Statistical and Possibilistic Methodology for the Evaluation of Classification Algorithms

Olgierd Hryniewicz

Systems Research Institute, Polish Academy of Sciences, Newelska 6, Warsaw, Poland
hryniewi@ibspan.waw.pl

Abstract. In the paper we consider the problem of the statistical evaluation and comparison of different classification algorithms. For this purpose we apply the methodology of statistical tests for testing independence in the case the multinomial distribution. We propose to use two-sample tests for the comparison of different classification algorithms. In the paper we consider only the case of the supervised classification when an external 'expert' evaluates the correctness of classification. The results of the proposed statistical tests are interpreted using possibilistic methodology based on indices of dominance introduced by [7].

Keywords: Classification, Accuracy, Statistical Tests of Independence, Multinomial Distribution, Comparison of Algorithms, Possibility and Necessity Indices.

1 Introduction

Statistical algorithms used for classification (discrimination) and clustering of observations (data points, data records) are considered as a part machine learning. Classification algorithms in machine learning are considered as the algorithms of supervised learning. On the other hand, data clustering algorithms in machine learning are used as the algorithms of unsupervised learning. In this paper we will discuss the problem of the evaluation of the quality of the algorithms used for classification, usually understood as the accuracy of classification, from a statistical point of view. A natural measure of such quality is the percentage of correctly classified objects, usually called *classification accuracy*. This measure is used by all authors of papers devoted to classification problems, both developers of new algorithms, and users of existing algorithms who apply them for solving practical problems.

Quality of classification measured by the *accuracy* index may not be sufficient for the comparison of algorithms. Consider for example a decision support system that classifies patients to different classes of illness. It is usually not unimportant if all false classifications are evenly distributed over all possible classes or if they are concentrated in one class. When we have only two classes or when this distinction is not important we can use indices whose background can be found in medical sciences, namely the indices of *sensitivity* and *specificity*. Let us assume that considered objects can be assigned to two disjoint classes called 'positive', and 'negative'. By *sensitivity* (also known in machine learning as *recall*) we understand the conditional probability that the object which should be classified to the 'positive'

M.J. Escalona, J. Cordeiro, and B. Shishkov (Eds.): ICSOFT 2011, CCIS 303, pp. 255–269, 2013.

class has been correctly assigned to this class. By *specificity* (also known in machine learning as *recall of negatives*) we understand the conditional probability that the object which should be classified to the ‘negative’ class has been correctly assigned to this class. For good classification rules the values of these indices should be both close to one. In machine learning some functions of these indices (e.g. *F-measures* or *ROC diagrams*) are used. For more information see e.g. Chapter 7 in [2].

When the number of possible classes is larger than two we have to take into account statistical relationship between errors of different kind. Some measures proposed for the evaluation of algorithms in the case of multiple classes, like e.g. the *error correlation EC,* have probabilistic interpretation, but the majority of them are based on some heuristics. For more information on this subject see e.g. Chapter 11 in [15]. The lack of statistical interpretation is of lesser importance if we deal with only one set of data. However, automatic classifiers may be used in situations when analyzed data sets may belong to different populations. For example, in automatic inspection of production processes classifiers are designed at the outset of the process using some training data, and then used using different set of ‘test’ data acquired prom a production process. In such cases possible classification algorithms should be compared using statistical methods. We believe that without statistical interpretation we are not able to present sound comparisons of different algorithms.

Application of statistical tests for the evaluation of classification algorithms has been proposed in [10]. In this paper, which presents an extended analysis of some problems considered in that paper, we propose to use some known statistical tests to evaluate and compare the quality of classification algorithms. In the second section of the paper we consider the problem of the comparison of algorithms. We consider two important practical cases. First is typical to the problem of supervised learning when the quality of classification of compared algorithms is evaluated using classification provided by an expert. In the second case, typical for algorithms related to unsupervised learning, we deal only with purely random data yielded by the compared algorithms.

The main problem with the application of different statistical tests is related to their interpretation. In the third section of the paper we propose a new application of *possibilistic measures* for the comparison of classification algorithms. This measures are based on the possibilistic interpretation of statistical tests proposed in [9], and provide the user with information about possibility or necessity of prefering one algorithm over another one. The paper is concluded in the fourth section where problems for future considerations are also formulated.

2 Statistical Tests for the Comparison of Classification Algorithms

Let us assume that we have to classify n objects into K disjoint classes using two algorithms, say A and B. In this paper we restrict ourselves to the case when the classification algorithm classifies each object to only one of possible classes. We do not impose any restriction on the type of the algorithm used for this purpose. This can be artificial neural network classifier, set of classification rules, vector supporting machine classifier, Bayes naïve classifier or any other algorithm that can be proposed for this purpose. An expert may act as one of these algorithms. In this case we are able to evaluate the correctness of the classification of each considered object by the second algorithm, as in the case of classical supervised learning. Thus, in this case we can use our statistical test to evaluate the quality of the classification algorithm. However, when we are not able to evaluate the correctness of the classification, we can only compare the performance of considered algorithms. This situation is typical when the algorithms are built using the methodology of unsupervised learning (e.g. using methods of data clustering).

