Rule Induction: Combining Rough Set and Statistical Approaches

Wojciech Jaworski

Faculty of Mathematics, Computer Science and Mechanics
Warsaw University, Banacha 2, 02-097 Warsaw, Poland
wjaworski@mimuw.edu.pl

Abstract. In this paper we propose the hybridisation of the rough set concepts and statistical learning theory. We introduce new estimators for rule accuracy and coverage, which base on the assumptions of the statistical learning theory. Then we construct classifier which uses these estimators for rule induction. These estimators allow us to select rules describing statistically significant dependencies in data. We test our classifier on benchmark datasets and show its applications for KDD.

Keywords: Rough sets, quality measures, accuracy, coverage, significance, rule induction, rule selection.

1 Introduction

Rough set theory [7] and statistical learning theory [11] provide two different methodologies for reasoning from data.

The rough set concept theory is a theoretical framework for describing and inferring knowledge. Examined knowledge is imperfect. It is imprecise due to vague concepts involved in knowledge representation and it is based on incomplete data. The central point of the theory is the idea of concept approximation by the sets of objects that certainly belongs to the concept and the set of those which may belong to the concept on the basis of possessed data.

The main goal of statistical learning theory is to provide a framework for studying the problem of inference. For this purpose, there are introduced statistical assumptions about the way the data is generated. A probabilistic model of data generation process, which is the core of the theory, establishes the formalisation of relationships between past and future observations.

While rough set theory provides an intuitive description of relationships in data, stereotypes that express general yet imprecise truths, statistical learning theory measures the significance and correctness of discovered dependencies.

The combination of both approaches provides us tools for building simple, human understandable classifiers, whose quality will be guaranteed by the statistical assumptions.

In this paper we propose the hybridisation of the rough set approach and statistical learning theory. We define the probabilistic model of data generation process. We recall rough set concepts in this new setting. Then we show how to

C.-C. Chan et al. (Eds.): RSCTC 2008, LNAI 5306, pp. 170–180, 2008.

extend set approximations from a sample to the set of all objects. Our attitude is similar to the idea of inductive extensions of approximation spaces presented, for example, in [8,9].

We introduce measures of approximation quality: accuracy and coverage. Taking advantage of the underlying probabilistic model we estimate values of the above indices on the set of all objects using a sample. We propose two estimators: one based on Hoeffding inequality [6], and second based on the optimal probability bound presented in [4,5].

The statistical nature of estimators leads us to the index, the measure called a significance. The significance measures how often sample-based accuracy and coverage estimations are correct. The trade-off relation between these three measures allow us to balance the approximation between fitting to the sample and generalisation.

The properties of accuracy and coverage were thoroughly studied in [10]. The author proposed the probabilistic definition of the indices, yet he neither defined an underlying probability model nor showed the trade-off between accuracy or coverage and significance. Quality measures are also examined from the statistical point of view in [3], but without placing them in the rough set context.

Gediga and Düntsch propose in [2] an application of statistical techniques in rough set data analysis, yet they did not incorporate the assumptions on the data generating process required by these techniques into the presented model.

In order to show how the estimators behave in practice we developed a simple rule based classifier. Estimated indices guarantee the quality of each rule, decide how accurate rules are acceptable and how many objects have to match the rule in order to make it significant. We test the classifier on benchmark datasets obtained from [1].

Test results revealed that the obtained classifier generates highly relevant rules. Each rule is assigned with its accuracy and coverage estimations. Rules cover only that part of universe for which it is possible to predict decision with high accuracy. As a consequence the classifier is able to judge whether it has enough knowledge to classify a certain object.

2 Probabilistic Model

We propose the following definition of the problem of induction. We are given a domain organised in terms of *objects* possessing *attributes*. Depending on the nature of domain, objects are interpreted as, e.g. cases, states, processes, patients, observations. Attributes are interpreted as features, variables, characteristics, conditions, etc.

Let $\mathbb{U}$ be a finite set of objects for a given domain. We denote $\mathbb{U}$ as *universe*. Let A be a non-empty finite set of attributes such that $a : \mathbb{U} \to V_a$ for every $a \in A$ and let $d \notin A$ such that $d : \mathbb{U} \to V_d$ be a decision attribute. We introduce a probability measure $P_{\#}$ on $2^{\mathbb{U}}$ according to the following formula:

$$\forall_{X \subseteq \mathbb{U}} \, P_{\#}(X) = \frac{|X|}{|\mathbb{U}|},$$

where $|\cdot|$ denotes the number of elements in a set.

