Int. J. Appl. Math. Comput. Sci., 2025, Vol. 35, No. 1, 33–43
DOI: 10.61822/amcs-2025-0003

EVIDENCE–THEORETICAL MODELING OF UNCERTAINTY INDUCED BY POSTERIOR PROBABILITY DISTRIBUTIONS

DANIEL KAŁUŻA [a,*], ANDRZEJ JANUSZ [a,b], DOMINIK ŚLĘZAK [a,c]

[a]Institute of Informatics
University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
e-mail: {d.kaluza,janusza,slezak}@mimuw.edu.pl

[b]School of Information Systems
Queensland University of Technology
Gardens Point Campus, Brisbane, Australia
e-mail: andrzej.janusz@qut.edu.au

[c]QED Software
ul. Mazowiecka 11/49, 00-052 Warsaw, Poland
e-mail: dominik.slezak@qedsoftware.com

We discuss how the posterior probability distributions produced by machine learning models for analyzed objects can be transformed into evidence-theoretical mass functions that model uncertainties associated with operating those distributions. We investigate the mathematical properties of the introduced mass functions and their corresponding belief functions. We also construct some uncertainty measures based on the functions considered and compare them with several classical uncertainty measures, both theoretically and practically, in the active learning scenarios.

Keywords: theory of evidence, posterior probabilities, measures of uncertainty, active learning.

1. Introduction

The Dempster–Shafer theory of evidence is a powerful mathematical tool that allows for reasoning and decision-making under uncertainty (Yager and Liu, 2008). It was used successfully in many practical applications (Bezerra *et al*., 2021). It was also combined in many interesting ways with other approaches such as, e.g., the theory of rough sets (Campagner *et al*., 2022).

The advantage of the theory of evidence is that it makes it possible to operate on the mass functions (also called the basic probability assignments) that allocate probabilities to subsets of values, not only singletons. This may be helpful in case we face incomplete or insufficiently convincing information. On the other hand, as discussed by Kałuża *et al*. (2023a), standard probability distributions can be a source of non-trivial mass functions. This may happen if some probabilities are too close to each other and, as a result, it is hard to choose among them while making the decisions.

Transforming probabilities into mass functions—and the evidence-theoretical belief and plausibility functions—opens us new opportunities for uncertainty modeling. This is important in machine learning, wherein the inference is often based on the posterior distributions produced by machine learning models. Although the machine learning approaches usually focus only on the most probable decisions, measuring the uncertainty associated with the whole distributions can be beneficial, e.g., for active learning (Settles, 2012) or diagnostics of the machine learning models (Janusz *et al*., 2023). In active learning, mass function-based quantification of model uncertainty can be used to suggest subsequently acquired targets. Those targets are then used in training the next generation of machine learning models, improving the quality of the final solution.

This paper extends our earlier work on transforming

*Corresponding author

probabilities into masses (Kałuża *et al*., 2023a). In Sections 2 and 3, we discuss the literature and compare it with our way of defining the mass functions. In Sections 4 and 5, we characterize the corresponding belief functions and uncertainty measures. In Sections 6 and 7, we share new insights about the active-learning-related evaluation of those measures and discuss further means of their analysis. In Section 8, we conclude the paper and describe possible future research areas.

2. Related work and our approach

Previously (Kałuża *et al*., 2023a), we investigated the literature on the evidence-theoretical approaches to uncertainty modeling in the area of active learning. This area reflects more general trends in machine learning, wherein the theory of evidence is adopted for specific types of models or their ensembles. Here are three such examples:

In the work of Vandoni *et al*. (2019), the evidence-theoretical combination rule is applied to aggregate the beliefs of multiple support vector machines and to express the uncertainty of such combinations. Hoarau *et al*. (2022) applied a modified k-nearest neighbor model to work on the mass functions instead of probabilities. In the work of Hemmer *et al*. (2020), an evidence-theoretical extension of the neural network architecture was utilized to measure the uncertainty of the network predictions. Still, in all those cases, the non-zero masses are assigned only to single decision classes or the sets of all classes; therefore they do not express any significantly richer information compared with standard posterior probabilities.

Our intuition is that the mass functions should reflect the differences between the most probable and the consecutive less probable decision classes. As a reference, let us consider the smallest margin measure (Nguyen *et al*., 2022; Scheffer *et al*., 2001), equal to $1-(p_1-p_2)$, where p_1 and p_2 denote respectively the posterior probabilities of the most probable decision class (let us call it Class 1) and the second most probable class (Class 2), obtained as a result of inference of a machine learning model about an analyzed object. This measure reflects the uncertainty of reasoning about Class 1 given that Class 2 is "right behind it." Some similar measures can be found in the literature as well (Agrawal *et al*., 2021; Zhang, 2021).

However, we claim that the analysis of p_1-p_2 is not enough, and it needs to be followed by further differences, as displayed in Figs. 1 and 2. While p_1-p_2 is regarded as our certainty about Class 1 against Class 2, p_2-p_3 can be interpreted as the one corresponding to reasoning about Classes 1 and 2 against Class 3, and so on. Consequently, we may attempt to assign non-zero masses to the subsets $\{\text{Class } 1\}$, $\{\text{Class } 1, \text{Class } 2\}$, $\{\text{Class } 1, \text{Class } 2, \dots\}$, taking into account the quantities p_i-p_{i+1}.

The remaining aspect is about equal probabilities. Imagine a minor change in the probability distribution in Figs. 1 and 2, wherein p_1 would be equal to p_2. (This happens a bit "later" for p_3 and p_4, anyway.) Then the certainty margin of pointing at Class 1 compared with Class 2 would disappear. These two decision classes would become in some sense indistinguishable from each other, although on the other hand, reasoning about the set $\{\text{Class } 1, \text{Class } 2\}$ would still make sense if the probability of Class 3 is significantly lower. This brings us to the final idea of operating with unique probability levels and sets of equally probable classes.

3. Mass functions based on $p_i - p_{i+1}$

We are now ready to formalize our approach. The idea is to transform a given n-dimensional posterior probability distribution $\overline{p}$ defined on a set of decision classes V ($n=|V|$) into the mass functions that assign probabilities to subsets of classes—some probability to the most probable class (or classes), some probability to the most probable and the second most probable classes, and so on. We want to pay special atlention to multiple classes with the same probability. Accordingly, let us change the meaning of p_i from the previous section and consider the descending sequence of all unique positive probability values in $\overline{p}$,

$$p_1 > p_2 > \dots > p_k > 0, \qquad k \le n, \tag{1}$$

together with the collection of sets $X^1 \subset X^2 \subset \dots \subset X^k$

$$X^i = \{j : \overline{p}[j] \ge p_i\}, \qquad i = 1, \dots, k, \tag{2}$$

where $\overline{p}[j]$ denotes the probability of the j-th class.

