Hybridization of Rough Sets and Statistical Learning Theory

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 hybridization 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. These estimators allow us to select rules describing statistically significant dependencies in data. Then we construct classifier which uses these estimators for rule induction. In order to make our solution applicable for information systems with missing values and multiple valued attributes, we propose axiomatic representation of information systems and we redefine the indiscernibility relation as a relation on objects characterized by axioms. Finally, we test our classifier on benchmark datasets.

Keywords: Rough sets, quality measures, accuracy, coverage, significance, rule induction, rule selection, missing values, multiple valued attributes.

1 Introduction

Rough set theory [1, 2] and statistical learning theory [3] 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 set of objects that certainly belong to the concept and the set of those which may belong to the concept on the basis of possessed data. Then these two sets are described in terms of available attributes.

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 and approximations for dependencies that cannot be defined in an exact way, 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 hybridization of the rough set approach and statistical learning theory. We define the probabilistic model of data generation process, which allow us to explain the process of data acquisition and to infer knowledge that would be applied for all existing objects, not only for the ones that are mentioned in data.

We recall rough set concepts in this new setting. However, we introduce axiomatic representation of information systems, which we developed in [4]. We define rough set concepts such as indiscernibility, definability and set approximations in this setting. In the case of complete information systems, the proposed approach is equivalent to the approach used so far in rough set theory [1, 2]. Yet, it allows us to incorporate information systems with missing values and multiple valued attributes into the theory 'seamlessly' — without the need of any modification of the rough set concepts.

Then we show how to 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 [5, 6].

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 [7], and second based on the optimal probability bound presented in [8, 9].

The statistical nature of estimators leads us to the index, the measure called significance. 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 a probabilistic definition of the indices, yet he neither defined any underlying probability model nor showed the trade-off between accuracy or coverage and significance. Quality measures were also examined from the statistical point of view in [11], but without placing them in the rough set context.

[12] propose 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, determine the required accuracy level for rule to be accepted and decide how many objects have to match the rule in order to make it significant. We test the classifier on benchmark datasets obtained from [13] and we apply it in analysis of Neo-Sumerian economic documents (for details see [14]).

Test results reveal 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. Let $\mathbb{U}$ be a finite set of objects for a given domain. We denote $\mathbb{U}$ as *universe*. We introduce a probability measure $P_{\#}$ on $2^{\mathbb{U}}$ according to the following formula:

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

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

Statistical learning theory [3] 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.:

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

We do not know $\mathbb{U}$ and $P_{\#}$.

Let $U \subseteq \mathbb{U}$ be a non-empty, finite set of observed objects called a *sample*. U is the only part of the domain $\mathbb{U}$ which is known to us. 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. The information which we posses about the domain is usually represented in terms of information system.

3 Complete information systems

In this section, we define information systems [15] and we recall their axiomatic representation, which we introduced in [4].

Information systems are based on the assumption that examined domain is 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 $U \subset \mathbb{U}$ be a sample — a non-empty, finite set of known objects. Let A be a non-empty finite set of known attributes. Each attribute $a \in A$ has its domain V_a. An information system defines attribute values for given objects. Let

$$m(u, a)$$

denote the set of values of the attribute a for the object u in the information system.

Usually information systems are presented in a form of tables whose rows represent objects and columns are labelled by attributes.

In case when each attribute has exactly one and known value for each object, i.e $m(u, a)$ contains one element for every u and a, information system is called *complete*.

We consider an information system as a set of axioms. The information system provides the structural information about the domains of the attributes. We represent this information by means of axioms that set constraints on a set of possible worlds. For each attribute a we state

$$\forall_{x,y}\ a(x, y) \Longrightarrow x \in \mathbb{U} \wedge y \in V_a.$$

The complete information system also states that every attribute has exactly one value for each object: for each attribute a we write the following axiom

$$\forall_{x \in \mathbb{U}}\ \exists!_y\ a(x, y).$$

We encode the contents of the information system as a set formulae in the following way: For each $u \in U$, for each $a \in A$ such that $v \in m(u, a)$ in the information system we add the following axiom:

$$a(u, v).$$

The above transformation treats both an object etiquette and an attribute value as constants. The attributes are considered as binary relations.

