Hodnocení výsledků klasifikace Evaluation of Classification Results

Hana Řezanková, Dušan Húsek

Abstrakt: Článek se zaměřuje na míry, které mohou být aplikovány pro hodnocení výsledků klasifikace objektů. K hodnocení jsou využity objekty, pro které je známo jejich zařazení do určitých skupin. Pomocí klasifikační metody uživatel získá odhady přiřazení objektů do skupin. Pro hodnocení úspěšnosti klasifikace již bylo navrženo mnoho koeficientů. Většina z nich je určena pro případ klasifikace do dvou skupin. Možnosti pro hodnocení klasifikace do více než dvou skupin jsou omezené. Cílem tohoto článku je podat přehled různých měr, diskutovat jejich původ a vzájemné vztahy. Pro klasifikace do tří nebo více skupin navrhujeme dvě nové míry, které jsou pro hodnocení vhodnější. První z nich bere v úvahu variabilitu četností na diagonále konfuzní matice. Druhá je založena na součtu čtvercových vzdáleností mezi maximálně možným správným zařazením objektů a reálně správně zařazených objektů v každé skupině.

Abstract: The paper focuses on measures which can be applied to evaluation of classification results. Objects with known assignment to certain groups are used for evaluation. Using a classification method the user obtains the assignment of objects to groups. Many coefficients have been proposed for evaluation of the success rate of classification. Most of them is determined for classification to two groups. Possibilities for classification to more groups are limited. The aim of this paper is to summarize different measures, discuss their origin and relationships. For classification to three or more groups we propose two novel measures which are more suitable. The first of them takes a variability of diagonal frequencies of the confusion matrix into account. The second one is based on the sum of squared differences between the maximum correctly assigned objects and real correctly assigned objects in each group.

Klíčová slova: míry podobnosti, míry souhlasu, úspěšnost klasifikace

Keywords: similarity measures, measures of agreement, success rate of classification

JEL classification: C38

1. Introduction

If a new method for object classification is developed or a model of object classification is estimated, evaluation of classification results is very useful and necessary. Many coefficients have been proposed for evaluation of the success rate of classification. Most of them is determined for classification to two groups. Possibilities for classification to more groups are limited. The aim of this paper is to summarize different measures, discuss their origin and relationships. For classification to three and more groups two novel measures, which better evaluate obtained results, are proposed. The first of them takes a variability of diagonal frequencies of the confusion matrix into account. The second one is based on the sum of squared differences between the maximum correctly assigned objects and real correctly assigned objects for each group. Measures evaluating classification are based on the confusion matrix including the frequencies n_{ij} . Each value n_{ij} expresses the number of objects observed in the *i*-th group and classified in the *j*-th group.

2. Classification to two groups

In case of classification to two groups, more approaches are available to the users in comparison with classification to more than two groups. The confusion matrix includes four frequencies n_{ij} , see Table 1, where *n* is the total number of objects, $n_{1+} = n_{11} + n_{12}$, $n_{2+} = n_{21} + n_{22}$, $n_{+1} = n_{11} + n_{21}$, and $n_{+2} = n_{12} + n_{22}$ are the marginal frequencies.

		Class		
		Group 1	Group 2	
Observed	Group 1	<i>n</i> ₁₁	<i>n</i> ₁₂	n_{1+}
Observed	Group 2	<i>n</i> ₂₁	<i>n</i> ₂₂	n_{2+}
		n_{+1}	n_{+2}	n

Table 1 Scheme of confusion matrix for two groups

Let us suppose that group 1 expresses a positive situation (a patient is cured, a client repaid the installments, a document contains relevant information) and group 2 expresses a negative situation. In the process of classification evaluation, the number of objects from group 1 which are classified correctly is usually denoted TP (true positive), i.e. $n_{11} = TP$. The number of objects from group 1 which are not classified correctly is usually denoted FN (false negative), i.e. $n_{12} = FN$. Similarly, the number of objects from group 2 which are classified correctly is usually denoted TN (true negative), i.e. $n_{22} = TN$, and the number of objects from group 2 which are not classified correctly is usually denoted FP (false positive), i.e. $n_{21} = FP$.