Statistical learning theory [11] assumes that the phenomena underlying generated data have statistical nature and the observed objects are independent, identically distributed random variables.

Formally we introduce a probability space $(\Omega, 2^{\Omega}, P)$. Observed objects $u_1, u_2, \ldots, u_i, \ldots$ are values of independent random variables $U_1, U_2, \ldots, U_i, \ldots$. Each U_i is a function $U_i : \Omega \to \mathbb{U}$. The distribution of U_i is identical to $P_{\#}$, i.e.:

$$\forall_i \forall_{X \subseteq \mathbb{U}} P_{\#}(X) = P(\{\omega \in \Omega \mid U_i(\omega) \in X\}) = P(U_i^{-1}(X))$$

Let $U \subseteq \mathbb{U}$ be a non-empty, finite set of observed objects called a *sample.* U, together with the values of attributes for elements of U, is our knowledge about the domain. We denote elements of U by $u_1, \ldots, u_n$, where u_i is a realisation (or value) of the random variable U_i. We represent this knowledge in terms of the triple $\mathcal{A} = (U, A, d)$, usually denoted as a *decision system.*

3 Set Approximations

Classification is the task of finding the dependence between the attribute values and the value of decision. The rough set theory [7] provides tools and methodology for performing classification.

The basic concept of rough set theory is the indiscernibility relation. Let $\mathcal{A} = (U, A, d)$ be a decision system and $B \subseteq A$.

$$IND_{\mathcal{A}}(B) = \{(u, u') \in U^2 | \forall a \in B \; a(u) = a(u')\}$$

is called the *B-indiscernibility relation.* The B-indiscernibility is an equivalence relation. We will denote its equivalence class generated by object u as $[u]_B$.

The notion of indiscernibility is used to define set approximations. A given set $X \subseteq U$ may be approximated using only the information contained in $B \subset A$ by constructing the *B-lower* and *B-upper approximations of* X, denoted $\underline{B}X$ and $\overline{B}X$ respectively, where $\underline{B}X = \bigcup\{[u]_B | [u]_B \subseteq X\}$ and $\overline{B}X = \bigcup\{[u]_B | [u]_B \cap X \neq \emptyset\}$.

In the case of classification, we approximate sets of objects that possess a given decision. Let $X_v = \{u \in U | d(u) = v\}$. The objects in $\underline{B}X_v$ can be with certainty classified as members of decision class v on the basis of knowledge represented by B, while the objects in $U \setminus \overline{B}X_v$ definitely are not members of decision class v on the basis of knowledge represented by B.

For a given set of attributes B, formulae of the form $a = v$, where $a \in B$ and $v \in V_a$ are called *descriptors* over B. The set of *conditional formulae* over B is defined as the least set containing all descriptors over B and closed with respect to the propositional connectives $\wedge$ (conjunction), $\vee$ (disjunction) and $\neg$ (negation).

Let φ be a conditional formula over B. $||\varphi||_{\mathcal{A}}$ denotes the meaning of φ in the decision system $\mathcal{A}$, which is the set of all objects in U with the property φ. These sets are defined as follows:

1. if φ is of the form $a = v$, then $||\varphi||_{\mathcal{A}} = \{x \in U | a(x) = v\}$;
2. $||\varphi \wedge \varphi'||_{\mathcal{A}} = ||\varphi||_{\mathcal{A}} \cap ||\varphi'||_{\mathcal{A}}$; $||\varphi \vee \varphi'||_{\mathcal{A}} = ||\varphi||_{\mathcal{A}} \cup ||\varphi||_{\mathcal{A}}$; $||\neg\varphi||_{\mathcal{A}} = U \setminus ||\varphi||_{\mathcal{A}}$.

Every indiscernibility class can be represented by means of conditional formulae composed out of conjunction of descriptors. Let $B = \{a_1, \ldots, a_n\}$, $u \in U$ and let $v_1, \ldots, v_n$ be such that $a_i(u) = v_i$. In such a case

$$[u]_B = \{u' \in U | \forall a \in B \; a(u) = a(u')\} = ||a_1 = v_1 \wedge \cdots \wedge a_n = v_n||_{\mathcal{A}}.$$

We express the lower approximation by means of a conditional formula $\underline{B}X = ||\varphi_1 \vee \cdots \vee \varphi_k||_{\mathcal{A}}$, such that $\varphi_1, \ldots, \varphi_k$ are formulae representing indiscernibility classes that compose $\underline{B}X$. Similarly, there exist $\psi_1, \ldots, \psi_l$ such that $\overline{B}X = ||\psi_1 \vee \cdots \vee \psi_l||_{\mathcal{A}}$.