When we compare two classification algorithms using *the same dataset* of n objects the results of the comparison may be presented in the form of a two-way contingency table, such as Table 1.

Table 1. Data for the comparison of algorithms using the same dataset

Alg.A/Alg.B	1	...	j	...	K	Total A
1	n_{11}	...	n_{1j}	...	n_{1K}	n_{A1}
...	...	...	...	...	...	...
i	n_{i1}	...	n_{ij}	...	n_{iK}	n_{Ai}
...	...	...	...	...	...	...
K	n_{K1}	...	n_{Kj}	...	n_{KK}	n_{AK}
Total B	n_{B1}	...	n_{Bj}	...	n_{BK}	n

By n_{ij}, $i=1,...,K;$ $j=1,...,K$ in this table we denote the number of observations that have been classified by the algorithm A to the ith class, and by the algorithm B to the jth class. We assume that the results of classification by the algorithm A are described by the set $(n_{A1}, n_{A2}, \ldots, n_{AK})$, and that the results of classification by the algorithm B are described by the set $(n_{B1}, n_{B2}, \ldots, n_{BK})$. The data can come from classifications performed on a test sample, combined results of cross-validation experiments or classification obtained in the learning process (training sample). However, in the latter case the results are of rather limited interest, as all good classification algorithms perform rather well on training data.

We are interested in the verification of the statistical hypothesis that the probability distributions are such that these sets of data are *strongly dependent*. This strong dependence means that both compared algorithms provide the same or nearly the same results of classification. When one of the compared algorithms is just an expert, the measure of such dependence is also the measure of the correctness of classification. Otherwise, the strength of dependence is the measure of the equivalence of the compared algorithms.

When we assume that the classification by the algorithm A and the classification by the algorithm B are *independent* then the data presented in Table 1 are distributed according to the *multiple hypergeometric* distribution. Probability of observing the two-way contingency table $\{n_{ij}\}$ with the fixed values of marginal observations $n_{Ai}, n_{Bj}, i = 1, \ldots, K, j = 1, \ldots, K$ is given by the formula:

$$P(\{N_{ij}\} = \{n_{ij}\}) = \frac{\prod_{i=1}^{K} n_{Ai}! \prod_{j=1}^{K} n_{Bj}! \cdot}{n! \prod_{i=1}^{K} \prod_{j=1}^{K} n_{ij}!} \tag{1}$$

The probability distribution given by (1) can be used for the construction of the test of independence. The general idea of this test, known as Fisher's exact test, is simple. We have to generate all possible contingency tables, such as Table 1, with the fixed margins equal to the margins observed for the considered table $n_{Ai}, n_{Bj}, i = 1,\ldots,K, j = 1,\ldots,K$. For all these tables we have to calculate, using (1), their probabilities. The sum of those probabilities whose values do not exceed the probability of the observed table is equal to the *p*-value (significance) of the test. Low values of this characteristics, say less than 0,05 (or 5%), indicate that the observed table does not support the hypothesis of the independence between classifications obtained using both compared algorithms, and thus, supports the alternative hypothesis of dependence.

Despite its simple and intuitive description the implementation of this algorithm is very difficult as the computational volume grows exponentially with the increasing values of *K* and *n*. Till the publication of the network algorithm in [13] computations were possible only for small tables. This algorithm, presented in the form of the FORTRAN code in [14] allows to compute *p*-values of Fisher's exact test for tables with larger values of *n*, provided that the table contains many cells with very low (i.e. equal to zero or close to zero) values. Fortunately, this is the case when we analyze good classification algorithms with a low percentage of false classifications.

In the case of large samples with significant percentage of false classifications we can use the well known Pearson's chi-square asymptotic test for independence. The chi-square statistic is given by

$$\chi_I^2 = \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}, \tag{2}$$

where

$$\hat{n}_{ij} = \frac{n_{Ai}n_{Bj}}{n} \tag{3}$$

is the expected number of observations in the *ij*th cell when both classifications, i.e. by the algorithm A and the algorithm B, are statistically *independent*. When the total number of observations *n* is large (greater than 100), and the expected number of observations in every cell is larger than 5, the chi-square statistic, defined by (2), is distributed according to the chi-square distribution with $(K\text{-}1)^2$ degrees of freedom. Thus, the *p*-value of this test is computed by solving, with respect to *p*, the following equation:

$$\chi_I^2 = \chi^2_{(K-1)^2,1-p}, \tag{4}$$

where $\chi^2_{(K-1)^2,1-p}$ is the quantile of the 1-*p* order from the chi-square distribution with $(K-1)^2$ degrees of freedom. When the total number of observations *n* is large we can consider the expectations calculated according to (3) as close to the theoretical expected values of observations. Then, we can use the rule proposed in [16] which states that if *r* is the number of cells with the expectations less than 5, then the lowest expectation could be as small as $5r/K^2$. When the chi-square test of independence is used for the evaluation of classification algorithms Yarnold's rule could be very useful in practice.

