INTERVAL-VALUED AGGREGATION AS A TOOL TO IMPROVE MEDICAL DIAGNOSIS

Andrzej Wójtowicz, Patryk Żywica, Anna Stachowiak and Krzysztof Dyczkowski
Faculty of Mathematics and Computer Science
Adam Mickiewicz University in Poznań
Umultowska 87, 61-614 Poznań, Poland
min@amu.edu.pl

Summary

In the paper we present experimental results on the problem of an effective decision making on incomplete data. In order to investigate this problem we examined a variety of interval aggregation methods. Exemplary results are based on a medical diagnosis support system. Our research shows that an application of the aggregation in this problem leads to promising results.

Keywords: interval-valued aggregation, missing data, incomplete information, decision making under uncertainty, supporting medical diagnosis.

1 INTRODUCTION

According to recent statistics, the annual numbers of deaths due to ovarian cancer in some countries are alarmingly high and still grow. The correct diagnosis became a serious problem that the medicine is trying to face. The correct classification of a tumor as malignant or benign is particularly important since a given type of the tumor determines whether the patient must undergo a surgery. Moreover, incorrect indication of malignant tumor as benign, in the longer term causes deterioration of the patient's health and results in a high risk of failure of the surgery.

For this reason, a wide range of preoperative diagnostic models have been developed, where the goal is to predict the type of malignancy. Both the sensitivity and specificity of the models rarely exceeds 90% in external evaluation [14, 8]. The Table 1 presents six most common ones (two based on scoring systems [1, 11] and four based on logistic regressions [13, 12, 7]) and a list of used attributes. The attributes are divided into two groups: objective medical history and the rest (which are subjective medical history, ultrasound and blood markers).

Our previous research indicated possible problems with collecting all the data by a physician during examinations [15, 10]. It is common that some examinations might be omitted by a gynaecologist, either due to the their unavailability or because of medical reasons. The possible lack of data can be due to e.g. the technical limitations of the health care unit, high costs of medical examination and high risk of patient's health deterioration after potential examination. Obviously, lack of data hinders making a final decision.

The main issue we investigate in this paper is how to overcome the problem of low-quality diagnosis in the presence of missing data. The approach presented in the following sections focuses on aggregating knowledge that comes from many diagnostic scales, in order to minimise the impact of incomplete data. In Section 2 we introduce notions of an interval-based model of a patient, a diagnostic scale and aggregation operator. In Section 3 we describe an aggregation strategy that allows to improve a diagnosis and we give a methodology of calculating, analysing and compering different aggregation methods. Section 4 concludes our results.

2 DESCRIPTION OF THE PROBLEM

2.1 INTERVAL-VALUED PATIENT MODEL

In a classical approach, a patient is modelled as a vector $\mathbf{p} = (p_1, p_2, ..., p_n)$ in a space $P = D_1 \times D_2 \times ... \times D_n$, where $D_1, D_2, ..., D_n$ are real closed intervals denoting domains of attributes that describe patients.

We extend this representation by introducing a possibility to model incomplete data. Each attribute D_i is substituted by its interval version $\hat{D}_i = \mathcal{I}_{D_i}$. Analo-

ISBN: 978–83–8012–519–3

Table 1: Attributes used by most common preoperative diagnostic models.