We will call X^i ($i=1,\dots,k$) the layered sets and denote their whole sequence as $X^{\overline{p}}$. Previously (Kałuża *et al*., 2023a), we proposed two ways of assigning the elements of this sequence the quantities of the form $p_i - p_{i+1}$ (we set $p_{k+1}=0$ to simplify notation):

$$m_{\blacktriangle}(X) = \begin{cases} (p_i - p_{i+1})|X^i| & \text{for } X = X^i, \\ & i = 1, \dots, k, \\ 0 & \text{for any other } X \subseteq V, \end{cases} \tag{3}$$

$$m_h(X) = \begin{cases} \frac{(p_i - p_{i+1})}{p_1} & \text{for } X = X^i, \\ & i = 1, \dots, k, \\ 0 & \text{for any other } X \subseteq V. \end{cases} \tag{4}$$

We call $m_{\blacktriangle} : 2^V \to [0,1]$ and $m_h : 2^V \to [0,1]$ the *pyramidal* and *height ratio* mass functions, respectively. As one can easily check, we have $\sum_{X \subseteq V} m_{\blacktriangle}(X) = \sum_{X \subseteq V} m_h(X) = 1$. Thus, according to the principles of the theory of evidence (Yager and Liu, 2008), $m_{\blacktriangle}$ and m_h can indeed be called the mass functions.

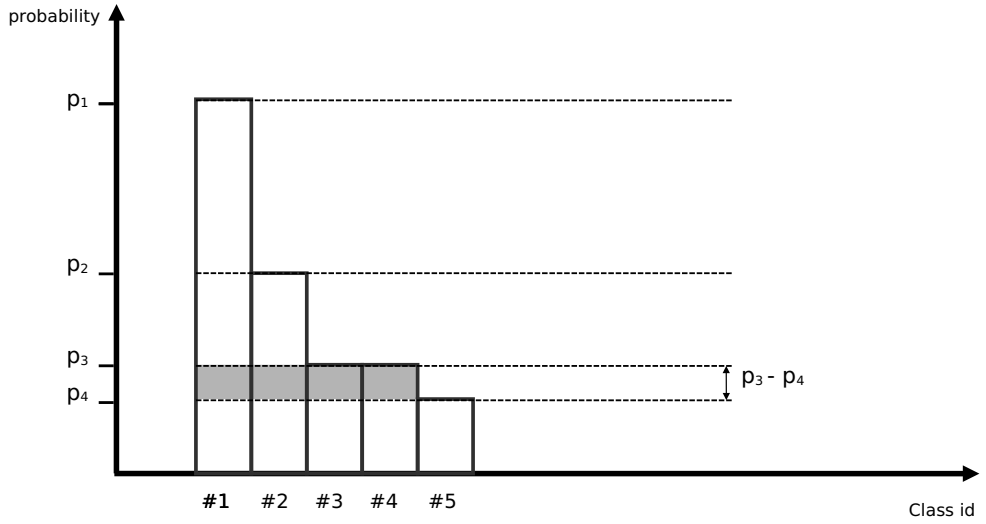

Fig. 1. Pyramidal mass function visualization. The marked area corresponds to $m_{\blacktriangle}(\{\#1, \#2, \#3, \#4\})$.

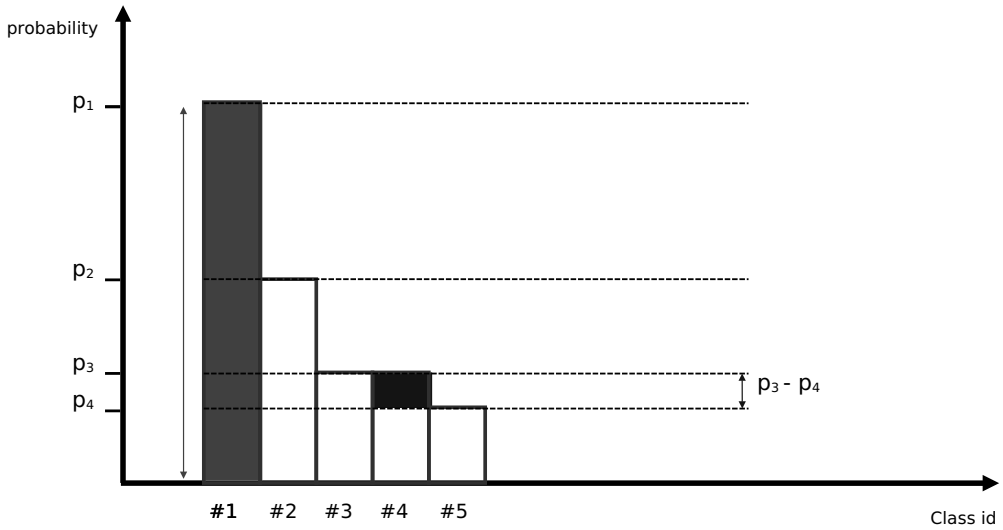

Fig. 2. Height ratio mass function. Division of the two marked heights $\updownarrow$ equals $m_h(\{\#1, \#2, \#3, \#4\})$.

Figure 1 shows that $m_{\blacktriangle}$ assigns the masses that correspond to horizontal slices of a probability histogram, which might be thought of as the layers of a *pyramid*. A mass from a given layer is assigned to a set of classes that belong to the corresponding *layered set*. In this example, the posterior distribution is defined over Classes 1–5 with the corresponding probabilities $0.4, 0.2, 0.15, 0.15, 0.1$. Accordingly, we have

$$
\begin{aligned}
&m_{\blacktriangle}(\{\#1\}) = 0.2,\\
&m_{\blacktriangle}(\{\#1, \#2\}) = 0.1,\\
&m_{\blacktriangle}(\{\#1, \#2, \#3, \#4\}) = 0.2,\\
&m_{\blacktriangle}(\{\#1, \#2, \#3, \#4, \#5\}) = 0.5.
\end{aligned}
$$

This example, firstly studied by Kałuża *et al*. (2023a), motivated us to consider m_h and compare it carefully with $m_{\blacktriangle}$. This is because $m_{\blacktriangle}$ tends to assign relatively large masses to full sets of probable decision classes. On the other hand, for m_h we obtain the following:

$$
\begin{aligned}
&m_h(\{\#1\}) = 0.5,\\
&m_h(\{\#1, \#2\}) = 0.125,\\
&m_h(\{\#1, \#2, \#3, \#4\}) = 0.125,\\
&m_h(\{\#1, \#2, \#3, \#4, \#5\}) = 0.25.
\end{aligned}
$$

Figure 2 visualizes the above calculation. In principle, both mass functions follow the same idea of operating with the consecutive "steps" of the ordered probability distribution (histogram), re-scaling them in two different ways—using cardinalities of the layered

sets X^i or simply normalizing all mass values by the dominant probability p_1. In the next sections, we will see that both $m_\blacktriangle$ and m_h may have advantages and disadvantages, considering their theoretical properties and practical usefulness. In particular, it will become clear that our original worries about the nature of $m_\blacktriangle$ were quite exaggerated.

