Probabilistic Rough Sets

Wojciech Ziarko

Computer Science Department,
University of Regina,
Regina, Saskatchewan, S4S 0A2, Canada

Abstract. The article introduces the basic ideas and investigates the probabilistic version of rough set theory. It relies on both classification knowledge and probabilistic knowledge in analysis of rules and attributes. One-way and two-way inter-set dependency measures are proposed and adopted to probabilistic rule evaluation. A probabilistic dependency measure for attributes is also proposed and demonstrated to have the monotonicity property. This property makes it possible for the measure to be used to optimize and evaluate attribute based-representation through computation of attribute reduct, core and significance factors.

1 Introduction

The rough set theory introduced by Pawlak [5] is concerned with finite universes and finite set cardinality-based evaluative measures. It lays out the foundations of the inspiring idea of classification knowledge, in the form of the approximation space, and of the notion of rough set and its approximations. Typical application scenario involves a partially known universe, represented by a set of samples, based on which rough set-based analysis is performed. The results are then considered to apply to the whole universe. This kind of approach is common in probabilistic reasoning, with the probability function used to represent relations among sets (events). The probability function values can be estimated from different sources, including assumed distribution functions and set frequencies in a sample. The set frequency estimators of probability theory correspond to set cardinality-based evaluative measures of rough set theory. This correspondence was observed quite early in the development of rough set methodology, leading to a succession of probabilistic generalizations [5-9,13-15] of the original rough set theory. The rough set theory methodologies provide additional instruments, originally not present in the probability theory, to conduct deeper analysis of experimental data and to construct adaptive models of the relations existing in the universe. The probability theory, on the other hand, contributes the basic notion of probability and its estimation, distribution evaluative measures, the notion of probabilistic independence and Bayes's equations, which together help to enhance the rough set theory to make it more applicable to real-life problems.

In what follows, the probabilistic version of rough set theory is presented and investigated, partially based on prior results of related research [7][13][14][9]. In the presentation, clear distinction is being made between classification knowledge

D. Ślęzak et al. (Eds.): RSFDGrC 2005, LNAI 3641, pp. 283–293, 2005.

and probabilistic knowledge. These two kinds of knowledge are defined in section 2. The probabilistic notion of event independence is generalized in section 3, to introduce one-way and two-way measures of set dependencies. One of the measures, the absolute certainty gain, is adopted as a probabilistic rule evaluative parameter. The probabilistic rules, their evaluation and their computation are discussed in section 4. In section 5, computation of rules satisfying predefined certainty requirements is discussed. Elements of the Bayesian rough set model [7] are introduced in section 6, as a prerequisite to the investigation of probabilistic attribute dependencies in section 8. In section 9, the monotonicity of the introduced probabilistic attribute dependency measure, called λ-*dependency*, is discussed. This leads to the definition of probabilistic reduct, core and significance factors for attributes. The characterization of unrelated, or independent attributes is also provided. Due to space restrictions, the proofs of theorems are omitted.

2 Classification and Probabilistic Knowledge

The rough set approaches are developed within the context of a universe of objects of interest U such as, for example, the collection of patients, sounds, web pages etc. We will assume here that the universe is infinite in general, but that we have access to a finite sample $S \subseteq U$ expressed by accumulated observations about objects in S. The sample represents available information about the universe U. We will say that a subset $X \subseteq U$ *occurred* if $X \cap S \neq \emptyset$, where $X \cap S$ is a *set of occurrences* of X.

