# Application of uncertainty-aware similarity measure to classification in medical diagnosis

Patryk Żywica
*Department of Artificial Intelligence, Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznań*
Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland
bikol@amu.edu.pl, ORCID: 0000-0003-3542-8982

***Abstract*—The uncertainty is a common problem in many areas of science. The best-studied variant is incomplete data where the values of some attributes are unknown. It turns out that even in this situation, many known classification methods fail. The objective of the work is to develop an effective classification method for uncertain data including pre-processing and uncertainty aware similarity-based algorithms. The problem of uncertainty is particularly important in medical diagnostics where incompleteness and uncertainty is often a natural and permanent feature of the data. The paper presents the evaluation of the proposed methods on two publicly available data sets and compare the effectiveness of the previous results.**

***Index Terms*—fuzzy, uncertainty, similarity, classification, imperfect information, decision support, nearest neighbours**

## I. Introduction

The problem of classification is to determine the class (category), to which one must assign a new, previously unknown object (instance). These objects are described using various attributes. The problem becomes significantly more complicated if we allow incompleteness or, more generally, data uncertainty. The simplest variant assumes missing values of selected attributes in the test set. Another option is the missing data in both the test and training set. Also, in both variants, not only uncertainty due to lack of data can be taken into account, but also more generally, any kind of epistemic uncertainty (see [1]). In such conditions, the design of an effective classifier using classical methods may prove to be very difficult or even impossible.

The main problems encountered when classifying uncertain data are:

- the inability to use all available data for model construction (due to removing objects or missing attributes),
- the inability to classify objects with missing values,
- the need to construct different classification methods for different data (e.g. a separate model for objects with and without missing values),
- a significant increase in computational complexity of both the classifier learning process and the classification itself,
- increase in the complexity of classification methods, which significantly impedes understanding of constructed models.

This work was supported by the National Science Centre, Poland, grant number 2016/21/N/ST6/00316.

The problem of data uncertainty is particularly important in medical diagnostics where incompleteness and uncertainty is often a natural and permanent feature of the data [2]. Hence, any attempt to neglect this problem leads to the models that do not reflect reality. Editing the data is generally not allowed in medical applications. Removing incomplete instances from often too small data set, significantly narrows the research possibilities. On the other hand, the imputation based approach can not be used to classify a specific medical case due to the high risk of the wrong diagnosis because imputed values in no way reflect the current patient condition.

On the other hand, sometimes even a high level of uncertainty may have no real impact on classification outcomes. We recall the example of a medical diagnosis problem, where we have found out that for some patients, the actual value of many medical markers or indicators does not impact final diagnosis [3]. In other words, in those situations all possible epistemic states were classified into the same class.

The classification methods proposed in this work attempt to solve these problems using *uncertainty-aware similarity measures* [4].

## II. Definitions

Let $U = \{u_1, u_2, \ldots, u_n\}$ be a crisp universal set. A mapping $A\colon U \to [0,1]$ is called a fuzzy set (FS) in $U$ [5]. For each $1 \leq i \leq n$, the value $A(u_i)$ ($a_i$ for short) represents the membership grade of $u_i$ in $A$. Any crisp set $X \subseteq U$ can be represented as a fuzzy set by its characteristic function $\mathbb{1}_X$. Let $\mathcal{F}(U)$ be the family of all fuzzy sets in $U$.

**Definition 1** (see [6])**.** A similarity measure of fuzzy sets is defined as a function on $\mathbb{E} \subset \mathcal{F}(U) \times \mathcal{F}(U)$

$$s\colon \mathbb{E} \to [0,1]\,, \tag{1}$$

where $\mathbb{E}$ needs to satisfy:
(S1) $(A,B) \in \mathbb{E}$ if and only if $(B,A) \in \mathbb{E}$,
(S2) $(A,B) \in \mathbb{E}$ if $(A, \mathbb{1}_U) \in \mathbb{E}$.

It is common to assume that the higher measure values indicate a higher similarity of arguments.

Any closed, nonempty subset $\widetilde{A}$ of $\mathcal{F}(U)$ will be called Fuzzy Membership Function Family (FMFF) (see [7]). Set $\widetilde{A}$ represents all the possible states that can hide behind uncertain information.

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

Let $\widetilde{\mathbb{E}} \subset \mathcal{FMFF}(U) \times \mathcal{FMFF}(U)$ be such that:

(E1) $(\widetilde{A}, \widetilde{B}) \in \widetilde{\mathbb{E}}$ if and only if $(\widetilde{B}, \widetilde{A}) \in \widetilde{\mathbb{E}}$,

(E2) $(\widetilde{A}, \widetilde{B}) \in \widetilde{\mathbb{E}}$ if $\big(\widetilde{A}, \{\mathbb{1}_U\}\big) \in \widetilde{\mathbb{E}}$,

(E3) $(\widetilde{A}, \widetilde{B}) \in \widetilde{\mathbb{E}}$ if and only if for any fuzzy sets $A \in \widetilde{A}$, $B \in \widetilde{B}$:

$$\big(\{A\}, \{B\}\big) \in \widetilde{\mathbb{E}}. \tag{2}$$

**Definition 2** (see [4])**.** A function $\widetilde{s}\colon \widetilde{\mathbb{E}} \to \mathcal{P}([0,1])$ is an *uncertainty–aware* similarity measure if:

(P1) For all $(\widetilde{A}, \widetilde{B}) \in \widetilde{\mathbb{E}}$,

$$\widetilde{s}(\widetilde{A}, \widetilde{B}) = \widetilde{s}(\widetilde{B}, \widetilde{A}). \tag{3}$$