Let us apply the methodology explained above for the analysis of the classical linear discrimination algorithm (LDA) applied for a well known benchmark test – the famous Fisher's Iris test. The results of classification using the LDA algorithm implemented in the statistical package STATISTICA and the Iris data set are displayed in Table 2.

Table 2. Classification of the Iris data with the LDA algorithm

Expert \ LDA	Iris-Setosa	Iris-Versicolor	Iris-Virginica
Iris-Setosa	50	0	0
Iris-Versicolor	0	48	2
Iris-Virginica	0	1	49

The probability of the observation of this table, when the hypothesis of independence is true, is extremely low (3,1E-65). Thus, the *p*-value for Fisher's exact test of independence in the case of these data is equal to 0. It means that the results of classification provided by the expert are, as expected, strongly dependent. This supports the opinion that the LDA algorithm for this data set is very efficient.

Now, let us consider the application of another algorithm, namely Classification Regression Tree (CRT). The results of the application of this algorithm implemented in the statistical package STATISTICA are presented in Table 3.

Table 3. Classification of the Iris data with a CRT algorithm

Expert \ LDA	Iris-Setosa	Iris-Versicolor	Iris-Virginica
Iris-Setosa	50	0	0
Iris-Versicolor	0	48	2
Iris-Virginica	0	4	46

The probability of the observation of this table, when the hypothesis of independence is true, is also extremely low (1,6E-61). Hence, the *p*-value for Fisher's exact test of independence in the case of these data is equal to 0. Therefore, from a statistical point of view both algorithms are fully efficient. We have to note, however, that the difference between the numbers of observed false classification (3 by the LDA algorithm, and 6 by the CRT algorithm) in the case of a relatively small sample (150 observations) may be considered as random.

The Iris data are well separable, and classification algorithms usually perform very well on this benchmark. Let us consider now another example, presented in the paper [4], where number of false classifications, even on a training data set, is quite large. The results of the classification using a proposed in this paper Complete Gradient Clustering Algorithm (CGCA) are presented in Table 4.

Table 4. Classification of the wheat kernel data with the CGCA algorithm

Expert \ CGCA	Kama	Rosa	Canadian
Kama	59	2	9
Rosa	3	67	0
Canadian	3	0	67

The application of Fisher's exact test gives in this case also a extremely low *p*-value (0,897E-74) showing the great strength of dependence between the results of classification provided by the expert and the evaluated algorithm.

The numerical examples presented above show that the proposed statistical methodology is computationally demanding, and its results are difficult to interpret. The ratio of observed *p*-values provides some information about the superiority of one algorithm over another one, but this interpretation does not have any sound statistical basis.

Consider now the situation when all false classifications are equally important. In this case we can put them together in one class of incorrectly classified objects. Let $(n_1, n_2, \ldots, n_K, n_{K+1})$ be the vector describing the results of the application of the considered classification algorithm. First *K* components of this vector represent the numbers of cases of the *correct* classification to *K* considered classes. The last component gives the total number of incorrectly classified objects.

Let us assume now that observed values of $(n_1, n_2, \ldots, n_K, n_{K+1})$ represent a *sample* from an unknown multinomial distribution, defined by the probability mass function

$$MB(p_1, \ldots, p_K, p_{K+1}) = \frac{n!}{n_1! \cdots n_{K+1}!} \prod_{i=1}^{K+1} p_i^{n_i} \tag{5}$$

where $\sum_{i=1}^{K+1} n_i = n$, and $\sum_{i=1}^{K+1} p_i = 1$, that describes a hypothetical population of objects classified in a similar way to that used for the classification of the considered sample.

Now, let us suppose that we have to compare *two* classification algorithms, whose results of application are given in the form of two vectors $(n_1, n_2, \ldots, n_K, n_{K+1})$, and $(m_1, m_2, \ldots, m_K, m_{K+1})$, respectively. First, let us consider the case that both algorithms are compared using *the same set* of observations. Thus, the sample sizes in both cases are equal and both observed vectors are statistically *dependent*. In such case in order to compare the considered algorithms we have to know the results of the classification of each object, and then to use statistical methods devised for the analysis of pair-wise matched data. Theoretically, it is possible if we construct Fisher's test of independence using three-dimensional contingency table. Taking into account computational problems with classical Fisher's exact test it seems to be rather impossible to propose a test which compares classification algorithms taking into account their efficiencies.