We will also assume the knowledge of an equivalence relation, called the *indiscernibility relation* on U[5], $IND \subseteq U \otimes U$ with finite number of equivalence classes called *elementary sets*. The pair (U, IND) is called the *approximation space*. The collection of elementary sets will be denoted by IND^*. The ability to form elementary sets reflects our *classification knowledge* about the universe U. In the context of this article, the classification knowledge means that each elementary set E is assigned a *description*, denoted as *des(E)*, which specifies a criterion distinguishing all elements of E from its complement. That is, $E = \{e \in U : des(e) = des(E)\}$. Any subset $X \subseteq U$ expressible as a union of some elementary sets is said to be *definable*. Otherwise, the set X is *undefinable*, or *rough*[5]. Any non-elementary definable set will be called a *composed set*. The classification knowledge is said to be *trivial* (and useless), if there is only one elementary set, corresponding to the whole universe U. The classification knowledge, in the framework of rough set theory, is normally used in the analysis of a *target set* $X \subseteq U$. The target set is usually undefinable. Typical objective of the rough-set analysis is to form an approximate definition of the target set in terms of some definable sets.

In the framework of the variable precision rough set model (VPRSM)[14], the classification knowledge is assumed to be supplemented with the *probabilistic knowledge*. It is assumed that all subsets $X \subseteq U$ under consideration in this article are measurable by a probabilistic measure function P with $0 < P(X) <$

1. That is, they are likely to occur but their occurrence is not certain. The probabilistic knowledge consists of three parts:

- For each equivalence class E of the relation IND, it is assumed that its probabilistic measure $P(E)$ is known;
- We assume that the conditional probability $P(X|E)$ of X, for each elementary set E, is also known;
- The *prior probability* $P(X)$ of the target set X is known.

All these probabilities can be estimated based on data in a standard way by taking ratios of cardinalities of sample data.

3 Probabilistic Dependencies Between Sets

In the presence of probabilistic knowledge, it is possible to evaluate the degree of dependencies between measurable subsets of the universe U. This is particularly of interest in context of evaluation of rules learned from data [12]. In what follows, we propose two kinds of measures to evaluate the degree of connection or dependency between any two sets. The measures can be seen as generalizations of the well-known notion of probabilistic independence of random events.

The first, *one-way dependency* measure is concerned with quantifying the degree of the one-way relation between sets, denoted as $Y \Rightarrow X$, where X and Y are arbitrary measurable subsets of U. For the one-way dependency measure, the use of function called *absolute certainty gain* (*gabs*), is proposed:

$$gabs(X|Y) = |P(X|Y) - P(X)|, \tag{1}$$

where $|*|$ denotes absolute value function. The one-way dependency represents the degree of change of the certainty of prediction of X as a result of the occurrence of the set Y. In an approximation space, if the set Y is definable then absolute certainty gain can be computed directly from the available probabilistic knowledge according to the following:

Proposition 1. *If Y is definable in the approximation space (U, IND), then the absolute certainty gain between sets X and Y is given by:*

$$gabs(X|Y) = \frac{|\sum_{E\subseteq Y} P(E)P(X|E) - P(X)\sum_{E\subseteq Y} P(E)|}{\sum_{E\subseteq Y} P(E)} \tag{2}$$

The values of the one-way dependency fall in the range $0 \leq gabs(X|Y) \leq max(P(\neg X), P(X)) < 1$. In addition, let us note that if sets X and Y are independent in probabilistic sense, that is if $P(X \cap Y) = P(X)P(Y)$ then $gabs(X|Y) = 0$. We may also note that $gabs(U|Y) = 0$ and $gabs(\phi|Y) = 0$, for any measurable subset Y such that $P(Y) > 0$.

The second, *two-way dependency* measure is concerned with measuring the degree of the two-way connection between sets, represented by $Y \Leftrightarrow X$, where X and Y are arbitrary measurable subsets. For the two-way measure, the function *dabs*, called *absolute dependency gain*, is suggested:

$$dabs(X,Y) = |P(X \cap Y) - P(X)P(Y)|. \tag{3}$$

The absolute dependency gain reflects the degree of probabilistic dependency between sets by quantifying the amount of deviation of $P(X \cap Y)$ from probabilistic independence between sets X and Y, as expressed by the product $P(X)P(Y)$. Similarly, $|P(\neg X \cap Y) - P(\neg X)P(Y)|$ is a degree of deviation of the $\neg X$ from total independence with Y. Since $P(\neg X \cap Y) - P(\neg X)P(Y) = -(P(X \cap Y) - P(X)P(Y))$, both target set X and its complement $\neg X$ are dependent in the same degree with any measurable set Y.