(P2) If $(\mathcal{F}(U), \mathcal{F}(U)) \in \widetilde{\mathbb{E}}$ then

$$\widetilde{s}(\mathcal{F}(U), \mathcal{F}(U)) = [0,1] \tag{4}$$

(P3) For all $(\widetilde{A}, \widetilde{B}), (\widetilde{A}, \widetilde{C}) \in \widetilde{\mathbb{E}}$ such that $\mathbb{1}_X \in \widetilde{A}$, $\mathbb{1}_X \in \widetilde{B}$ and $\mathbb{1}_{X^c} \in \widetilde{C}$ for some $X \subset U$,

$$1 \in \widetilde{s}(\widetilde{A}, \widetilde{B}), \tag{5}$$

$$0 \in \widetilde{s}(\widetilde{A}, \widetilde{C}). \tag{6}$$

(P4) For all fuzzy sets $A, B \in \mathcal{F}(U)$ such that $(\{A\}, \{B\}) \in \widetilde{\mathbb{E}}$,

$$\widetilde{s}(\{A\}, \{B\}) = \{a\}, \text{ for some } a \in [0,1]. \tag{7}$$

(P5) For any $(\widetilde{A}, \widetilde{C}), (\widetilde{B}, \widetilde{D}) \in \widetilde{\mathbb{E}}$ such that $\widetilde{A} \subset \widetilde{B}$ and $\widetilde{C} \subset \widetilde{D}$,

$$\widetilde{s}(\widetilde{A}, \widetilde{C}) \subset \widetilde{s}(\widetilde{B}, \widetilde{D}). \tag{8}$$

(P6) For any $(\widetilde{A}, \widetilde{D}), (\widetilde{B}, \widetilde{C}) \in \widetilde{\mathbb{E}}$ and for all $A \in \widetilde{A}, B \in \widetilde{B}, C \in \widetilde{C}, D \in \widetilde{D}$ such that $A \subset B \subset C \subset D$ we have

$$s_{ad} \leq s_{bc} \tag{9}$$

where

$$\widetilde{s}\big(\{A\}, \{D\}\big) = \{s_{ad}\} \text{ and } \widetilde{s}\big(\{B\}, \{C\}\big) = \{s_{bc}\}. \tag{10}$$

In (P6) it should be noted that inclusion relation plays a purely technical role – it only guarantees the proper ordering of the membership functions.

**Definition 3.** Let $s\colon E \to [0,1]$ be a similarity measure of fuzzy sets. Function $\widetilde{s} : \widetilde{\mathbb{E}} \to \mathcal{P}([0,1])$ can be defined as an image of a product of two families of fuzzy sets via similarity measure $s$:

$$\widetilde{s}(\widetilde{A}, \widetilde{B}) = \Big\{ s(A,B) : A \in \widetilde{A}, B \in \widetilde{B} \Big\}, \tag{11}$$

where

$$\widetilde{\mathbb{E}} = \Big\{ (\widetilde{A}, \widetilde{B}) \in \mathcal{FMFF}(U) \times \mathcal{FMFF}(U) : \widetilde{A} \times \widetilde{B} \subset \mathbb{E} \Big\}. \tag{12}$$

Let $G$ be a convex subset of $[0,1] \times [0,1]$ and the Equality Value $\Psi : G \to [0,1]$ be continuous and such that:

($\Psi$1) for any $(a,b) \in G$, $\Psi(a,b) = \Psi(b,a)$

($\Psi$2) for any $(a,d), (b,c) \in G$ such that $a \leq b \leq c \leq d$ we have that $\Psi(a,d) \leq \Psi(b,c)$

TABLE I: Co-implication operators obtained from nine basic implication operators.

| Implication operator | Fuzzy equivalence, $\Psi(x,y)$ |
|---|---|
| Łukasiewicz, $I_{\text{ŁK}}$ | $1 - \lvert x - y \rvert$ |
| Gödel, $I_{\text{GD}}$ | $\begin{cases} 1, & x = y \\ \min(x,y), & x \neq y \end{cases}$ |
| Reichenbach, $I_{\text{RC}}$ | $\begin{cases} \min(x,y), & 1-x \leq y \\ \min(1-x, 1-y), & 1-x > y \end{cases}$ |
| Goguen, $I_{\text{GG}}$ | $\begin{cases} 1, & x = y = 0 \\ \frac{\min(x,y)}{\max(x,y)}, & x \neq 0, y \neq 0 \end{cases}$ |
| Rescher, $I_{\text{RS}}$ | $\begin{cases} 1, & x = y \\ 0, & x \neq y \end{cases}$ |
| Yager, $I_{\text{YG}}$ | $\begin{cases} 1, & x = y = 0 \\ \min(x^y, y^x), & x, y > 0 \end{cases}$ |
| Weber, $I_{\text{WB}}$ | $\begin{cases} 1, & x, y < 1 \\ \min(x,y), & x = 1 \text{ or } y = 1 \end{cases}$ |
| Fodor, $I_{\text{FD}}$ | $\begin{cases} 1-x, & y \leq \min(x, 1-x) \\ y, & 1-x < y < x \\ 1-y, & x < y < 1-x \\ x, & y > \max(x, 1-x) \end{cases}$ |
| Kleene-Dines, $I_{\text{KD}}$ | $\begin{cases} \min(x,y), & 1-x \leq y \\ \min(1-x, 1-y), & 1-x > y \end{cases}$ |

($\Psi$3) $\Psi(0,1) = 0$, $\Psi(1,1) = 1$ and $\Psi(0,0) = 1$.

