FORUM STATISTICUM SLOVACUM Vol XV (2019) No 2 pp 31-43

Návrh modifikovaného přístupu k hodnocení klasifikace
A proposal of a modified approach to classification evaluation

Hana Řezanková
Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky, nám. W. Churchilla 4, 130 67 Praha, Česká republika
University of Economics in Prague, Faculty of Informatics and Statistics, W. Churchill Sq. 4, 130 67 Prague, Czech Republic
hana.rezankova@vse.cz

Abstrakt: *V článku jsou diskutovány vybrané koeficienty, které mohou být použity pro hodnocení výsledků získaných v klasifikačních úlohách. Je uvažována klasifikace objektů do tří nebo více skupin. K tomuto účelu mohou být aplikovány následující koeficienty: celková míra úspěšnosti, skupinová míra úspěšnosti, koeficienty souhlasu kappa, kappa-n a lambda-gama. K těmto koeficientům je navíc navržena míra SRGM (Success Rate for Groups / Mutability), která rozšiřuje skupinovou míru úspěšnosti o penalizaci v případech, kdy se liší variabilita souhlasů (pozorovaného a predikovaného přiřazení objektů do skupin) a variabilita vysvětlované proměnné. Jako míra variability je aplikován nominální rozptyl (mutabilita). V porovnání s pěti hodnotícími měrami diskutovanými v tomto článku míra SRGM hodnotí výsledky predikce nejkomplexněji.*

Abstract: *In the paper, selected coefficients for evaluation of results obtained in classification tasks are discussed. Classification of objects into three or more known groups is considered. For this purpose, the following coefficients can be applied: the total success rate, the group success rate, the kappa coefficient of agreement, and kappa-n and lambda-gamma coefficients. In addition to these coefficients, we propose the SRGM (Success Rate for Groups / Mutability) measure which extends the group success rate by its penalties in the cases when the variability of agreements (observed and predicted assignments of objects into groups) and the variability of the target variable differ. The nominal variance (mutability) is applied as a variability measure. In comparison with the other five evaluation measures discussed in this paper, the SRGM measure evaluates the prediction results most comprehensively.*

Klíčová slova: *klasifikační úloha, míra úspěšnosti, míra souhlasu.*
Key words: *classification task, success rate, measure of agreement.*

1 Introduction

One group of multivariate statistical methods for classification is created by methods for data analysis with the known assignment of objects into classes (groups). The assignment is expressed by a target nominal variable with codes of groups. The aim of the classification task is to estimate a model or rules for the assignment of new objects into the defined groups and to estimate the values of the target variable for new objects.

Different methods can be used in the classification task described above, e.g. discrimination analysis, logistic regression (binary or multinomial), and classification trees. Each approach is more suitable for a certain type of explanatory variables (quantitative or qualitative), but generally, the quantitative variable can be categorized and the qualitative variable can be transformed into a set of binary variables. Thus, several methods can be used for the analysis of the same data.

The analyst should compare results obtained by several methods and choose the most suitable model or rules for the prediction of unknown values of the target variable. Different approaches to such a comparison have been proposed and some of them are implemented in software systems. The simplest way is to use the percentage of successfully assigned objects (the success rate for prediction) according to a certain method.

If only two groups exist, there are many measures for evaluation and comparison of classification methods (see, e.g. Sokolova and Lapalme, 2009; Řezanková and Húsek, 2015). In addition, the area under the receiver operating characteristic (ROC) curve (AUC) is also often used as a criterion for choosing the best method in the case of two groups (see, e.g. Fawcett, 2006).

For the case of three or more groups, there are two basic approaches to the evaluation. The first one (two-step) consists in the application of the measures for two groups to the individual groups considering one selected group and all other groups joined into the second group. A final measure is based on these individual measures (Sokolova and Lapalme, 2009). The success rates can be computed in this way. The second approach is the computation of the evaluation coefficients in one step. The coefficients of agreement can be mentioned, e.g. Cohen's kappa (see Cohen, 1960).