A *decision rule* for $\mathcal{A}$ is any expression of the form $\varphi \rightarrow d = v$, where φ is a conditional formula, $v \in V_d$ and $||\varphi||_{\mathcal{A}} \neq \emptyset$. A decision rule $\varphi \rightarrow d = v$ is *true* in $\mathcal{A}$ if, and only if, $||\varphi||_{\mathcal{A}} \subseteq ||d = v||_{\mathcal{A}}$. A decision rule describes the dependence between a decision class and its approximation.

4 Extended Approximations

In the above section we considered set approximations that described the dependence between the attribute values and the value of decision for objects in U. Now, we extend set approximations on the whole universe $\mathbb{U}$. The extended approximations of all decision classes will compose a classifier.

The assumption that past and future observations are both sampled independently from the same distribution provide us with tools for extending the approximations. However, the extension will be correct only with some probability.

We represented approximations by means of conditional formulae which are interpreted in the decision system. For a given set of attributes B, extended approximations are represented by means of conditional formulae over B interpreted in the universe $\mathbb{U}$. Let φ be a conditional formula over B and let $||\varphi||_{\mathbb{U}}$ denote its meaning in the universe of all objects. The meaning is defined as follows:

1. if φ is of the form $a = v$ then $||\varphi||_{\mathbb{U}} = \{u \in \mathbb{U} | a(u) = v\}$;
2. $||\varphi \wedge \varphi'||_{\mathbb{U}} = ||\varphi||_{\mathbb{U}} \cap ||\varphi'||_{\mathbb{U}}$; $||\varphi \vee \varphi'||_{\mathbb{U}} = ||\varphi||_{\mathbb{U}} \cup ||\varphi||_{\mathbb{U}}$; $||\neg\varphi||_{\mathbb{U}} = \mathbb{U} \setminus ||\varphi||_{\mathbb{U}}$;

For every U_i we obtain from its definition[1]

$$P_{\#}(||a = v||_{\mathbb{U}}) = P_{\#}(\{u \in \mathbb{U} | a(u) = v\}) =$$

$$= P(\{\omega \in \Omega | a(U_i(\omega)) = v\}) = P(a(U_i) = v).$$

This correspondence may be easily extended on all conditional formulae.

Now, we define extended approximations using conditional formulae interpreted in the universe $\mathbb{U}$:

[1] The latter equality introduces a standard probabilistic notation in which 'ω', '{' and '}' are omitted in expressions with random variables.

Definition 1. *Let $X \subseteq \mathbb{U}$ and B be a set of attributes and let $Y \subseteq \mathbb{U}$ be such that*

$$Y = ||\varphi||_{\mathbb{U}},$$

*where φ is a conditional formula over B. The set $Y \subseteq \mathbb{U}$ is called B-α-κ-*approximation *of X when*

$$P_{\#}(X \mid Y) \geq \alpha \text{ and } P_{\#}(Y \mid X) \geq \kappa.$$

We denote α as the approximation accuracy *and we denote κ as the approximation* coverage.

On the contrary to the standard approximations defined in a decision system this definition does not construct a set Y, it only states whether a given set possesses a property of being an α-κ-approximation.

Accuracy and coverage are indices of the approximation quality. Accuracy measures the probability that an object belonging to the approximation belongs also to the approximated set. Coverage measures the percent of objects in a set that are included in its approximation. When the approximation accuracy is equal to 1 and the coverage is maximised the approximation may be considered as *lower* and when the approximation coverage is equal to 1 and the accuracy is maximised the approximation may be considered as *upper*.

Accuracy and coverage are defined by means of the underlying probability distribution, according to which the sample is drawn. Since we are given only a sample and we do not know the probability distribution, we must estimate values of the indices using the sample and probabilistic inequalities of the form

$$P\big(P_{\#}(X \mid Y) \geq f_n(U_1, \ldots, U_n)\big) \geq \gamma_n.$$

The above inequality may be interpreted in the following way: if we draw $\{(u_1^i, u_2^i, \ldots, u_n^i)\}_{i=1}^{\infty}$, an infinite sequence of n-element samples, then according to the law of large numbers

$$P\big(P_{\#}(X \mid Y) \geq f_n(U_1, \ldots, U_n)\big) =$$
$$= \lim_{k \to \infty} \frac{1}{k} \cdot |\{i \leq k \mid P_{\#}(X \mid Y) \geq f_n(u_1^i, \ldots, u_n^i)\}|.$$