**Definition 4.** The logic-based similarity measure is defined as an aggregation of the Equality Values $\Psi : G \to [0,1]$ over all elements of the universe

$$s_\Psi(A,B) = \bigoplus_{u \in U} \Psi(\mu_A(u), \mu_B(u)). \tag{13}$$

The domain of such similarity measure is the following:

$$\mathbb{E}_G = \big\{ (A,B) \in \mathcal{F}(U)^2 : \forall_{u_i \in U} \ (A(u_i), B(u_i)) \in G \big\} \tag{14}$$

where $G \subset [0,1]^2$ and

($G$1) $(a,b) \in G$ if and only if $(b,a) \in G$,

($G$2) $(a,b) \in G$ if $(a,1) \in G$.

This definition generalizes the concept of *additive similarity measure* recently proposed by Couso [8], [9]. It should be noted that replacing arithmetic mean by any other aggregation operator does not affect any of the assumptions or properties. Moreover, this approach allows for direct integration of the weights of individual elements of the universe into a similarity measure. From a practical point of view, two families of aggregation operators are noteworthy: weighted means and weighted medians. The first one because of its simplicity and widespread use. Second, because of much better computational complexity when applied to *uncertainty-aware* similarity measure [4].

All continuous Fuzzy Equivalences defined by Fodor and Roubens [10] can be used to define Equality Value $\Psi$. Examples are given in Table. I.

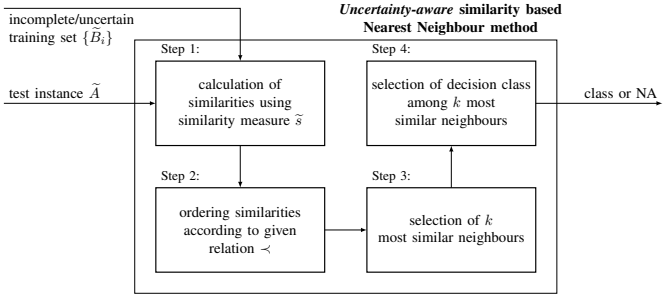


Fig. 1: A scheme of operation of the classifier based on the kNN method.

## III. Proposed method

This work will propose a classification method based on the *uncertainty-aware* similarity measure presented in the previous section. The feature that distinguishes it is full support for data uncertainty. It can easily deal with epistemic uncertainty of any type, both in the training and test sets.

In expert systems and, more generally, in the problem of decision support, very often, apart from the effectiveness itself, the interpretability of the mathematical models used is also important. Too complicated and difficult to understand methods discourage their use. Such a problem occurs, for example, in the field of medical diagnostics, where doctors are very skeptical about the methods they would use as a black box [2]. For this reason, one of the basic assumptions was to maintain simplicity and low conceptual complexity. The chosen solution is based on the nearest neighbor rule, which is very similar to the case study method used in medicine.

In this work, we use the variant of the $k$ nearest neighbors (kNN) method, where instead of the distance between objects their similarity is used (see [11]). The basic requirement was the possibility of classifying data subject to epistemic uncertainty (e.g. resulting from incomplete data). However, *uncertainty-aware* similarity measure cannot easily be used to construct the distance function for kNN. Hence, we add steps responsible for dealing with uncertainty. Figure 1 presents the main steps of the proposed classification method.

The basic difference between the distance function and *uncertainty-aware* similarity measure is the set-theoretic character of similarity value. The main problem concerns the ordering of sets (Step 2), which, unlike real numbers, is not clearly defined. Standard methods for interval ranking are:

1) ordering by center, lower or upper bound of the interval
2) ordering based on comparing bipolar values [12]

$$\begin{aligned}&[a_1,b_1]\prec[a_2,b_2],\text{ if}\\&\quad a_1+b_1<a_2+b_2\text{ or}\\&\quad a_1+b_1=a_2+b_2\text{ and }a_1-b_1<a_2-b_2\,,\end{aligned}\tag{15}$$

3) interval dominance

$$[a_1,b_1]\prec[a_2,b_2],\text{ if }b_1<a_2\,,\tag{16}$$

4) lattice partial order

$$[a_1,b_1]\prec[a_2,b_2],\text{ if }a_1<a_2\text{ and }b_1<b_2\,.\tag{17}$$

These relations can be generalized to case of $[0,1]$ subsets.

For the third and fourth methods, the order received is only partial. For this reason, it is not possible to unambiguously order the intervals. From a theoretical point of view, any linear extension of a partial order is sufficient. However, in practical applications, it is better to use a deterministic algorithm that tries to extrapolate the order such as the Local Partial Order Model (LPOM, [13]).

Also, the remaining steps of the proposed approach allow you to adjust the parameters affecting the quality of the classification. The most important parameter is the selection of the similarity measure in Step 1. In Step 2, it is necessary to indicate the order relation on the set of similarity measure values. The problem of choosing the value of the $k$ parameter from Step 3 is analogous to the classical the Nearest Neighbor method.

The last step is to choose the class for the instance based on $k$ the most similar neighbors. The simple solution is to choose the dominant class. The inclusion of uncertainty in the classification process allows the use of another class selection method. It consists of choosing the class of the neighbor whose similarity carries the least uncertainty (i.e. the set that is the value of the similarity measure has the smallest cardinality). The next method can only be used if the classifier is allowed not to make a decision. In this method, the class for the object is the one to which all $k$ most neighbors belong. In the case of different classes, the decision is not made. No decision can also be made for the even values of the $k$ parameter and the dominant class selection method. Such a solution may be particularly useful in expert systems supporting medical diagnostics, wherein the case of insufficient premises it is possible to give no recommendation, thus indicating that more data should be collected.

The most important advantages of the proposed method include the simplicity of its concept, ease of implementation and a homogeneous method of operating on complete, incomplete and general uncertain data. The third advantage is particularly important because there are few models with this property.