In this paper, we discuss the selected coefficients for comparison of prediction results in the case of three or more groups. These coefficients are the total success rate, the group success rate, the kappa coefficient of agreement, and kappa-*n* and lambda-gamma coefficients. In addition to these coefficients, we propose a new coefficient with some better properties than the other five measures of interest have. This proposed *SRGM* (Success Rate for Groups / Mutability) measure extends the group success rate by its penalties in the cases when the variability of agreements (observed and predicted assignments of objects into groups) and the variability of the target variable differ. The nominal variance (mutability) is applied as a variability measure.

The total success rate and the group success rate can be computed using both the two-step and one-step approaches mentioned above. So, in this paper, all

coefficients are presented using the notation based on one classification table for all individual groups.

2 Approaches to classification evaluation

The classification table is often prepared for the purpose of prediction evaluation. Let us have such a table with frequencies (absolute N_{ij} or relative p_{ij}) of cases in which a certain object of the ith group was predicted into the jth group, where $i = 1, 2, \ldots, n$ (the same is valid for the j index), n is the number of rows and also the number of columns, i.e. the number of groups. The total number of objects is denoted as N. The agreement occurs if the object of the ith group is predicted into the ith group. The counts of such cases are expressed by the frequencies on the main diagonal (absolute N_{ii} or relative p_{ii}), N_{i+}, N_{+i} are the corresponding marginal absolute totals, similarly p_{i+}, p_{+i} are the corresponding marginal relative totals.

The total *success rate* (*SR*) for predictions expresses the ratio of the number of the correctly predicted objects and the total number of the objects, i.e.

$$SR = \frac{\sum_{i=1}^{n} N_{ii}}{N} . \tag{1}$$

It means that the value of *SR* is the sum of the relative frequencies on the main diagonal and it takes on the values from the interval $\langle 0; 1 \rangle$. It is usually presented as a percentage. Sometimes (e.g., in some statistical software products), it is called *overall percentage* or *classification accuracy*. This measure can be also expressed by a formula based on true positive and false positive counts (used for the measure *precision*) applied to the individual groups separately. Sokolova and Lapalme (2009) call this measure as *precision*$_\mu$.

However, the total percentage does not consider the differences among the percentages obtained for individual groups, when for one group the success rate can be close to 100% and for some other group close to 0%. For this reason, the use of other measures or coefficients is a better solution. In each group, the number of correctly predicted objects should be as high as possible. The success rates for the individual groups can be expressed in summary using the average. This *group success rate* or *success rate for groups* (*SRG*) is calculated as

$$SRG = \frac{1}{n} \sum_{i=1}^{n} \frac{N_{ii}}{N_{i+}} . \tag{2}$$

This measure can be also expressed as an average of the precisions calculated for the individual groups separately. Sokolova and Lapalme (2009) call this

measure as *precision*$_M$. If the sizes of all groups are the same, then $N_{i+} = N/n$ for $i = 1, 2, \dots, n$, and the values of *SR* and *SRG* are equal.

The observed sizes of individual groups can be expressed relatively by row marginal relative totals p_{i+}. The counts of agreements for individual groups can be expressed relatively within the total sum of agreements. For the good prediction, the variability of agreements should be the same (or similar) as the variability of the row marginal totals. The nominal variance (i.e. the mutability) proposed by Gini (1912), reviewed in (Light and Margolin, 1971) can be applied as a variability measure.

The mutability of the row marginal totals is defined as

$$M_{RMT} = \sum_{i=1}^{n} \left(\frac{N_{i+}}{N} \left(1 - \frac{N_{i+}}{N} \right) \right) = \sum_{i=1}^{n} \left(p_{i+}(1 - p_{i+}) \right). \tag{3}$$

Generally, the mutability takes on the value zero for a constant. However, we consider prediction into three or more groups, it means that the mutability of the row marginal totals cannot be zero. It reaches its maximum, i.e. $(n-1)/n$, when all row marginal totals are equal. By dividing the maximal value, we obtained the normalized mutability from the interval (0; 1⟩.

Let us define – for the case where at least one $N_{ii} \neq 0$ – the mutability of agreements as

$$M_A = \sum_{i=1}^{n} \left(\frac{N_{ii}}{\sum_{j=1}^{n} N_{jj}} \left(1 - \frac{N_{ii}}{\sum_{j=1}^{n} N_{jj}} \right) \right). \tag{4}$$

It takes on values from the interval ⟨0; $(n-1)/n$⟩. By dividing by its maximal possible value, we obtained the normalized mutability from the interval ⟨0; 1⟩. Zero indicates that the objects are correctly assigned only into one group, which means it makes no sense for the method to be used for prediction.