Real-life data is frequently incomplete , i.e. values for some attributes are missing (see e.g. [16–19]). We will assume three different interpretations of missing values:

- missing attribute values that are *lost*, i.e they are specified, yet their values are unknown
- attributes *not applicable* in a certain case, e.g. the date of death of a person who is still alive.
- *do not care values*: the attribute may have any value from its domain.

We will extend the definition of $m(u, a)$. $m(u, a) =?$ will mean that the value of attribute a for object u is lost, $m(u, a) = \star$ that it is 'do not care' and $m(u, a) = -$ that it is not applicable.

We express the various types of missing value semantics using axioms:

- for each $u \in U$, for each $a \in A$ we state

  $$a(u, v),$$

  where $v \in m(u, a)$ in the information system.
- 'lost' values are defined as follows: for each $u \in U$, for each $a \in A$ we state

  $$a(u, v_1) \vee \cdots \vee a(u, v_n),$$

  where $v_1, \ldots, v_n$ are all possible values of attribute a.

– for each $u \in U$, for each $a \in A$ whose value is not applicable we state

$$\forall_x \neg a(u, x),$$

– for each $u \in U$, for each $a \in A$, for each v from the domain of a we state

$$a(u, v),$$

when the value of a is 'do not care' for object u.

Multiple valued attributes (introduced in [15] and studied in [20]) may reflect our incomplete knowledge about their values, what makes them similar to 'lost' missing values. The may also represent attributes that have a few values simultaneously, in which case the are like 'do not care' missing values.

– 'lost' multiple values we define as follows: for each $u \in U$, for each $a \in A$ we state

$$a(u, v_1) \vee \cdots \vee a(u, v_n),$$

where $v_1, \ldots, v_n$ are all possible values of attribute a for object u mentioned in the information system.

– for each $u \in U$, for each $a \in A$, for each value v of attribute a for object u in information system

$$a(u, v),$$

when the value of a is 'do not care' multiple value for object u.

4 Rough set theory

The rough set theory [1, 2] is based on the idea of an indiscernibility relation. In this section, we define indiscernibility and set approximations.

In this and the following sections we assume that we are given an information system $\mathcal{A}$. U denote the finite set of objects described in $\mathcal{A}$, A is the finite set of attributes in $\mathcal{A}$ and $\mathbb{A}$ is a set of axioms derived from $\mathcal{A}$.

Let B be a nonempty subset of A. The indiscernibility relation $IND(B)$ is a relation on objects in a complete information system defined for $x, y \in U$ as follows

$$(x, y) \in IND(B) \text{ iff } \forall a \in B\big(m(x, a) = m(x, a)\big).$$

IND is an equivalence relation. We will denote its equivalence class generated by object u as

$$[u]_{IND(B)}.$$

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

$$\overline{B}X = \bigcup\{[u]_B | [u]_B \cap X \neq \emptyset\}.$$

The above theory was designed or complete information systems. However, in [4], we proved that the concepts of indiscernibility and set approximations in a way that they could cover also information systems with missing values and multivariate attributes.

Now we recall this definition. First we introduce auxiliary concepts of descriptor, query and conditional formula.

Definition 1. *For a given set of attributes $B \subseteq A$, formulae of the form*

$$a(x, v),$$

where $a \in A$, $v \in V_a$ and x is a free variable, are called descriptors *over B.*

Definition 2. *By a* query *over the set of attributes B we denote any formula*

$$\bigwedge_{i=1}^{n} \varphi_i(x),$$

where each φ_i is a descriptor over B and $n \leq |B|$. x is a free variable ranging over objects.

Definition 3. *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).*

Note that every query is a conditional formula.

Definition 4. *Let $\varphi(x)$ be a conditional formula. By $||\varphi(x)||_{U,\mathbb{A}}$ we will denote the set of all elements from U for which φ is a semantic consequence of $\mathbb{A}$, i.e.:*

$$||\varphi(x)||_{U,\mathbb{A}} = \{x \in U \mid \mathbb{A} \models \varphi(x)\}.$$

We postulate the following definition of indiscernibility:

Definition 5. *Let $\varphi(x)$ be a query with free variable x. Let u_1 and u_2 be constants. We say that u_1 and u_2 are* indiscernible by the query $\varphi(x)$ *if*

$$\big(\mathbb{A} \models \varphi(u_1)\big) \Longleftrightarrow \big(\mathbb{A} \models \varphi(u_2)\big).$$

Theorem 1. *Let $\mathcal{A}$ be a complete information system. Let B be a subset of A. Objects $u_1 \in U$ and $u_2 \in U$ are indiscernible with respect to attribute set B iff they are indiscernible with respect to every query over the set of attributes B.*

Proof. See [4].