4. Belief, plausibility, pignistic probability

Given the foundation provided by $m_\blacktriangle$ and m_h, we can now consider the *belief* and *plausibility* functions defined as follows, for an arbitrary $m : 2^V \to [0, 1]$:

$$Bel(X) = \sum_{Y:Y\subseteq X} m(Y), \tag{5}$$

$$Pl(X) = \sum_{Y:X\cap Y\neq\emptyset} m(Y). \tag{6}$$

We get the following characteristics for m_h and $m_\blacktriangle$:

$$\begin{aligned} Pl_h(X) &= \sum_{i:X\cap X^i\neq\emptyset} m_h(X^i) \\ &= \sum_{i:p_X\geq p_i} m_h(X^i) = \frac{p_X}{p_1}, \end{aligned} \tag{7}$$

$$\begin{aligned} Bel_h(X) &= \sum_{i:X^i\subseteq X} m_h(X^i) \\ &= 1 - Pl_h(X') = 1 - \frac{p_{X'}}{p_1}, \end{aligned} \tag{8}$$

$$\begin{aligned} Pl_\blacktriangle(X) &= \sum_{i:p_X\geq p_i} m_\blacktriangle(X^i) \\ &= \|\overline{p}_{\downarrow X}\|_1, \end{aligned} \tag{9}$$

$$\begin{aligned} Bel_\blacktriangle(X) &= 1 - Pl_\blacktriangle(X') \\ &= 1 - \|\overline{p}_{\downarrow X'}\|_1, \end{aligned} \tag{10}$$

where

- $X' = V \setminus X$ is the set-theoretic complement of X,
- the coefficients p_X and $p_{X'}$ denote the probabilities of the most probable elements of X and X':

$$p_X = \max_{j\in X} \overline{p}[j], \quad p_{X'} = \max_{j\in X'} \overline{p}[j], \tag{11}$$

- the coefficients $\|\overline{p}_{\downarrow X}\|_1$ and $\|\overline{p}_{\downarrow X'}\|_1$ denote the sums of the coordinates of the vectors obtained by cutting the distribution $\overline{p}$ down to the level of p_X and $p_{X'}$:

$$\overline{p}_{\downarrow X} = \min(\overline{p}, p_X), \quad \overline{p}_{\downarrow X'} = \min(\overline{p}, p_{X'}). \tag{12}$$

The above derivations are new compared to those of Kałuża *et al*. (2023a), although they are quite straightforward. As the non-zero masses are assigned only to the sets $X^1 \subset X^2 \subset \cdots \subset X^k$, the condition $Y : X \cap Y \neq \emptyset$ in (6) can be replaced by $i : X \cap X^i \neq \emptyset$, and further by $i : p_X \geq p_i$ in (7) and (9). Further computations can be followed based on visual interpretations in Figs. 1 and 2. In particular, the union of "horizontal" layered set areas in Fig. 1 can be regrouped into the union of "vertical" histogram bars cut down to the level of the most probable class in X. This is caused by the fact that X will always have a nonempty intersection with all nonzero mass assignments with lower probabilities. This is because they all contain the most probable class in X.

An important property of the above functions is that classes with the same probabilities are interchangeable. In particular, when referring to the example studied in Section 3, if a given X contains Class 3, then we can replace it with Class 4 without changing the values of $Bel_\blacktriangle(X)$, $Pl_\blacktriangle(X)$, $Bel_h(X)$, and $Pl_h(X)$.

Another observation is that it is hard for X to have a high value of $Bel_\blacktriangle(X)$ or $Bel_h(X)$. If X does not contain the most probable classes, then the values of belief functions drop to 0. Even if X is a singleton containing a decision class with the highest probability, but there is one more class with the same probability, then $Bel_\blacktriangle(X)$ and $Bel_h(X)$ equal to 0. In such a case, the mass functions $m_\blacktriangle$ and m_h focus on a pair of classes as they are equally probable. In other words, there is no difference between their probabilities so our certainty about pointing at one of them against the other cannot be higher than 0.

In the rest of this section, we investigate how well the introduced functions approximate the original probabilities. For every $j = 1, \ldots, n$, we have

$$Bel_\blacktriangle(\{j\}) \leq \overline{p}[j] \leq Pl_\blacktriangle(\{j\}), \tag{13}$$

$$\overline{p}[j] \leq Pl_h(\{j\}). \tag{14}$$

Again, this result is new compared to our previous work (Kałuża *et al*., 2023a). For the plausibility functions, due to Eqns. (7) and (9), we have $Pl_\blacktriangle(\{j\}) = \|\min(\overline{p}, \overline{p}[j])\|_1$ and $Pl_h(\{j\}) = \overline{p}[j]/p_1$. In the first case, $\overline{p}[j]$ is just one of the coordinates that are summed up in $\|\cdot\|_1$. In the second case, $\overline{p}[j] \leq \overline{p}[j]/p_1$. For beliefs, we have already noticed that the only chance for them to be positive is when $\overline{p}[j]$ is the unique highest probability, i.e., $\overline{p}[j] = p_1$ and $p_{\{j\}'} = p_2$. Then we obtain $Bel_\blacktriangle(\{j\}) = p_1 - p_2 \leq p_1 = \overline{p}[j]$. However, $Bel_h(\{j\}) \leq \overline{p}[j]$ does not hold. Let us return to the example from Section 3 one more time. Therein, $Bel_h(\{\#1\}) = 1 - 0.2/0.4 = 0.5$, higher than 0.4.