As in the case of one-way dependency, if the set Y is definable then the absolute dependency gain can be computed directly from the available probabilistic knowledge, according to the following:

Proposition 2. *If Y is definable in the approximation space (U, IND), then the absolute dependency gain between sets X and Y is given by:*

$$dabs(X,Y) = |\sum_{E \subseteq Y} P(E)P(X|E) - P(X) \sum_{E \subseteq Y} P(E)| \tag{4}$$

The one-way and two-way dependencies are connected by $dabs(X,Y) = P(Y)gabs(X|Y)$. It follows that the values of the two-way dependency fall in the range $0 \leq dabs(X,Y) \leq P(Y)max(P(\neg X), P(X)) < P(Y) < 1$. Also $0 \leq dabs(X,Y) \leq P(X)max(P(\neg Y), P(Y)) < P(X) < 1$ i.e. $0 \leq dabs(X,Y) < min(P(X), P(Y))$. In addition, let us note that if sets X and Y are independent in probabilistic sense, that is if $P(X \cap Y) = P(X)P(Y)$ then $dabs(X,Y) = 0$. We may also note that $dabs(U,Y) = 0$ and $dabs(\phi|Y) = 0$, for any arbitrary subset Y such that $P(Y) > 0$.

4 Probabilistic Rules

The inter-sets dependency measures introduced in previous section can be used to evaluate the quality of probabilistic rules [14][12]. In the context of probabilistic approach to rough set theory, probabilistic rules are formal linguistic expressions representing relationships between subsets of the universe U. For any definable subset Y and an arbitrary subset X of the universe U, the *probabilistic rule* is a statement $des(Y) \rightarrow s(X)$, denoted shortly by $r_{X|Y}$, where $s(X)$ is a string of characters used to refer the set X and $des(Y)$ is a description of the set Y. The set Y is referred to as *rule support set*. As opposed to the description of a set, $s(X)$ is just a *reference* to a possibly undefinable set, whose description might be unknown. Since rules of this kind are normally used to determine, or to guess, the membership of an object in the set X based on knowing that it belongs to the definable set Y, for obvious reason it does not make much sense dealing with rules in which X is definable. Consequently, we will assume that the conclusion part $s(X)$ of the rule $r_{X|Y}$ corresponds to an undefinable set X.

Traditionally, the probabilistic rules are assigned two probabilistic parameters characterizing the relation between sets X and Y:

- The rule $r_{X|Y}$ *certainty* parameter defined as the conditional probability $cert(r_{X|Y}) = P(X|Y)$;
- The rule $r_{X|Y}$ *generality* (also called *support*) parameter defined as the probability $gen(r_{X|Y}) = P(Y)$;

Certainty and generality parameters can be equivalently replaced by certainty and *strength* measures, where the strength is defined as $str(r_{X|Y}) = P(X \cap Y)$. However, rule certainty and generality, or the certainty and strength, do not completely capture the intuitive perception of rule quality. For example, a rule with high certainty $P(X|Y)$ may not be very useful if the prior probability of X is also high. On the other hand, if the prior probability of X is low, a high certainty rule will represent a significant increase in the ability to predict X. Intuitively, such a rule will be very valuable.

To properly represent the degree of *certainty increase* attributed to a probabilistic rule $r_{X|Y}$, relative to the prior probability $P(Y)$, the use of the absolute certainty gain parameter $gabs(r_{X|Y}) = gabs(X|Y)$ is proposed. The absolute certainty gain represents the degree of increase of the certainty of prediction of X, as a result of the occurrence of the set Y. As the absolute certainty gain cannot be derived from certainty and generality parameters, we propose that probabilistic rules be evaluated in terms of the following three parameters: generality (or strength), certainty and certainty gain instead of generality and certainty only.

