

# Fuzzy Dempster-Shafer Modelling and Decision Rules

Zenon A. Sosnowski<sup>(✉)</sup> and Jarosław S. Walijewski

Faculty of Computer Science, Białystok University of Technology,  
Wiejska 45A, 15-351 Białystok, Poland  
z.sosnowski@pb.edu.pl

**Abstract.** In this study, we discuss the use of Dempster-Shafer theory as a well-rounded algorithmic vehicle in the construction of fuzzy decision rules. The concept of fuzzy granulation realized via fuzzy clustering is aimed at the discretization of continuous attributes. Detailed experimental studies are presented concerning well-known medical data sets available on the Web.

**Keywords:** Dempster-Shafer theory · Fuzzy modelling · Fuzzy decision rules

## 1 Introduction

Fuzzy modeling is regarded to be one of the possible classification architecture of machine learning and data mining. There have been a significant number of studies devoted to generating fuzzy decision rules from sample cases or examples. These include attempts to extend many classical machine learning methods to learn fuzzy rules. One very popular approach is decision trees [10]. Since the inception of this concept, it has been extended for the construction and interpretation of more advanced decision trees [3, 5–7, 9, 13, 15, 18]. Although the decision trees based methods can extract a set of fuzzy rules which works well, a problem is that the lack of backtracking in splitting the node leads to lower learning accuracy comparing to other machine learning methods. Another widely used machine learning method is artificial neural network. In recent years enormous work has been done in attempt to combine the advantages of neural network and fuzzy sets [14]. Hayashi [4] has proposed to extract fuzzy rules from trained neural net. Lin [8], on the other hand, introduced a method of directly generating fuzzy rules from self-organized neural network. The common weakness of neural network, however, is a problem of determination of the optimal size of a network configuration, as this has a significant impact on the effectiveness of its performance.

The objective of this paper is to employ the Dempster-Shafer theory (DST) as a vehicle supporting the generation of fuzzy decision rules. More specifically, we concentrate on the role of fuzzy operators, and on the problem of discretization of continuous attributes. We show how they can be effectively used in the quantization of attributes for the generation of fuzzy rules.

The material is arranged in the following way. First, we summarize the underlying concepts of the Dempster-Shafer theory and briefly discuss the nature of the underlying

construction. By doing so, the intension is to make the paper self-contained and help identify some outstanding design problems emerging therein. In Sect. 4 we explain essentials of our model. Finally, in Sect. 5, we report exhaustive experimental studies.

This paper is a continuation of our earlier work [12]. Here we apply theoretical vehicle, introduced in previous research, to the new input data in order to find possible area of application. Our important objective here is to reveal a way in which this approach becomes essential to a more comprehensive treatment of continuous attributes.

## 2 Dempster-Shafer Theory

The Dempster-Shafer theory starts by assuming a Universe of Discourse  $\Theta$  also called **Frame of Discernment**, which is a finite set of mutually exclusive alternatives. The frame of discernment may consist of the possible values of an attribute. For example, if we are trying to determine the disease of a patient, we may consider  $\Theta$  being the set consisting of all possible diseases.

For each subset S of  $\Theta$  it is associated:

- a basic probability assignment  $m(S)$
- a belief  $Bel(S)$
- a plausible belief  $Pla(S)$

$m(S)$ ,  $Bel(S)$  and  $Pla(S)$  have value in the interval  $[0,1]$ , and  $Bel(S)$  is not greater than  $Pla(S)$ .

In particular,  $m$  represents the strength of some evidence. For example, in rule-based expert system,  $m$  may represent the effect of applying of a rule.  $Bel(S)$  summarizes all our reasons to believe S.  $Pla(S)$  expresses how much we should believe in S if all currently unknown facts were to support S. Thus the true belief in S will be somewhere in the interval  $[Bel(S), Pla(S)]$ . More formally, a map

$$m : 2^\Theta \rightarrow [0, 1] \tag{1}$$

such that for each  $A \in 2^\Theta$  (where  $2^\Theta$  is set of all subsets of  $\Theta$ )

1.  $m(\emptyset) = 0$
2.  $\sum_{A \subseteq \Theta} m(A) = 1$

is called a **basic probability assignment** for  $\Theta$ .