The basic disadvantage of the proposed method is high computational complexity. Although the learning phase of the classifier, in this case, is not associated with a large computational effort, the cost of classifying a single instance is quite high. Assuming that the computational complexity of the similarity measure is $O(f(n))$, the first step requires $O(mf(n))$ operations, where $m$ is the number of instances in the training set. The ordering of the values in the second step, depending on the method used, can be done in $O(m\log m)$, if the selected order is a total order, or $O(m^2)$, for partial order sorted using topological sort or LPOM algorithms. Steps three and four require $O(k)$ operations. Thus, in total, $O(mf(n)+m\log m)$ or $O(mf(n)+m^2)$ operation is required to classify a single test instance. This complexity, although large, still allows effective classification, even with several

thousand items in the training database. It is important to note, that due to the complex nature of data being handled, it is hard to replace the brute force approach with well-known data structures such as KD or ball trees [14], [15].

## IV. Evaluation

It is difficult to carry out a fair comparison of the developed methods with other known classifiers. Most of them cannot operate on uncertain (interval or set) data. On the other hand, comparing methods created to deal with information uncertainty on complete and certain data significantly favors classic solutions.

Another problem is the availability of well-documented, incomplete data sets that could serve such a comparison. To overcome this problem, the KEEL project created an archive of the data sets [16]. Unfortunately, it contains only one real data set (in various variants) that can be used to perform this task.

For this section, implementations of the similarity measures and classification methods were provided under the *open source* license [17]. The recognized and free programming language *R* was chosen as the platform.

This section is devoted to the evaluation of the proposed classification methods on a *dyslexic* data set from the KEEL archive and OvaExpert data set. The effectiveness will be compared with the previous methods proposed by the authors of those data sets.

### A. *The* dyslexic *dataset*

Dyslexia can be defined as an impairment of the learning process in people with the correct intelligence quotient and without other physical or psychological problems that could explain this condition. The *dyslexic* dataset [16] contains the values of twelve indicators (attributes) to assess whether reading, writing and calculating skills are developed according to the age of the child. The data include 65 children between the ages of 6 and 8, residing in the province of Asturias in Spain. It is estimated that 4-5% of children suffer from this problem.

The data set is subject to the uncertainty resulting from the decision of the psychologist. Each attribute takes values from the range of $[0, 10]$. In the event of difficulties in determining the exact value, the psychologist may use any interval. Also, the dataset is not complete and some attribute values are not available. The descriptions of all instances were normalized and converted to a Fuzzy Membership Familly (FMFF). The missing values have been replaced by the set of all possibilities, i.e. $[0, 1]$.

During the examination, the child is classified into the following four groups: *without dyslexia*, *for control*, *with dyslexia* and *other disorder*. Each child as a result of the examination can be assigned to many groups at the same time.

The evaluation was carried out using 10-fold cross-validation, the same as in [18], [19], thanks to which it will be possible to compare the obtained results.

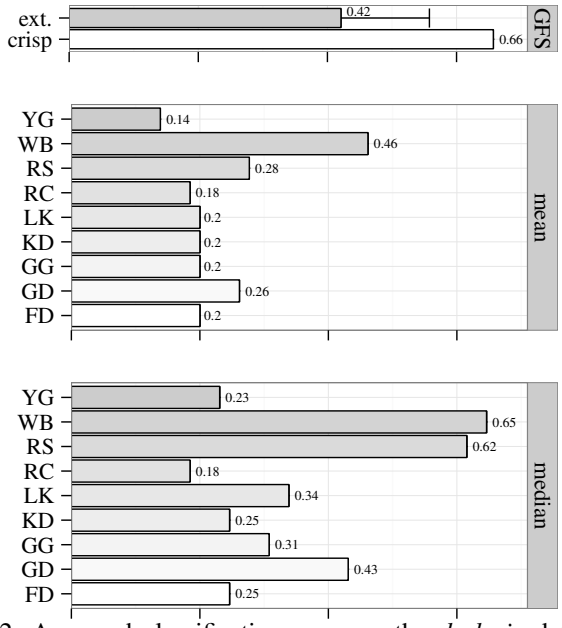


Fig. 2: Averaged classification error on the *dyslexic* dataset. The results of classifiers based on the GFS method are also marked on the left.

*1) Results and discussion:* Figure 2 shows the average error rate in the four-class classification. For comparison, the results obtained by both variants of the GFS (Genetic Fuzzy System) method, which were proposed by the authors of the *dyslexic* dataset, are also presented. The basic variant (crisp) does very poorly with uncertain data, obtaining a classification error of $0.657$. Its extended version performs much better, obtaining an error rate in the range $[0.421, 0.558]$.

However, both of these methods are significantly inferior to most variants based on the similarity measure. A classifier based on Yager Fuzzy Equality and mean aggregation obtained a classification error of $0.14$. This is not an isolated result, as for many other Fuzzy Equalities the error is not higher than $0.2$.

The lower part of Fig. 2 presents results of median-based similarity measures. The first observation is that on this dataset mean aggregation proved to be more robust. Median still can achieve good performance but the error rate is higher, except Reichenbach Fuzzy Equality.

Another observation is that Rescher and Weber Fuzzy Equalities yield higher errors, than others. This is because these two operators are extreme cases that in most cases assign $0$ and $1$, respectively. This makes them ignore all uncertainty in the data, hence error rate is comparable with the crisp GFS method. The last observation is that $0.14$ error is a very good result. It overcomes the results obtained by the generalized Jaccard index extended to uncertain data proposed in [4], [20].

### B. *The* OvaExpert *dataset*

This subsection is devoted to the problem of differential diagnosis of ovarian tumors. This was the starting point for research on the classification of incomplete and uncertain data.