The situation is different if we want to compare two algorithms without taking into account their efficiencies understood as probabilities of yielding correct classifications. In such a case the comparison can be done relatively easily when the data from a classification experiment performed on the *same* sample are given in the form presented in Table 5.

Table 5. Comparison of dependent test data

	Alg.1 -correct	Alg.1 - false
Alg. 2-correct	k_{11}	k_{12}
Alg. 2 - false	k_{21}	k_{22}

In this table k_{11} denotes the number of objects classified correctly by both algorithms, k_{12} denotes the number of objects classified correctly by the Algorithm 1 but incorrectly by the Algorithm 2, k_{21} is the number of objects classified correctly by the Algorithm 2 but incorrectly by the Algorithm 1, and k_{22} is the number of objects classified incorrectly by both algorithms.

Let us notice that the only information about the differences between both algorithms are contained in k_{12} and k_{21}. We can verify two hypotheses related to these values. First hypothesis is that the probabilities of incorrect classification that generate observations k_{12} and k_{21} are the for both compared algorithms are the same, and this hypothesis is tested against the alternative that they are simply different. In this case we have to apply the so called two-sided statistical test. We may also consider testing this statistical hypothesis against the alternative hypothesis that one particular algorithm is better than a second one. In this case we have to apply the so-called one-sided statistical test.

When both compared probabilities are equal it is known, see e.g. [1] for more information, that the number of incorrect classifications by only one algorithm k_{21} (or k_{12}) is described by the Binomial probability distribution with the parameters k=k_{12}+k_{21} and p=0,5. Let us assume now that we observe k_{12}^* and k_{21}^* incorrectly classified (only by one algorithm!) objects. The probability of observing k_{21} false classification given $k^* = k_{12}^* + k_{12}^*$ can be calculated from the following formula

$$P\left(k_{21} \mid k^*\right) = \begin{pmatrix} k^* \\ k_{21} \end{pmatrix} \left(\frac{1}{2}\right)^{k_{21}} \left(\frac{1}{2}\right)^{k^* - k_{21}}. \tag{6}$$

In order to verify the hypothesis of equal probabilities of misclassification we have to calculate, according to (6), the probabilities of all possible pairs $\left(k_{21}, k^*\right)$. In the case of the two-sided test the sum of those probabilities that do not exceed the probability of the observed pair $\left(k_{21} = k_{21}^*, k^*\right)$ gives the value of the significance (known also as the p-value) of the tested hypothesis. When this value is greater than 0,05 it is usually assumed that the hypothesis of the equal probabilities should not be rejected. In the case of the one-sided test we consider only these pairs $\left(k_{21}, k^*\right)$ that are less or equally probable that the observed pair $\left(k_{21}^*, k^*\right)$, and support the one-sided alternative. Thus, the p-value in the case of the one-sided alternative is smaller than in the case of the two-sided alternative. Hence, it is easier to reject the hypothesis that one algorithm is not worse than the other one than to reject the hypothesis that they are statistically equivalent.

When the number of objects k^* that are incorrectly classified only by one algorithm is sufficiently large (in practice it is required that the inequality k^*>10 must be fulfilled) the following statistic

$$T = \frac{\left(k_{12} - k_{21}\right)^2}{k_{12} + k_{21}} \tag{7}$$

is approximately distributed according to the chi-square distribution with 1 degree of freedom. This statistic is used in the well known McNemar’s test of the homogeneity of proportions for pair-wise matched data.

Let us consider again the example of Fisher's Iris data. We use this benchmark set for the comparison of two algorithms: LDA (Linear Discrimination Analysis) and CRT (Classification Regression Tree) – both implemented in a popular statistical software such as e.g. STATISTICA. For more information about these algorithms see e.g. [11]. Close examination of the classifications given by both algorithms results in the data presented in Table 6.

Table 6. Comparison of algorithms (LDA vs. CRT) – Iris data set

CRT\LDA	LDA – correct	LDA - false
CRT - correct	143	1
CRT - false	4	2

The *p*-value for these data, computed according to the algorithm given above, is equal to 0,375. Therefore, the obtained statistical data do not let us to reject the hypothesis that the probabilities of incorrect classification are in case of these two algorithms the same despite the fact that the CRT algorithm gives twice as many false classification in comparison to the LDA classifier.

Now, let us use the data that are less separable that the Iris data set. This situation is in the case of wheat kernel data considered in [4]. We will use these test data for the comparison of two algorithms: the Bayesian algorithm proposed in [12] and the classical Quadratic Discrimination Algorithm (QDA) algorithm described in [11]. The results of the comparison are presented in Table 7.