Subset A is called a **focal element** of  $m$  if  $m(A) > 0$ .

For a given **basic probability assignment**  $m$ , the **Belief** of a subset A of  $\Theta$  is the sum of  $m(B)$  for all subsets B of A, so

$$Bel : 2^\Theta \rightarrow [0, 1] \tag{2}$$

such that  $Bel(A) = \sum_{B \subseteq A} m(B)$ .

The **Plausibility** of a subset A of  $\Theta$  is defined as  $Pla(A) = 1 - Bel(A')$ , where  $A'$  is the complement of A in  $\Theta$ .

If we are given two basic probability assignments  $m_1$  and  $m_2$ , we can combine them into a third basic probability assignment  $m : 2^\Theta \rightarrow [0, 1]$  in the following way.

Let us consider frame of discernment  $\Theta$  and two belief functions  $Bel_1$  and  $Bel_2$ . We denote focal elements of  $Bel_1$  as  $A_1 \dots A_K$  and focal elements of  $Bel_2$  as  $B_1 \dots B_L$  respectively and the basic probability assignments as  $m_1$  and  $m_2$ . Then we can show graphically this combination as an orthogonal sum of  $m_1$  and  $m_2$ .

The mass of probability of the interval  $A_i \cap B_j$  expressed as a measure  $m_1(A_i) \cdot m_2(B_j)$  is illustrated in Fig. 1.

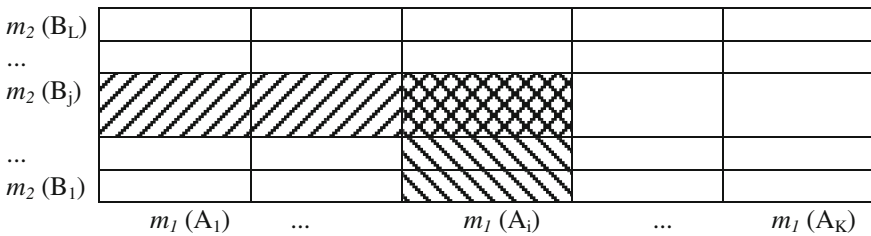


Fig. 1. Orthogonal sum of  $m_1$  and  $m_2$

Of course, more intersections can give the same focal element A. In general the mass of probability of set A is defined as:

$$m(A) = \sum_{\substack{i,j \\ A_i \cap B_j = A}} m_1(A_i) \cdot m_2(B_j) \tag{3}$$

Then we find a problem with an empty set. It is possible that sets with empty intersection exist. We can meet this normal situation in many combinations. Then the mass of  $\emptyset$ , according to above definition, will be greater than zero, but according to the definition of basic probability assignment, it is not possible.

We assume that

$$\sum_{\substack{i,j \\ A_i \cap B_j = \emptyset}} m_1(A_i) \cdot m_2(B_j) < 1 \tag{4}$$

to define the orthogonal sum  $m_1$  and  $m_2$ , and denote it as  $m_1 \oplus m_2$ .

Then it is necessary to change the definition (3) of the formula of the basic probability assignment of the combination as follows:

1.  $m(\emptyset) = 0$
2.  $m(A) = \frac{\sum_{i,j} m_1(A_i) \cdot m_2(B_j)}{1 - \sum_{i,j} m_1(A_i) \cdot m_2(B_j)}$  for non-empty  $A \subset \Theta$   
 $A_i \cap B_j = \emptyset$

We call this as an orthogonal sum of  $Bel_1$  and  $Bel_2$ , and denote as  $Bel_1 \oplus Bel_2$ .

This is the **Dempster Rule for Combining** of Beliefs [11].

For  $\Theta$  we have  $m(\Theta) = \sum_{A \subseteq \Theta} m(A) = 1$  and for combination if

$$\sum_{i,j} m_1(A_i) \cdot m_2(B_j) = 0$$

$$A_i \cap B_j = \emptyset$$

we have

$$m(\Theta) = \sum_{A \subseteq \Theta} m(A) = \sum_{A \subseteq \Theta} \sum_{i,j} m_1(A_i) \cdot m_2(B_j) = \sum_{i,j} m_1(A_i) \cdot m_2(B_j) = 1$$

$$A_i \cap B_j = A$$

### 3 Fuzzy Modelling

Fuzzy set theory is widely known and we do not introduce its underlying concepts essential to understand this framework. Readers interest themselves we refer to [16] and [17].