group	attribute	SM [11] m_1	Alc. [1] m_2	LR1 [13] m_3	LR2 [13] m_4	Timm. [12] m_5	RMI1 [7] m_6
objective medical history	age	-	-	✓	✓	-	✓
	menopausal status	✓	-	-	-	✓	✓
	ovarian cancer in family	-	-	✓	-	-	-
	hormonal therapy	-	-	✓	-	-	-
	hysterectomy	-	-	-	-	-	✓
other	pain during examination	-	-	✓	-	-	-
	lesion volume	✓	-	✓	-	-	-
	internal cyst walls	✓	-	✓	✓	-	-
	septum thickness	✓	-	-	-	-	-
	echogenicity	✓	✓	-	-	-	-
	localisation	✓	-	-	-	-	✓
	ascites	✓	-	✓	✓	-	✓
	papillary projections	-	✓	-	-	✓	-
	solid element size	-	✓	✓	✓	-	✓
	blood flow location	-	✓	✓	✓	-	-
	resistance index	-	✓	-	-	-	-
	acoustic shadow	-	-	✓	✓	-	-
	amount of blood flow	-	-	✓	-	✓	-
	CA-125 blood marker	-	-	-	-	✓	✓
	lesion quality class	-	-	-	-	-	✓

gously as before we define $\hat{P} = \hat{D}_1 \times \hat{D}_2 \times ... \times \hat{D}_n$. Consequently, for each vector $\mathbf{p}^* \in P^*$ we can define its interval equivalent $\hat{\mathbf{p}} \in \hat{P}$ that has a form $\hat{\mathbf{p}} = \Big([\underline{p}_1, \overline{p}_1], ..., [\underline{p}_n, \overline{p}_n]\Big)$, where

$$\underline{p}_i = \begin{cases} p_i & \text{if } p_i \neq \emptyset \\ \min\limits_{d \in D_i} d & \text{if } p_i = \emptyset \end{cases}, \overline{p}_i = \begin{cases} p_i & \text{if } p_i \neq \emptyset \\ \max\limits_{d \in D_i} d & \text{if } p_i = \emptyset \end{cases}. \quad (1)$$

2.2 INTERVAL-VALUED DIAGNOSTIC SCALES

Diagnostic scale can be formalised as a function $m : P \to [0,1]$. Values returned by a function indicate malignancy of a tumor and are interpreted in the following way:

- $m(\mathbf{p}) > 0.5$ – diagnosis towards malignant;
- $m(\mathbf{p}) < 0.5$ – diagnosis towards benign;
- $m(\mathbf{p}) = 0.5$ – indicates the impossibility of determining the nature of malignancy.

We construct an extended diagnostic scale $\hat{m} : \hat{P} \to \mathcal{I}_{[0,1]}$ defined as:

$$\hat{m}(\hat{\mathbf{p}}) = \Big\{ m(\mathbf{p}) : \forall_{1 \le i \le n} \quad \underline{p}_i \le p_i \le \overline{p}_i \Big\} = \left[\min_{\mathbf{p} \in \hat{\mathbf{p}}} m(\mathbf{p}), \max_{\mathbf{p} \in \hat{\mathbf{p}}} m(\mathbf{p}) \right] \quad (2)$$

where by $\mathbf{p} \in \hat{\mathbf{p}}$ we denote that $\mathbf{p}$ is an embedded vector of $\hat{\mathbf{p}}$.

Such extended diagnostic scale is able to operate on interval-valued representation of a patient. The resultant interval represents all the possible diagnoses that can be made basing on a patient description, in which every missing value was substituted with all possible values for that attribute. The more incomplete description, the more uncertain the diagnosis. However, worth noticing is that in many cases it is still possible to make a proper diagnosis, since some amount of missing values is acceptable and would not affect the final result significantly.

2.3 INTERVAL-VALUED AGGREGATION

A diagnosis in a form of an interval (2) has its advantages and drawbacks. An advantage is that such model gives a diagnosis even in the presence of missing data. A drawback is that the diagnosis is often uncertain and not so easy to apply by a physician. A main problem is thus how to efficiently support a physician in making a final diagnosis under incomplete information.

In order to solve this problem we make a following observation. As presented in Table 1 different scales denoted by $m_1, \dots m_n$ use different attributes describing the patient, and therefore are subjected to different levels of uncertainty. The main idea is thus to improve the final diagnosis by taking advantage of the diversity of diagnostic scales. Given n scales $\hat{m}_1, \ldots \hat{m}_n$ we construct a function $\text{Agg} : {\mathcal{I}_{[0,1]}}^n \to \mathcal{I}_{[0,1]}$. Its result $\text{Agg}(\hat{m}_1, \ldots \hat{m}_n)$ is an interval that gathers and integrates information from the input sets. Thanks to this interpretation we immediately see the relationship with the issue of group decision making and information aggregation [3].

