based rules there exist some other rules which are derived from Dezert-Smarandache theory [10] and correspond to a non-Bayesian reasoning approach, e.g., the PCR5 [11]. Additionally, Smarandache et al. give several counterexamples where the DSR fails to provide coherent results (or provides no results at all) [12].

In summary, most of all known DST based rules do not fulfill—as they were not designated to—the additional fifth requirement. Because of this fact Andrade et al. [5] propose an IDM based rule (IDMR) and show that this rule accomplishes all of the properties concerning the above requirements. At this point, it should be mentioned that none of the mentioned known rules considers (1) the class order and (2) available uncertain labels for similar samples (measured in an appropriate feature space). We overcome these points with our EIDMR.

B. Methodical Foundations: IDM and IDMR

To motivate the IDM, we assume a set $\Omega = 1, \ldots, C$ of C mutually exclusive events, in our case a number of C different classes. Let the probability θ_c for the choice of a class c during a labeling process for a sample be denoted as element of a vector $\boldsymbol{\theta}$. Usually, these probabilities θ_c are unknown. If we suppose that an object is labeled by altogether N_E experts, the result can be denoted by the classes' occurrence frequencies $\boldsymbol{n} = (n_1, \ldots, n_C)^{\mathrm{T}}$ with $\sum_{c=1}^{C} n_c = N_E$. Furthermore, we model the likelihood of a result $\boldsymbol{n}$ with [13]:

$$p(\boldsymbol{n}|\boldsymbol{\theta}) = \prod_{c=1}^{C} (\theta_c)^{n_c}. \tag{1}$$

Because of the fact that $\boldsymbol{\theta}$ is usually unknown, we now use the Dirichlet distribution, which is the prior distribution in a Bayesian parameter estimation approach with hyperparameters h and $\boldsymbol{t} = (t_1, \ldots, t_C)^{\mathrm{T}}$, to model a distribution of vector $\boldsymbol{\theta}$ which is assumed to be a random variable itself [14]:

$$p(\boldsymbol{\theta}) = \frac{\Gamma(h)}{\prod\limits_{c=1}^{C} \Gamma(ht_c)} \cdot \prod_{c=1}^{C} (\theta_c)^{ht_c - 1} \tag{2}$$

with $h > 0$, $0 < t_c < 1$ for $c = 1, \ldots, C$, and $\sum_{c=1}^{C} t_c = 1$. Γ is the gamma function [15]. The vector $\boldsymbol{t}$ represents the prior knowledge about $\boldsymbol{\theta}$. In a next step, the combination of (1) and (2) according to $p(\boldsymbol{\theta}|\boldsymbol{n}) \propto p(\boldsymbol{n}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})$ yields the following posterior probability:

$$p(\boldsymbol{\theta}|\boldsymbol{n}) = \frac{\Gamma\left(\sum_{c=1}^{C}(n_c + ht_c)\right)}{\prod_{c=1}^{C}\Gamma(n_c + ht_c)} \cdot \prod_{c=1}^{C}(\theta_c)^{n_c + ht_c - 1}. \quad (3)$$

We get a Dirichlet distribution again, because a Dirichlet is the conjugate distribution of a multinomial distribution. The influence of the prior knowledge $\boldsymbol{t}$ on the posterior probability can be controlled with the hyperparameter h [13]. In a real classification application, there is often no information about the hyperparameters h and $\boldsymbol{t}$ available.

The idea of the IDMR is now, not only to consider a single one, but to start with a set of Dirichlet distributions for a fixed value of h [13]. Thereby, a set of Dirichlet distributions is chosen in such a way that $\sum_{c=1}^{C} t_c = 1$ is adhered [16]. Under these conditions, the following upper and lower bounds for the probability that a sample is assigned to class c can be determined by a maximization and minimization of t_c, respectively [14]. That is, for each $\boldsymbol{t}$, the corresponding Dirichlet distribution is updated using (1) [13]. If there is no prior information on $\boldsymbol{\theta}$ available, the bounds are determined with the one-sided limits $t_c \to 0$ and $t_c \to 1$ (because of the linear updating step) and result according to [16] in:

$$\underline{p}(c|n_c, N_E) = \frac{n_c}{N_E + h}, \quad (4)$$

and

$$\overline{p}(c|n_c, N_E) = \frac{n_c + h}{N_E + h} \quad (5)$$

where c is regarded here as the random variable for the combined statement (i.e., the fused class label). The hyperparameter h determines how quickly these upper and lower probabilities converge with an increasing number of observations. Because of this reason, Walley defined h as a number of observations needed to reduce the imprecision to half its initial value [16]. Typically, h is set to either 1 or 2 [16]. Using the derived bounds, the vector $\boldsymbol{t}$ needs not to be specified.

IDMR (not related to DST, cf. [5], [16]) is an alternative to the fusion rules based on DST. Andrade et al. used the probability boundaries from (4) and (5) to combine a collection of N_E classification statements c_j ($c_j \in \Omega$ is the statement of expert j, $j = 1, \ldots, N_E$) among which some experts vote for class c.

Above, n_c is the number of experts that assign a given sample to class c. With N_E experts, we have $0 \leq n_c \leq N_E$. This approach does not consider a difficulty assessment of the experts. To account for this, we introduce weights w_j for each classification statement c_j of an expert (with $w_j \geq 1$, as suggested in [5]). Then, we need an indicator function $I_{j,c}$ which is 1 if expert j assigns label c to the sample under

consideration (i.e., $c_j = c$), and 0 otherwise. We replace the number of experts that assign a given sample to a class c by

$$n_c = \sum_{j=1}^{N_E} w_j \cdot I_{j,c}. \quad (6)$$

Thus, the higher the weight w_j of an expert, the more influence has the corresponding statement. With this n_c, we may determine the boundaries according to (4) and (5). The boundaries can be interpreted similar to belief and plausibility in DST. Their difference depends on the number of classification statements with their weights and can be compared to the degree of ignorance in DST.