Fuzzy Modeling is applied in those areas where the model of the system cannot be described precisely because of many reasons. The input data received by the system may not be completely reliable, may contain noise, or may be inconsistent with other data or with expectation about these data. The system is described by the set of linguistic rules. Let  $D$  denotes an output variable of the system, and  $X_1, X_2, \dots, X_n$  denote an input variables. The linguistic rules have the following format:

$$\text{If}(X_1 \text{ is } A_{k,1,j_1}) \text{ And } \dots \text{ And } (X_n \text{ is } A_{k,n,j_n}) \text{ Then } (D \text{ is } S_{k,p}) \tag{5}$$

where  $(X_i \text{ is } A_{k,i,j_i})$  are the fuzzy antecedents,  $A_{k,i,j_i}$  ( $1 \leq j_i \leq |A_i|$ ) are values of the  $i$ -th input variable, and  $S_{k,p}$  ( $1 \leq p \leq |S|$ ) is the value of output variable in  $k$ -th rule.

The rules are implemented as fuzzy relation according to the formula:

$$R_k = A_{k,1,j_1} \times A_{k,2,j_2} \times \dots \times A_{k,n,j_n} \times S_{k,p} \tag{6}$$

where  $\times$  denotes the fuzzy Cartesian product.

Then all rules are aggregated to relation  $R$  described as:

$$R = \bigcup_{k=1}^M R_k \tag{7}$$

where  $M$  is the number of rules.

The conclusion is based on the compositional rule of inference

$$S' = \left( A'_{1,j_1} \times A'_{2,j_2} \times \dots \times A'_{n,j_n} \right) \circ R \tag{8}$$

where  $A'_{1,j_1}, A'_{2,j_2}, \dots, A'_{n,j_n}$  are input values,  $S'$  is a conclusion (decision class), and  $\circ$  denotes the composition of fuzzy relation.

In fuzzy modeling we can assume that expert defines the rule set, or we can automatically generate them from the set of samples describing the behavior of the system being modeled.

### 4 Fuzzy Dempster-Shafer Model

In Fuzzy Dempster-Shafer (FDS) model [2] we consider rules  $R_r$  as:

$$\text{If } (X_1 \text{ is } A_{r,1,j_1}) \dots \text{ And } \dots (X_n \text{ is } A_{r,n,j_n}) \text{ Then } (D \text{ is } m_r) \tag{9}$$

where  $X_i$  and  $D$  stand for input and output respectively, and  $m_r$  is a fuzzy belief structure, that is a standard belief structure with focal elements  $S_{r,p}$  as fuzzy subset of frame of discernment  $\Theta$  with basic probability assignment  $m_r(S_{r,p})$ , and  $m_r(S_{r,p})$  is the believe that the conclusion should be represented as class  $S_{r,p}$ .

#### 4.1 Learning – Rules Construction

In antecedent construction, let us assume that we have  $n$  features (attributes) in antecedents of testing example. We consider a collection of  $m$  generic linguistic terms characterized by membership functions defined in a universe of discourse being a domain of each attribute. The conclusion belongs to decision class  $S$ .

For each element of data  $t$  we build a collection:

$$\begin{matrix} A_{1,1,t} & A_{2,1,t} & \dots & A_{n,1,t} \\ A_{1,2,t} & A_{2,2,t} & \dots & A_{n,2,t} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,m,t} & A_{2,m,t} & \dots & A_{n,m,t} \end{matrix} \tag{10}$$

where  $A_{i,j,t}$  are the values of  $j$ -th membership function for  $i$ -th feature and for  $t$ -th element of data.