In this paper, we propose a new measure for the classification evaluation based on the *SRG* measures and the mutability ratio. We will denote it as the *SRGM* measure. If the mutability of agreements is the same as the mutability of the row marginal totals then the mutability ratio is one and *SRG* and *SRGM* are equal. All other cases are “penalized” by the mutability ratio which has to take on values less than one (the lower value is divided by the greater value). If objects are assigned only into one group, then the mutability is zero and *SRGM* is also zero (some objects are assigned correctly but the applied method cannot be used for prediction).

We can express the calculation of the *SRGM* measure in the following way:

$$SRGM = 0 \qquad \text{for } \sum_{i=1}^{n} N_{ii} = 0\,, \tag{5}$$

$$SRGM = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{N_{ii}}{N_{i+}}\right)\frac{M_A}{M_{RMT}} \qquad \text{for } \sum_{i=1}^{n} N_{ii} \neq 0 \text{ and } M_A \leq M_{RMT}\,, \tag{6}$$

$$SRGM = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{N_{ii}}{N_{i+}}\right)\frac{M_{RMT}}{M_A} \qquad \text{for } \sum_{i=1}^{n} N_{ii} \neq 0 \text{ and } M_A > M_{RMT}\,. \tag{7}$$

This measure takes on values from the interval from 0 (no agreement or correct assignments only to one group) to 1 (complete agreement for all groups).

Řezanková and Húsek (2015) proposed the *RH coefficient* for prediction evaluation. The authors consider only a special case with the same sizes of groups when the values of *SR* and *SRG* are equal. The mutability of agreements is expressed in the normalized form with the values from the interval $\langle 0; 1\rangle$. Then this measure is expressed by the formula

$$RH = \frac{\sum_{i=1}^{n} N_{ii}}{N}\frac{n}{n-1}\sum_{i=1}^{n}\left(\frac{N_{ii}}{\sum_{j=1}^{n} N_{jj}}\left(1 - \frac{N_{ii}}{\sum_{j=1}^{n} N_{jj}}\right)\right), \tag{8}$$

where the ratio $n/(n-1)$ normalizes the mutability. It is a special case of the general *SRGM* measure because, in the case of the same sizes of groups, the normalized mutability of the row marginal totals is one (in Eqs. (6) and (7), the ratio $n/(n-1)$ is omitted because there would be the same value in the numerator and denominator).

The well-known coefficient of agreement, which is used for the prediction evaluation, is the *kappa coefficient of agreement*. It was proposed by Cohen (1960) for the purpose when two psychologists independently categorize psychological test protocols into a certain number of categories. However, its use is wider. Its application for the prediction evaluation is often described in the literature (e.g. Powers, 2012). This coefficient is implemented in some software systems, usually as a part of the contingency table analysis for the square tables.

The basic features of kappa are the following. If all predicted groups correspond to observed groups, the value of this coefficient is one. If the counts in the classification table are the same as the frequencies expected under the condition of independence (according to the Pearson chi-square test considering marginal totals for predictions the same as they were really predicted), the value of this coefficient is zero. The values of kappa can be also negative. The lower limit depends on the marginal distributions, e.g. for two groups with the same group sizes it is −1.

In this paper, the sample coefficients based on the observed frequencies in the classification table are considered. Using the relative or absolute frequencies, kappa (κ) is expressed as

$$\kappa = \frac{\sum_{i=1}^{n} p_{ii} - \sum_{i=1}^{n} p_{i+}p_{+i}}{1 - \sum_{i=1}^{n} p_{i+}p_{+i}} = \frac{\sum_{i=1}^{n} N_{ii} - \frac{1}{N}\sum_{i=1}^{n} N_{i+}N_{+i}}{N - \frac{1}{N}\sum_{i=1}^{n} N_{i+}N_{+i}} . \quad (9)$$

If all frequencies in the contingency table except for the diagonal ones are zero, then the sum of all diagonal relative frequencies $\sum p_{ii}$ is one and also kappa is one. If the diagonal frequencies correspond to the frequencies expected under statistical independence, then $\sum p_{ii} = \sum p_{i+}p_{+i}$ and kappa is zero. Many modifications of the kappa coefficient were proposed; for their survey, see e.g. (Warens, 2013).