The basic characteristics of classification to two groups are *sensitivity* (*true positive rate* or *recall*) and *specificity* (*true negative rate*). The former is defined as n_{11}/n_{1+} , the latter is expressed as n_{22}/n_{2+} . Further, the *precision* (*positive predictive value*) is defined as n_{11}/n_{1+} and the *false positive rate* is expressed as (1 - specificity), i.e. n_{21}/n_{2+} . The list of characteristics based on the frequencies in the confusion matrix and usually used for classification evaluation is shown in Table 2. Some terms comes from the area of information retrieval.

Name 1	Name 2 etc.	Equation	
true positive rate (TPR)	sensitivity, recall	n_{11}/n_{1+}	(1)
true negative rate (TNR)	specificity (SPC)	n_{22}/n_{2+}	(2)
positive predictive value (PPV)	precision	n_{11}/n_{+1}	(3)
negative predictive value (NPV)		n_{22}/n_{+2}	(4)
false positive rate (FPR)	fall-out	n_{21}/n_{2+}	(5)
false negative rate (FNR)	miss rate	n_{12}/n_{1+}	(6)
false discovery rate (FDR)		n_{21}/n_{+1}	(7)
false omission rate (FOR)		n_{12}/n_{+2}	(8)
accuracy (ACC)		$(n_{11} + n_{22})/n$	(9)
F1 score		$2n_{11}/(n_{1+}+n_{+1})$	(10)
Matthews correlation coef. (MCC)		$\frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}$	(11)

Table 2 Definitions of classification characteristics

The confusion matrix is a contingency table in which different kinds of frequencies can be displayed, e.g. relative frequencies within the total table $(p_{ij} = n_{ij}/n)$ with the marginal relative frequencies , $p_{1+} = p_{11} + p_{12}$, $p_{2+} = p_{21} + p_{22}$, $p_{+1} = p_{11} + p_{21}$, and $p_{+2} = p_{12} + p_{22}$ (see Table 3), the row relative frequencies (see Table 4) or column relative frequencies (see Table 5) which provide different views on the relationships between observed and suggested assignment of objects to groups.

		Clas	Classified		
		Group 1	Group 2		
Observed	Group 1	p_{11}	p_{12}	p_{1+}	
	Group 2	p_{21}	p_{22}	p_{2+}	
		p_{+1}	p_{+2}	1	

Table 3 Scheme of confusion matrix for two groups with total relative frequencies

Table 4	Scheme o	of confusion	matrix for two	groups with row	, relative frequencies
I uvic 7	Scheme 0	η τυπμισιοπ		groups wan row	remaine frequencies

		Classified		
		Group 1	Group 2	
Observed	Group 1	$TPR = n_{11}/n_{1+}$	$FNR = n_{12}/n_{1+}$	1
Observed	Group 2	$FPR = n_{21}/n_{2+}$	$TNR = n_{22}/n_{2+}$	1

Table 5 Scheme of confusion matrix for two groups with column relative frequencies

		Classified		
		Group 1	Group 2	
Observed	Group 1	$PPV = n_{11}/n_{+1}$	$FOR = n_{12}/n_{+2}$	
Observed	Group 2	$FDR = n_{21}/n_{+1}$	$NPV = n_{22}/n_{+2}$	
		1	1	

The rates, predictive values, accuracy and F1 score have values from the interval [0; 1]. The true rates (*TPR* and *TNR*), predictive values (*PPV* and *NPV*), accuracy and *F1* score should be close 1, false rates should be close 0.