Table 7. Comparison of algorithms (Bayes vs. QDA)– Wheat kernels data set

QDA\Bayes	Bayes - correct	Bayes - false
QDA - correct	85	9
QDA - false	5	6

The *p*-value in this case is equal to 0,424. Therefore, the obtained statistical data do not let us to reject the hypothesis that the probabilities of incorrect classification are in the case of these two algorithms the same despite the fact that one of the compared algorithms (QDA) seems to be significantly better (nearly 30% lower probability of incorrect classification).

When we do not have an access to individual results of classification we can compare algorithms using *independent* samples described by the multinomial distributions. Let the data be described by (5), and $\sum_{i=1}^{K+1} n_i = n$ and $\sum_{i=1}^{K+1} m_i = m$ be the sample sizes which in general, as we compare the classifications of different samples, do not have to be equal. Moreover, note that in the case when one of these algorithms is a perfect classifier (e.g. a domain expert) we have $n_{K+1} = 0$ (or $m_{K+1} = 0$). If the results of the application of the first algorithm are described by the multinomial distribution $MB(p_1,\ldots,p_K,p_{K+1})$, and the results of the application of the second algorithm are described by the multinomial distribution $MB(q_1,\ldots,q_K,q_{K+1})$ their performance can be compared by testing the statistical hypothesis

$$H_0 : p_1 = q_1,\ldots,p_K = q_K, p_{K+1} = q_{K+1} . \qquad (8)$$

To test this hypothesis we may apply the methodology of two-way contingency tables. Test data in the case of the accumulation of all falsely classified object into one (K+1) class are presented in Table 8.

Table 8. Independent test data

Alg./Class	1	…	j	…	K	K+1	Total
Alg. 1	n_{11}	…	n_{1j}	…	n_{1K}	n_{1K+1}	N
Alg. 2	n_{21}	…	n_{2j}	…	n_{2K}	n_{2K+1}	M
Total	c_1	…	c_j	…	c_K	c_{K+1}	$N+M$

When the hypothesis H_0 given by (8) is true, the conditional distribution of observed random vectors $(n_1,n_2,\ldots,n_K,n_{K+1})$, and $(m_1,m_2,\ldots,m_K,m_{K+1})$, given the vector of their sum $(c_1,c_2,\ldots,c_K,c_{K+1})$, is given by the multivariate hypergeometric distribution [5]

$$P(\mathbf{n};\mathbf{m}\,|\,\mathbf{c},H_0)=\frac{m!n!}{N!}\prod_{i=1}^{K+1}\binom{c_i}{n_i} \tag{9}$$

This probability function is used for the construction of the multivariate generalization of Fisher's exact test that is used for the verification of (4). Let $\mathbf{n}^*$, $\mathbf{m}^*$ and $\mathbf{c}^*$ be the observed data vectors. The *p*-value (significance) of the test is computed from the formula [5]

$$(p-value)=\sum_{\Gamma}P(\mathbf{n},\mathbf{m}\,|\,\mathbf{c}^*,H_0), \tag{10}$$

where

$$\Gamma=\{(\mathbf{n},\mathbf{m}):P(\mathbf{n},\mathbf{m}\,|\,\mathbf{c}^*,H_0)\leq P(\mathbf{n}^*,\mathbf{m}^*\,|\,\mathbf{c}^*,H_0)\} \tag{11}$$

The *p*-values of this test can be computed by the tools of statistical packages such as SPSS or SAS. However, in the case of many categories and large (or even moderate) samples the computation time may be prohibitively long.

It can be shown that the test of the equality of two sets of multinomial probabilities is formally equivalent to the test of independence of categorical data, considered in the first part of this section. Hence, in the case of sufficiently large sample sizes with all cells having at least 5 observations, for testing (8) one can use Pearson's chi-square test of independence. These assumptions are usually fulfilled in testing classification algorithms, except for situations when tested data allows building perfect or nearly perfect classifiers. However, in such cases the problem of choice of the best classifiers does not exist.

The χ^2 statistic in the considered case can be written as

$$\chi^2=\sum_{i=1}^{K+1}\frac{(n_i-\hat{n}_i)^2}{\hat{n}_i}+\sum_{i=1}^{K+1}\frac{(m_i-\hat{m}_i)^2}{\hat{m}_i} \tag{12}$$

where

$$\hat{n}_i = \frac{nc_i}{N}, \tag{13}$$

and

$$\hat{m}_i = \frac{mc_i}{N}. \tag{14}$$

The p-value for this test is obtained by solving, with respect to p, the equation

$$\chi^2 = \chi^2_{K,1-p}, \tag{15}$$

where $\chi^2_{K,1-p}$ is the quantile of order 1-p in the chi-square distribution with K degrees of freedom. Also in this case the p-values of Pearson's chi-square test of independence can be computed using the tools available in various statistical packages.