One of the several modifications of the kappa coefficient is *kappa-n* (κ_n), based on free marginal proportions when the random assignment of objects into groups can be characterized by value $1/n$ for all relative marginal totals. *Kappa-n* is derived as

$$\kappa_n = \frac{\sum_{i=1}^{n} p_{ii} - n\frac{1}{n}\cdot\frac{1}{n}}{1 - n\frac{1}{n}\cdot\frac{1}{n}} = \frac{\sum_{i=1}^{n} p_{ii} - \frac{1}{n}}{1 - \frac{1}{n}} = \frac{\sum_{i=1}^{n} N_{ii} - N\cdot\frac{1}{n}}{N - N\cdot\frac{1}{n}} . \quad (10)$$

This coefficient was firstly proposed by Bennet et al. (1954) as the *S coefficient*. Several other authors later proposed this coefficient under different names. The well-known name kappa-*n* was introduced by Brennan and Prediger (1981). For the history of this coefficient, see e.g. (Zwick, 1988; Hsu and Field, 2003).

The lower limit of kappa-*n* depends on the number of groups. It corresponds to the situation when the sum of frequencies on the main diagonal is zero. In the case of two groups, it is −1. However, in classification evaluation, row marginal proportions p_{i+} are known. Considering the known row marginal totals and the free column marginal total, we can write kappa as

$$\kappa_n = \frac{\sum_{i=1}^{n} p_{ii} - \frac{1}{n}}{1 - \frac{1}{n}} = \frac{\sum_{i=1}^{n} p_{ii} - \frac{1}{n}\sum_{i=1}^{n} p_{i+}}{1 - \frac{1}{n}\sum_{i=1}^{n} p_{i+}} \quad (11)$$

Both kappa and kappa-*n* have their negatives. Kappa considers the marginal totals for prediction the same as they were really predicted, it does not consider the free proportions. On the other hand, kappa-*n* is not dependent on the observed marginal totals.

Before Cohen's kappa, the *lambda-gamma* (λ_γ) coefficient was proposed by Goodman and Kruskal (1954). It is expressed as

$$\lambda_\gamma = \frac{\sum_{i=1}^{n} p_{ii} - \frac{1}{2}\max(p_{i+}+p_{+i})}{1 - \frac{1}{2}\max(p_{i+}+p_{+i})} = \frac{\sum_{i=1}^{n} N_{ii} - \frac{1}{2}\max(N_{i+}+N_{+i})}{N - \frac{1}{2}\max(N_{i+}+N_{+i})} . \qquad (12)$$

For obtaining the values greater than zero, the sum of the frequencies on the main diagonal must be greater than half of the sum of the two largest marginal totals.

The value −1 characterizes a situation where non-zero frequencies are only either in the first row and in the first column or in the last row and in the last column of the classification table, and the frequency in the corner cell is equal or greater than other frequencies. However, the value −1 can be a result of some corner frequencies lower than other frequencies.

Table 1 shows a summary of the properties of the six coefficients described above. The coefficients achieve their maximum in the case of complete agreement. They achieve their minimum if the sum of agreement frequencies for all groups is zero.

3 Discussion of prediction evaluation

Let us consider a classification task with three groups with frequencies 12, 9 and 15. For *SR* equal to one, we obtain the classification table denoted as Table 2 (the frequencies of agreements are in bold). For this situation, the values of all coefficients mentioned in this paper are equal to one.

If the predicted marginal totals are free (the same group sizes are considered), as it is shown in Table 3, *SR* = 0.333, *SRG* = 0.333, *SRGM* = 0.333, κ = 0, κ_n = 0 and $\lambda_\gamma = -0.067$.

If a certain method assigns (predicts) all objects only into one group, then the success rate can be promising but kappa cannot be computed. The user can obtain the zero value by some software product because the observed and expected frequencies are the same. If all 36 objects considered in Tables 2 and 3 are predicted only into group 3, then *SR* = 0.417, *SRG* = 0.333, *SRGM* = 0, κ = 0, κ_n = 0.125 and $\lambda_\gamma = -1$.