The *accuracy* is known as the *simple matching coefficient (SMC)* in the area of similarity measures for binary variables, applied e.g. in hierarchical cluster analysis. It is the weighted arithmetic average of sensitivity and specificity:

$$SMC = \frac{\frac{n_{11}}{n_{1+}} n_{1+} + \frac{n_{22}}{n_{2+}}}{n_{1+} + n_{2+}} = \frac{n_{11} + n_{22}}{n}.$$
(12)

The *F1 score* is a harmonic mean of sensitivity and precision. It is well known similarity measure for two asymmetric dichotomous variables, called *Dice* or *Czekanowski* or *Sorenson*. Usually, it is expressed in the form

$$DICE = \frac{2n_{11}}{2n_{11} + n_{12} + n_{21}}.$$
(13)

The weighted harmonic mean of sensitivity and precision is the *F*-measure:

$$F_{\alpha} = \frac{(1+\alpha) \cdot \frac{n_{11}}{n_{+1}} \cdot \frac{n_{11}}{n_{1+}}}{\alpha \cdot \frac{n_{11}}{n_{+1}} + \frac{n_{11}}{n_{1+}}} = \frac{(1+\alpha) \cdot n_{11}}{(1+\alpha) \cdot n_{11} + \alpha n_{12} + n_{21}},$$
(14)

where α may be varied between 0 and 1. It is the significance level, see (Billinger et al., 2012). The arithmetic average of sensitivity and precision gives the *Kulczynski similarity* measure 2, i.e.

$$K2 = \frac{\frac{n_{11}}{n_{1+}} + \frac{n_{11}}{n_{+1}}}{2}.$$
(15)

The geometric mean of sensitivity and precision gives the Ochiai similarity measure in the form

$$OCHIAI = \sqrt{\frac{n_{11}}{n_{1+}}} \frac{n_{11}}{n_{+1}}.$$
 (16)

It is the special equation for the cosine similarity measure determined for quantitative variables.

The average of *TPR*, *TNR*, *PPV* and *NPV* is called *Sokal and Sneath similarity measure* 4, i.e.

$$SS4 = \frac{\frac{n_{11}}{n_{1+}} + \frac{n_{11}}{n_{+1}} + \frac{n_{22}}{n_{2+}} + \frac{n_{22}}{n_{+2}}}{4}.$$
(17)

The average of the values from the interval [0; 1] is the value from the same interval.

The false positive rate and the true positive rate are a basis for a construction of the ROC (Receiver Operator Characteristic) curve. The values of the former creates the x axis and the values of the letter creates the y axis. For evaluation of classification, the AUC (Area Under the receiver operating characteristic Curve) characteristic is used. It is the total measure of ability of the model to predict assignment of objects to groups, see (Hosmer and Lemeshow, 2000).

Some suitable measures of association known from the contingency table analysis can be used for evaluation of classification. Further, some similarity measures for binary variables used in hierarchical cluster analysis can be also applied.

The Matthews correlation coefficient (Mathews, 1975), see Table 2, Eq. (11), is the classical *Pearson correlation coefficient* expressed for two binary variables by frequencies from the contingency tables. In a case of two binary variables, the equation for the Pearson correlation coefficient is the same as the formulas for the Spearman and Kendall rank correlation coefficients (Kendall's tau-b). That means there is only one correlation coefficient for measurement of association of two dichotomous variable which can be used for classification evaluation. This coefficient has values from the interval [-1; 1]. Three examples of frequencies which give values 1, 0 and -1 are shown in Tables 6–8. The absolute values of the correlation coefficient are the same as the values of the *phi coefficient* of association based on the Pearson chi-squared statistic.