In order to illustrate the application of the proposed tests in the evaluation of classification algorithms tested on samples of N=100 objects each which are classified into K=*3* classes. Suppose that we want to compare three algorithms A, B, and C, together with a "perfect" algorithm represented by an expert E. All compared algorithms have their basic and 'improved' versions indexed by subscripts 1 and 2, respectively. In order to make the evaluation simple we assume that all incorrect (false) classifications are assigned to the additional fourth class. Suppose that the results of this experiment are presented in Table 9.

Algorithms A, B and C in their both versions are characterized by the *same* total percentages of incorrect classification equal to 10% and 5%, respectively. However, the distribution of incorrectly classified objects depends upon the used algorithm. We face this situation when the algorithms are "aimed" at correct classification of chosen classes (e.g. Bayes classifiers).

Table 9. Results of an experiment with independent samples

Alg.\Class	1	2	3	4
Expert	20	30	50	0
A_1	18	27	45	10
A_2	19	29	47	5
B_1	10	30	50	10
B_2	15	30	50	5
C_1	20	30	40	10
C2	20	30	45	5

In the case of the algorithm A incorrectly classified objects are distributed proportionally to the actual sizes of classes. For the algorithm B all incorrectly classified objects are assigned to the class with the lowest number of actual observations. Finally, in the case of the algorithm C all incorrectly classified objects are assigned to the class with the highest number of actual observations.

In Table 10 we present the *p*-values of both considered tests when the performance of each classification algorithm is compared to the classification given by the expert.

Table 10. Comparison of different algorithms with the expert

	Fisher's	Chi-square
A_1 vs. E	0,008	0,015
B_1 vs. E	0,002	0,004
C_1 vs. E	0,006	0,011
A_2 vs. E	0,177	0,162
B_2 vs. E	0,132	0,126
C_2 vs. E	0,165	0,154

In the case of basic versions of all algorithms the results of classification are statistically significantly different than the classification provided by the expert. The worse classification is provided by the algorithm A. In the sample analyzed by this algorithm all falsely classified objects are evenly distributed over all classes. The best performance is observed in case of the algorithm B characterized by the largest percentage-wise differences between levels of the accuracy of classification in different classes. In the case of the 'improved' versions of the considered algorithms the data do not let us to reject the hypothesis that the results of classification are statistically equivalent to the results of classification provided by the expert.

Now, let us apply the proposed methodology for the comparison of basic and 'improved' versions of our hypothetical algorithms. The results of this comparison are presented in Table 11.

Table 11. Comparison of different versions of algorithms

	Fisher's	Chi-square
A_1 vs. A_2	0,640	0,613
B_1 vs. B_2	0,470	0,446
C_1 vs. C_2	0,599	0,581

The results of this comparison are somewhat unexpected for a non-statistician. Despite seemingly large improvement (reduction of the percentage of incorrect classifications from 10% to 5%) the compared results statistically do not differ. The reason for this behavior is, of course, a small sample size. What is also interesting that the difference is the least significant (the highest *p*-value in the test of equality) in the case of evenly distributed misclassifications. The lowest *p*-value (but still very high using statistical standards) is for the case of algorithm B which assigns all incorrectly classified objects to the class with the smallest number of observations.

Now, let us consider an example of the application of this methodology to real data. Suppose, that we have been provided with two algorithms for the classification of vehicle silhouettes data (data provided by Turing Institute, Glasgow, and available at the UCI web-site). One of these algorithms implements the Bayesian algorithm proposed in [12], and the second one implements a classical CRT algorithm described in [3]. The algorithms have been tested on two *independent* samples, and the results of this comparison are presented in Table 12.

Table 12. Comparison of algorithms - Vehicle Silhouettes data set

Alg.\Class	1	2	3	4	5
Bayes	55	48	112	90	141
CRT	46	55	86	84	175

The p-value obtained as the solution of (15) for these data is equal to 0,079. According to the classical statistical approach this result does not let us claim that the Bayes algorithm is better than the CRT. Note however, that similar results obtained on the *same* sample would probably indicate the superiority of the Bayes algorithm.

3 Possibilistic Evaluation of Test Results

In the previous section we have proposed statistical tests for the evaluation of classification procedures. The results of the proposed test procedures have been expressed in terms of significance, known also as the test volume or the p-value. Examples given in this section show that the results of statistical tests interpreted in a traditional way are not well suited for finding if one classification algorithm is better than the other one. Therefore, there is a need to present an additional indicator that can be used to show to what extent one algorithm is better than the other one despite the fact that they are statistically equivalent. This goal can be achieved using the methodology proposed in the *theory of possibility*. In the papers [8] and [9] the possibilistic interpretation of statistical tests has been proposed. This interpretation gives a decision maker the evaluation of test's result using notions of *possibility* or *necessity* of making certain decisions.