**Example 1.** We demonstrate the calculations on the set of synthetic data presented in Table 1.

**Table 1.** Sample data set

Item	F1	F2	F3	F4	Class
L1	1	1	1	1	1
L2	9	8	8	9	2
L3	1	1	3	4	1
L4	2	1	2	2	1
L5	2	2	2	2	2
L6	5	6	7	8	2
T1	3	3	3	2	1
T2	1	2	2	1	1
T3	4	7	7	9	2
T4	2	8	7	8	2

First six rows (L1–L6) will constitute learning data, while the remaining ones (T1–T4) will form testing data. All the features are numbers from the  $<0; 9>$  interval. The last column represents the decision class equal to 1 or 2. We will consider four membership quadratic functions uniformly distributed along the space of all attributes. Other membership functions will be discussed in the next section.

According to (10) for row T1 we have:

$$\begin{matrix}
 1.0 & 1.0 & 0.0625 & 0.0156 \\
 0 & 0 & 0.9375 & 0.9844 \\
 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0
 \end{matrix}$$



On the base of (10), for a given data point  $t$  we can calculate two vectors:

$$A_{\mu,t} : A_{1,\max_{1,t}} \quad A_{2,\max_{2,t}} \quad \dots \quad A_{n,\max_{n,t}} \tag{11}$$

and index of membership functions

$$I_{c,t} : I_{1,\max_{1,t}} \quad I_{2,\max_{2,t}} \quad \dots \quad I_{n,\max_{n,t}} \tag{12}$$

where  $A_{i,\max_{i,t}}$  is a maximum value of all membership functions designed for the feature  $i$  and  $I_{i,\max_{i,t}}$  is the number of the best membership function for feature  $i$ .

Then we have the following candidate for a rule

$$R_t : I_{1,\max_{1,t}} \quad I_{2,\max_{2,t}} \quad \dots \quad I_{n,\max_{2,t}} \tag{13}$$

The firing level of the rule is calculated according to the following formula

$$\tau_t = \bigoplus_{i=1}^n (A_{i,\max_i,t}) \tag{14}$$

where  $\phi$  means the operator of fuzzy matching. See Sect. 5.2. for details.

The rule candidate is added to rules set if  $\phi[\tau_r, m_r] \geq Th$  (where  $Th$  threshold value, and  $\phi$  matching operator). This can help to eliminate bad rules from the final rule set.

More ten one rule can have the same antecedent part but it is also possible that conclusion of these rules are different. Then we have to use appropriate counters  $c_{t,1}, \dots, c_{t,|S|}$ , where  $|S|$  denotes the power of decision class set. These counters can show us how many data, according to rule pattern, vote for each decision class.

**Example 2.** In our sample (T1) the vectors are:

$A_{\mu,1}$ : 1.0000, 1.0000, 0.9375, 0.9844  
 $I_{c,1}$ : 1, 1, 2, 2 with counters vector 1, 0

In our sample matching value equals to 0.9229, were multiplication was used as the matching operator ( $1.0000 * 1.0000 * 0.9375 * 0.9844 = 0.9229$ ). For the threshold set on 0.75, we obtain a new rule. ■

The product is a new belief structure on  $X$

$$\hat{m}_r = \tau_r \wedge m_r \tag{15}$$

Focal elements are fuzzy subset given as

$$F_{r,p}(x) = \tau_r \wedge S_{r,p}(x) \tag{16}$$

and appropriate distributions of new focal elements are defined as:

$$\hat{m}_r(F_{r,p}) = m_r(S_{r,p}) \tag{17}$$

So we can build an aggregate:

$$m = \bigcup_{r=1}^R \hat{m}_r \tag{18}$$

Then for each collection

$$\mathfrak{S} = \{F_{r_1,p_1}, F_{r_2,p_2}, \dots, F_{r_R,p_R}\} \tag{19}$$

where  $F_{r_i,p_i}$  are focal elements of  $\hat{m}_r$  we have focal element  $E$  of  $m$  described as

$$E = \bigcup_{t=1}^R F_{r_t,p_t} \tag{20}$$

with appropriate probability distribution

$$m(E) = \prod_{t=1}^R m(F_{r_t, p_t}) \tag{21}$$

At this point, the rule generation process is complete.