Table 6 Example of frequencies which give the value 1 of the correlation coefficient

		Class		
		Group 1	Group 2	
01 1	Group 1	90	0	90
Observed	Group 2	0	90	90
		90	90	180

		Class		
		Group 1	Group 2	
Observed	Group 1	45	45	90
Observed	Group 2	45	45	90
		90	90	180

 Table 7 Example of frequencies which give the value 0 of the correlation coefficient

Table 8 Example of frequencies which give the value -1 of the correlation coefficient

		Class		
		Group 1	Group 2	
	Group 1	0	90	90
Observed	Group 2	90	0	90
		90	90	180

The correlation coefficient is a measure of linear dependence. There are two other coefficients proposed for linear dependence measurement of two ordinal variables which can be applied to binary variables – symmetric Somers's d and Goodman and Kruskal's gamma. *Somers's d* is a harmonic mean of two asymmetric coefficients. For the 2×2 contingency table it is defined as

$$d = \frac{2(n_{11}n_{22} - n_{12}n_{21})}{n_{1+}n_{2+} + n_{+1}n_{+2}}.$$
(18)

Its values are very close to values of the correlation coefficient, because generally Kendall's tau-b can be expressed as a geometric mean of two asymmetric Somers's coefficients. For classification we can consider a directional measure, it means if classification (C) with frequencies placed in columns of the confusion matrix is dependent on observed assignment of objects to groups placed in rows (R) or not. In this case

$$d_{C|R} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21} + n_{11}n_{21} + n_{12}n_{22}}.$$
(19)

If the confusion matrix is symmetric, then the values of all three Somers's coefficients are the same.

Goodman and Kruskal's gamma is called as *Yule's Q* for the 2×2 contingency table. It is defined by the equation

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}.$$
(20)

This coefficient has also values from the interval [-1; 1]. Values 1, 0 and -1 correspond with situations in which correlation coefficient has the same values, see Tables 6–8. Yule's Q is used as a similarity measure in multivariate statistical methods.

Another coefficient with similar properties is *Yule's coefficient of colligation*. It can be expressed by the formula

$$Y = \frac{\sqrt{n_{11}n_{22}} - \sqrt{n_{12}n_{21}}}{\sqrt{n_{11}n_{22}} + \sqrt{n_{12}n_{21}}}.$$
(21)

Yule's coefficients assess the association between items as the predictability of one given the other.

The example in Table 6 is an ideal situation and the best classification. Beside of the measures mentioned above there are some different measure with this property, e.g. measures of agreement. One of them is *Kohen's kappa* which can be used for any square contingency table (two variables have the same number of categories which correspond with their senses). For the 2×2 table it is calculated according to the formula

$$\kappa = \frac{(n_{11} + n_{22}) - \left(\frac{n_{1+}n_{+1}}{n} + \frac{n_{2+}n_{+2}}{n}\right)}{n - \left(\frac{n_{1+}n_{+1}}{n} + \frac{n_{2+}n_{+2}}{n}\right)}.$$
(22)

This coefficient has values from the interval [-1; 1]. The value 0 means that the frequencies in the table are the same as the frequencies expected under the hypothesis of independence. The value -1 can be achieve only for the symmetric confusion matrix.

If all non-zero values are in the diagonal (see Table 6) then $n_{11} + n_{22} = n$ and $\kappa = 1$. If frequencies in the contingency table correspond to independence (see Table 7), then $\kappa = 0$. If the value in the diagonal are zero (see Table 8), then $n_{12} = n_{1+} = n_{+2}$ and $n_{21} = n_{2+} = n_{+1}$, i.e.

$$\kappa = \frac{-\left(\frac{n_{1+}n_{+1}}{n} + \frac{n_{2+}n_{+2}}{n}\right)}{n - \left(\frac{n_{1+}n_{+1}}{n} + \frac{n_{2+}n_{+2}}{n}\right)} = \frac{-\left(\frac{n_{1+}n_{+1}}{n} + \frac{n_{2+}n_{+2}}{n}\right)}{\left(\frac{n_{1+}n_{+2}}{n} + \frac{n_{2+}n_{+1}}{n}\right)} = -\frac{2n_{12}n_{21}}{n_{12}n_{12} + n_{21}n_{21}}.$$
(23)

If $n_{12} = n_{21}$ then $\kappa = -1$.