Statistical decision problems are described by setting a certain hypothesis H (usually called the null hypothesis), and an alternative hypothesis K. In the context of decision-making we usually choose this hypothesis which is better supported by statistical data. Hryniewicz [9] proposes to consider these two hypotheses separately. Suppose that the significance of H is given by the p-value of the test, and is equal to p_H. The value of p_H shows to what extent the statistical evidence supports the null hypothesis.

In [9] it was proposed to evaluate the null hypothesis H by a fuzzy set $\tilde{H}$ with the following membership function

$$\mu_H(x) = \begin{cases} \min[1, 2p_H] & x = 0 \\ \min[1, 2(1 - p_H)] & x = 1 \end{cases}. \tag{16}$$

This membership function may be interpreted as a *possibility distribution* of the truth of H. If $\mu_H(1) = 1$ holds, it means that it is quite *plausible* that the considered hypothesis is not true. On the other hand, when $\mu_H(0) = 1$, we would not be surprised if H were true.

The same can be done for the alternative hypothesis K. The statistical test of this hypothesis may be described by another p-value, denoted by p_K. When K= *not H* we have p_K=1-p_H. However, in a general setting this equality usually does not hold. The alternative hypothesis K is now represented by a fuzzy set $\tilde{K}$ with the following membership function

$$\mu_K(x) = \begin{cases} \min[1, 2p_K] & x = 0 \\ \min[1, 2(1 - p_K)] & x = 1 \end{cases}. \tag{17}$$

In order to choose an appropriate decision, i.e. to choose either H or K, Hryniewicz [9] proposes to use three measures of possibility defined in [6].

For two fuzzy sets $\tilde{A}$ and $\tilde{B}$, described by their membership functions $\mu_A(x)$ and $\mu_B(y)$, respectively, the *Possibility of Dominance* (*PD*) measure is defined in [6] in the following way

$$PD\left(\tilde{A} \geq \tilde{B}\right) = \sup_{x, y: x \geq y} \min[\mu_A(x), \mu_B(y)]. \tag{18}$$

The second index is called the *Possibility of Strict Dominance* (*PSD*), and for two fuzzy sets $\tilde{A}$ and $\tilde{B}$ is given by the expression

$$PSD\left(\tilde{A} > \tilde{B}\right) = \sup_x \left\{ \inf_{y: x \leq y} [\min(\mu_A(x), 1 - \mu_B(y))] \right\}. \tag{19}$$

Positive, but smaller than 1, values of this index indicate certain weak evidence that $\tilde{A}$ strictly dominates $\tilde{B}$.

Third measure is named the *Necessity of Strict Dominance*, and for two fuzzy sets $\tilde{A}$ and $\tilde{B}$ has been defined in [6] as:

$$NSD\left(\tilde{A} > \tilde{B}\right) = 1 - \sup_{x, y: x \leq y} [\min(\mu_A(x), \mu_B(y))]. \tag{20}$$

The *NSD* index represents a *necessity* that the fuzzy set $\tilde{A}$ strictly dominates the set $\tilde{B}$.

In the considered statistical problem of testing a hypothesis H against an alternative K these indices have been calculated in [8], and are given by the following formulae

$$PD\left(\tilde{H} \geq \tilde{K}\right) = \max[\mu_H(0), \mu_K(1)], \tag{21}$$

$$PSD\left(\tilde{H} > \tilde{K}\right) = \min[\mu_H(0), 1 - \mu_K(0)], \tag{22}$$

$$NSD\left(\tilde{H} > \tilde{K}\right) = 1 - \max[\mu_H(1), \mu_K(0)]. \tag{23}$$

The value of *PD* represents the *possibility* that according to the observed statistical data the choice of the null hypothesis is not a worse decision than choosing its alternative. The value of *PSD* gives the measure of *possibility* that the data support rather the null hypothesis than its alternative. Finally, the value of *NSD* gives the measure of *necessity* that the data support the null hypothesis rather than its alternative.

It has been proved that

$$PD \geq PSD \geq NSD. \tag{24}$$

It means that according to the practical situation we can choose the appropriate measure of the correctness of our decision. If the choice between H and K leads to serious consequences we should choose the *NSD* measure. In such a case p_H>0,5 is required to have *NSD*>0. When these consequences are not so serious we may choose the *PSD* measure. Finally, the PD measure, which is always positive, gives us the information of the possibility that choosing H over K is not a completely wrong decision.