**Example 3.** Our sample data produce the following rule set.

	I <sub>1max</sub>	I <sub>2max</sub>	I <sub>3max</sub>	I <sub>4max</sub>	C <sub>1</sub>	C <sub>2</sub>	m
R1 :	1	1	2	2	1	0	2.5000
R2 :	1	1	4	1	2	1	2.0833
R3 :	4	4	4	4	0	1	1.2500
R4 :	2	3	3	4	0	1	1.2500

The first four elements are numbers of the best membership function for proper features, the next two are counters for decision classes and the last one is a probability distribution.

Let us observe that rule R2 covers the data L1, L4 and L5. L1 and L4 produce decision class C<sub>1</sub> but L5 decision class C<sub>2</sub>. ■

Now we can move to the testing of new rules.

### 4.2 Test

In testing we ignore the value from the last column in Table 1, that is decision class number, because our goal is to calculate it.

To compute the firing level of a rule *k* for a given data

$$X_k : X_{1,k} \ X_{2,k} \ \dots \ X_{n,k} \ D_k \tag{22}$$

where *X<sub>i,k</sub>* – feature’s value, *D<sub>k</sub>* – conclusion decision class that we have to compare with the result of inference; we build a rule matrix

$$\mu_{k,t} = \prod_{i=1}^n (A_{i,l,k}(X_{i,t})), \quad l = I_{i,max,k} \tag{23}$$

We are interested only in active rules i.e. rows with matching value  $\mu_{k,t} > 0$ .

**Example 4.** In test we will demonstrate calculations on

L5.	2	2	2
T1.	3	3	2

For sample data L5 we have two active rules:

R1:	1	1	2	2	1	0	0.859375	0.859375	0.609375	0.609375	0.274242
R2:	1	1	1	1	2	1	0.859375	0.859375	0.859375	0.859375	0.54542



The first four elements are the rule pattern, the next two are the counters for decision classes. The next four numbers are the values of appropriate membership function. The number 0.859375 is the value of the first membership function, according to the first number in the rule, on the first feature. The next three numbers are calculated in the same way.

The last numbers in the above rows are the matching value for the rule. It has been calculated by matching operator for the values of membership function.

We focused only on the rows with matching value greater than zero.

For sample data T1 we have:

R1: 1 1 2 2	1 0	0.4375 0.4375 0.9375 0.6094 0.1093
R2: 1 1 1 1	2 0	0.4375 0.4375 0.4375 0.8594 0.0720

■

For each collection of  $F_{r_i,p_i}$  focal elements  $\hat{m}_r$  we define an aggregate

$$E = \cup_{i=1}^R F_{r_i,p_i} \tag{24}$$

with basic probability assignment

$$m(E) = \prod_{i=1}^R m(F_{r_i,p_i}) \tag{25}$$

The results of classification are  $D$  is  $m$ , with focal elements  $E_k(k = 1, \dots, R^{|S|})$  and distribution  $m(E_k)$ . That results are calculated using focal elements and appropriate counters  $c_{t,1}, \dots, c_{t,|S|}$ .

**Example 5.** For sample point L5 and T1 the counters are 3, 1, and 3, 0 respectively

■

Then we perform defuzzification according to COA method [1].

$$\bar{y} = \sum_{k=1}^{R^{|S|}} \bar{y}_k m(E_k) \tag{26}$$

where  $\bar{y}_k$  are defuzzified values for focal element  $E_k$  defined as

$$\bar{y}_k = \frac{\sum_{1 \leq t \leq n} x_t \mu_{k,t}(x_t)}{\sum_{1 \leq t \leq n} \mu_{k,t}(x_t)} \tag{27}$$

In the next step, the rules structure is simplified to

$$\text{If Antecedent}_r \text{ Then } (D \text{ is } H_r),$$

where  $H_r = \left\{ \frac{1}{\gamma_r} \right\}$  is a singleton fuzzy set for factor  $\gamma_r = \sum_{p=1}^{|S|} \bar{y}_p m_r(S_{r,p})$ .