Tab. 1 Summary of properties of the selected coefficients for classification evaluation *(Source: the author)*

Coefficient	Maximum	Minimum	Disadvantages	Advantages
SR	1	0	*SR* is commonly available in statistical software systems.	The success rates in individual groups are not taken into account.
SRG	1	0	The success rates in individual groups are taken into account.	The variability of the agreements and the variability of the row marginal totals are not taken into account.
SRGM	1	0	The variability of the agreements and the variability of the row marginal totals are taken into account.	*SRGM* is not available in the software.
kappa	1	It depends on the marginal distributions (−1 for two groups with the same group sizes).	Kappa is commonly available in statistical software systems.	Kappa considers the marginal totals for predictions the same as they were really predicted.
kappa-*n*	1	It depends on the number of groups (−1 for two groups).	Kappa-*n* is based on the free marginal proportions.	Kappa-*n* does not take into account the observed group sizes.
lambda-gamma	1	It depends on the marginal distributions.	The assumptions for value −1 are defined.	Lambda-gamma is not available in commercial software systems.

For an illustration of coefficient value changes depending on the changes of frequencies in the classification table, let us consider marginal frequencies both for rows and for columns the same as they are in Table 2, i.e., 12, 9 and 15. In Table 4, there are frequencies in individual cells (*i*, *j*) of the classification table for nine different situations. These situations are considered from the completely correct prediction to the completely wrong prediction. The values of coefficients *SR*, *SRG*, *SRGM*, kappa, kappa-*n,* and lambda-gamma are included.

Tab. 2 A classification table for the 100% success rate *(Source: the author)*

	Prediction into group 1	**Prediction into group 2**	**Prediction into group 3**	**Marginal total**
Observed group 1	**12**	0	0	12
Observed group 2	0	**9**	0	9
Observed group 3	0	0	**15**	15
Marginal total	12	9	15	36

Tab. 3 A classification table for the uniform assignment of objects into groups *(Source: the author)*

	Prediction into group 1	**Prediction into group 2**	**Prediction into group 3**	**Marginal total**
Observed group 1	**4**	4	4	12
Observed group 2	3	**3**	3	9
Observed group 3	5	5	**5**	15
Marginal total	12	12	12	36

The *SR*, *SRG*, *SRGM* coefficients, take on values from the interval 0 to 1. Kappa, kappa-*n,* and lambda-gamma have their maximum also in 1 but the lower limits are different. For these coefficients, the definitions of lower limits may be interesting from a mathematical point of view, but from a practical point of view (for prediction evaluation) they have no meaning.

The *SR* coefficient equals 0.5 (the middle of the value interval) for Situation 4 in Table 4. The values of *SRG* and *SRGM* are lower for this situation. Kappa-*n* equals 0 for Situation 6 and lambda-gamma equals 0 for Situation 5. Kappa cannot have a value equal 0 for the considered marginal totals because some expected values under statistical independence are not integers.

In the illustration example mentioned above, the values of all coefficients are ordered from 1 to the last one. However, in some other cases, the orders are not the same, see below. For practical use, the *SRG* and *SRGM* coefficients should be greater than 0.5. There are only the first three situations in the example. In the case of the kappa coefficient, the situations with kappa greater than 0.4 are recommended. The first three situations also satisfy this condition.

In Table 5, there are frequencies in individual cells (*i*, *j*) of the classification table for different nine situations for the 36 objects. The values of *SR* and *SRG* coefficients are greater than 0.5. We can compare them with the values of the other coefficients studied in this paper.