The special measure of agreement proposed for the 2×2 table is *Hamann's coefficient* (*HC*) defined by the equation

$$HC = \frac{(n_{11} + n_{22}) - (n_{12} + n_{21})}{n}.$$
 (24)

This coefficient has values from the interval [-1; 1]. It is applied as a similarity measure in multivariate methods. The examples of frequencies which give the values 1, 0 and -1 are in Tables 6–8.

Beside of the simple matching coefficient with a range from 0 to 1, some other similarity measures with this range can be applied. We can mention *Sokal and Sneath similarity measure 5* calculated as

$$SS5 = \frac{n_{11}n_{22}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}.$$
(25)

Some similarity measures favor the correct classification. The example are the *Sokal* and *Sneath similarity measure 1*

$$SSI = \frac{2(n_{11} + n_{22})}{2(n_{11} + n_{22}) + n_{12} + n_{21}}.$$
(26)

and the Rogers and Tanimoto similarity measure

$$RT = \frac{n_{11} + n_{22}}{n_{11} + n_{22} + 2(n_{12} + n_{21})}.$$
(27)

These coefficients have also values from the interval [0; 1].

There are also similarity measures which have not values either from the interval [0; 1] nor [-1; 1]. These measures have a minimum value of 0 and have no upper limit. We will not evaluate them in this study.

Moreover, there are several coefficients which represent only the proportion of correct classification to the first group as the F1 score (or the Dice similarity measure) or the K2 or

Ochiai similarity measures express. They have values from the interval [0; 1]. There are Russel and Rao (*RR*), Jaccard (*JACC*), and Sokal and Sneath (*SS2*) similarity measures. The *Russel and Rao similarity measure* is defined as

$$RR = \frac{n_{11}}{n},\tag{28}$$

the Jaccard coefficient as

$$JACC = \frac{n_{11}}{n_{11} + n_{12} + n_{21}} = \frac{n_{11}}{n - n_{22}},$$
(29)

and the Sokal and Sneath similarity measure 2 as

$$SS2 = \frac{n_{11}}{n_{11} + 2(n_{12} + n_{21})}.$$
(30)

In Table 9, there are examples of frequencies and corresponding values of coefficients. Different frequencies for all marginal frequencies equal 90 are considered (as in Tables 6–8). It means that the confusion matrix is symmetric. For the reason that given marginal frequencies and the frequencies in one cell determine frequencies in three other cells, only the value of n_{11} is included in the table. The coefficients are ordered according their values (from higher to lower values in the group of a certain coefficient type). We can see that for the symmetric confusion matrix some coefficients give the same results. Graphical representation of coefficient values is shown in Figures 1 and 2.



Figure 1 Dependence of coefficient values on selected frequencies in the first cell of the confusion matrix with marginal frequencies 90, 90, 90 90 (coefficients with range from 0)



Figure 2 Dependence of coefficient values on selected frequencies in the first cell of the confusion matrix with marginal frequencies 90, 90, 90 90 (coefficients with range [-1; 1])