**Example 6.** For both L5 and T1 we calculate decision class 1. It is correct in case of T1, but wrong for L5. The values of  $H_r$  are 0.4283 and 0.4800 respectively. ■

## 5 Empirical Learning for FDS Model

In this section we compare and analyze the performance of several membership functions and matching operators. We start from a standard solution used in the introduction to fuzzy modeling, then we consider more complicated models. We compute results for the following membership function: *Linear*, *Quadratic*, *Gaussian*, and *FCM*. We concentrate on *Minimum*, *Multiply* and *Implication* as a matching operators. The most valuable is comparing the results of all calculations. In the end of this section we show some results of experimental research.

### 5.1 Membership Functions

The membership function makes possible the division of data into  $n$  intervals. It is a way of discretization of the input data. Hence, we get the best result for continuous data or for data with several discrete (nominal) values. If we have discrete or binary data then results of the proposed model are not good enough.

The choice of membership function has a great influence on the quality of rules. Although the quantity of rules is different, the quality of classification is comparable.

The most interesting membership function was generated by Fuzzy c-Means (FCM) algorithm [1]. The results of experimental research with membership functions have been summarized in Table 2.

**Table 2.** Membership functions

	Function	Formula	Features
1	<i>Linear triangle</i>	F1: $Y = (X - x_{i+1}) / (x_i - x_{i+1})$ F2: $Y = (X - x_i) / (x_{i+1} - x_i)$	The simplest, The result are not so good;
2	<i>Quadratic</i>	$Y = 1 - (X - x_i)^2 / (x_{i+1} - x_i)^2$	Still simply calculated function, The result are better than in the case of Linear function;
3	<i>Gaussian</i>	$y = e^{-\left(\frac{x-x_i}{x_{i+1}-x_i}\right)^2}$	Exponential function, Using it leads in general to proper conclusions especially in learn sample;
4	<i>FCM</i>	FCM algorithm	The complex algorithm – values of function have to be previously generated. In general proper conclusion

## 5.2 Matching Operators

It was shown in our experiments that a matching operator applied to data sample with existing rules plays the very important role in the accuracy of diagnoses. It occurred as early as the rules were generated. A matching operator influences the quality of the generated rules. Of course, this quality has secondary means, but in general, the more rules the better accuracy.

From the analysis of the results of experiments in Table 4, we can infer that the most powerful operator is implication. This is not all the true, because Table 4 shows the results only for one fixed threshold value. It is not optimal in all instances, especially in multiply operator. The change of the threshold value (e.g. to 0.25) gives almost the same results as for implication operator. Anyway, the choice of the threshold value is of minor importance here, but it can have influence on the result of the receiving of rules. Of course, we cannot analyze the threshold value without keeping in mind the features of the membership function and the number of intervals. The choice of the threshold value will be subject of future works.

The results of the investigation of various matching operators are collected in Table 3.

**Table 3.** Matching operators

	Name	Formula	Features
1	Minimum	$\text{Max}(0, \min(x, y))$	The simplest case In many experiments the results are good
2	Multiply	$\text{Max}(0, x*y)$	It's case of <i>implication</i> , with $\alpha = -1$ . The results are better than in the case of Minimum operator. Important setting of threshold value
3	Implication	$\text{Max}(0, (\alpha+1)* (x + y-1) - \alpha*x*y)$	Function seems to be complex. The best results. Setting constant $\alpha$ - direction for future research

## 6 Experimental Studies

Some results of experimental research are shown in Table 4. We fixed here count of membership functions on 6 and threshold value on 0.75. All data sets have been divided into two parts: learning (training) data (about 2/3 of the entire data set) and testing data (remaining 1/3). The learning data has been used to generate the rule set. Testing data has been applied to test the produced rule set. To obtain reliable results, we carried out the experiment several times.

That formulation does not show the most favorable case but it is allowed to see the part of real results with using different methods for generalizing rules. These results can be comparable.