In an application of IDMR, the boundaries can be used in different ways. If we want to come to a sharp decision for a certain class, we may act carefully and consider the lower boundary $\underline{p}(c|n_c, N_E)$. Alternatively, we may choose class

$$c' = \arg\max_c n_c, \quad (7)$$

i.e., the highest number of weighted observations. Then, the uncertainty of the decision can be determined by analyzing the differences $\overline{p}(c|n_c, N_E) - \underline{p}(c|n_c, N_E)$. The uncertainty values can be used, e.g., for a comparison of the uncertainty for different samples or for a comparison to another expert group in a second labeling approach.

C. EIDMR

In a next step, we propose a new combination rule EIDMR which is based on the same idea as IDMR. But, the key idea is to decrease the uncertainty of the fused result by means of additional information which is available from similar samples. Here, "similar" means that their feature vectors are similar, determined with an appropriate similarity measure (which may be based on a metric). To choose these additional samples, the IDMR is extended by a k-nearest neighbor technique ($k$nn) applied in the feature space. Then, the expert statements for the sample under consideration and the statements of its k nearest neighbors are considered. The application of $k$nn ensures that only samples which have a characteristic that is similar to the one of the considered sample are involved in a weighted combination. Assuming that N_E statements are available for each sample, $(k+1) \cdot N_E$ statements are considered by EIDMR to assess one specific sample. EIDMR weights the statements not only depending on the experts' difficulty assessment as described above, but also depending on differences between the class assignments for the sample under consideration and its k neighbors. EIDMR also considers the fact that we have ordinal classes: Larger differences in class numbers lead to lower weighting factors.

To describe the approach in a formal way, we first need an additional index i for several variables as we have to consider not only one sample, but also its k neighbors. Thus, the sample under consideration gets the index 0, and its neighbors $i = 1, \ldots, k$. Then, for a sample i ($i = 0, \ldots, k$) we have N_E expert statements (assignments to classes) summarized in $\boldsymbol{c}_i = (c_{i,1}, \ldots, c_{i,N_E})^{\mathrm{T}}$ with corresponding difficulty assessments

$\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,N_E})^{\mathrm{T}}$. As we have ordinal classes, the dissimilarity between a vector $\boldsymbol{c}_i$ and the vector $\boldsymbol{c}_0$ can be measured using a standard metric, e.g., the Euclidean distance $\|\boldsymbol{c}_i - \boldsymbol{c}_0\|$. Then, we are able to define weights for samples g_i (with $i = 0, \ldots, k$) that consider differences in their class assignments as follows:

$$g_i = \frac{1}{k} \cdot \left(1 - \frac{\|\boldsymbol{c}_i - \boldsymbol{c}_0\|}{\sum_{l=1}^{k} \|\boldsymbol{c}_l - \boldsymbol{c}_0\|} \right). \tag{8}$$

To avoid dividing by zero we have to exclude the very special case where all class vectors $\boldsymbol{c}_i$ are equal (in this case we would proceed with $g_0 = 1$ and $g_i = 0$ for $i = 1, \ldots, k$). These weights have the property $\sum_{i=0}^{k} g_i = 1$.

With this spadework, we formally extend the two probability boundaries in the following way:

$$\underline{p}(c|n_{c,0}, \ldots, n_{c,k}, N_E) = \frac{\sum_{i=0}^{k} g_i \cdot n_{c,i}}{\sum_{c=1}^{C} \sum_{i=0}^{k} g_i \cdot n_{c,i} + h}, \tag{9}$$

$$\overline{p}(c|n_{c,0}, \ldots, n_{c,k}, N_E) = \frac{\sum_{i=0}^{k} g_i \cdot n_{c,i} + h}{\sum_{c=1}^{C} \sum_{i=0}^{k} g_i \cdot n_{c,i} + h} \tag{10}$$

with

$$n_{c,i} = \sum_{j=1}^{N_E} w_{i,j} \cdot I_{i,j,c}, \tag{11}$$

where now $I_{i,j,c}$ indicates whether sample i has been assigned to class c by expert j (cf. 6 above).

Such as with IDMR, the probability boundaries can be directly used or a sharp decision for a single class can be made by choosing c' with

$$c' = \arg\max_c \sum_{i=0}^{k} g_i \cdot n_{c,i} \tag{12}$$

That is, class c with the highest number of weighted observations is chosen as the combined label. The uncertainty accompanied by this decision can be determined with $\overline{p}(c|n_{c,0}, \ldots, n_{c,k}, N_E) - \underline{p}(c|n_{c,0}, \ldots, n_{c,k}, N_E)$. The uncertainty values may vary more likely between different samples using EIDMR because of the additional k samples and the additional weights g_i.

The EIDMR procedure is set out in Algorithm 1.

D. Illustrative Example

To illustrate EIDMR and differences to IDMR, we now elaborate a simple example with three samples in a feature space as shown in Figure 1. We assume that each sample was classified by two experts into one of the three ordinal classes 1, 2, and 3. Furthermore, we suggest that a difficulty assessment is known for each statement (either "easy" (with weight $w_{i,j}$ = 1.5), "medium" ($w_{i,j}$ = 1.25), or "hard" ($w_{i,j}$ = 1.0). In

---

Algorithm 1: EIDMR algorithm for one sample.

Input: set of samples with feature vectors, experts' classification statements, and difficulty assessment; sample 0 that has to be labeled.

1 Search k nearest neighbors of sample 0 in the feature space with a Euclidean distance measure;
2 **for** $i = 0$ *to* k **do**
3   Compute similarity weights g_i;
4 **for** $c = 1$ *to* C **do**
5   compute lower boundary $\underline{p}(c|n_{c,0}, \ldots, n_{c,k}, N_E)$;
6   compute upper boundary $\overline{p}(c|n_{c,0}, \ldots, n_{c,k}, N_E)$;
7 choose class c' with $c' = \arg\max_c \sum_{i=0}^{k} g_i \cdot n_{c,i}$;

Output: class c' and upper and lower boundaries $\overline{p}$ and $\underline{p}$.