Two scenarios for using the proposed classification methods will be presented. In both cases, the extensive evaluations were carried out on actual medical data.

Malignant ovarian tumors are one of the most difficult problems of modern gynecological oncology both in terms of diagnostics and treatment [21]. Not fully understood etiology, as well as increasing morbidity and mortality, further increase the importance of this issue. The high mortality rate in women is due to the fact that ovarian cancers are difficult to detect in the early stages [21].

In recent years, several studies have been conducted to identify patient-describing parameters that would enable the preoperative prediction of ovarian tumor type. In addition to the classic features collected during the medical interview, such as menopausal status or the occurrence of cancer in the patient's family [22], attention was paid to information obtained during the ultrasound examination [23]. Cancer markers CA-125 and HE4 are a separate category, the level of which is also one of the most important premises during differentiation [24].

The OEA model, which was created on the OvaExpert dataset especially to deal with ovarian tumor diagnosis under data uncertainty should be treated as a state of the art solution for this problem [25]. In this paper we proposed another approach, that is more universal since it can be applied to any field and the results show that efficacy is close to original results.

*1) Scenario 1: diagnostics based on raw patient data:* The first scenario assumes the use of all available medical data to make a diagnosis. This is the simplest and most direct method since the only transformation that the data undergoes is normalization to FMFF. The similarity measure is used directly to compare the medical data of two patients.

The research group consists of 388 patients treated with ovarian tumors at the Clinic of Operational Gynecology at the Medical University of Karol Marcinkowski in Poznań, Poland in the years 2005 - 2015. Of these, 61% were diagnosed with benign and 39% were malignant. In addition, 56% of patients had no missing data and 40% were incomplete in less than 50%. The distribution of missing data is shown in Fig. 3. A significant subset of this dataset, covering the majority of patients with a complete description, was presented in a medical context in the work [3].

The evaluation procedure is based on the classic division into training (optimization) and test data sets. The initial data set does not have a homogeneous distribution of the missing data. If such data were divided evenly, this could lead to a situation in which some levels of missing data would not be available at the optimization and/or testing stage. This situation is very undesirable because the goal is to develop a method that works for all levels of missing data. The test set consists of instances with missing data and some of those with a complete data set. On the other hand, the optimization set is built from instances with a complete description, and the incompleteness is simulated. In the simulations, it was assumed that the incompleteness of data occurs randomly because it is not possible to accurately simulate the diagnostic

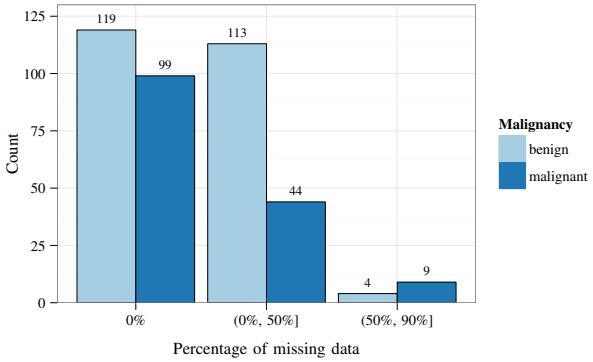

Fig. 3: Division of patients in terms of data completeness level.

process. The actual distribution of the levels of lack of data is not known, hence in the optimization phase, an assumption was made about its homogeneity. Thanks to this, instances with different levels of data incompleteness occur in both the optimization and test sets.

The description of each instance is converted to a FMFF. Let the $U$ be the universe composed of all attributes describing the patient. The degrees of membership of a given attribute $u_i \in U$ to the FMFF $\widetilde{A}$, representing the patient, carry information about the extent to which the value of this attribute for a given instance is large. Importantly, if the value of a certain attribute is not available, the entire unit interval is used as the set of possible membership degrees. The more information regarding the normalization procedure and semantics is given in [7].

The optimization set consists of 200 complete instances, while the test set contains the remaining 18 complete patient descriptions and all descriptions with a data incompleteness level below 50% - a total of 175 instances. Instances incomplete in over 50% were excluded from the study.

Reducing the number of instances, or the removal of outliers can significantly impact the effectiveness and efficiency of the classification algorithm. However, the problem of data preprocessing seems to be neglected in the available literature on uncertain data classification. The first approaches to this problem were taken in [26], [27]. Although results are promising, authors assess the quality of instance by the single numerical value. Such an approach neglects the epistemic nature of the data because this uncertainty score is external to the data itself.

Two main approaches to data reduction for NN rule can be distinguished: condensation (CNN, [28]) and editing (ENN, [29]). The first try to build a minimal consistent subset of the training set. The second reduces the dataset by removing noisy instances. In this study, we focus on the second approach because CNN based methods are not able to find any consistent set when data is uncertain in an epistemic manner.

Because the removal of a noisy data point might lead to a new source of noise, Repeated Editing Nearest Neighbour rule (RENN) repeatedly removes noisy data until no noise of this kind is found [30]. This approach was used in our medical dataset evaluation.