Let us present one more result which seems to put the pyramidal approach in favor of the height ratio one. It refers to the concept of *pignistic probabilities* that have been proposed to model the behavior of a rational person

while betting at particular classes (Smets, 2005):

$$BetP(j) = \sum_{X:j\in X} \frac{1}{|X|}\frac{m(X)}{1-m(\emptyset)}. \tag{15}$$

Denote by $BetP_{\blacktriangle}$ the pignistic probabilities derived from $m_{\blacktriangle}$. Immediately from Eqn. (3), given that $m_{\blacktriangle}(\emptyset) = 0$, the coefficients $|X^i|$ get reduced, and the differences $p_i - p_{i+1}$ collapse while summing them up, we obtain the following equivalence, for $j = 1, \ldots, n$:

$$BetP_{\blacktriangle}(j) = \overline{p}[j]. \tag{16}$$

The analogous characteristics cannot be considered for $BetP_h$. In summary, given the findings (13)–(14) and (16), one may claim that the mass function $m_{\blacktriangle} : 2^V \to [0,1]$ provides stronger theoretical foundations than in the case of $m_h : 2^V \to [0,1]$. Nevertheless, as displayed in Eqns. (7)–(10), the belief and plausibility functions associated with both $m_{\blacktriangle}$ and m_h have elegant and intuitive forms. Moreover, we will see in the next sections that both $m_{\blacktriangle}$ and m_h can validly model the uncertainty, and finally, they can be efficiently applied in the machine learning processes, in particular, in active learning.

5. Measures of uncertainty revisited

In this section, we discuss some examples of uncertainty measures that are built upon the proposed mass functions. But first, let us recall some already-mentioned classical measures (Agrawal *et al*., 2021; Nguyen *et al*., 2022; Zhang, 2021). Actually, let us write them down using the notation introduced in Section 3:

(i) entropy

$$Entr(\overline{p}) = -\sum_{j:\overline{p}[j]>0} \overline{p}[j]\log(\overline{p}[j]), \tag{17}$$

(ii) the smallest margin

$$SMar(\overline{p}) = \begin{cases} 1-(p_1-p_2) & \text{if } |X^1| = 1, \\ 1 & \text{otherwise,} \end{cases} \tag{18}$$

(iii) the ratio of confidence

$$RCon(\overline{p}) = \begin{cases} p_2 \,/\, p_1 & \text{if } |X^1| = 1, \\ 1 & \text{otherwise,} \end{cases} \tag{19}$$

(iv) the least confidence

$$LCon(\overline{p}) = 1 - p_1. \tag{20}$$

The formulas (17) and (20) are obvious. In particular p_1 in (20) denotes the highest posterior probability according to the notation (1). The formulas (18) and (19) seem more interesting. In both cases, the condition $|X^1| = 1$ means that there is a unique decision class with the highest probability; therefore, we can reason about it with nonzero certainty compared with less probable classes. If $|X^1| > 1$, then there is no way to point at the most probable class with non-zero certainty because there is at least one other class with the same probability. Thus, the above new way of expressing the particular uncertainty measures is consistent with their meaning in the literature.

Another fresh observation compared with those of Kałuża *et al*. (2023a) is that the smallest margin and the ratio of confidence can be rewritten using functions considered in Section 4. Namely, it is easy to get that

$$SMar(\overline{p}) = Pl_{\blacktriangle}(\{\#1\}) - Bel_{\blacktriangle}(\{\#1\}), \tag{21}$$

$$RCon(\overline{p}) = Pl_h(\{\#1\}) - Bel_h(\{\#1\}). \tag{22}$$

This simple fact gives us an additional reason to claim that the mass functions $m_{\blacktriangle}$ and m_h are thoughtfully defined. Moreover, it supports a more general intuition that there is a potential in expressing uncertainty in the language of differences between the plausibility and belief functions. On the other hand, the above derivations illustrate that classical uncertainty measures usually touch just a small fraction of information about such differences.

Ślęzak (2002) proved that the average value of the differences between Pl and Bel for arbitrary subsets of events can be expressed in terms of the cardinalities of sets assigned with nonzero masses:

$$\frac{1}{2^{|V|}}\sum_{X\subseteq V}(Pl(X) - Bel(X)) = 1 - \sum_{X\subseteq V}\cdot\frac{m(X)}{2^{|X|-1}} \tag{23}$$

The above result motivates us to investigate the examples of measures that attempt to model uncertainty as being proportional to cardinalities and masses of positively weighted sets in $X^{\overline{p}}$. Below $*$ can stand for $\blacktriangle$ or h:

(i) exponent evidence

$$Exp_*(\overline{p}) = 1 - \sum_{i=1}^{k}\frac{m_*(X^i)}{2^{|X^i|-1}}, \tag{24}$$

(ii) large exponent evidence

$$LEx_*(\overline{p}) = 1 - \sum_{i=1}^{k}\frac{m_*(X^i)}{4^{|X^i|-1}}, \tag{25}$$

(iii) log-plus evidence

$$Log_*(\overline{p}) = \sum_{i=1}^{k} m_*(X^i)\log(|X^i|+1). \tag{26}$$

The measures $Exp_{\blacktriangle}$ and Exp_h are inspired directly by (23). As for LEx_*, its motivation is to empirically check what the best mathematical relationship between $m_*(X^i)$ and $|X^i|$ could be. Finally, log-plus evidence is a slight modification of one of the known entropy counterparts in the theory of evidence (Dubois and Prade, 1987), wherein $+1$ under the logarithm softens the difference between the weights of the values corresponding to larger sets.

We might also consider some other formulas for measures. However, many of them turn out equivalent to the previous ones. For instance, one could think of a more linear relationship. It would look as follows:

$$Lin_*(\overline{p}) = 1 - \sum_{i=1}^{k} \frac{m_*(X^i)}{|X^i|}. \tag{27}$$

However, for a similar reason as in (16), we obtain

$$Lin_{\blacktriangle}(\overline{p}) = LCon(\overline{p}). \tag{28}$$

Therefore, although Lin_h would not correspond to the other measures so trivially, we decided to skip Lin_* in our empirical analysis reported in the next section.

6. Experiments with active learning

As pointed out in Section 1, modeling the uncertainty of the machine learning models' outcomes may be useful in many practical scenarios. One such scenario refers to active learning, where intermediate models are utilized, in the loop, to analyze not-yet-selected objects, and uncertainties of their inference outcomes imply which of those objects will be added to the training data sets in the next iterations of learning. Given that active learning is at the heart of our research interests (Kałuża *et al.*, 2023b), we empirically examined our framework in this context. We were particularly interested in investigating the efficiency of the active learning process when using classical uncertainty measures (17)–(20) in comparison with the examples of evidence-theoretical measures (24)–(26) derived from our mass functions $m_{\blacktriangle}$ and m_h.