**Table 4.** Experimental results

		Linear			Quadratic			Gaussian			FCM			Decision Trees	
		Rules	Learn	Test	Rules	Learn	Test	Rules	Learn	Test	Rules	Learn	Test	Learn	Test
<b>Minimum</b>															
	Iris	15	93.80	97.50	41	<b>97.50</b>	97.50	41	93.75	<b>100.00</b>	-	-	-	94.00	91.30
	Ulcers	43	66.25	9.52	79	<b>93.75</b>	28.57	79	87.50	52.38	76	78.75	57.14	-	-
	Diabetes	0	0.00	0.00	58	<b>94.29</b>	37.84	23	87.14	<b>56.76</b>	9	47.14	24.32	-	-
	BCW	2	59.66	73.73	203	91.63	85.71	203	40.56	24.42	201	95.28	<b>91.71</b>	96.00	90.30
	Derm.	0	0.00	0.00	244	<b>100.00</b>	86.89	244	<b>100.00</b>	50.82	241	<b>100.00</b>	73.77	94.00	87.50
	EKG	1	2.27	5.56	44	<b>100.00</b>	38.89	44	<b>100.00</b>	66.67	43	<b>100.00</b>	72.22	76.00	59.00
<b>Multiply</b>															
	Iris	0	0.00	0.00	18	95.00	97.50	19	95.00	<b>100.00</b>	-	-	-	-	-
	Ulcers	0	0.00	0.00	2	35.00	2.38	2	70.00	2.38	74	65.00	54.76	-	-
	Diabetes	0	0.00	0.00	35	82.86	29.73	25	80.00	54.05	9	47.14	24.32	-	-
	BCW	2	59.66	73.73	30	81.33	79.72	33	95.06	97.24	17	60.30	82.95	-	-
	Derm.	0	0.00	0.00	230	95.08	86.89	235	99.18	51.64	13	9.84	24.59	-	-
	EKG	1	2.27	5.56	6	45.45	5.56	6	50.00	11.11	35	56.82	55.56	-	-
<b>Implication</b>															
	Iris	35	83.75	97.50	39	<b>97.50</b>	97.50	40	96.25	<b>100.00</b>	-	-	-	-	-
	Ulcers	43	61.25	9.52	75	90.00	28.57	75	85.00	52.38	75	80.00	<b>59.52</b>	-	-
	Diabetes	31	0.00	0.00	58	<b>94.29</b>	37.84	58	87.14	<b>56.76</b>	9	47.14	24.32	-	-
	BCW	15	40.34	46.08	203	91.63	85.71	103	62.02	70.51	203	<b>97.21</b>	93.55	-	-
	Derm.	0	0.00	0.00	233	99.18	86.89	243	99.59	50.82	200	95.90	<b>90.16</b>	-	-
	EKG	22	70.45	0.00	41	<b>100.00</b>	38.89	43	<b>100.00</b>	66.67	39	<b>100.00</b>	77.78	-	-

The methods of automatically generated decision rules that are described in this paper have the best results on *Iris* data set. In a few points, reach a destination 100 % of the decision accuracy. This set consists of all data as continuous values. In other cases, the construction of the data discretization caused a little worse results. In spite of the fact that in the case of *Ulcers* data set, for which over the half features was discrete, the results on the learning data was nearly the same like in the case of *Iris* data.

*Diabetes* data set is a sample of testing proposed algorithms on discrete data. The results of rules accuracy are not satisfactory. It has shown the case when the method of generating and verifying decision rules ineffectual as only one.

Another observation is that, the smaller number of binary data in antecedent, then the better accuracy of our rules. If the features have no binary data or if number of it is strongly less, then others then our rules can be applicable. We can see that during analyzing *Breast Cancer Wisconsin* and *Dermatology* data sets. In *Echocardiogram* data set, when number of binary data is equal to three the results are worse.

In all the instances, intermediate reports were stored in disk data files. It was used to compute results with the FCM algorithm. It also made possible to connect described method with others.

In our research, the proper choice of threshold value gave us information if rule is valuable or not. The importance of this value has been shown in the case of *Multiply* operator. The result presented in Table 4 could suggest less “weight” of this operator, but it is not true. If we change threshold value, we notice that it has almost the same occurrence as more complied in calculations *Implication* operator. *Implication* is a sample of a very interesting and powerful operator of fuzzy relation.

We compare the results of our research with standard decision trees algorithm [10]. For all data sets, we get better results using *Gaussian* function or *FCM* algorithm [1], and in a few points of *Quadratic* function, we obtained also better accuracy.