		Frequencies – examples							
	1	2	3	4	5	6	7	8	9
n_{11}	0	10	20	30	45	60	70	80	90
				Values	s of coe	fficients			
Measures for	or symme	etric bina	ary variał	oles with r	ange [0	; 1]			
SS1	0	0.20	0.36	0.50	0.67	0.80	0.88	0.94	1
SS4	0	0.11	0.22	0.33	0.50	0.67	0.78	0.89	1
ACC (SMC)	0	0.11	0.22	0.33	0.50	0.67	0.78	0.89	1
RT	0	0.06	0.13	0.20	0.33	0.50	0.64	0.80	1
SS5	0	0.01	0.05	0.11	0.25	0.44	0.60	0.79	1
Measures for	or asymn	netric bin	nary varia	ıbles					
F1 score (Dice)	0	0.11	0.22	0.33	0.50	0.67	0.78	0.89	1
<i>F-0.7</i>	0	0.11	0.22	0.33	0.50	0.67	0.78	0.89	1
K2	0	0.11	0.22	0.33	0.50	0.67	0.78	0.89	1
Ochiai	0	0.11	0.22	0.33	0.50	0.67	0.78	0.89	1
JACC	0	0.06	0.13	0.20	0.33	0.50	0.64	0.80	1
SS2	0	0.03	0.07	0.11	0.20	0.33	0.47	0.67	1
RR	0	0.06	0.11	0.17	0.25	0.33	0.39	0.44	0.5
Measure with range [-1; 1]									
Q	-1	-0.97	-0.85	-0.60	0	0.60	0.85	0.97	1
Y	-1	-0.78	-0.56	-0.33	0	0.33	0.56	0.78	1
MCC(r)	-1	-0.78	-0.56	-0.33	0	0.33	0.56	0.78	1
Som. d	-1	-0.78	-0.56	-0.33	0	0.33	0.56	0.78	1
НС	-1	-0.78	-0.56	-0.33	0	0.33	0.56	0.78	1
kappa	-1	-0.78	-0.56	-0.33	0	0.33	0.56	0.78	1

Table 9 Examples of frequencies and corresponding values of coefficients

Číslo 1/2015, Roč. 3.



Figure 3 Dependence of coefficient values on selected frequencies in the first cell of the confusion matrix with marginal frequencies 90, 90, 60, 120 (coefficients with range [0; 1])



Figure 4 Dependence of coefficient values on selected frequencies in the first cell of the confusion matrix with marginal frequencies 90, 90, 60 120 (coefficients with range [-1; 1])

3. Classification to three and more groups

If we consider more than two groups, then the number of possibilities for evaluation of classification is considerably less. We can evaluate classification to individual groups separately (assignment to a certain groups and to all other groups) and use the approaches mentioned above. However, for evaluation of assignment to all groups simultaneously only the simple matching coefficient and the kappa coefficient are usually applied. The *simple matching coefficient* is defined as

$$SMC = \frac{\sum_{i=1}^{K} n_{ii}}{n},$$
(31)

where K is the number of groups. This coefficient has values from the interval [0; 1]. The *SMC* measure is a weighted arithmetic average of individual sensitivities for each group:

$$SMC = \sum_{i=1}^{K} \frac{n_{ii}}{n_{i+}} \cdot \frac{n_{i+}}{n} = \sum_{i=1}^{K} \frac{n_{ii}}{n} = \frac{\sum_{i=1}^{K} n_{ii}}{n}.$$
(32)

This coefficient does not distinguish variability of diagonal frequencies.

For K groups the kappa coefficient is expressed as

$$\kappa = \frac{\sum_{i=1}^{K} n_{ii} - \sum_{i=1}^{K} \frac{n_{i+} n_{+i}}{n}}{n - \sum_{i=1}^{K} \frac{n_{i+} n_{+i}}{n}}.$$
(33)

It has values from the interval [-1; 1]. If the frequencies in the confusion matrix are the same as the frequencies expected under the hypothesis of independence, then the value of kappa is 0. However, if the value is 0, then random frequencies are only one from several possibilities. For this reason, the kappa coefficient is the suitable measure of agreement, but it is not a suitable measure for evaluation of classification results.

The modified *Hamann's coefficient* for K groups can be applied for evaluation of classification results in the form

$$HC = \frac{\sum_{i=1}^{K} n_{ii} - \sum_{i=1}^{K} \sum_{j=1; j \neq i}^{K} n_{ij}}{n}.$$
 (34)

This coefficient has values from the interval [-1; 1]. The value 0 means that the number of correct assigned objects and the number of incorrect assigned objects are the same. However, we obtain similar information as using the SMC measure, only in the different interval. The comparison of three mentioned coefficients applied for the symmetric confusion matrix for 90 objects is shown in Figure 5.