In our experiments, we considered four well-known data sets (see Table 1) and we applied the following standard active learning procedure for each of them:

1. Split the data set into a training pool and a holdout test set. (For each data set, the split was half-to-half. Moreover, it was stratified concerning the number of objects supporting particular decision classes.)

2. Choose an initial training set at random. (We tested two strategies for the size of the initial training set: for *vowel* and *car* we selected simply 100 objects, while for *pendigits* and *letter* we wanted to work more ambitiously with just 0.1% of the training pool. As the initial set may influence the results, we repeated all experiments 10 times with different random choices of initial sets. Within each repetition, we used the same initial set for each of the uncertainty measures to compare them better.)

3. Train an initial machine learning model on the initial training set. (We used *RandomForestClassifier* from scikit-learn (Pedregosa *et al.*, 2011), with default parameters and the fixed initial random seed.)

4. Repeat the following till a desired number of training objects is added to the initial set (in all cases we wanted to add 100 more objects from the pool):

   (a) Use the current machine learning model to infer a posterior probability distribution over decision classes for every object that belongs to the training pool but does not belong to the current training set. (One can easily derive such probabilities from *RandomForestClassifier*.)

   (b) Out of all such objects, select the one with the highest value of uncertainty measure associated with its posterior distribution. (If there are more such objects, choose one randomly.)

   (c) Add the selected object to the training set and retrain the model. (*RandomForestClassifier* was applied throughout the whole process.)

The results of our experimental evaluation are associated with the balanced accuracy of the machine learning models calculated on the holdout test sets averaged over 100 sample selection iterations. For every data set, we created models for the following 11 methods:

- Entropy (17), Smallest Margin (18), Ratio of Confidence (19), Least Confidence (20) used as the uncertainty measure in Step 4(b) above.

- Exponent Evidence (24), Large Exponent Evidence (25), Log-Plus Evidence (26) used as the uncertainty measure, all of them in two modes, i.e, [▲] and [h].

- $Rand$—replacing Step 4(b) above with a fully random choice of an object. (Technically, this means we immediately add 100 objects to the initial training set and then train the final *RandomForestClassifier*.)

Table 2 summarizes our empirical findings. For each of the above 11 scenarios, the obtained balanced accuracy was averaged over 10 repetitions. Then we transformed those average scores into ranks (the higher the average accuracy compared to the others, the better the rank), so it makes sense to report the average rank over four considered data sets. These average scores have already been reported in our previous work (Kałuża *et al.*, 2023a) but one can find herein the following additions:

Table 1. Tabular data sets considered in the active learning experiments. In particular, the # classes column includes the number of decision classes occurring in the data. Each data set was split: 50% of objects in the training pool (the source of objects being added to the training data set) and 50% in the holdout test set. For example, the split of *letter* was 10000 vs. 10000.

Data set name (with references)	Description of decision problem	# Classes	# Attributes	# Objects
letter (Frey and Slate, 1991)	Recognition of letters obtained from rectangular pixel displays using attributes such as e.g. edge counts	26	16	20000
pendigits (Alpaydin and Alimoglu, 1998)	Tabular-attribute-based recognition of digits that were written down using a pressure-sensitive tablet	10	16	10992
car (Bohanec, 1997)	Car purchase evaluation based on a hierarchical structure that describes some car properties	4	6	1728
vowel (Deterding, 1990)	Recognition of steady state vowels of British English based on attributes extracted from speech	11	12	990

Table 2. Balanced accuracy ranks delivered by particular uncertainty measures on particular data sets, and on average, in the active learning process. The lower ranks the better. An average of the consecutive ranks was taken in case of a tie. The best average rank is marked in bold ($Exp_{\blacktriangle}$ and $LEx_{\blacktriangle}$). The cases when the measures considered performed worse than a random approach are marked in italics ($Entr$ on *letter* and *pendigits*, as well as $LCon$, LEx_h, Log_h, and $Log_{\blacktriangle}$ on *letter*).

Data set name	$Entr$	$SMar$	$RCon$	$LCon$	Exp_h	LEx_h	Log_h	$Exp_{\blacktriangle}$	$LEx_{\blacktriangle}$	$Log_{\blacktriangle}$	$Rand$
letter	*9*	1.5	1.5	*11*	4	*8*	*7*	3	5	*10*	6
pendigits	*11*	1.5	1.5	8	6	3	5	7	4	9	10
car	3	9	10	4	7	8	5	2	6	1	11
vowel	8	9	10	4	1	3	6	5	2	7	11
Average of all	7.75	5.25	5.75	6.75	4.5	5.5	5.75	**4.25**	**4.25**	6.75	9.5

- Separate ranks are reported first for every data set.
- Previously (Kałuża *et al.*, 2023a), we considered one more uncertainty measure (called *log-divide evidence*). However, its mathematical properties turned out to be questionable. We therefore decided to exclude it from the analysis. On the other hand, we include the ranks of random sampling (Rand) as the baseline.

The summary of ranks obtained by each uncertainty measure across all data sets is available in Table 2. The results of applying $Entr$ are quite disappointing and it was even worse than $Rand$ on two data sets. $SMar$ and $RCon$ seem to be a bit unstable. They are the best (delivering exactly the same balanced accuracy) on two data sets but also the worst, not counting $Rand$, on the other two. $LCon$ is on average slightly better than $Entr$, although it is worse than $Rand$ on the *letter* data set.

Compared with the above, the evidence-theoretical uncertainty measures lead toward quite good outcomes, except (26). We are satisfied especially with the efficiency of $Exp_{\blacktriangle}$ given its strong mathematical foundations. Actually, $Exp_{\blacktriangle}$ and $LEx_{\blacktriangle}$ deliver the best ranks on average. Exp_h is not much worse and yields the best score on *vowel*. On the other hand, LEx_h (like several other measures) behaves worse than $Rand$ on *letter*.

7. Further study of measure properties

Although the results described in the previous section are promising, certainly more research is needed to validate the performance of the mass functions $m_{\blacktriangle}$ and m_h, embedded into multiple uncertainty measures, examined for different data sets and machine learning tasks (Janusz *et al.*, 2023; Nguyen *et al.*, 2022).