In *Dermatology* and *Echocardiogram* data sets, we removed records with missing input features, because we concentrate only on complete data.

## 7 Conclusions

The study has focused on the use of Fuzzy Dempster-Shafer model for generating of fuzzy decision rules. Fuzzy sets are useful in discretization of continuous attributes. The approach is discussed in the concrete applications of two real medical data sets (especially to problems of identification of diseases) and several well-known data sets available on the Web. The results are used to classify objects. It has been found through series of experiments that this approach outperforms the results of the C4.5 algorithm in terms of higher classification accuracy.

**Acknowledgement.** This work was supported by the grant S/WI/1/2013 from Bialystok University of Technology founded by Ministry of Science and Higher Education.

## References

1. Bezdek, J.C., Sabin, M.J., Tucker, W.T.: Convergence theory for fuzzy c-means: counterexamples and repairs. *IEEE Trans. Syst. Man Cybern.* **smc-17**(5), 873–877 (1987)
2. Binaghi, E., Gallo, I., Madella, P.: A neural model for fuzzy Dempster-Shafer classifiers. *Int. J. Approx. Reason.* **25**, 89–121 (2000)
3. Chiang, I.J., Hsu, J.Y.J.: Integration of fuzzy classifiers with decision trees. In: *Proceedings of 1996 Asian Fuzzy Systems Symposium*, pp. 266–271 (1996)
4. Hayashi, Y., Imura, A.: Fuzzy neural expert system with automated extraction of fuzzy if-then rules from a trained neural network. In: *1st International Symposium on Uncertainty Modeling and Analysis*, pp. 489–494, December 1990
5. Hayashi, A., Maeda, T., Bastian, A., Jain, L.C.: Generation of fuzzy decision trees by fuzzy ID3 with adjusting mechanism of AND/OR operators. In: *Proceedings of 1988 International Conference on Fuzzy Systems (FUZZ-IEEE 1988)*, pp. 681–685 (1998)
6. Ichihashi, H., Shirai, T., Nagasaka, K., Miyoshi, T.: Nero-fuzzy ID3: a method of inducing fuzzy decision trees with linear programming for maximizing entropy and an algebraic method for incremental learning. *Fuzzy Sets Syst.* **81**(1), 157–167 (1996)
7. Janikow, C.Z.: Fuzzy decision trees: issues and methods. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **28**(1), 1–14 (1998)
8. Lin, C.T., Lee, C.S.: Neural-network-based fuzzy logic control and decision system. *IEEE Trans. Comput.* **12**, 1320–1336 (1991)
9. Pedrycz, W., Sosnowski, Z.A.: C-fuzzy decision trees. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **35**(4), 498–511 (2005)
10. Quinlann, J.R.: Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986)
11. Shafer, G.: *Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton (1976)
12. Sosnowski, Z.A., Walijewski, J.S.: Generating fuzzy decision rules with the use of Dempster-Shafer theory. In: *ESM 1999*, pp. 419–426, Warsaw (1999)
13. Umano, M., Okamoto, H., Hatono, I., Tamura, H., Kawachi, F., Umedzu, S., Kinoshita, J.: Generation of fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis by

- gas in oil. In: Proceedings of 1994 Japan-USA Symposium on Flexible Automation, pp. 1445–1448 (1994)
14. Viharos, Z.J., Kis, K.B.: Survey on neuro fuzzy systems and their applications in technical diagnostics. *Measurement* **67**, 126–136 (2015)
  15. Weber, R.: Fuzzy ID3: a class of methods for automatic knowledge acquisition. In: Proceedings of 2nd International Conference on Fuzzy Logic and Neural Networks, Iizuka, Japan, pp. 265–268 (1992)
  16. Yager, R., Filev, D.: *Essentials of Fuzzy Modeling Control*. Wiley, New York (1994)
  17. Yager, R., Filev, D.: Including probabilistic uncertainty in fuzzy logic controller modeling using Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**(8), 1221–1230 (1994)
  18. Yun, J., won Seo, J., Yoon, T.: The new approach on fuzzy decision trees. *Int. J. Fuzzy Logic Syst.* **4**(3) (2014)