An n-argument interval–valued aggregation operator is a mapping $\text{Agg} : {\mathcal{I}_{[0,1]}}^n \to \mathcal{I}_{[0,1]}$ with the following properties [6]:

1. if $\hat{x}_i \preceq \hat{y}_i$ for all $i \in 1, ..., n$, then $\text{Agg}(\hat{x}_1, ..., \hat{x}_n) \preceq \text{Agg}(\hat{y}_1, ..., \hat{y}_n)$,
2. $\text{Agg}([1,1], ..., [1,1]) = [1,1]$,
3. $\text{Agg}([0,0], ..., [0,0]) = [0,0]$,

where relation $\preceq$ is defined as follows:

$$[x_1, x_2] \preceq [y_1, y_2] \iff x_1 \leq y_1 \text{ and } x_2 \geq y_2 .$$

Recent research has led to the construction of many interval–valued aggregation methods [16, 5, 6, 2]. The most commonly used aggregation methods in group decision making are based on the weighted arithmetic average [3]. We propose to use various aggregation methods to improve the quality of diagnosis as well as minimise the impact of the lack of data and uncertainty in decision making.

In medical decision making problem, final diagnosis obtained from aggregation must indicate whether tumor is malignant or not. However, supporting decision in the case when there is not enough information may led to wrong diagnosis. Thus we accept situation when no diagnosis recommendation is made. The conversion of interval diagnosis into final diagnosis (binarization) is very important and may influence overall efficacy.

3 AGGREGATION PROCESS

3.1 METHODOLOGY DESCRIPTION

The proposed methodology is aimed at evaluating aggregation operators for coping with the lack of data. For this purpose, an essential element of the methodology is to simulate different levels of missing data. In order to better reflect the reality we have divided the attributes that describe the patient into two separate groups: those that were subjected to obscuration and those that were not. This separation naturally exists in many problems, including the problem of medical diagnosis because some data about the patient, such as age and other objective data from the medical history, are always available to the physician.

The results of all diagnostic scales are represented as intervals. The list of those intervals forms an input to the aggregation operators. Each operator synthesise input diagnoses in accordance with its principle of operation. The result of the aggregation is an interval representing the synthesised diagnosis. In order to make the final diagnosis it is required to perform the binarization process. Resulting diagnoses are compared to reference values in order to calculate the necessary statistics. The final statistics for a given level of missing data are calculated by averaging the results of all iterations.

Our study group was 268 women diagnosed and treated due to ovarian tumor in the Division of Gynaecological Surgery, Poznań University of Medical Sciences between 2005 and 2012. Among them, 62% was diagnosed with a benign tumor and 38% with a malignant one. In each iteration we chose 50 patients for positive and negative groups. All patients had no missing values in attributes required by diagnostic scales. Whole dataset is described in details in [8].

In evaluation we used six different diagnostic scales $\hat{m}_1, ..., \hat{m}_6$ obtained from basic scales listed in Table 1.

3.2 AGGREGATION STRATEGY

For the experiment we chose the simplest methods of aggregation, which base on weighted average, sum and intersection of sets, and majority vote. Such methods are most often used in the problem of group decision making [3]. However, the authors are aware that these methods do not cover recent research in that field (e.g. [4]).

A construction of a certain method of aggregation consists in choices of aggregation strategy and binarization strategy. Moreover, methods which base on weighted average require definition of weight of intervals. We chose 10 methods of aggregation presented in Table 2.