We also need to continue studying the mathematical properties of the proposed measures. Previously (Kałuża *et al.*, 2023a), we conducted a deepened analysis of the dynamics of the considered uncertainty measures subject to changes in the input distributions. Figure 3, prepared using Matplotlib (Hunter, 2007) and mpltern (Ikeda, 2024) libraries, refers to some of our findings. An analogous study could be done using combinatorial tools that were recalled in the work of Cattaneo (2023).

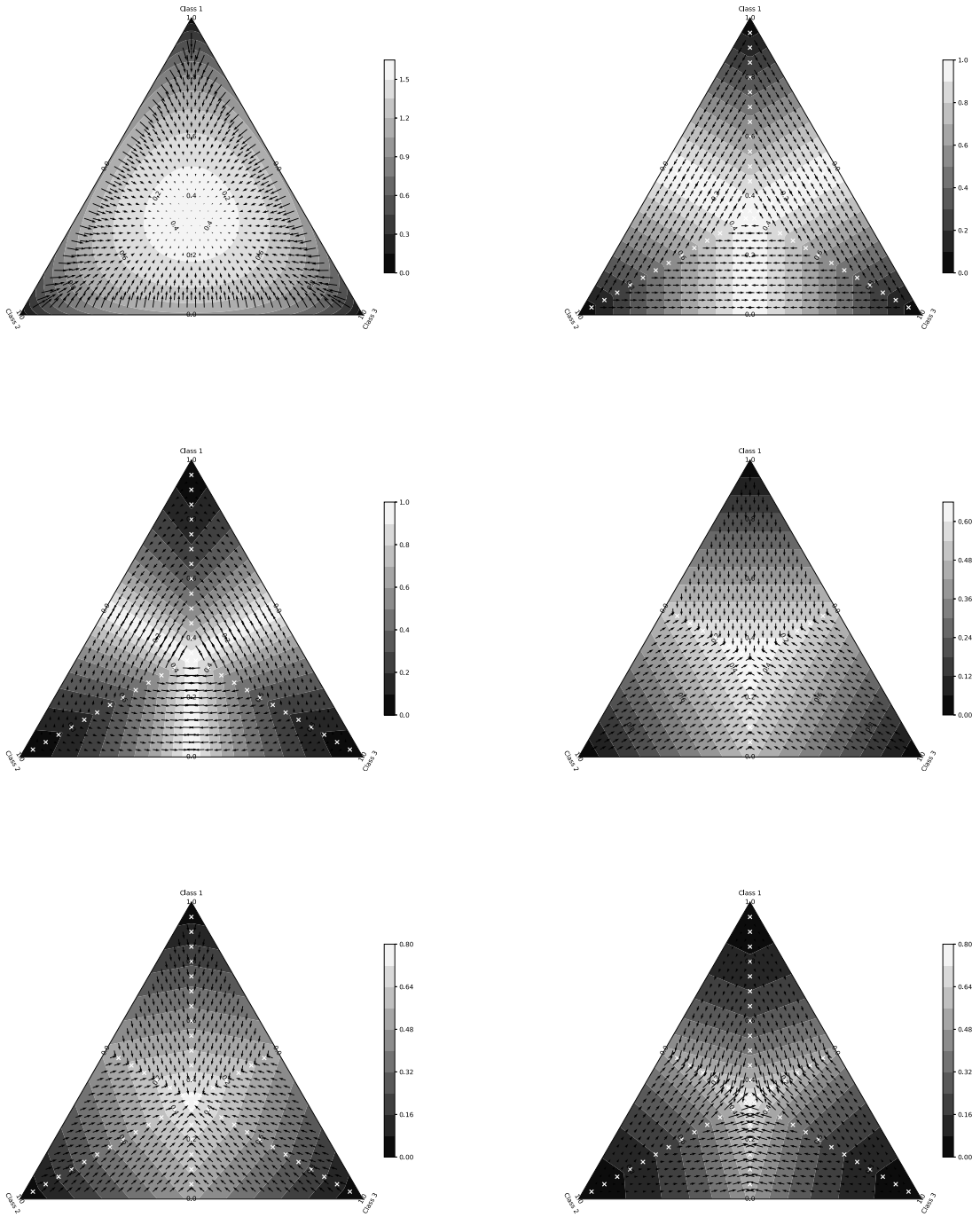

Fig. 3. Simplex visualization of six uncertainty measures in the three-decision-class scenario. Top left: $Entr$, top right: $SMar$, mid left: $RCon$, mid right: $LCon$, down left: $Exp_{\blacktriangle}$, down right: Exp_h. High uncertainty distributions are light and lower uncertainty distributions become darker. The areas around uniform distributions are the most uncertain but the "shapes" of the uncertainty decrease toward deterministic distributions vary from measure to measure. Small black arrows show the gradients of descent. Some figures contain white cross artifacts—the places where the arrow directions cannot be uniquely specified.

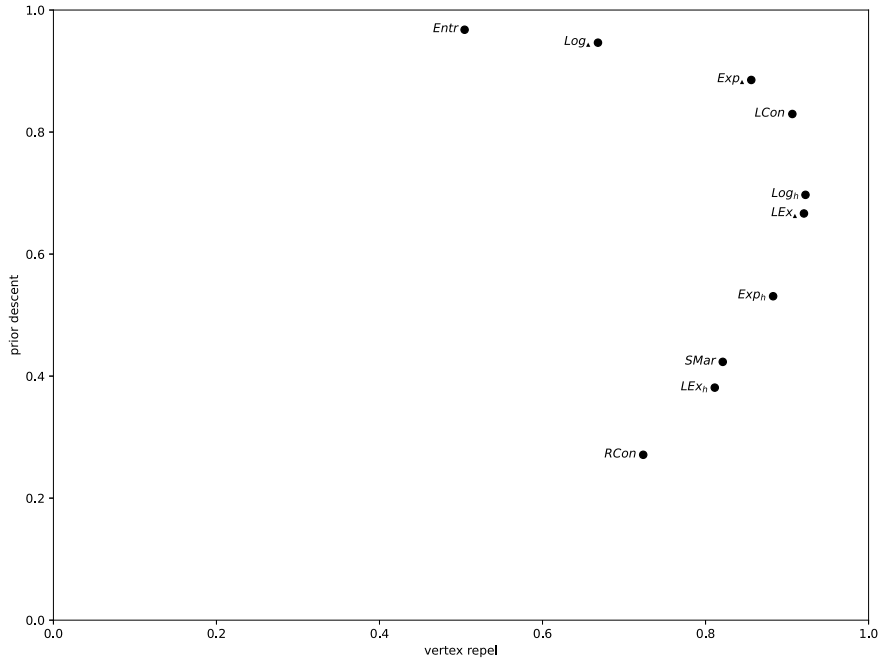

Fig. 4. Comparison of the gradient properties of the investigated measures, averaged over the points (distributions) located on a uniform grid over the 2D simplex. The axes correspond to the X/Y characteristics introduced in Section 7. The scale is between 0 and 1, where the cosine similarity equal to 1 means a total alignment of the uncertainty gradient with the vector connecting the nearest simplex vertex with the given point (X) or the given point with the simplex center (Y), while 0 indicates perpendicular vectors.