Hence γ_n describes how frequent it is true that $P_{\#}(X \mid Y) \geq f_n(u_1^i, \ldots, u_n^i)$ or, in other words how likely $P_{\#}(X \mid Y) \geq f_n(u_1^i, \ldots, u_n^i)$ is to happen in one occurrence. γ_n is a measure called *significance.*

We propose two methods of deriving estimators of the accuracy and the coverage on the basis of sample. The first bases on the Hoeffding inequality [6]:

Theorem 1. *Let $Z_1, \ldots, Z_n$ be identically distributed independent random variables. Assume that each $Z_i \in [0, 1]$. Then, for every $\varepsilon > 0$, the following inequality takes place:*

$$P(EZ_1 \leq \frac{1}{n} \sum_{i=1}^{n} Z_i + \varepsilon) \geq 1 - e^{-2n\varepsilon^2}. \tag{1}$$

Assume that Y is an α-κ-approximation for the set X. Let U be a sample and let $\{U_1, \ldots, U_n\} = U \cap Y$. For the purpose of accuracy estimation we declare that

$$Z_i = \begin{cases} 0, & \text{when } U_i \in X \\ 1, & \text{when } U_i \notin X \end{cases} .$$

Since

$$EZ_1 = P(Z_1 = 1) = P(U_1 \notin X \mid U_1 \in Y) = 1 - P_{\#}(X \mid Y),$$

we obtain the following inequality

$$P((1 - P_{\#}(X \mid Y)) \leq \frac{1}{n} \sum_{i=1}^{n} Z_i + \varepsilon) \geq 1 - e^{-2n\varepsilon^2}$$

Now, we take the advantage of the law of large numbers and the fact that we know the realisation of the sample U. We calculate a realisation for each Z_i in the following way

$$z_i = \begin{cases} 0, & \text{when } u_i \in X \\ 1, & \text{when } u_i \notin X \end{cases} ,$$

where u_i is i-th u_k such that $u_k \in Y$. The statement

$$(1 - P_{\#}(X \mid Y)) - \frac{1}{n} \sum_{i=1}^{n} z_i \leq \varepsilon$$

is likely to happen with significance $1 - e^{-2n\varepsilon^2}$.

n denotes the number of variables Z_i. It is equal from the definition to the number of elements in the sample that belong to Y. On the other hand $Z_i = 1$ if and only if the corresponding U_i does not belong to X. Since U_i have to belong to U and Y we obtain

$$n = |U \cap Y| \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} z_i = \frac{|(U \cap Y) \setminus X|}{|U \cap Y|} = 1 - \frac{|U \cap Y \cap X|}{|U \cap Y|}.$$

If we assume that significance is equal to γ we obtain

$$\varepsilon = \sqrt{\frac{\ln(1 - \gamma)}{-2|U \cap Y|}}$$

and the approximation accuracy is estimated from (1) with the significance γ according to the formula

$$P_{\#}(X \mid Y) \geq \frac{|U \cap Y \cap X|}{|U \cap Y|} - \sqrt{\frac{\ln(1 - \gamma)}{-2|U \cap Y|}}.$$

The coverage estimator is developed in the analogous way from (1), and the following estimator is obtained

$$P_{\#}(Y \mid X) \geq \frac{|U \cap Y \cap X|}{|U \cap X|} - \sqrt{\frac{\ln(1 - \gamma)}{-2|U \cap X|}}.$$

Table 1. Exemplary decision system

	a	d
u_0	1	1
u_1	0	0
u_2	0	0
$\vdots$	$\vdots$	$\vdots$
u_{100}	0	0

We illustrate the trade-off between these three numerical factors using the following example. Consider decision system presented in Table 1. We obtain the following lower approximation for the objects in the system:

$$\underline{\{a\}}||d=0||_{\mathcal{A}} = ||a=0||_{\mathcal{A}}, \quad \underline{\{a\}}||d=1||_{\mathcal{A}} = ||a=1||_{\mathcal{A}}.$$

Yet we cannot state that $||a=0||_{\mathbb{U}}$ is an approximation of $||d=0||_{\mathbb{U}}$ with a 100% accuracy, since there may exist an object u_{101} in $\mathbb{U} \setminus U$ such that $a(u_{101}) = 0$ and $d(u_{101}) = 1$. The given decision system suggests that such an occurrence is unlikely, yet still it is possible.