---

the figure, the classification statements and weights are set out as vectors $\boldsymbol{c}_i$ and $\boldsymbol{w}_i$ ($i = 0, 1, 2$). It can be seen that the weights of the experts' statements are lower for the data sample 0 under consideration (red) compared to the weights of its two nearest neighbors (green). This simulates the case when a sample is hard to classify and there exist some similar samples for which the classification task is easier.
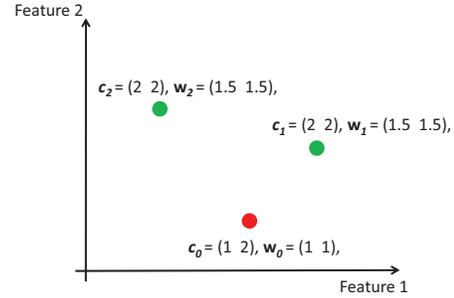

Fig. 1. Sample 0 and its $k = 2$ neighbors in the feature space.

Using IDMR for sample 0 results in $n_1 = 1$, $n_2 = 1$, and $n_3 = 0$. Obviously, a clear decision concerning the combined class label can not be made. The values for the lower and upper boundaries are $\underline{p}(1) = \underline{p}(2) = 1/3$ and $\overline{p}(1) = \overline{p}(2) = 2/3$ for the classes 1 and 2, as well as $\underline{p}(3) = 0$ and $\overline{p}(3) = 1/3$ for class 3 if we use $h = 1$. That is, the uncertainty regarding the class decision—estimated by the difference of the boundaries—is $1/3$ for each class.

Now, we proceed with EIDMR. First, the three dissimilarities of samples are computed with the Euclidean distance: $\|\boldsymbol{c}_0 - \boldsymbol{c}_0\| = 0$, $\|\boldsymbol{c}_1 - \boldsymbol{c}_0\| = 1$, and $\|\boldsymbol{c}_2 - \boldsymbol{c}_0\| = 1$. In a next step, the similarity weights are computed: $g_0 = \frac{1}{2}(1 - 0) = \frac{1}{2}$, $g_1 = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}$, $g_2 = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}$. Then, we calculate the $n_{c,i}$ using (11): $n_{1,0} = 1$, $n_{2,0} = 1$, $n_{2,1} = 3$, $n_{3,2} = 3$. All other $n_{c,i}$ are 0 in this example. These values finally lead to the class decision $\arg\max_c \sum_{i=0}^{k} g_i \cdot n_{c,i} = \arg\max_c \left\{ \frac{1}{2}, 2, 0 \right\} = 2$. Hence, the additional information of the two nearest neighbors considered in EIDMR leads

to the rather clear decision that class 2 should be taken as label. The computation of the boundaries with $h = 1$ yields $\underline{p}(1) = 1/7$, $\overline{p}(1) = 3/7$, $\underline{p}(2) = 4/7$, $\overline{p}(2) = 6/7$, and $\underline{p}(3) = 0$, $\overline{p}(3) = 2/7$. In comparison to IDMR, the uncertainty is reduced to $2/7$.

Altogether, EIDMR leads to a clearer class decision with reduced uncertainty in this very simple example.

III. Evaluation on Artificial Data Sets

In this section, the accuracy of the combination rules proposed in Section II-A is investigated and compared to two further well-known rules, DSR and MR, which are derived from DST, on three artificial data sets for which the true classes are known. The generation of the artificial data sets that we used in our experiments is described in Section III-A. Then, the simulation of artificial experts is sketched in Section III-B. The accuracy of the four combination rules on these artificial data is investigated in Section III-C.

A. Generation of Artificial Data Sets

Before we apply the four combination rules to a real data set, we want to assess these rules with regard to their ability to detect the true classes under different conditions. To do this, we generated three data sets which contain 1000 samples with five ordinal classes. The feature space consists of two dimensions which allows for an illustrative presentation of the data sets. The samples were generated with the use of Gaussian mixture models (GMM) with five components, each assigned to one of the five ordinal classes. All mixture coefficients of the GMM were set to 0.2. Thus, each class label occurred approximately 200 times in a data set. The three data sets were generated with similar GMM but differ regarding the location of the samples in the feature space, because the distances of the expectations (mean values) are highest in data set I and lowest in data set III. That is, the overlap of the component densities increases from data set I to III and, thus, the samples belonging to different classes become more difficult to separate. In Figure 2, the data sets with highest and lowest overlap are set out.

B. Generation of Artificial Expert Statements

Having prepared the data sets, we generated 30 artificial, ordinal expert statements (class labels) for every sample by altering the true class labels with a random process according to the influences identified in Section I. Each of the influences was modeled separately, and, hence, could be chosen individually for every simulated expert. Particularly, every simulated expert was allotted an individual level of experience (the form of the day was not modeled separately), an individual notion of strictness, and an individual tendency not to opt for extremes. Each of these influences leads to a probability that a simulated expert labels a sample with an altered (i.e., wrong) class. An expert with a distinct notion of strictness will, e.g., have a high probability to choose a class label which is lower than the true class. To illustrate the results of this process, the statements

for the samples in data sets I and III are shown for one of the experts in Figure 3.

Additionally, the generation of a difficulty statement reflecting the certainty of an expert concerning the label for a specific sample was implemented with three levels "easy", "medium", and "difficult" (with numerical values of 1.00, 1.25, and 1.50, respectively, cf. also the example in Section II-D). For the application of the DSR and MR, simple support functions (cf. [5]) were used. We had to transform the difficulty values to an appropriate interval by multiplying with 0.6 (i.e., to 0.60, 0.75, and 0.90).

C. Properties of the Combination Rules

In this section, we apply each of the four combination rules DSR, MR, IDMR, and EIDMR to the artificial data sets to combine the experts' label statements. To measure the degree of agreement between the results of the combination rules and the true labels, the *inter-rater agreement* is used as an accuracy measure. *Cohen's weighted kappa statistic* κ_w is a standard measure for this purpose [17]. It allows for the use of weights to reflect the extent of similarity between ordinal classes. Hence, it incorporates the magnitude of each disagreement and provides partial credit for disagreement when agreement is not perfect [18]. The two most widely used weighting schemes are symmetric linear weights and symmetric quadratic weights [19]. For our analysis, we use symmetric linear weights. Cohen's weighted kappa statistic κ_w can be interpreted as a chance-corrected index of agreement. It yields 1 for a perfect agreement. If κ_w is 0, the agreement is equal to that expected under independence. A negative value indicates that the agreement is less than expected by chance [20].