First group of aggregation operators (A–C) is based on arithmetic mean with use of interval arithmetic:

$$\text{Agg}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = \frac{\sum_{i=1}^{n} \omega(\hat{x}_i) \times \hat{x}_i}{\sum_{i=1}^{n} \omega(\hat{x}_i)} .$$

Second group of the operators (D–G) is based on weighted mean which is calculated with reference to a representative (rep) of the interval. Selected strategies in choosing representatives are minimum, maximum and centre of a interval:

$$\text{Agg}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = \frac{\sum_{i=1}^{n} \omega(\hat{x}_i) \cdot \text{rep}(\hat{x}_i)}{\sum_{i=1}^{n} \omega(\hat{x}_i)} .$$

Next two operators (H–I) are based on sum and in-

Table 2: Selected aggregation methods.

ID	aggregation	strategy of weight calc.	binarization
A	Interval avg.	width	margin
B	Interval avg.	entropy	no margin
C	Interval avg.	constant	no margin
D	Lower bound avg.	width	margin
E	Upper bound avg.	width	margin
F	Center avg.	width	margin
G	Center avg.	entropy	margin
H	Intersection	-	no margin
I	Sum	-	no margin
J	Majority vote	-	margin

tersection from the set theory:

$$\mathrm{Agg}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = \bigcup_{i=1}^{n} \hat{x}_i$$

and

$$\mathrm{Agg}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = \bigcap_{i=1}^{n} \hat{x}_i \,.$$

Last method (J) differs from former in a way that it firstly binarizes input intervals, and after that it chooses a diagnose which appeared more frequently. In case of draw, the diagnosis is not taken.

3.3 WEIGHT CALCULATION STRATEGY

In our evaluation we selected three strategies for choosing weights:

- constant value: $\omega([a, b]) = 1$,
- interval length: $\omega([a, b]) = b - a$,
- normalised interval distance from 0.5 (entropy):

$$\omega([a,b]) = \begin{cases} 0 & \text{if } a \le 0.5 \le b \\ 2(a - 0.5) & \text{if } a \ge 0.5 \\ 2(0.5 - b) & \text{otherwise}\,. \end{cases}$$

3.4 BINARIZATION STRATEGY

In our research we chose the simplest variant of interval binarization:

$$\tau_\epsilon([a,b]) = \begin{cases} 0 & \text{if } b < 0.5 + \epsilon \\ 1 & \text{if } a \ge 0.5 - \epsilon \\ \text{NA} & \text{otherwise}\,. \end{cases} \quad (3)$$

In this approach, an instance is classified as a positive when whole interval is greater than 0.5 with respect to ϵ margin. The negative case is defined similarly. In case when first or second conditions are not met, it is not possible to make a decision. For example, diagnosis $[0.1, 0.3]$ will be classified as benign, but for interval $[0.1, 0.6]$ it is not possible to make a decision, when margin is set to $\epsilon = 0.025$.

In our evaluation we arbitrarily chose two values for ϵ: 0 (no margin) and 0.025.

3.5 EVALUATION

Statistical evaluation as well as implementation of proposed methodology were performed using R software, version 3.1.1 [9]. We set levels of missing data to vary from 0% to 50% with 5% step. For each level we made 1000 repeats of random data obscuration with other calculations. With such number of repeats, the averaged results are stable, so that it is possible to reliably analyse them. We set a baseline to 69% accuracy that is achieved by a classifier based only on menopausal status – all methods of aggregation should be better than a baseline classifier and single diagnostic scales.

The most significant results are presented on Fig. 1. The figure presents how the aggregators and single diagnostic scales perform with increasing level of missing data. Sub-figure (a) presents diagnostic accuracy (ACC) and sub-figure (b) presents percentage of patients for whom the decision could be made. Upper and lower bounds of the shaded regions correspond to the biggest and the smallest values achieved among diagnostic scales.

The diagrams show that preserving high diagnosability frequently prevent models from achieving high accuracy – and *vice versa*.