A given set $X \subset U$ is either *definable* or *indefinable* by attributes in the information system depending on the existence of conditional formula that recognizes its elements:

Definition 6. *Let X be a subset of U. We say that X is* definable *by $\mathbb{A}$ iff there exist queries $\varphi_1(x), \ldots, \varphi_n(x)$ such that*

$$X = ||\varphi_1(x) \vee \cdots \vee \varphi_n(x)||_{U,\mathbb{A}}$$

Each definable set is a sum of objects that satisfy at least one of a given queries.

Any set $X \subset U$ may be approximated by two definable sets. The first one is called the *lower approximation* of X, denoted by $\underline{\mathbb{A}}X$, and is defined by

$$\bigcup\{Y \mid Y \subset X \wedge Y \text{ is definable by } \mathbb{A}\}.$$

The second set is called the *upper approximation* of X, denoted by $\overline{\mathbb{A}}X$, and is defined by

$$\bigcap\{Y \mid X \subset Y \wedge Y \text{ is definable by } \mathbb{A}\}.$$

$\overline{\mathbb{A}}X \subset U$ because every definable set is a subset of U.

Theorem 2.

$$\underline{A}X = \underline{\mathbb{A}}X \text{ and } \overline{A}X = \overline{\mathbb{A}}X.$$

For the proof of the above theorem and further results concerning comparison of our concept of set approximations with the one proposed by other authors see [4].

Classification is a process of finding dependencies between values of attributes. Let $\mathbb{A}$ be a given set of axioms which define attributes A for objects from the set U. We select one of attributes from A which we denote as d — decision attribute. Let $B = A \setminus \{d\}$. Our goal is to estimate the value of attribute d on the basis of other attribute values for a given object. For each value v of the decision attribute, there exist conditional formulae over B that define the lower and upper approximation of $||d(x, v)||_{U,\mathbb{A}}$. We denote them $\underline{\varphi}_v(x)$ and $\overline{\varphi}_v(x)$ respectively.

$$||\underline{\varphi}_v(x)||_{U,\mathbb{A}} \subseteq ||d(x, v)||_{U,\mathbb{A}} \subseteq ||\overline{\varphi}_v(x)||_{U,\mathbb{A}}$$

Set approximations for all decision values compose a classifier.

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

Inductive reasoning is based on the assumption that the definition generated for the sample data is still valid in the general case. 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 $||\varphi||_{\mathbb{U},\mathbb{A}}$ denote the subset of elements of the universe $\mathbb{U}$ that satisfy the formula.

For every U_i we obtain from its definition[1]

$$P_{\#}(||a(x,v)||_{\mathbb{U},\mathbb{A}}) = P_{\#}(\{x \in \mathbb{U} \mid \mathbb{A} \models a(x,v)\}) =$$

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

This correspondence may be easily extended on all conditional formulae.

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

Definition 7. *Let $X \subseteq \mathbb{U}$ and B be a set of attributes and let $Y \subseteq \mathbb{U}$ be such that*

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

where φ is a conditional formula over B. Let $\alpha, \kappa \in [0,1]$. 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 call α as the approximation accuracy *and we denote κ as the approximation* coverage.

As opposed 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 fraction 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* one and when the approximation coverage is equal to 1 and the accuracy is maximised the approximation may be considered as *upper* one.

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, where u_j^i is a realisation of U_j^i, then according to the law of large numbers

$$P\big(P_{\#}(X \mid Y) \geq f_n(U_1, \ldots, U_n)\big) = P\big(P_{\#}(X \mid Y) \geq f_n(U_1^i, \ldots, U_n^i)\big) =$$

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

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

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 [7]:

Theorem 3. *Let $Z_1, \ldots, Z_n$ be identically distributed independent random variables. Assume that each $Z_i : \Omega \to [0, 1]$. Then, for every $\varepsilon > 0$, the following inequality holds:*

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

□

We derive estimator from this theorem as follows: assume that Y is an α-κ-approximation for a 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, by 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

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

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

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

following lower approximation for the objects in the system:

$$\underline{\{a\}}||d(x,0)||_{U,\mathbb{A}} = ||a(x,0)||_{U,\mathbb{A}}, \quad \underline{\{a\}}||d(x,1)||_{U,\mathbb{A}} = ||a(x,1)||_{U,\mathbb{A}}.$$

Yet we cannot state that $||a(x,0)||_{\mathbb{U},\mathbb{A}}$ is an approximation of $||d(x,0)||_{\mathbb{U},\mathbb{A}}$ 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 event is unlikely, yet still it is possible.