To apply EIDMR, the data had to be scaled using a z-transform because features with larger values would dominate those with smaller values in the $k$nn approach [21]. This scaling is not needed for the other three combination rules. In our first experiment, the number of additionally considered samples in EIDMR was set to $k = 3, 5, 7$, or 9. In order to investigate the influence of the number of experts on the accuracy, we increased this number step-by-step. To obtain statistically significant results, we repeated the generation of the artificial class labels and the application of the combination rules ten times for each data set. The fact that the true classes are known for the generated data sets enables us to evaluate the accuracy of the rules.

In Figure 4, the mean values $\mu(\kappa_w)$ and the standard deviations $\sigma(\kappa_w)$ for the ten repetitions of the experiment are outlined for data set I. Different values of k have been used applying EIDMR. It can be seen that EIDMR outperforms the other combination rules for every considered number of experts and k with regard to $\mu(\kappa_w)$, and, thus, with regard to the accuracy in detecting the true classes. This is due to the additional statements in the $k$nn approach. Because of the low overlap of the Gaussian components there are no differences in the combination accuracy no matter which of the considered values is used for k if there is more than one expert statement
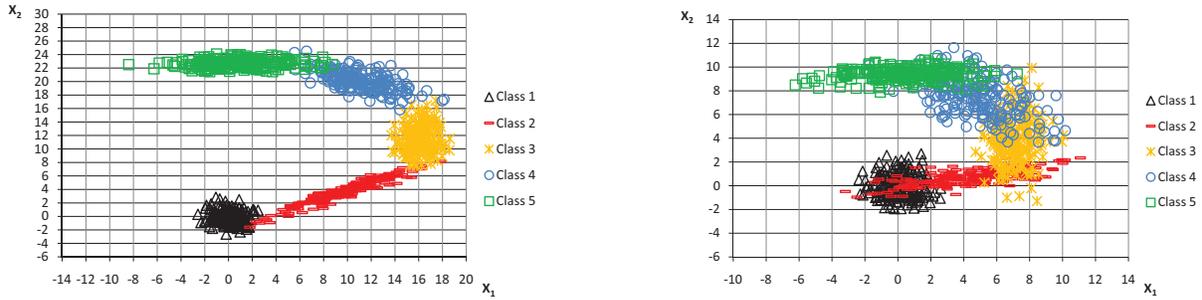
Fig. 2. Two of the data sets, I (left) and III (right), generated with different GMM (samples with true labels).
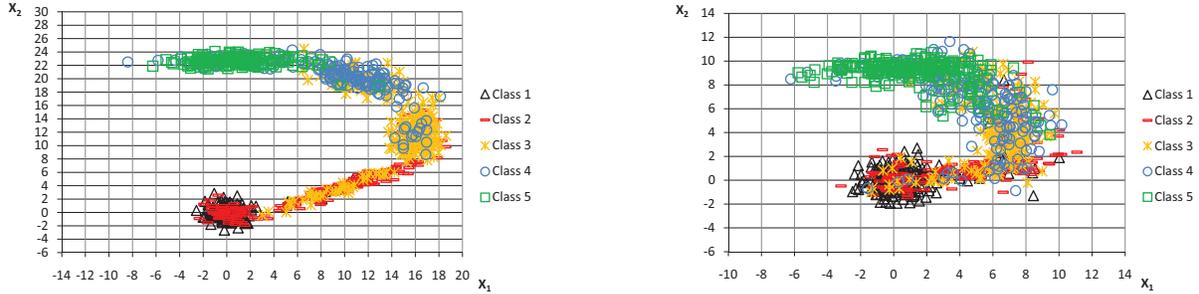

Fig. 3. Statements of an artificial expert for samples in data set I (left) and III (right).
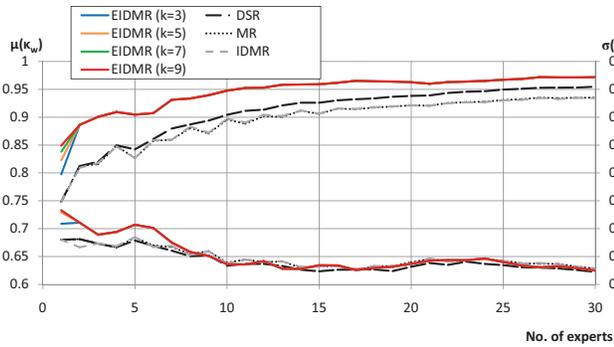

Fig. 4. Combination accuracy for data set I: mean (top) and standard deviation (bottom).
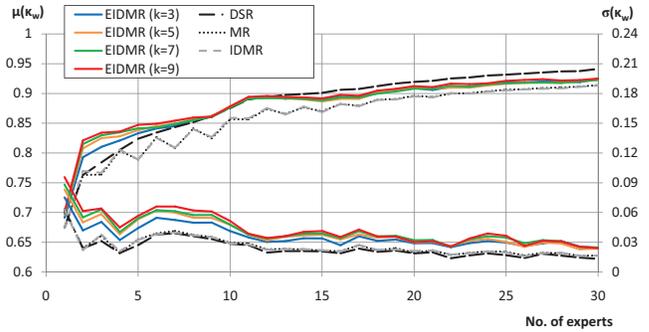

Fig. 5. Combination accuracy for data set II: mean (top) and standard deviation (bottom).

available for a sample. In comparison to the other rules, the standard deviations of EIDMR are higher if the number of experts is lower than 8. This can be reasoned with the influence of the often differing expert statements gathered with the $k$nn approach. The results of MR and IDMR coincide strongly, only minor differences can be stated. This is reasoned by the combination process using MR where an average belief function is computed and then combined $N_E - 1$ times using DSR. This yields very similar results compared to IDMR.