The goal of the optimization phase is to select, based

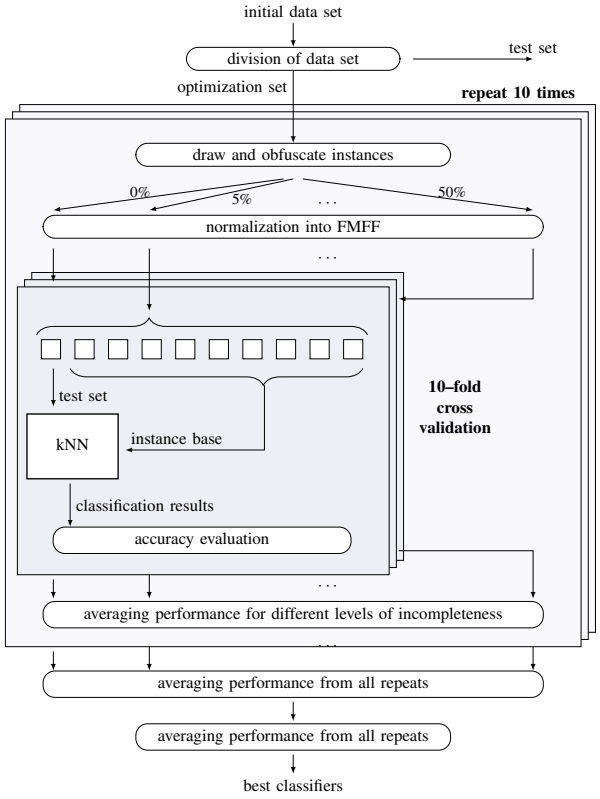


Fig. 4: Diagram presenting the optimization phase of the evaluation.

on simulated incompleteness, optimal parameter values for each of the proposed methods. The simulated incompleteness level was between 0% and 50% with 5% steps. For each level of incompleteness, 10 repeats of the following procedure were performed. First of all, 75 instances of each class were drawn from the optimization set (malicious, benign). Then the determined percentage of attributes describing selected patients is obfuscated (erased). Such data is used in a 10-fold cross-validation procedure of all the variants of the proposed classification methods. The classifiers are assessed based on performance averaged for all repeats and levels of data incompleteness. The single step of the optimization phase is shown in Fig. 4.

The result of the optimization phase is a set of classifiers that did best on simulated incomplete data. In the test phase, they are run on data from the test set, for which the source of uncertainty is the actual diagnostic process. The entire optimization set is used as the instance base for the proposed classifiers. To examine how the developed classification methods behave when the instance database is incomplete, its various levels were simulated. The classification quality assessment obtained is then averaged for all replicates and data incompleteness levels.

*2) Scenario 2: diagnostics based on diagnostic models:* A common problem with many distance- or similarity-based classification methods is a decrease in efficiency as a number of dimensions increase. This is because distances in multidimensional space become large regardless of the actual proximity or similarity of the compared objects. Of course, the strength of this undesirable effect depends on the selection of the original similarity measure. For example, while using Euclid's distance much better efficiency should be expected with a reduced number of dimensions.

The second scenario involves reducing the number of attributes describing the patient. Many methods are known for reducing the dimensionality of data sets, e.g. principal component analysis. However, another method will be used in this case. There are many models for the problem of differential diagnosis of ovarian tumors. Most of them used on complete data describing the patient return a numerical value from a certain range indicating a suggested diagnosis. The values returned by several models can be used to construct new attributes describing the patient.

The evaluation uses the same data set as the first scenario. Six diagnostic models were selected for the reduction (see [25]): two based on scoring systems (SM, Alcazar) and four on logistic regression (LR1, LR2, Timmerman, and RMI).

Selected diagnostic models require complete data. This raises the problem of how to calculate the return value of the model for instance, which is not complete. This problem was solved using uncertaintification [25]. The basic idea is to replace a single value in the $[0, 1]$ interval (0 means a benign change, while 1 - a malignant one) with the set of all values that can be obtained by completing the incomplete description of the patient in any way. The data obtained in this way is subject to epistemic uncertainty, which makes it perfectly fit into the developed classification methods. Each instance is represented by a FMFF. In this scenario, we define the universe as a set of diagnostic models and the degree of membership $\mu_{\widetilde{A}}(u)$ describes the similarity of the value returned by a given model to a malignant diagnosis.

The evaluation procedure is identical to the previous one, except that in the step *normalization into FMFF* instead of normalizing the data, a reduction is performed.

*3) Results and discussion:* The quality of classification algorithms can be expressed through many different indicators, such as accuracy, sensitivity, and specificity. In the considered problem of medical diagnostics, the best classifier should provide very high sensitivity as well as slightly lower, but still high specificity. Moreover, for some instances, data may not indicate which decision to make. In this case, the classifier should not indicate the diagnosis and the patient should be sent to a reference treatment center. For this reason, it is acceptable that some instances will not have a class assigned (less than 100% decisiveness). Since choosing one quality indicator that would meet all criteria is a very difficult task [31], the total cost method was used, where the sum of costs assigned to individual decisions is taken as a measure of the quality of the classification. Correct classification as *true positive* or *true negative* is not associated with any cost. The highest cost is attributed to the *false negative* diagnosis, when the classifier indicates benignity, while in fact, the patient has a malignant

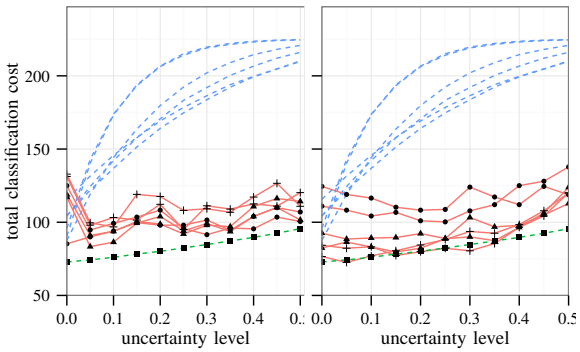


Fig. 5: Evaluation results for the proposed classification methods depending on the level of uncertainty. The effectiveness of the reference models was also plotted.

tumor. The cost of *false positive* was set as twice lower than the cost of *false negative* because unnecessary surgery still poses some danger to the patient's health. Also, if the classifier fails to make a decision, there is a difference in the assigned cost. It is lower than in the case of *false positive* because the patient is sent to a reference center, where a good diagnosis can still be made. However, the cost in the case when the tumor is malignant is twice as high as for a benign tumor.