In some cases considered in this paper the alternative hypothesis has been formulated as the complement of the null hypothesis, Thus, we have the equality $p_K = 1 - p_H$. In this case we have

$$PD\left(\tilde{H} \geq \tilde{K}\right) = \mu_H(0) = \min(1, 2p_H), \tag{25}$$

$$PSD\left(\tilde{H} > \tilde{K}\right) = NSD\left(\tilde{H} > \tilde{K}\right) = \max(2p_H - 1, 0). \tag{26}$$

Let us apply these results for the comparison of different algorithms using the test result presented in Table 6 for the comparison of the LDA and CRT algorithms used for the classification of the Iris data. For this statistical test we have p_H=0,375, and p_K=0,625. Hence, we have *PD*=0,750, and *PSD*=*NSD*=0. Therefore, there is only a certain possibility that these two algorithms are equivalent, but the measure of the necessity of such claim is equal to zero.

The possiblilistic comparisons are not necessary when null and alternative hypotheses are, as in the cases considered above, complementary. In such case strong evidence in favor of the null hypothesis means automatically weak support of its complementary alternative.

4 Conclusions

In the paper we have considered the problem of the evaluation and comparison of different classification algorithms. For this purpose we have applied the methodology of statistical tests for the multinomial distribution. We restricted our attention to the case of the supervised classification when an external ‘expert’ evaluates the correctness of classification. The results of the proposed statistical tests are interpreted using the possibilistic approach introduced in [9]. The results presented in this paper can be extended to the case of imprecise data. In this case the applicability of the proposed possibilistic measures is even much stronger when we omit, for example, the assumption that there exists an ‘expert’ who indicates only one ‘true’ class. In such cases we have to use the methodology of fuzzy statistics, whose overview can be found e.g. in [7]. We will face such problems, for example, when we will adapt the methodology presented in this paper for the case of the evaluation of fuzzy classifiers.

Acknowledgements. The author expresses his thanks to Dr. P. A. Kowalski and Dr. S. Łukasik for providing solutions for some practical examples of classification problems.

References

1. Agresti, A.: Categorical Data Analysis, 2nd edn. J. Wiley, Hoboken (2006)
2. Berthold, M., Hand, D.J. (eds.): Intelligent Data Analysis. An Introduction, 2nd edn. Springer, Berlin (2007)
3. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. CRC Press, Boca Raton (1984)
4. Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Łukasik, S., Żak, S.: Complete Gradient Clustering Algorithm for Features Analysis of X-Ray Images. In: Piętka, E., Kawa, J. (eds.) Information Technologies in Biomedicine. AISC, vol. 69, pp. 15–24. Springer, Heidelberg (2010)
5. Desu, M.M., Raghavarao, D.: Nonparametric Statistical Methods for Complete and Censored Data. Chapman & Hall, Boca Raton (2004)
6. Dubois, D., Prade, H.: Ranking Fuzzy Numbers in the Setting of Possibility Theory. Information Science 30, 183–224 (1983)
7. Gil, M.A., Hryniewicz, O.: Statistics with Imprecise Data. In: Meyers, R.A. (ed.) Encyclopedia of Complexity and Systems Science, pp. 8679–8690. Springer, Heidelberg (2009)
8. Hryniewicz, O.: Possibilistic Interpretation of the Results of Statistical Tests. In: Proceedings of Eight International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU 2000, Madrid, pp. 215–219 (2000)
9. Hryniewicz, O.: Possibilistic decisions and fuzzy statistical tests. Fuzzy Sets and Systems 157, 2665–2673 (2006)
10. Hryniewicz, O.: Possibilistic methodology for the evaluation of classification algorithms. In: Proceedings of the 6th International Conference on Software and data Technology, ICSOFT 2011, Seville (July 2011)
11. Krzanowski, W.J.: Principles of Multivariate Analysis: A User's Perspective. Oxford University Press, New York (1988)
12. Kulczycki, P., Kowalski, P.A.: Bayes classification of imprecise information of interval type. Control and Cybernetics 40, 101–123 (2011)
13. Mehta, C.R., Patel, N.R.: Network algorithm for performing Fisher's exact test in r × c contingency tables. Journ. Amer. Stat. Assoc. 78, 427–434 (1983)
14. Mehta, C.R., Patel, N.R.: ALGORITHM 643: FEXACT: a FORTRAN subroutine for Fisher's exact test on unordered r × c contingency tables. ACM Transactions on Mathematical Software (TOMS) 12, 154–161 (1986)
15. Nisbet, R., Elder, J., Miner, G.: Statistical Analysis and Data Mining. Applications. Elsevier Inc., Amsterdam (2009)
16. Yarnold, J.K.: The Minimum Expectation in X^2 Goodness of fit test and the Accuracy of Approximations for the Null Distribution. Journ. Amer. Stat. Assoc. 70, 864–886 (1970)