We estimate the approximation accuracy with significance 95%:

$$P_{\#}(||d=0||_{\mathbb{U}} \mid ||a=0||_{\mathbb{U}}) \geq \frac{|\,||d=0 \wedge a=0||_{\mathcal{A}}|}{|\,||a=0||_{\mathcal{A}}|} - \sqrt{\frac{\ln(1-0.95)}{-2|\,||a=0||_{\mathcal{A}}|}} =$$

$$= \frac{100}{100} - \sqrt{\frac{\ln(0.05)}{-200}} = 0.88.$$

Hence, the accuracy of the approximation of the set $||d=0||_{\mathbb{U}}$ by means of $||a=0||_{\mathbb{U}}$ is greater than 88% with significance 95%. On the other hand, for the approximation $\underline{\{a\}}||d=1||_{\mathbb{U}} = ||a=1||_{\mathbb{U}}$, we do not obtain any significant accuracy estimation.

Hoeffding inequality provides us with a simple analytic formula for the approximation accuracy, yet the obtained estimator is not optimal. That is why we propose the second estimator based on the bound proposed in [4]. It results in an optimal estimator.

Theorem 2. *Let $Z_1, \dots, Z_n$ be identically distributed independent random variables such that $Z_i \in \{0,1\}$, $i = 1, \dots, n$. Then, the following inequality takes place:*

$$P\Big(EZ_1 > g_{n,\gamma}(\frac{1}{n}\sum_{i=1}^{n} Z_i)\Big) < \gamma,$$

where, for a given $k < n$, $g_{n,\gamma}$ satisfies the equation

$$\sum_{i=0}^{k} \binom{n}{i} g_{n,\gamma}(\frac{k}{n})^i (1 - g_{n,\gamma}(\frac{k}{n}))^{n-i} = \gamma$$

and $g_{n,\gamma}(1) = 1$. $g_{n,\gamma}$ provides the optimal (most sharp) bound of EZ_1.

The second estimator does not provide any analytic formula for an estimator value, yet $g_{n,\gamma}(\frac{k}{m})$ may be calculated using an algorithm proposed in [4].

According to the second estimator the accuracy of the approximation of the set $||d = 0||_{\mathbb{U}}$ by means of $||a = 0||_{\mathbb{U}}$ is greater than 97% with significance 95%.

5 Rule Induction Algorithm

Extended approximations of all decision classes compose a classifier. Unfortunately an extended approximation for a given set is not uniquely defined. Many algorithms for calculating approximations were developed. Often the approximations are represented by means of decision rules.

In order to illustrate the link of theory with practical results we propose a simple algorithm for rule induction. The algorithm generates a classifier calculating extended approximations for all decision classes. Each approximation is represented as a set of decision rules whose predecessors are conjunctions of descriptors. For each rule, the accuracy, the coverage and the significance are calculated. The algorithm is parametrised by minimal levels of significance and accuracy and it induces all the rules that satisfy these minimal levels of indices. As a consequence induced rules do not cover all objects, and the classifier has not enough knowledge to recognise some objects. On the other hand all the classified objects are certified to be classified correctly with a very high probability.

The algorithm works as follows: In the k-th step the algorithm tries to induce rules whose predecessors possess k descriptors In the 0th step it checks using the estimator whether there is a decision value v such that the rule with empty predecessor and decision value v would have the desired accuracy and significance. If the answer is positive, then the rule is generated and the rule induction process ends. Otherwise, all the possible rule predecessors with one selector are generated and checked using the estimator. Then the second selector is added, and so on.

The algorithm uses two heuristics that speed it up: it does not try to generate a rule that is more specific than any existing rule and it checks whether there is enough objects matching to the rule predecessor to make it significant.

The algorithm ends when no more rules may be created.

In the case when during classification several rules may be applied to a given object, we choose the rule with the greatest accuracy.

Many more effective algorithms for rule generation that the one described above were developed (for example, in RSES system). However, our objective was to illustrate the theory with a practical application and to show the link between set approximations and induced rules only.

6 Tests

To evaluate the performance of the algorithm, 3 benchmark data sets were selected: *chess*, *nursery*, *census94*. The data sets are obtained from the repository of University of California at Irvine [1].

Each data set is split into a training and a test set. For *census94* data sets the original partition available in the repository was used in the experiments. The remaining data sets (*chess* and *nursery*) ware randomly split into a training and a test part with the split ratio 2 to 1.

All the selected sets are the data sets from UCI repository that have data objects represented as vectors of attributes values and have the size between a few thousand and several tens thousand of objects.