Tab. 4 Frequencies for the individual cells (described by the row and column orders) of the classification table for 9 situations with marginal totals 12, 9 and 15 (both for rows and for columns) and values of the selected measures *(Source: the author)*

Cell	**Sit. 1**	**Sit. 2**	**Sit. 3**	**Sit. 4**	**Sit. 5**	**Sit. 6**	**Sit. 7**	**Sit. 8**	**Sit. 9**
(1,1)	**12**	**10**	**8**	**6**	**5**	**4**	**2**	**0**	**0**
(1,2)	0	1	2	3	3	4	5	6	6
(1,3)	0	1	2	3	4	4	5	6	6
(2,1)	0	1	2	3	4	4	4	3	0
(2,2)	**9**	**7**	**5**	**3**	**2**	**1**	**0**	**0**	**0**
(2,3)	0	1	2	3	3	4	5	6	9
(3,1)	0	1	2	3	3	4	6	9	12
(3,2)	0	1	2	3	4	4	4	3	3
(3,3)	**15**	**13**	**11**	**9**	**8**	**7**	**5**	**3**	**0**
Coefficient									
SR	1.000	0.833	0.667	0.500	0.417	0.333	0.194	0.083	0.000
SRG	1.000	0.826	0.652	0.478	0.391	0.304	0.167	0.067	0.000
SRGM	1.000	0.818	0.635	0.447	0.351	0.252	0.104	0.000	0.000
kappa	1.000	0.745	0.489	0.234	0.106	−0.021	−0.234	−0.404	−0.532
kappa-*n*	1.000	0.750	0.500	0.250	0.125	0.000	−0.208	−0.375	−0.500
lambda-gamma	1.000	0.714	0.429	0.143	0.000	−0.143	−0.381	−0.571	−0.714

In Situations 1, 2 and 3, the number of correctly predicted objects is 27, in Situation 4 it is 21 and for the rest situations, it is 19. The values of the *SR* coefficient are 0.75, 0.783 and 0.528 for these three different numbers. The values of the *SRG* coefficient are equal for Situations 3 and 4 because all objects from two groups and none object from the third group were correctly predicted. The different values for Situations 1 and 2 were obtained only for kappa and lambda-gamma because these coefficients take into account the column marginal totals, compared to the other studied coefficients.

The order of the values of the *SRGM* coefficient is similar to the order of the *SRG* values. Situation 5 was evaluated as the worst by both the *SRG* and *SRGM* coefficients. However, in some cases (Situations 3, 4 and 5), the *SRGM* values are lower by more than 0.1 due to the greater difference between the mutability of

agreements and the mutability of the row marginal totals, in comparison with other situations.

Tab. 5 Frequencies for the individual cells (described by the row and column orders) of the classification table for 9 situations with different marginal totals and values of the selected measures *(Source: the author)*

Cell	Sit. 1	Sit. 2	Sit. 3	Sit. 4	Sit. 5	Sit. 6	Sit. 7	Sit. 8	Sit. 9
(1,1)	**12**	**12**	**12**	**12**	**12**	**10**	**8**	**6**	**6**
(1,2)	0	0	0	0	0	1	3	3	0
(1,3)	0	0	0	0	0	1	1	3	6
(2,1)	0	0	4	0	0	0	0	0	0
(2,2)	**9**	**9**	**0**	**9**	**1**	**2**	**3**	**5**	**8**
(2,3)	0	0	5	0	8	7	6	4	1
(3,1)	0	9	0	15	0	2	4	6	6
(3,2)	9	0	0	0	9	6	3	1	4
(3,3)	**6**	**6**	**15**	**0**	**6**	**7**	**8**	**8**	**5**
Coefficient									
SR	0.750	0.750	0.750	0.583	0.528	0.528	0.528	0.528	0.528
SRG	0.800	0.800	0.667	0.667	0.504	0.507	0.511	0.530	0.574
SRGM	0.787	0.787	0.504	0.500	0.385	0.448	0.486	0.529	0.573
kappa	0.640	0.629	0.597	0.394	0.282	0.277	0.277	0.277	0.292
kappa-*n*	0.625	0.625	0.625	0.375	0.292	0.292	0.292	0.292	0.292
lambda-gamma	0.538	0.500	0.514	0.000	0.209	0.190	0.190	0.190	0.244

As concerns the coefficients of agreements, kappa-*n* takes on the values 0.625, 0.375 and 0.292, depending only on the number of correctly predicted objects, as well as the *SR* coefficient. The values of kappa and kappa-*n* are the same in the case when the column marginal totals are equal (Situation 9). The lowest value was found by lambda-gamma for Situation 4. For classification evaluation, the coefficients of agreement are not as suitable as the SRG and *SRGM* measure; they evaluate the obtained results in a very different way.