In general, a "good" uncertainty measure should increase whenever we move the input distribution $\overline{p}$ toward the uniform distribution and decrease whenever we move it toward the deterministic (one-zero) distribution which is nearest to $\overline{p}$. However, we realize that this intuition is not so easy to formalize. Figure 3 seems to support it for the classical measures (17)–(20), as well as for Exp, which is our favorite evidence-theoretical uncertainty measure because of the result (23). On the other hand, all these measures differ in more complicated scenarios, e.g., when we get closer to the distributions for which only one of the probabilities is zeroed.

Let us approach the above discussion from a new perspective compared with our previous work (Kałuża *et al.*, 2023a). For a given $\overline{p}$, we can consider two characteristics:

X Draw the arrow from the nearest simplex vertex to $\overline{p}$ and calculate its cosine similarity to the gradient-based direction of the fastest decrease of the values of the considered uncertainty measure.

Y Draw the arrow from $\overline{p}$ to the simplex center and calculate the same for the uncertainty increase.

Figure 4 illustrates the average similarities X and Y calculated for a uniformly drawn sample of distributions, for all 10 uncertainty measures considered in our experiments, i.e., (17)–(20), as well as (24)–(26) computed for both $m_{\blacktriangle}$ and m_h. As we are interested in maximizing both X and Y, the five cases forming the Pareto front, i.e., $Entr$, $Log_{\blacktriangle}$, $Exp_{\blacktriangle}$, $LCon$, and Log_h (which is almost the same as $LEx_{\blacktriangle}$), are of special importance, with $Exp_{\blacktriangle}$ looking as the most balanced "middle point."

Up to now, the measure Exp (24) in combination with the mass function $m_{\blacktriangle}$ (3) seemed to be the most convincing basis for uncertainty modeling. This is because of the mathematical properties of $m_{\blacktriangle}$ and Exp, the experimental evaluation in Section 6, and just-discussed gradient characteristics. However, everyone knows that there are no perfect measures and perfect approaches. Consequently, let us admit that the proposed uncertainty measure $Exp_{\blacktriangle}$ did not obtain the best results for many of the datasets. Therefore, it may be treated as a safe choice, but it may not produce as an good results as uncertainty measure attuned to the particular problem.

8. Conclusions and future directions

Our goal in this paper was to study new formulations of the evidence-theoretical mass functions together with the corresponding belief functions and uncertainty measures that allow for analyzing posterior probability distributions produced by the machine learning models. This research is a continuation of our previous publication (Kałuża

et al., 2023a), with significant extensions related to the mathematical properties and applications in active learning (Nguyen *et al*., 2022; Settles, 2012).

The proposed evidence-theoretical functions are constructed by transforming the posterior probabilities into masses depending on the differences between the consecutive probability levels, assigned to the subsets of decision classes exceeding those levels. In this way we obtain the mass functions that reflect the uncertainty of choosing between decisions with too similar (or equal) posterior probabilities. The uncertainty grows when relatively larger subsets of decisions are labeled with larger masses. Such situations are known to be correlated with large differences between the evidence-theoretical plausibility and belief functions (Ślęzak, 2002). That was also the motivation for us to consider new uncertainty measures that pay attention to both the masses and the cardinalities of subsets labeled with those masses.

There are several interesting directions for further research to consider. First of all, based on the results gathered in this paper, it is clear that there is no uniquely best uncertainty measure for every data set and every task. Even if we narrow ourselves down to the field of active learning, the experiments presented in Section 6 need to be continued to understand better which measures and mass functions fit best the data sets of different characteristics. Moreover, the conducted experiments show the results for one classification quality metric. It would be valuable to further investigate properties of the presented uncertainty measures and consider additional metrics commonly used in the field, such as the ROC AUC or F1-score. Going broader, it may turn out that the uncertainty measures that seem relatively good for active learning, are not so well-applicable in the other practical scenarios such as, e.g., the machine learning model diagnostics (Janusz *et al*., 2023) or the elimination of redundant attributes (Pięta and Szmuc, 2021).

Another question is how to infer the posterior probabilities and mass functions from modern models. In our experiments, we used the *RandomForestClassifier* ensemble (Pedregosa *et al*., 2011) to generate distributions, and the masses $m_{\blacktriangle}$ or m_k were derived from them as the next step. Instead, we can transform to $m_{\blacktriangle}$ or m_k the distributions obtained from separate ensemble components and use the Dempster–Shafer combination rule to get the final representation. Such an approach would be analogous to some earlier works using the combination rule for the machine learning ensembles (Vandoni *et al*., 2019). On the other hand, as mentioned in Section 2, our way of inducing mass functions from posterior distributions is more informative so we can anticipate higher efficiency of the final models.

References

Agrawal, A., Tripathi, S. and Vardhan, M. (2021). Active learning approach using a modified least confidence sampling strategy for named entity recognition, *Progress in Artificial Intelligence* **10**(2): 113–128, DOI: 10.1007/s13748-021-00230-w.

Alpaydin, E. and Alimoglu, F. (1998). Pen-based recognition of handwritten digits, UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/81pen+based+recognition+of+handwritten+digits, DOI: 10.24432/C5MG6K.

Bezerra, E.D.C., Teles, A.S., Coutinho, L.R. and da Silva e Silva, F.J. (2021). Dempster–Shafer theory for modeling and treating uncertainty in IoT applications based on complex event processing, *Sensors* **21**(5): 1863, DOI: 10.3390/s21051863.

Bohanec, M. (1997). Car evaluation, UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/19/car+evaluation, DOI: 10.24432/C5JP48.

Campagner, A., Ciucci, D. and Denoeux, T. (2022). Belief functions and rough sets: Survey and new insights, *International Journal of Approximate Reasoning* **143**: 192–215, DOI: 10.1016/j.ijar.2022.01.011.