Results for data set II are shown in Figure 5. Compared to the other two data sets, this data set is based on a medium overlap of the five components in the feature space that are assigned to different classes. Comparing the three combination rules DSR, MR, and IDMR, there are no significant differences in their accuracies if the number of experts is lower than five. As for data set I, the curves of MR and IDMR coincide

strongly. Due to additional information gathered with the $k$nn approach, EIDMR outperforms the other rules if the number of fused statements (i.e., experts) is lower than 12. However, the standard deviations are higher, too. The highest accuracy using EIDMR is reached with $k = 9$. If the number of experts exceeds 12, DSR performs best in detecting the true classes. Thus, the additional benefit of the $k$nn approach decreases with an increasing number of experts.

Figure 6 outlines the results for data set III which is the data set with the highest overlap of components. In the cases with a low number of expert statements (i.e., where the number of experts is at most 3), EIDMR outperforms the other rules. In comparison to the other data sets, a variation of k has the highest influence on the combination accuracy. The highest accuracy using EIDMR is reached for $k = 3$. For an increasing number of experts (> 3), the influence of the $k$nn approach leads to a lower accuracy of EIDMR compared to the other
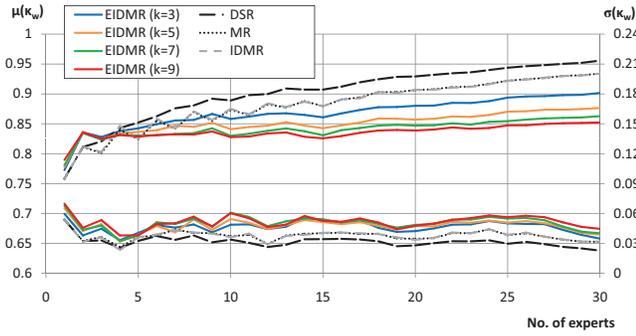
Fig. 6. Combination accuracy for data set III: mean (top) and standard deviation (bottom).

combination rules. Both effects can be reasoned by the high overlap of the components of the mixture model underlying data set III. Then, DSR has the highest average accuracy and also the lowest standard deviation of all combination rules.

We now investigate some properties of EIDMR in two additional experiments.

In our second experiment, we varied the number of samples in data set III to examine the influence of the sample density on the combination accuracy of EIDMR. This is a very interesting aspect as we can expect that the sample density influences the values of κ_w. Figure 7 outlines the results for 250, 750, and all 1000 samples contained in data set III which result from using EIDMR with $k = 3$ and 9. It can be seen that the mean $\mu(\kappa_w)$ (ten repetitions again, see above) actually increases significantly with the number of samples. The increasing sample density will lead to a selection of more similar k samples in the $k$nn approach.
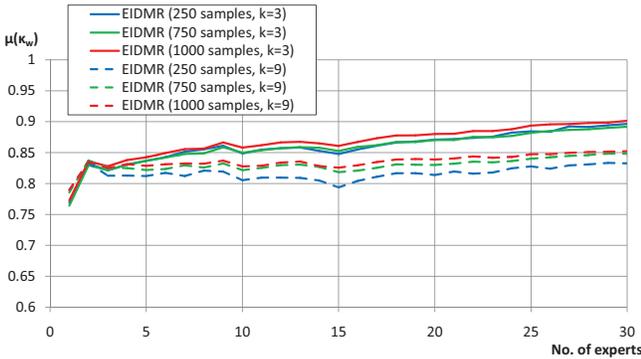

Fig. 7. Combination accuracy for different sample numbers (data set III).

In our third experiment, we analyzed the influence of the difficulty weights $w_{i,j}$ and the use of information about the class order in EIDMR. To do this, we may set the value of each weight $w_{i,j}$ to one which is equivalent to ignoring the experts' difficulty assessments for each sample. To ignore the class order information contained in the class labels, we may set the value of each similarity g_i to $1/(k + 1)$. In this experiment, we investigate both cases and also their combination. The latter case leads to a simple majority decision considering the

$k + 1$ considered samples with EIDMR. The results, outlined in Figure 8, show that both, the difficulty weights and the class order information, have a noticeable positive effect on the combination accuracy.
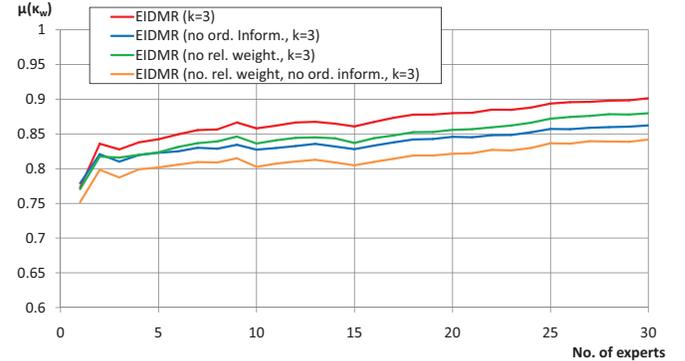

Fig. 8. Influence of class order information and difficulty weights on combination accuracy for data set III.

Considering the results of our experiments for all three data sets, the DSR has a strong performance with regard to the detection of the true class labels for every data set. However, the combination accuracy can be further improved using EIDMR in many cases where the the number of available expert statements is rather low. The actual benefit from EIDMR depends on the particular data set. The accuracy of EIDMR increases with more available samples in a data set. Furthermore, the use of the ordinal information given by the class labels and the consideration of reliability weights lead to an additional improvement in the application of EIDMR.

IV. CLASSIFICATION OF LOW-VOLTAGE GRIDS

In the preceding two sections we formally elaborated a new combination rule EIDMR and compared it to IDMR, DSR, and MR on three artificial data sets. To further investigate and analyze the combination approaches with regard to a real application, we now present a case study where we apply the combination rules to classify 300 low-voltage grids. In Section IV-A, we describe the background of our case study as well as the collection of the grid data. Additionally, our expert knowledge based approach to gather labels for the grids under investigation is briefly presented. The previously discussed combination rules are compared with regard to their accuracy in an application of SVM classifiers in Section IV-B.