Figure 5 presents the characteristics of the total cost variability in both scenarios (scenario 1 – left, scenario 2 – right plot) depending on the simulated level of uncertainty in the training set. For comparison, the values obtained by the most commonly used medical diagnostic models for ovarian tumors are also presented (blue, upper dashed lines). We also plot the cost of *OEA* model (green, lower dashed line) which was achieved with 1000 repeats compared to 10 in this study.

Unlike the original diagnostic models, the increase in uncertainty level does not result in a significant increase in the cost of classification for the proposed methods. This feature allows concluding that the proposed classification methods can work effectively even at a 50% level of incompleteness.

Classic diagnostic models are based on very simple mathematical methods, so it should not come as a surprise that the proposed methods obtained a much lower average total classification cost. The comparison also includes the OEA model, for which the total cost is generally slightly lower (72) than the results obtained by the proposed methods. However, this difference is not large, especially considering that the classifiers tested are more versatile because they do not use external medical knowledge. The average total classification cost for both scenarios is presented in Figure 6.

In the first Scenario, the overall total cost is small. For a Goguen Fuzzy Equality based classifier, it is 91.4 using the mean as aggregation operator. It is a model based solely on raw data without additional medical knowledge. Adding weights to the model (based on medical knowledge) did not significantly improve the efficiency of classification, from which one can conclude about the good ability to differentiate input data based on the uncertainty level. On the other hand, the use

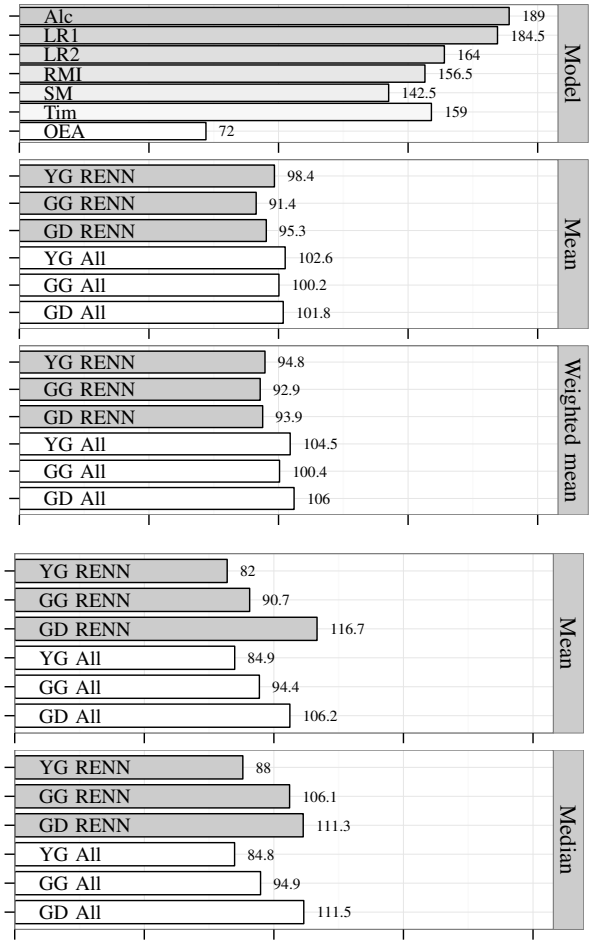


Fig. 6: The average total classification cost for diagnostic models and proposed classification methods. Scenario 1 and 2 (upper and lower parts, respectively).

of instance selection techniques to detect outliers (RENN) significantly improves classification efficiency (cost reduction by almost 10%).

In the second scenario, the results are very similar to the OEA model, but at the expense of using external medical knowledge to reduce the data set. However, this reduction significantly weakened the positive impact of instance selection (RENN). It can be concluded that the simultaneous reduction using diagnostic models and selection of instances can lead to the loss of relevant information. One may observe that the median achieves results comparable to the arithmetic mean but have better computational complexity.

## V. Conclusions and Further Work

The presented results form the basis for the adoption of a new approach to the problem of medical data classification. Rather than ignoring or artificially improving the quality of the input data, it should be explicitly addressed in the data model, and then in the classifier during both the learning and testing. This approach offers many advantages comparing to data editing. The classifier with more accurate (though incomplete)

description of reality, will be able to learn the greater number of dependencies, and then make a better classification.

In this paper, we showed that the use of *uncertainty aware* similarity measure in uncertain data classification leads to easy-to-interpret models thanks to the proposed Nearest Neighbour based method. Moreover, pre-processing of uncertain data improves the classification effectiveness.

We also showed that developed classification methods achieve better results for incomplete data than classical diagnostic models applied to the data set. Moreover, results are comparable with state of the art diagnostic model – OEA. The advantage of the proposed method is that is general and does not require any domain-specific assumptions. This allows to conclude that utilization of all available knowledge (including uncertain one) improves efficiency in medical decision support systems.

For further research, we recommend a deeper study of *uncertainty aware* similarity measures and their practical as well as computational properties. As was indicated in the presented results, the selection of appropriate Equality Value has a significant impact on the classification quality. Finding and investigating new equality values and considering equivalence relationships for interval data can result in a further increase in classification quality [32], [33].

Proposed methods should be also evaluated on different, non-medical data sets. This should also include the optimization of classification computational complexity. As mentioned in the paper, currently due to the complex nature of data handled, there is no known way to speed up classification using tree data structures. This problem needs to be solved to handle big data sets.

## REFERENCES


[1] D. Dubois and H. Prade, "Gradualness, uncertainty and bipolarity: Making sense of fuzzy sets," *Fuzzy Sets and Systems*, vol. 192, pp. 3–24, 2012.