4 Conclusion

In the paper, selected coefficients for result evaluation in classification tasks with three or more groups of objects were discussed – five already proposed

coefficients and one newly proposed coefficient. Although some coefficients of this type were proposed more than 50 years ago (the first of them more than 100 years ago), they are still discussed in the literature (Powers, 2012; Warrens, 2013). In recent years, the interest in these coefficients again increases with the increasing need to evaluate results obtained in classification tasks.

We can distinguish two groups of these coefficients: success rates and coefficients of agreement. The success rates discussed in this paper (the total success rate, the group success rates, and the new *SRGM* measure) take on the values from the interval 0 to 1. Coefficients of agreement (kappa, kappa-*n*, and lambda-gamma) have the upper limit also equal to 1 but they can also take on negative values. The lower limit depends on the marginal distributions and the value of zero is achieved in different situations.

The total success rate is well known and commonly available in the software. However, it does not take into account the success rates in individual groups. Hence, the average of the success rates for the individual groups (the group success rate) characterizes the predictions better. In the paper, a new measure is proposed which extends the group success rate by its penalties in the cases when the variability of agreements (observed and predicted assignments of objects into groups) and the variability of the target variable differ. The nominal variance (mutability) is applied as a variability measure. Thus, this *SRGM* measure evaluates the prediction results most comprehensively. Moreover, the correct assignments of objects only to one group are evaluated by the value zero.

Coefficients of agreement are represented by coefficients kappa, kappa-*n* and lambda-gamma. The kappa coefficient of agreement is well-known and commonly available in the software. However, it considers the marginal totals for predictions the same as they were really predicted. Kappa-*n* is based on the free marginal proportions but it does not take into account the observed group sizes. Lambda-gamma evaluates the correct assignments of objects only to one group by low values.

As shown in this paper, the application of coefficients of agreement is not suitable for classification evaluation in some cases. These coefficients were proposed for other purposes and they do not take into account the specifics of the prediction. For that reason, using success rates is a better way. The total success rate does not distinguish some different situations. Thus, we can recommend either the group success rate or the newly proposed *SRGM* measure for prediction evaluation.

This paper was supported by Grant IGA F4/44/2018 of the University of Economics in Prague

5 References

Cohen, J. (1960). *A coefficient of agreement for nominal scales*. Educational and Psychological Management, 20(1), 37-46.

Bennet, E. M., Alpert, R., Goldstein, A. C. (1954). *Communications through limited-response questioning.* Public Opinion Quarterly, 18(3), 303-308.

Brennan, R.L., Prediger, D.J. (1981). *Coefficient kappa: Some uses, misuses, and alternatives*. Educational and Psychological Management, 41, 687-699.

Fawcett, T. (2006). *An introduction to ROC analysis*. Pattern Recognition Letters, 27(8), 861-874.

Gini, C. W. (1912). *Variability and mutability. Contribution to the study of statistical distributions and relations*. Studi Economico-Giuridici della R. Universita de Cagliari, 1912.

Goodman, L. A., Kruskal, W. H. (1954). *Measures of association for cross classifications*. Journal of the American Statistical Association, 49, 1732-1769.

Hsu, L. M., Field, R. (2003). *Interrater agreement measures: Comments on Kappan, Cohen's kappa, Scott's π, and Aickin's α.* Understanding Statistics, 2(3), 205-219.

Light, R. J., Margolin, B. H. (1971). *An analysis of variance for categorical data*. Journal of the American Statistical Association, 66, 534-544.

Powers, D. M. W. (2012). *The problem with kappa.* In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon: Association for Computational Linguistics, 2012, s. 345-355.

Řezanková, H., Húsek, D. (2015). *Measures for Classification Results Evaluation*. *Technical Report No. 1220*. Praha: Institute of Computer Science, the Czech Academy of Science, 2015. Dostupné na: http://hdl.handle.net/11104/0249608.

Sokolova, M., Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks*. Information Processing and Management, 45, 427-437.

Warrens, M. J. (2013). *A comparison of Cohen's kappa and agreement coefficients by Corrado Gini*. International Journal of Research and Reviews in Applied Sciences, 16(3), 345-351.

Zwick, R. (1988). *Another look at interrater agreement*. Psychological Bulletin, 103, 374-378.