A. Background and Collection of Empirical Data

The low-voltage distribution level is the one at which most of the end users—for example households—are connected to the electric power system. From the beginning of the supply with electricity, the power system was designed in a hierarchical way to transport electric energy from central power plants to consumers. With changing political circumstances, upcoming new market trends, increasing environmental loading, and decreasing availability of fossil energy sources, a paradigm change can be recognized [22], [23], [24], [25]. Especially

in rural and suburban areas—because of their high potential for Renewable Energies (RE)—the installation of distributed generators (DG) has been forced in the past decade. Without appropriate counteraction, the high amount of installed DG in the low-voltage grids—especially Photovoltaic Generators (PV)—may cause overloading of the electrical equipment and violation of voltage limits. The emergence of such bottlenecks is highly dependent on the grid structure and the configuration of the DG within the grid. Regarding these problems, the responsible distribution system operator (DSO) has to consider specific enhancement and long-acting development of low-voltage grids. But regarding the limited financial possibilities in regulated markets (e.g., incentive regulation), the decision in which low-voltage grids an investment is placed becomes increasingly important. The discrimination of low-voltage grids into different ordinal classes with regard to their hosting capacity for DG can support the investment decision but it is a difficult task, because various and complex grid structures exist. This is due to the fact that low-voltage grids have historically grown structures with local and geographic dependencies (e.g., rivers, composition of the ground). To cope with the challenge of classifying grids into different ordinal classes, expert knowledge can be elicited. Our aim is to build a system which is based on expert knowledge and allows for an automatic classification of a grid.

In our grid survey, the ten grid parameters shown in Table I were gathered for 300 real rural and suburban low-voltage grids. In order to label these data, we accomplished an expert based procedure. Admittedly, a real ground truth to validate the experts' classification results could only be obtained by an actual increase of the DG power in these grids which is obviously not possible. A number of five experts from distribution grid planning practice (DSO staff) was chosen as a trade-off between reliability of the combined statements and costs of the inquiry. The classification was intended to yield information on the DG capacity of the low-voltage grids under consideration. For that, we used five distinct ordinal classes with an ascending order describing the strength of the grid structure: (1) "very weak", (2) "weak", (3) "average", (4) "strong", and (5) "very strong".

The design of the questionnaire was oriented to hold the quality criteria validity, reliability, and objectivity for the measurement. Hence, we provided the gathered grid parameters and a plan for every grid as well as some supplementary information for optional usage by the experts. The first supplementary information consisted of five prototype grids, one for each class, which we selected from the 300 real grids. To provide more information, we divided the range of the parameters into five intervals so that every interval contains 20% of the grids. This results in indicator functions assigning one of the five grid classes for the manifestation of every particular parameter (e.g., if the percentage of intermeshing lies between 0% and 31% the corresponding indicator function yields class 1 for this parameter). Additionally, the experts were asked for a global ranking and weighting of the grid parameters concerning their importance for the classification

TABLE I
GRID PARAMETERS OF TWO SAMPLE GRIDS.

Characteristic Grid Parameter	Grid I	Grid II
1 No. of transformer stations	2	2
2 Sum of rated transformer power [kVA]	880	1260
3 No. of cable distribution boxes	5	9
4 Sum of wired line length [m]	7475	5375
5 Sum of intermeshed line length [m]	1764	3105
6 Portion of intermeshing [%]	23,6	57,8
7 Portion of new line type NAYY 150 [%]	50,3	62,8
8 Max. straight-line dist. to transf. [m]	1082	354
9 Avg. straight-line dist. to transf. [m]	840	344
10 No. of house connections	92	100

decision at the beginning of the inquiry. With regard to the ranking and weighting of the parameters, we provided a classification indicator for every grid using the first five parameters and their weighting. The indicator represents the rounded weighted average of the first five indicator functions in the ranking.

Although we provided supplementary information to the five experts they, unsurprisingly, often made conflicting statements. We also asked every expert for a difficulty statement according to three difficulty levels ("easy", "medium", "hard") while assessing a grid. This information was considered by EIDMR. The numeric values for the weights were set as described in Section III-B.

B. Classification with Support Vector Machines

We now apply the four combination rules IDMR, EIDMR, DSR, and MR to the data of our case study. That is, we aim for a combination of five experts' statements into one combined class label for every sample using each of the four combination rules. Using EIDMR, the number of considered additional samples was set to $k = 5$ and the features were z-transformed. The accuracy of the combination results cannot be evaluated such as in Section III-C. This is due to the fact that the true classes are not known for the data of our case study. We train a classifier on training data, evaluate it on test data, and do so in a cross-validation approach using the combined class labels. Thus, we aim to compare the classification accuracy regarding the combined class labels of each combination rule. Our assumption is that a high accuracy of a combination rule concerning the detection of the true classes is accompanied by a high accuracy of the trained classifier.

In our experiments, we evaluate two different classification concepts predicting the classes. In the first concept, we treated the expert statements individually by extending the training data in the following way: Each feature vector occurs five times and as targets we used the five expert statements. This extension of the training data can be seen as an interpretation of the set of the five expert statements as one gradual label. Only the testing is done with the combined label in this concept. The second concept already uses the combined label in the training phase. The testing is, such as in the first concept, done with the combined label. To implement the classifiers, the LIBSVM library with a standard Gaussian kernel [26] was

TABLE II
CLASSIFICATION RESULTS OF THE FIRST SVM CLASSIFICATION CONCEPT
PREDICTING EIDMR-COMBINED CLASS LABELS.

	#	$\kappa_{w,train}$	$\kappa_{w,test}$	e_{train} [%]	e_{test} [%]
		Accuracy Measure			
	1	0.456	0.736	42.8	23.7
	2	0.456	0.744	43.1	23.7
	3	0.449	0.737	43.3	21.7
	4	0.459	0.713	42.8	24.7
	5	0.459	0.768	42.9	20.7
Results	6	0.457	0.752	43.1	22.7
	7	0.459	0.767	42.8	20.7
	8	0.453	0.738	43.1	23.3
	9	0.457	0.763	43.0	20.7
	10	0.453	0.755	43.1	22.0
	μ	0.456	0.747	43.0	22.4
	σ	0.003	0.017	0.2	1.5

TABLE III
CLASSIFICATION RESULTS OF THE SECOND SVM CLASSIFICATION
CONCEPT PREDICTING EIDMR-COMBINED CLASS LABELS.