[2] S. Hatch, *Snowball in a Blizzard: A Physician's Notes on Uncertainty in Medicine*. New York: Basic Books, 2016.

[3] R. Moszyński, P. Żywica *et al.*, "Menopausal status strongly influences the utility of predictive models in differential diagnosis of ovarian tumors: An external validation of selected diagnostic tools," *Ginekologia Polska*, vol. 85, no. 12, pp. 892–899, 2014.

[4] P. Żywica and M. Baczyński, "An effective similarity measurement under epistemic uncertainty," *Fuzzy sets and systems*, 2020, in review.

[5] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[6] V. V. Cross and T. A. Sudkamp, *Similarity and Compatibility in Fuzzy Set Theory. Assessment and Applications*. Heidelberg: Physica-Verlag, 2002.

[7] P. Żywica, "Modelling medical uncertainties with use of fuzzy sets and their extensions," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, Cham, 2018, pp. 369–380.

[8] I. Couso and L. Sánchez, "Additive similarity and dissimilarity measures," *Fuzzy Sets and Systems*, vol. 322, pp. 35–53, 2017.

[9] ——, "A note on "similarity and dissimilarity measures between fuzzy sets: A formal relational study" and "additive similarity and dissimilarity measures"," *Fuzzy Sets and Systems*, 2019, in press.

[10] J. C. Fodor and M. Roubens, *Fuzzy preference modelling and multicriteria decision support*. Springer Science & Business Media, 1994, vol. 14.

[11] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-based classification: Concepts and algorithms," *Journal of Machine Learning Research*, vol. 10, no. Mar, pp. 747–776, 2009.

[12] D. H. Hong and C.-H. Choi, "Multicriteria fuzzy decision-making problems based on vague set theory," *Fuzzy Sets and Systems*, vol. 114, no. 1, pp. 103–113, 2000.

[13] R. Brüggemann and L. Carlsen, "An improved estimation of averaged ranks of partial orders," *MATCH Commun. Math. Comput. Chem*, vol. 65, pp. 383–414, 2011.

[14] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[15] S. M. Omohundro, *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989.

[16] J. Alcalá, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2010.

[17] P. Żywica, "Uncertain classification project source code," (online) http://github.com/bikol/uncertain-classification.

[18] A. M. Palacios, L. Sánchez, and I. Couso, "Extending a simple genetic cooperative-competitive learning fuzzy classifier to low quality datasets," *Evolutionary Intelligence*, vol. 2, no. 1-2, pp. 73–84, 2009.

[19] ——, "Diagnosis of dyslexia with low quality data with genetic fuzzy systems," *International Journal of Approximate Reasoning*, vol. 51, no. 8, pp. 993–1009, 2010.

[20] P. Żywica and A. Stachowiak, "Uncertainty-aware similarity measures – properties and construction method," in *2019 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)*. Atlantis Press, 2019, pp. 512–519.

[21] D. Szpurek, R. Moszynski, S. Szubert, and S. Sajdak, "Urban and rural differences in characteristics of ovarian cancer patients," *Annals of Agricultural and Environmental Medicine*, vol. 20, no. 2, 2013.

[22] D. W. Cramer, G. B. Hutchison, W. R. Welch, R. E. Scully, and K. J. Ryan, "Determinants of ovarian cancer risk. I. Reproductive experiences and family history," *Journal of the National Cancer Institute*, vol. 71, no. 4, pp. 703–709, 1983.

[23] D. Timmerman, L. Valentin, T. Bourne, W. Collins, H. Verrelst, I. Vergote *et al.*, "Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group," *Ultrasound in Obstetrics and Gynecology*, vol. 16, no. 5, pp. 500–505, 2000.

[24] R. G. Moore, A. K. Brown, M. C. Miller, S. Skates, W. J. Allard, T. Verch, M. Steinhoff, G. Messerlian, P. DiSilvestro, C. O. Granai *et al.*, "The use of multiple novel tumor biomarkers for the detection of ovarian carcinoma in patients with a pelvic mass," *Gynecologic Oncology*, vol. 108, no. 2, pp. 402–408, 2008.

[25] A. Wójtowicz, P. Żywica, A. Stachowiak, and K. Dyczkowski, "Solving the problem of incomplete data in medical diagnosis via interval modeling," *Applied Soft Computing*, vol. 47, pp. 424–437, 2016.

[26] B. Wang *et al.*, "Distance-based outlier detection on uncertain data," in *Ninth IEEE International Conference on Computer and Information Technology 2009*, 2009, pp. 293–298.

[27] C. C. Aggarwal and P. S. Yu, "Outlier detection with uncertain data," in *Proceedings of the 2008 SIAM International Conference on Data Mining*, 2008, pp. 483–493.

[28] P. Hart, "The condensed nearest neighbor rule," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 515–516, 1968.

[29] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, pp. 408–421, 1972.

[30] G. Li, N. Japkowicz, T. J. Stocki, and R. K. Ungar, "Instance selection by border sampling in multi-class domains," in *International Conference on Advanced Data Mining and Applications*. Springer, 2009, pp. 209–221.

[31] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. New York: Cambridge University Press, 2011.

[32] B. Pekala, U. Bentkowska, H. Bustince, J. Fernandez, and M. Galar, "Operators on intuitionistic fuzzy relations," in *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015, pp. 1–8.

[33] M. Elkano, J. A. Sanz, M. Galar, B. Pekala, U. Bentkowska, and H. Bustince, "Composition of interval-valued fuzzy relations using aggregation functions," *Information Sciences*, vol. 369, pp. 690–703, 2016.