Figure 5 Dependence of coefficient values on selected diagonal frequencies

For the reason that the coefficients mentioned above do not express a success rate of classification in a suitable way, we propose a new coefficient including variability of frequencies. Sensitivities are relative frequencies. However, the sum of them usually is not 1. If these relative frequencies are recalculated for the sum equal 1, then we can express variability of corrected classification.

We suggest to apply nominal variance (i.e. mutability) proposed by Gini (1912) as a variability measure. So each sensitivity is divided by the sum of sensitivities and then the normalized nominal variance of these proportions are calculated. The *SMC* coefficient is multiplied by the obtained value, i.e.

$$RH = \frac{1}{n} \sum_{i=1}^{K} n_{ii} \cdot \left(\frac{K}{K-1} \sum_{i=1}^{K} \frac{\underline{n_{ii}}}{\sum_{i=1}^{K} \underline{n_{i+}}} \left(1 - \frac{\underline{n_{ii}}}{\sum_{i=1}^{K} \underline{n_{i+}}} \right) \right).$$
(35)

If all proportions are the same, then the normalized nominal variance is 1 and RH = SMC. If only one class is classified correctly, then variance is 0 and RH = 0.

If there are the same numbers of objects in all groups, then variability of diagonal frequencies instead variability of sensitivities can be calculated. Then the RH coefficient can be expressed as

$$RH = \frac{1}{n} \sum_{i=1}^{K} n_{ii} \cdot \left(\frac{K}{K-1} \sum_{i=1}^{K} \frac{n_{ii}}{\sum_{i=1}^{K} n_{ii}} \left(1 - \frac{n_{ii}}{\sum_{i=1}^{K} n_{ii}} \right) \right).$$
(36)

We can illustrate taking variability into account with the following example. Let us suppose that there are 3 groups and 9 objects. In each group there are 3 objects correctly. The values of the *SMC* and *RH* coefficients are in Table 10.

Another way to take into account different sensitivities in individual groups is to compute the sum of squared differences between the maximum correctly assigned objects and real correctly assigned objects for each group:

$$Dif2 = \sum_{i=1}^{K} (n_{i+} - n_{ii})^2 .$$
(37)

For obtaining values from the interval from 0 to 1, the *Dif2Norm* measure can be expressed in the form

$$Dif2Norm = \frac{\sum_{i=1}^{K} n_{i+}^2 - Dif2}{\sum_{i=1}^{K} n_{i+}^2}.$$
(38)

The values of *Dif2* and *Dif2Norm* for the example mentioned above are in Table 10. In this table the rows are ordered according to values of the *RH* and *Dif2Norm* measures. We can see that both measures give the same order for the given marginal frequencies. Both measures give for two cases of frequencies the same values but the pairs of frequencies are different.

Diagonal frequencies	SMC	Normalized mutability	RH	Dif2	Dif2Norm
0,0,0,	0	0	0	27	0
1,0,0	0.111	0	0	23	0.148
2,0,0	0.222	0	0	19	0.296
3,0,0,	0.333	0	0	18	0.333
1,1,0	0.222	0.750	0.167	17	0.370
2,1,0	0.333	0.667	0.222	14	0.481
3,1,0	0.444	0.563	0.250	13	0.519
1,1,1	0.333	1.000	0.333	12	0.556
2,2,0	0.444	0.750	0.333	11	0.593
3,2,0	0.556	0.720	0.400	10	0.630
2,1,1	0.444	0.938	0.417	9	0.667
3,3,0	0.667	0.750	0.500	9	0.667
2,2,1	0.556	0.960	0.533	6	0.778
3,2,1	0.667	0.917	0.611	5	0.815
2,2,2	0.667	1.000	0.667	3	0.889
3,2,2	0.778	0.980	0.762	2	0.926
3,3,2	0.889	0.984	0.875	1	0.963
3,3,3	1	1.000	1	0	1

Table 10 Examples of diagonal frequencies and values of SMC, RH and Dif2Norm measures

According to our opinion the *RH* and *Dif2Norm* measures are more suitable for evaluation of classification results than usually used measures because the successful classification to all groups is preferred over the total number of correctly classified objects.