Chess and *nursery* have only nominal attributes. *Census94* possess both nominal and numeric attributes. The numeric attributes were discretised.

Table 2 presents test results obtained using the estimator based on Thm. 1. Table 3 presents test results obtained using the estimator based on Thm. 2. In both cases rules were induced with significance 95%.

The tests results show that the algorithm generates a small number of highly relevant rules which makes it useful for knowledge discovery. The fact that it estimates accuracy and coverage for each rule provide us with an insight into

Table 2. Test results obtained using the estimator based on Thm. 1

dataset	min accuracy	number of rules	classifier accuracy	classifier coverage
nursery	0.900000	42	0.985617	0.778395
chess	0.900000	80	0.952963	0.954944
census94	0.950000	32	0.951100	0.502610
census94	0.900000	83	0.899346	0.758307
census94	0.800000	107	0.812987	0.998894

Table 3. Test results obtained using the estimator based on Thm. 2

dataset	min accuracy	number of rules	classifier accuracy	classifier coverage
nursery	0.900000	112	0.989269	0.884722
chess	0.900000	310	0.957419	0.968085
census94	0.950000	92	0.951274	0.590873

Table 4. Part of 53 rules induced from *census94* dataset with significance 0.95 and minimal accuracy 0.85

Accuracy	Coverage	Rule
0.874541	0.388500	sex=Female → class=<=50K
0.938863	0.417531	marital-status=Never-married → class=<=50K
0.883111	0.310253	relationship=Not-in-family → class=<=50K
0.958077	0.198552	relationship=Own-child → class=<=50K
0.967552	0.195818	age=17-23 → class=<=50K
0.893863	0.143788	age=24-28 → class=<=50K
0.899943	0.064978	hours-per-week=18-24 → class=<=50K
0.940021	0.168487	capital-gain=7000-99999 → class=>50K
0.843932	0.050181	occupation=Machine-op-inspct, hours-per-week=40 → class=<=50K
0.879734	0.052272	occupation=Handlers-cleaners → class=<=50K
0.827732	0.040772	occupation=Adm-clerical, education=Some-college → class=<=50K
0.901273	0.048653	education=11th → class=<=50K

the internal structure of data. Table 4 illustrates the above statements presenting a part of rules induced from *census94* dataset.

7 Conclusions

The hybridisation of roughs sets and statistical learning theory resulted in the concept of extended approximation and statistical estimators for rule accuracy and coverage.

These estimators may be used with any rule induction algorithm. They guarantee the relevance of induced rules.

Extended approximations create a theoretical background for the classification. They indicate the connection between lower and upper approximations and rules induced from sample.

The theory and algorithms may be further developed to make them suitable for handling missing values, numerical attributes and other types of data.

Acknowledgment. The research has been partially supported by grants N N516 368334 and N N206 400234 from Ministry of Science and Higher Education of the Republic of Poland and by the grant Innovative Economy Operational Programme 2008-2012 (Priority Axis 1. Research and development of new technologies) managed by Ministry of Regional Development of the Republic of Poland.

References

1. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
2. Gediga, G., Düntsch, I.: Statistical techniques for rough set data analysis. In: Polkowski, L., et al. (eds.) Rough set methods and applications: New developments in knowledge discovery in information systems, pp. 545–565. Physica Verlag, Heidelberg (2000)
3. Guillet, F., Hamilton, H.J. (eds.): Quality Measures in Data Mining. Studies in Computational Intelligence, vol. 43. Springer, Heidelberg (2007)
4. Jaworski, W.: Model Selection and Assessment for Classification Using Validation. In: Ślęzak, D., et al. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 481–490. Springer, Heidelberg (2005)
5. Jaworski, W.: Bounds for Validation. Fundamenta Informaticae 70(3), 261–275 (2006)
6. Hoeffding, W.: Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association 58, 13–30 (1963)
7. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
8. Pawlak, Z., Skowron, A.: Rough sets: Some extensions. Information Sciences 177(1), 28–40 (2007)

9. Skowron, A., Swiniarski, R., Synak, P.: Approximation spaces and information granulation. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets III. LNCS, vol. 3400, pp. 175–189. Springer, Heidelberg (2005)
10. Tsumoto, S.: Accuracy and Coverage in Rough Set Rule Induction. In: Alpigini, J.J., Peters, J.F., Skowron, A., Zhong, N. (eds.) RSCTC 2002. LNCS (LNAI), vol. 2475, pp. 373–380. Springer, Heidelberg (2002)
11. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)