4 RESULTS AND CONCLUSIONS

The developed methodology led us to conclusion that in the case of our medical diagnosis problem, aggregation is useful as a tool to solve the problem of missing data. Even the simplest methods presented in this paper received efficacy which exceed individual diagnostic scales, both in terms of accuracy and the number of diagnosed patients, despite missing data. There are three interesting cases:

1. result of an aggregation is an achievement of very high and stable accuracy (over 95%, regardless to level of missing data) at the expense of small number of patients, in which it was possible to make a diagnose (below 50%, less than individual diagnostic scales) – see e.g. I aggregation operator,
2. result of an aggregation is an achievement of very high diagnosability (over 90%) regardless of the

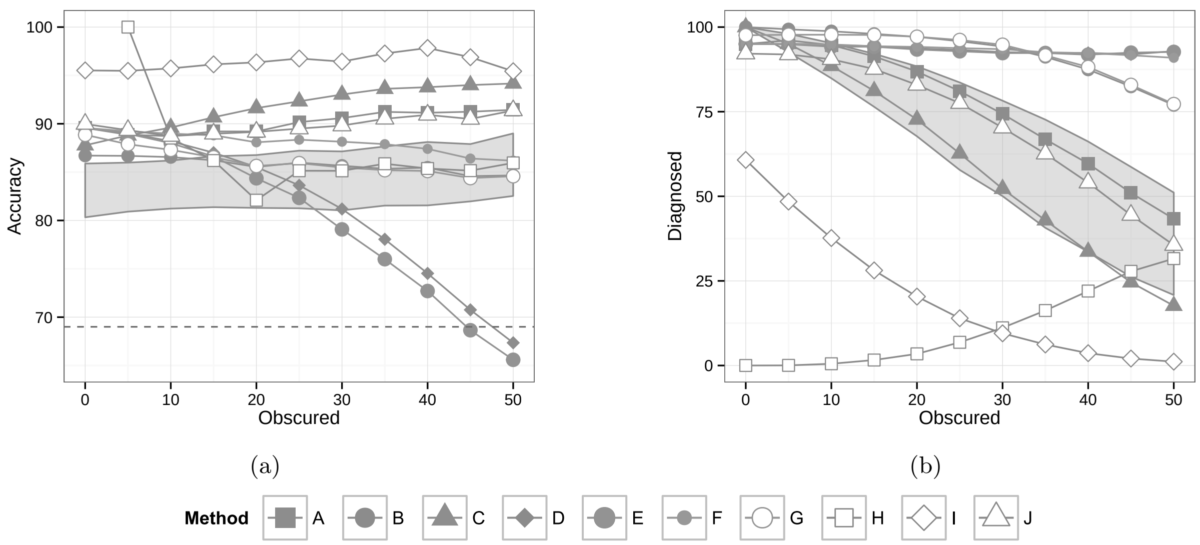

Figure 1: Results of evaluation of selected aggregation methods. Shaded region indicates bounds of single diagnostic methods. A dashed horizontal line in sub-figure (a) indicates accuracy of the baseline classifier.

level of missing data, at the expense of decreasing accuracy with respect to increasing level of missing data (accuracy might be even lower than these achieved by individual diagnostic scales) – see e.g. E and F aggregation operators,

3. result of an aggregation is an achievement of persistent high accuracy which is comparable to those achieved by individual diagnostic scales, with simultaneous high level of diagnosability (significantly higher than those achieved by individual diagnostic scales) – see e.g. F i G.

In the problem of ovarian tumor diagnosis, it seems that the most promising results were obtained by aggregation operator F. It is capable of maintaining high accuracy and diagnosability. Its little sensitivity to the lack of data makes it a promising candidate in the search for robust aggregation operators.

An interesting result is obtained in the case of the aggregation operator H, which is based on a set intersection. In the case of complete data it is not able to make any decision. This is due to the fact that in such situation the intervals are degenerated to a single points, and the intersection of such intervals is mostly an empty set.