	#	$\kappa_{w,train}$	$\kappa_{w,test}$	e_{train} [%]	e_{test} [%]
		Accuracy Measure			
	1	0.912	0.799	8.2	18.0
	2	0.914	0.805	8.0	17.7
	3	0.910	0.786	8.3	19.4
	4	0.909	0.818	8.5	16.6
	5	0.917	0.783	7.7	19.7
Results	6	0.910	0.780	8.3	19.6
	7	0.905	0.807	8.8	17.3
	8	0.912	0.789	8.2	19.3
	9	0.905	0.819	8.8	16.3
	10	0.908	0.814	8.5	16.7
	μ	0.910	0.800	8.3	18.1
	σ	0.004	0.015	0.4	1.3

used. The feature space was built by the characteristic grid parameters set out in Table I.

To estimate good parameter values for the SVM classifiers and to prevent them from overfitting, we used a stratified 3-fold cross-validation. The stratification is essential here because the 300 combined class labels were not equally distributed no matter which of the four combination rules is applied. Thus, the process of randomly rearranging the data into 3 folds has to ensure that each fold represents the distribution of the class labels. Furthermore, the selection of the SVM parameters C and γ was done using a grid search. Because the test data of the 3-fold cross-validation process must not be used to find the parameters, we implemented another 3-fold cross validation within the actual validation procedure to robustly search the parameters on the training data. One problem is that the class labels have an ordinal character which are interpreted as nominal classes by the SVM. Thus, to estimate the model performance, we do not make use of the classification error e (amount of incorrectly classified samples) only but optimized the model parameters with regard to κ_w (which we use to measure the classification accuracy, thus, $e \neq 1 - \kappa_w$) between the test data and the output of the model (cf. Section III-B). The resulting parameter combination which we used for all implemented SVM is $C = 10$ and $\gamma = 0.03$.

Having found good parameter values for the SVM, we made

ten 3-fold cross-validations for each classification concept. In each of these cross-validations, the data is randomly rearranged into 3 stratified folds. As a consequence, the samples which are assigned to the folds will differ in each repetition. A number of ten repetitions of the cross-validations is chosen to get statistically significant results and to consider the influence of the random rearrangement of the folds.

In Table II, the classification error and the classification accuracy κ_w of the first classification concept predicting EIDMR-combined labels are set out for each of the ten cross-validation repetitions. The results of all single cross-validations are outlined to illustrate the influence of the random rearrangement of the folds on the classification error and classification accuracy. The last two columns show the classification error for the training samples and the test data. In the preceding two columns, the values of κ_w are outlined for the training samples as well as the test data. We show both, training and test accuracies, to illustrate the differences between training and test. In addition to the single values of the cross-validations, the overall mean values (μ) and the overall standard deviations (σ) are presented for each accuracy measure. It can be seen, that the overall test accuracy represented by the mean value $\mu(\kappa_{w,test})$ is approximately 0.75. The corresponding overall test error $\mu(e_{test})$ is 22.4%. The overall training accuracy and the overall training error attain significantly worse values of $\mu(\kappa_{w,train}) \approx 0.46$ and $\mu(e_{test}) = 43.0\%$, respectively. This is reasoned by the fact that in the first classification concept the same feature vector of a training set is contained multiple times (here, five times) in the training data (often with different, i.e., conflicting expert statements).

Table III shows the results of the second classification concept which directly uses the combined labels to train the SVM. Because of the direct use of the combined labels, the overall training accuracy as well as the overall training error are noticeably better compared to the first classification concept. The test values result in $\mu(\kappa_{w,train}) = 0.80$ and $\mu(e_{test}) = 18.1\%$ and are also better than the respective values of the first classification concept.

In summary, the second classification concept outperforms the first one, and, thus, is more advantageous to predict the combined labels in this experiment. Furthermore, our results show that the use of EIDMR has a high potential to yield a good classification accuracy. But is this accuracy significantly better compared to the use of the other three combination rules? To investigate the influence of the combination rule on the classification results, we conducted another experiment. We realized the better performing second classification concept for each combination rule. The overall classification results are set out in Figure 9. Using the labels combined with the three well-known approaches DSR, MR, and IDMR yields an overall classification accuracy of approximately $\mu(\kappa_{w,test}) \approx 0.61$ for each of these three rules. These values are significantly worse compared to the discussed results based on EIDMR. Using EIDMR labels leads to an enhancement of the overall classification accuracy to about 0.8 on test data. Additionally, we observe that the use of EIDMR reduces the standard
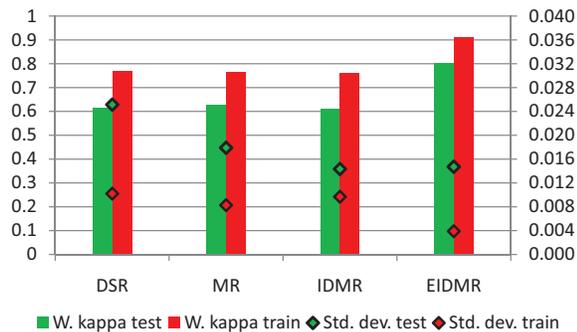
Fig. 9. Overall classification accuracy $\mu(\kappa_w)$ for each combination rule using the second classification concept.

deviations of the classification accuracy.

V. CONCLUSION AND FUTURE RESEARCH

In this article, we proposed a novel combination rule for expert statements and compared it to the well-known combination rules DSR, MR, and IDMR. The results show that especially if there are only a small number of expert statements available (e.g., due to high elicitation costs), the additional information gathered from similar samples by means of a $k$nn approach leads to significantly better combination results with the new EIDMR. The use of the ordinal information given by the class labels and the consideration of reliability weights leads to an additional improvement in the application of EIDMR.

If desired, EIDMR could be developed further by considering the distance (measured in the feature space) of similar samples in the $k$nn approach by means of additional weights.