Cattaneo, G. (2023). Abstract approach to entropy and co-entropy in measurable and probability spaces, *in* M. Ganzha *et al.* (Eds), *Proceedings of FedCSIS 2023*, Annals of Computer Science and Information Systems, Vol. 35, Warsaw, pp. 63–64, DOI: 10.15439/2023F0004.

Deterding, D.H. (1990). *Speaker Normalisation for Automatic Speech Recognition*, PhD thesis, University of Cambridge, Cambridge.

Dubois, D. and Prade, H. (1987). Properties of measures of information in evidence and possibility theories, *Fuzzy Sets and Systems* **24**(2): 161–182, DOI: 10.1016/0165-0114(87)90088-1.

Frey, P.W. and Slate, D.J. (1991). Letter recognition using holland-style adaptive classifiers, *Machine Learning* **6**(2): 161–182, DOI: 10.1007/BF00114162.

Hemmer, P., Kühl, N. and Schöffer, J. (2020). DEAL: Deep evidential active learning for image classification, *Proceedings International Conference on Machine Learning and Applications (ICMLA)*, pp. 865–870, DOI: 10.1109/ICMLA51294.2020.00141, (virtual event).

Hoarau, A., Martin, A., Dubois, J. and Gall, Y.L. (2022). Imperfect labels with belief functions for active learning, *Proceedings of BELIEF 2022*, Lecture Notes in Computer Science, Vol. 13506, Springer, Cham, pp. 44–53, DOI: 10.1007/978-3-031-17801-6_5.

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment, *Computing in Science & Engineering* **9**(3): 90–95, DOI: 10.1109/MCSE.2007.55.

Ikeda, Y. (2024). yuzie007/mpltern: 1.0.4., https://zenodo.org/records/11068993, DOI: 10.5281/zenodo.11068993.

Janusz, A., Zalewska, A., Wawrowski, Ł., Biczyk, P., Ludziejewski, J., Sikora, M. and Ślęzak, D. (2023). BrightBox—A rough set based technology for diagnosing mistakes of machine learning models, *Applied Soft Computing* **141**: 110285, DOI: 10.1016/j.asoc.2023.110285.

Kałuża, D., Janusz, A. and Ślęzak, D. (2023a). On several new Dempster–Shafer-inspired uncertainty measures applicable for active learning, *in* A. Campagner *et al.* (Eds), *Proceedings of IJCRS 2023*, Lecture Notes in Computer Science, Vol. 14481, Springer, Cham, pp. 479–494, DOI: 10.1007/978-3-031-50959-9_33.

Kałuża, D., Janusz, A. and Ślęzak, D. (2023b). Robust assignment of labels for active learning with sparse and noisy annotations, *Proceedings of ECAI 2023*, *in* K. Gal *et al.* (Eds), *Frontiers in Artificial Intelligence and Applications*, IOS Press, Amsterdam, pp. 1207–1214, DOI: 10.3233/FAIA230397.

Nguyen, V., Shaker, M.H. and Hüllermeier, E. (2022). How to measure uncertainty in uncertainty sampling for active learning, *Machine Learning* **111**(1): 89–122, DOI: 10.1007/s10994-021-06003-9.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**(85): 2825–2830.

Pięta, P. and Szmuc, T. (2021). Applications of rough sets in big data analysis: An overview, *International Journal of Applied Mathematics and Computer Science* **31**(4): 659–683, DOI: 10.34768/amcs-2021-0046.

Scheffer, T., Decomain, C. and Wróbel, S. (2001). Active hidden Markov models for information extraction, *in* F. Hoffmann *et al.* (Eds), *Proceedings of IDA 2001*, Lecture Notes in Computer Science, Vol. 2189, Springer, Berlin/Heidelberg, pp. 309–318, DOI: 10.1007/3-540-44816-0_31.

Settles, B. (2012). *Active Learning*, Morgan & Claypool, San Rafael, DOI: 10.2200/S00429ED1V01Y201207AIM018.

Ślęzak, D. (2002). *Approximate Decision Reducts*, PhD thesis, University of Warsaw, Warsaw, (in Polish).

Smets, P. (2005). Decision making in the TBM: The necessity of the pignistic transformation, *International Journal of Approximate Reasoning* **38**(2): 133–147, DOI: 10.1016/j.ijar.2004.05.003.

Vandoni, J., Aldea, E. and Le Hégarat-Mascle, S. (2019). Evidential query-by-committee active learning for pedestrian detection in high-density crowds, *International Journal of Approximate Reasoning* **104**: 166–184, DOI: 10.1016/j.ijar.2018.11.007.

Yager, R.R. and Liu, L. (2008). *Classic Works of the Dempster–Shafer Theory of Belief Functions*, Springer, Berlin/Heidelberg.

Zhang, G. (2021). Four uncertain sampling methods are superior to random sampling method in classification, *Proceedings of ICAIE 2021, Dali, China*, pp. 209–212.

Daniel Kałuża received his MS in computer science degree from the University of Warsaw in 2020, where he is currently pursuing a PhD. His main research topics involve active learning and uncertainty in machine learning models. His work has so far been featured at several international conferences such as *ECAI* or *IEEE BigData* and in leading scientific and industry journals.

Andrzej Janusz is an active academic scholar and scientist in fields related to data exploration, machine learning, and artificial intelligence. In 2014, he received his PhD in computer science from the University of Warsaw, and in 2024, he was awarded a habilitation degree in technical computer science by the Systems Research Institute, Polish Academy of Sciences. Currently, he is a lecturer in the School of Information Systems at the Queensland University of Technology. His research projects focused on safety monitoring in hazardous environments, video game data analytics, active learning, and explainable artificial intelligence. He is a co-founder of `KnowledgePit.ai`—an online data science platform, where he regularly organizes international competitions.

Dominik Ślęzak received his PhD in computer science in 2002 from the University of Warsaw, where he works as a full professor. He had also worked at the Polish-Japanese Academy of Information Technology and the University of Regina in Canada. In 2020, he was awarded the professorial title by the President of Poland. He has co-authored over 200 articles in the fields of data mining, databases, and rough sets. He has co-invented over 20 US patents and chaired over 20 international conferences. He serves as the vice-president of the Polish AI Society. He is known for his work at the edge of academia and industry. He is a co-founder of `KnowledgePit.ai` and such companies as QED Software, Infobright, and OnstageAI.

Received: 5 June 2024
Accepted: 9 January 2025