4. Conclusion

Many different measures have been proposed for evaluation of classification results. For classification to two groups, the possibilities are varied. If the numbers of objects observed in two groups are the same and the numbers of objects predicted to two groups are also the same, then the obtained values of some coefficients are the same. In this paper we point out which coefficients have the same dependence on the frequencies in the confusion matrix.

For evaluation of assignment to three or more groups simultaneously, only the simple matching coefficient and the kappa coefficient are usually applied. If the value of the kappa coefficient is zero, it can mean that the frequencies in the confusion matrix are the same as the frequencies expected under the hypothesis of independence, however it is only one from many possibilities. The simple matching coefficient, which is a weighted arithmetic average of individual sensitivities for each group, does not distinguish variability of diagonal frequencies.

For this reason we proposed two novel coefficients, the *HR* and *Dif2Norm* coefficients. The former takes a variability of diagonal frequencies into account. The normalized nominal variance is used as a variability measure in this case. The *HR* coefficient is a product of the simple matching coefficient and the normalized nominal variance. The *Dif2Norm* coefficient is based on the sum of squared differences between the maximum correctly assigned objects

and real correctly assigned objects for each group. The obtained value is normalized to the interval [0; 1]. We believe that the proposed coefficients evaluate the results of classification suitably.

5. References

BILLINGER, M. – DALY, I. – KAISER, V. – JIN, J. – ALLISON, B. Z. – MÜLLER-PUTZ, G.R. – BRUNNER, C. 2012. Is it significant? Guidelines for reporting BCI performance. In: ALLISON, B. Z. et al. (Eds.): Towards Practical Brain Computer Interfaces, Biological and Medical Physics, Biomedical Engineering. Berlin Heidelberg: Springer-Verlag, 2012.

CHEETHAM, A.H. – HAZEL, J.E. 1969. Binary (presence-absence) similarity coefficients. In: Journal of Paleontology, vol. 43, no. 5, 1969, p. 1130–1136.

GINI, C.W. 1971. Variability and mutability. Contribution to the study of statistical distributions and relations. Studi Economico-Giuridici della R. Universita de Cagliari (1912). Reviewed in: LIGHT, R.J., MARGOLIN, B.H. An Analysis of Variance for Categorical Data. J. American Statistical Association, vol. 66, 1971, p. 534–544.

HOSMER, D.W. – LEMESHOW, S. 2000. Applied Logistic Regression. 2nd ed. New York: John Wiley and Sons, 2000.

MATTHEWS, B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. In: Biochimica et Biophysica Acta (BBA) – Protein Structure, vol. 405, no. 2, 1975, p. 442–451.

SOKAL R.R. – MICHENER C.D. 1958. A statistical method for evaluating systematic relationships. In: The University of Kansas Scientific Bulletin, vol. 38, 1958, s. 1409–1438.

SNEATH, P.H. – SOKAL, R.R. 1973. Numerical Taxonomy. San Francisco: W. H. Freeman and Company, 1973.

SOKAL, R.R. – SNEATH, P.H. 1963. Principles of Numerical Taxonomy. San Francisco: W. H. Freeman and Company, 1963.

Acknowledgment

This research has been partly funded by long-term strategic development financing of the Institute of Computer Science (RVO:67985807).

Author address:

HANA ŘEZANKOVÁ, prof. Ing., CSc. Vysoká škola ekonomická v Praze nám. W. Churchilla 4, 130 67 Praha 3 hana.rezankova@vse.cz

DUŠAN HÚSEK, Ing., CSc. Ústav informatiky AV ČR, v.v.i. Pod Vodárenskou věží 2, 182 07 Praha 8 dusan@cs.cas.cz Ústav informatiky AV ČR, v.v.i. Pod Vodárenskou věží 2, 182 07 Praha 8