The authors are aware that since the evaluation was performed on the whole dataset with arbitrarily chosen binarization margins, general conclusions on the performance of the presented aggregation operators should not be drawn. To make the results more reliable, the performance should be validated on a separate dataset with optimised aggregation parameters.

In the future work, we are going to study broader range of aggregation methods and determine guidelines on their applicability to various problems.

The presented results are promising and show that the competent selection and use of aggregation methods can significantly improve the quality of decisions taken by a diagnostic system. The problem is particularly significant when the knowledge is based on incomplete information. Proper selection of the method of aggregation is essential for weakening the negative impact of the incomplete data on the quality of decisions. Because the design of aggregation method depends on the particular problem, each time its extensive evaluation is needed. It can be done using our proposed method.

References

[1] J. L. Alcázar, L. T. Mercé, et al. A new scoring system to differentiate benign from malignant adnexal masses. *Obstetrical & Gynecological Survey*, 58(7):462–463, 2003.

[2] G. Beliakov, H. Bustince, et al. Aggregation for atanassov's intuitionistic and interval valued fuzzy sets: The median operator. *Fuzzy Systems, IEEE Transactions on*, 20(3):487–498, 2012.

[3] G. Beliakov, A. Pradera, et al. *Aggregation functions: A guide for practitioners*. Springer, 2008.

[4] L. De Miguel, H. Bustince, J. Fernandez, E. Induráin, A. Kolesárová, and R. Mesiar. Construction of admissible linear orders for interval-valued atanassov intuitionistic fuzzy sets with an application to decision making. *Information Fusion*, 2015, doi:10.1016/j.inffus.2015.03.004.

[5] G. Deschrijver and E. Kerre. Implicators based on binary aggregation operators in interval-valued fuzzy set theory. *Fuzzy Sets and Systems*, 153(2):229–248, 2005.

[6] G. Deschrijver and E. Kerre. Aggregation operators in interval-valued fuzzy and Atanassov's intuitionistic fuzzy set theory. In *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*, pages 183–203. Springer, 2008.

[7] I. Jacobs, D. Oram, et al. A risk of malignancy index incorporating ca 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *BJOG: An International Journal of Obstetrics & Gynaecology*, 97(10):922–929, 1990.

[8] 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*, 85(12):892–899, 2014.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[10] A. Stachowiak, P. Żywica, et al. An interval-valued fuzzy classifier based on an uncertainty-aware similarity measure. In P. Angelov et al., editors, *Intelligent Systems'2014*, volume 322 of *Advances in Intelligent Systems and Computing*, pages 741–751. Springer International Publishing, 2015.

[11] D. Szpurek, R. Moszyński, et al. An ultrasonographic morphological index for prediction of ovarian tumor malignancy. *European Journal of Gynaecological Oncology*, 26(1):51–54, 2005.

[12] D. Timmerman, T. H. Bourne, et al. A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: the development of a new logistic regression model. *American Journal of Obstetrics and Gynecology*, 181(1):57–65, 1999.

[13] D. Timmerman, A. C. Testa, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. *Journal of Clinical Oncology*, 23(34):8794–8801, 2005.

[14] C. Van Holsbeke, B. Van Calster, et al. External validation of mathematical models to distinguish between benign and malignant adnexal tumors: a multicenter study by the International Ovarian Tumor Analysis Group. *Clinical Cancer Research*, 13(15):4440–4447, 2007.

[15] A. Wójtowicz, P. Żywica, et al. Dealing with Uncertainty in Ovarian Tumor Diagnosis. In K. Atanassov et al., editors, *Modern Approaches in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics*, pages 151–158. IBS PAN – SRI PAS, Warsaw, 2014.

[16] R. Yager. Owa aggregation of intuitionistic fuzzy sets. *International Journal of General Systems*, 38(6):617–641, 2009.