We estimate the approximation accuracy with significance 95%:

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

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

Hence, the accuracy of the approximation of the set $||d = 0||_{\mathbb{U},\mathbb{A}}$ by means of $||a = 0||_{\mathbb{U},\mathbb{A}}$ is greater than 88% with significance 95%. On the other hand, for the approximation $\underline{\{a\}}||d(x,1)||_{\mathbb{U},\mathbb{A}} = ||a(x,1)||_{\mathbb{U},\mathbb{A}}$, 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 [8]. It results in an optimal estimator.

Theorem 4. *Let $Z_1, \ldots, Z_n$ be identically distributed independent random variables such that $Z_i : \Omega \to \{0,1\}$, $i = 1, \ldots, n$. Then, the following inequality holds:*

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

where, for a given $k < n$, $g_{n,\gamma}$ satisfies the equation and $g_{n,\gamma}(1) = 1$. $g_{n,\gamma}$ provides the optimal bound of EZ_1.

The second estimator does not provide any analytic formula for the estimator value, yet $g_{n,\gamma}(\frac{k}{m})$ may be calculated using the algorithm proposed in [8].

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

6 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 have been developed. Often approximations are represented by means of decision rules.

A *decision rule* for a given decision system is any expression of the form $\varphi(x) \to d(x,v)$, where φ is a conditional formula, d is a decision attribute, $v \in V_d$ and $||\varphi(x)||_{U,\mathbb{A}} \neq \emptyset$. A decision rule $\varphi(x) \to d(x,v)$ is *true* in the decision system if, and only if, $||\varphi||_{U,\mathbb{A}} \subseteq ||d = v||_{U,\mathbb{A}}$. A decision rule describes the dependence between a decision class and its approximation.

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 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 the algorithm moves to the 1st step.
- In the 1st step the set P_1 of all the possible rule predecessors with one descriptor are generated. Each element of P_1 is checked using the estimator. If the answer is positive, then the rule is generated. Then we remove from P_1 all elements used to generate rules and we denote the remaining set as P_1'.
- In the k-th step, $k > 1$, the generates the set P_k on the basis of P_{k-1}' in the following way: each element $\varphi(x)$ of P_{k-1}' and for each descriptor $a(x, v)$ such that a does not appear in $\varphi(x)$ we add add $\varphi(x) \wedge a(x, v)$ to P_k. Each element of P_k is checked using the estimator. If the answer is positive, then the rule is generated. Then we remove from P_k all elements used to generate rules and we denote the remaining set as P_k'.
- 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 sufficiently many objects matching the rule predecessor to make it significant.
- The algorithm ends when no more rules may be created.

The algorithm generates short and relevant rules that cover only a part of universe.

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 [21] 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.

7 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 [13].

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%.

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

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

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

8 Decision rules extraction

We used our methodology in practical task of decision rules extraction from Neo-Sumerian economic documents [14]. We extracted decision rules from the database of animal transfer transactions, which was obtained as a result of parsing of documents. These rules bring to light dependencies between Sumerian

officials, animal types and transaction dates. The following table presents a few rules (generated with significance at least 95%):

Accuracy	Coverage	Rule
0.762606	0.895833	Receiver=lu2-{d}gesz-bar-e3 → Kiszib=a-kal-la sipa
0.866110	1.000000	MuDu=lum-ma → Kiszib=ensi2 u3-da
0.888281	0.873563	MuSze=ur-ra, Giri={d}en-lil2-la2 → Kiszib={d}szul-gi-a-a-mu
0.863621	0.739130	Year=SS09, Supplier=du-du → Kiszib=u4-de3-nig2-sag10
0.799641	0.226601	Year=SZ41, Animal=udu → Maszkim=en-{d}nansze-ki-ag2
0.775528	0.928571	Year=AS07, Kiszib=ab-ba-sa6-ga → Maszkim=du-du
0.770440	0.857143	MuSze=ensi2 zimbir{ki} → Maszkim=ur-{d}gubalag nar ta2-hi-isz-a-tal
0.827830	0.273504	Receiver=lu2-mah → MuDu={d}szara2

The first rule states that if lu2-{d}gesz-bar-e3 is a receiver of goods in transaction then with a probability 76% a-kal-la sipa seals the document. And this rule covers 89% of cases when a-kal-la sipa seals any document. The remaining rules are interpreted analogically.