To further validate our new combination rule, we presented a comprehensive case study by applying all combination rules to the data of 300 real low-voltage grids. The grids were assessed with ordinal labels by five experts from a regional distribution system operator. In this case study, the combination rules were compared with regard to the prediction accuracy in a classification approach with SVM. Our results show that EIDMR noticeably outperforms the other rules concerning the prediction of the combined labels.

Our future research activities in the field of low-voltage grid classification will further consolidate the concluded results and we will investigate other ways to predict the combined class labels such as, e.g., ensemble learning. Furthermore, we will improve the classification accuracy by integrating additional features and using classification results from stochastic load flow simulations [27].

REFERENCES

[1] S. A. Sandri, D. Dubois, and H. W. Kalfsbeek, "Elicitation, Assessment, and Pooling of Expert Judgments Using Possibility Theory," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 313–335, 1995.
[2] S. S. Stevens, "On the use of Expert Judgement on Complex Technical Problems," *IEEE Transactions on Engineering Management*, vol. 36, no. 2, pp. 83–86, 1989.
[3] C. Murphy, "Combining Belief Functions When Evidence Conflicts," *Decision Support Systems*, vol. 29, no. 1, pp. 1–9, 2000.
[4] K. Sentz and S. Ferson, "Combination of Evidence in Dempster-Shafer-Theory," *Sandia National Laboratories, Technical Report*, 2002.
[5] D. Andrade, T. Horeis, and B. Sick, "Knowledge Fusion Using Dempster-Shafer Theory and the Imprecise Dirichlet Model," *IEEE Conference on Soft Computing in Industrial Applications, Muroran*, pp. 142–148, 2008.
[6] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.
[7] G. Hua-wei, S. Wen-kang, and D. Yong, "Evidential Conflict and its 3D Strategy: Discard, Discover and Disassemble," *Systems Engineering and Electronics*, vol. 29, no. 6, pp. 890–898, 2007.
[8] L. A. Zadeh, "Review of Books: A Mathematical Theory of Evidence," *The AI Magazine*, vol. 5, no. 3, pp. 81–83, 1984.
[9] D. Yi-jie, W. She-liang, Z. Xin-dong, L. Miao, and Y. Xi-qin, "A Modified Dempster-Shafer Combination Rule Based on Evidence Ullage," *International Conference on Artificial Intelligence and Computational Intelligence, Shanghai*, pp. 387–391, 2009.
[10] F. Smarandache and J. Dezert, *Advances and Applications of DSmT for Information Fusion: Collected Works*. American Research Press, Rehoboth, 2009.
[11] A. Tchamova and J. Dezert, "On the Behavior of Dempster's Rule of Combination and the Foundations of Dempster-Shafer Theory," *IEEE International Conference on Intelligent Systems, Sofia*, pp. 108–113, 2012.
[12] F. Smarandache, J. Dezert, and V. Kroumov, "Examples Where the Conjunctive and Dempster's Rules are Insensitive," *Proceedings of the 2013 International Conference on Advanced Mechatronic Systems, Luoyang*, pp. 7–9, 2013.
[13] M. C. M. Troffaes and F. Coolen, "On the Use of the Imprecise Dirichlet Model With Fault Trees," *Technical Report, Durham University.*, pp. 1–8, 2007.
[14] L. V. Utkin, "Extensions of Belief Functions and Possibility Distributions by Using the Imprecise Dirichlet Model," *Fuzzy Sets and Systems*, vol. 154, no. 3, pp. 413–431, 2005.
[15] E. T. Whitaker and G. N. Watson, *A Course Of Modern Analysis*. Cambridge University Press, Cambridge, 1996.
[16] P. Walley, "Inferences From Multinomial Data: Learning About a Bag of Marbles," *Journal of Royal Statistical Society, Series B (Methodological)*, vol. 58, no. 1, pp. 3–57, 1996.
[17] M. J. Warrens, "Equivalences of Weighted Kappas for Multiple Raters," *Statistical Methodology*, vol. 9, no. 3, pp. 407–422, 2012.
[18] M. Maclure and W. C. Willett, "Misinterpretation and Misuse of the Kappa Statistic," *American Journal of Epidemiology*, vol. 126, no. 2, pp. 161–169, 1987.
[19] P. W. Mielke and K. J. Berry, "A Note on Cohen's Weighted Kappa Coefficient Of Agreement With Linear Weights," *Statistical Methodology*, vol. 6, no. 5, pp. 439–446, 2009.
[20] M. J. Warrens, "Cohen's Quadratically Weighted Kappa is Higher Than Linearly Weighted Kappa for Tridiagonal Agreement Tables," *Statistical Methodology*, vol. 9, no. 3, pp. 440–444, 2012.
[21] C.-C. Chang and C.-J. Lin, "A Practical Guide to Support Vector Classification," *Technical Report, National Taiwan University.*, pp. 1–16, 2010.
[22] G. W. Ault, J. R. McDonald, and G. M. Burt, "Strategic Analysis Framework for Evaluating Distributed Generation and Utility Strategies," *IEEE Proceedings on Generation, Transmission and Distribution*, vol. 150, no. 4, pp. 475–481, 2003.
[23] H. B. Puttgen, P. R. MacGregor, and F. C. Lambert, "Distributed Generation: Semantic Hype or the Dawn of a new era?" *IEEE Power and Energy Magazine*, vol. 1, no. 1, pp. 1–5, 2003.
[24] P. Chiradeja, "Benefit of Distributed Generation: A Line Loss Reduction Analysis," *Transmission and Distribution Conference and Exhibition: Asia and Pacific, Yokohama*, pp. 1–5, 2005.
[25] J. Driesen and R. Belmans, "Distributed Generation: Challenges and Possible Solutions," *Power Engineering Society General Meeting, Montreal*, pp. 1–5, 2006.
[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011, software available at http://www.csie.ntu.edu.tw/cjlin/libsvm (13.01.2015).
[27] S. Breker, A. Claudi, and B. Sick, "Capacity of Low-Voltage Grids for Distributed Generation: Classification by Means of Stochastic Simulations," *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 689–700, 2015.