We extracted 12841 rules. These rules help to direct Sumerological research pointing out interesting dependencies. Sumerologists may, for example, try to deduct from documents the reasons for a given dependence. Also, analysis of the whole set of rules is interesting, because it provide broad picture of Sumerian economy.

9 Conclusions

The hybridization 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 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.

References

1. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences **11**(5) (1982) 341–356
2. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
3. Vapnik, V.N.: Statistical Learning Theory. John_Wiley (1998)
4. Jaworski, W.: Generalized indiscernibility relations: Applications for missing values and analysis of structural objects. Transactions of Rough Sets **8** (2008) 116–145

5. Pawlak, Z., Skowron, A.: Rough sets: Some extensions. Inf. Sci **177**(1) (2007) 28–40
6. Skowron, A., Świniarski, R., Synak, P.: Approximation spaces and information granulation. LNCS Transactions on Rough Sets **3400**(3) (2005) 175–189
7. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association **58** (1963) 13–30
8. Jaworski, W.: Model selection and assessment for classification using validation. In Ślzak, D., Wang, G., Szczuka, M.S., Düntsch, I., Yao, Y., eds.: Proc. of Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 10th International Conference (RSFDGrC 2005). Volume 3641 of Lecture Notes in Computer Science., Springer (2005) 481–490
9. Jaworski, W.: Bounds for validation. Fundam. Inform **70**(3) (2006) 261–275
10. Tsumoto, S.: Accuracy and coverage in rough set rule induction. Lecture Notes in Computer Science **2475** (2002) 373–380
11. Guillet, F., Hamilton, H.J., eds.: Quality Measures in Data Mining. Volume 43 of Studies in Computational Intelligence. Springer (2007)
12. Gediga, G., Düntsch, I.: Rough Set Data Analysis — A Road to Non-Invasive Knowledge Discovery. Methodos Publishers, UK (2000)
13. Asuncion, A., Newman, D.J.: UCI machine learning repository. Technical report, University of California, Irvine, School of Information and Computer Sciences (2007)
14. Jaworski, W.: Contents modelling of Neo-Sumerian Ur III economic text corpus. In: Proc. of the 22nd International Conference on Computational Linguistics (COLING 2008), Manchester, UK, Coling 2008 Organizing Committee (2008) 369–376
15. Pawlak, Z.: Information systems — theoretical foundations. Information Systems **6**(3) (1981) 205–218
16. Grzymaa-Busse, J.W., Grzymaa-Busse, W.J.: An experimental comparison of three rough set approaches to missing attribute values. T. Rough Sets **6** (2007) 31–50
17. Grzymala-Busse, J.W.: A rough set approach to data with missing attribute values. In Wang, G., Peters, J.F., Skowron, A., Yao, Y., eds.: Proc. of Rough Sets and Knowledge Technology, First International Conference (RSKT 2006). Volume 4062 of Lecture Notes in Computer Science., Springer (2006) 58–67
18. Kryszkiewicz, M.: Rough set approach to incomplete information systems. Inf. Sci **112**(1-4) (1998) 39–49
19. Kryszkiewicz, M.: Properties of incomplete information systems in the framework of rough sets. In Polkowski, L., Skowron, A., eds.: Rough Sets in Knowledge Discovery 1. Methodology and Applications. Studies in Fuzziness and Soft Computing. Physica-Verlag, Heidelberg (1998) 422–450
20. Lipski, W.J.: On Databases with Incomplete Information. Journal of the Association of Computing Machinery **28**(1) (1981) 41–70
21. Bazan, J., Szczuka, M.: RSES and RSESlib - a collection of tools for rough set computations. In Ziarko, W., Yao, Y., eds.: Proc. of the Second International Conference on Rough Sets and Current Trends in Computing (RSCTC 2000). Volume 2005 of Lecture Notes in Computer Science., Heidelberg, Springer (2001) 106–113