Any elementary set $E \in IND^*$ corresponds to an *elementary rule* $des(E) \rightarrow s(X)$. The strength, certainty and the absolute certainty gain of elementary rules can be simply obtained from the available probabilistic knowledge. It was shown in the Proposition 1 that the absolute certainty gain can be computed from the probabilities associated with the elementary sets. The following Proposition 3 demonstrates that strength and certainty of any probabilistic rule $des(Y) \rightarrow s(X)$ can also be computed in similar way.

Proposition 3. *The strength, certainty and absolute certainty gain of the rule* $r = des(Y) \rightarrow s(X)$ *are respectively given by* $str(r_{X|Y}) = P(Y) = \sum_{E \subseteq Y} P(E)$ *and* $cert(r_{X|Y}) = P(X|Y) = \frac{\sum_{E \subseteq Y} P(E)P(X|E)}{\sum_{E \subseteq Y} P(E)}$.

The practical implication from the Propositions 1 and 3 is that once the basic probabilistic knowledge is estimated from data, there is no need to refer to the data set again to compute any kind of probabilistic rules and attribute dependencies.

5 Probabilistic Approximation Regions

In applications related to data mining and machine learning, a common objective is finding rules that meet predefined level of quality. We show in this section that rules computed within the context of VPRSM have the quality level in the form of the certainty gain level requirement imposed through settings of model parameters. In the VPRSM, the probabilistic knowledge represented by

the probability estimates associated with elementary sets is used to construct generalized rough approximations of the target subset $X \subseteq U$. The defining criteria are expressed here in terms of conditional probabilities and of the prior probability $P(X)$ of the target set X. Two *certainty control* criteria parameters are used to control degree of required certainty gain in the lower approximations of the set X or its complement $\neg X$.

The first parameter, referred to as the *lower limit* l, satisfying the constraint $0 \leq l < P(X) < 1$, represents the highest acceptable degree of the conditional probability $P(X|E)$ to include the elementary set E in the negative region of the set X, i.e. in the positive region of its complement $\neg X$.

The second parameter, referred to as the *upper limit* u, satisfying the constraint $0 < P(X) < u \leq 1$, defines the *positive region* of the set X. The upper limit reflects the least acceptable degree of the conditional probability $P(X|E)$ to include the elementary set E in the positive region.

The VPRSM is called *symmetric* if $l = 1 - u$ [13][14]. In this case, with the precision control parameter denoted as $\beta = u = 1 - l$, the *negative* and *positive* regions of the set X, are defined respectively by $NEG_\beta(X) = \cup\{E : P(\neg X|E) \geq \beta\}$ and $POS_\beta(X) = \cup\{E : P(X|E) \geq \beta\}$. Because $\beta > P(X)$, then both positive and negative regions can be expressed in terms of absolute certainty gain: $NEG_\beta(X) = \cup\{E : gabs(\neg X|E) \geq \beta - P(X)\}$ and $POS_\beta(X) = \cup\{E : gabs(X|E) \geq \beta - P(X)\}$. Consequently, we can define the positive region $POS(X, \neg X) = NEG(X) \cup POS(X)$ of the classification $(X, \neg X)$ by a single formula as $POS_\beta(X, \neg X) = \cup\{E : gabs(X|E) \geq \beta - P(X)\}$

Clearly, the approximation regions for the *asymmetric* VPRSM [14] can be also expressed in terms of the absolute gain function. The positive region of the classification $(X, \neg X)$ represents the area of desired absolute certainty gain, as expressed by the parameter β. Based on the positive region, probabilistic rules can be computed using any lower approximation-based techniques [8][2][15]. All these rules will satisfy the imposed minimum absolute certainty gain requirement $\beta - P(X)$.

The boundary area is a definable subset of U where the minimum certainty gain requirement is not satisfied, that is: $BND_\beta(X, \neg X) = \cup\{E : gabs(X|E) < \beta - P(X)\}$ No probabilistic rule computed from $BND(X, \neg X)$ will meet the minimum absolute certainty gain threshold of $\beta - P(X)$.

The definable area of the universe U characterized by the total lack of relationship to the target set $X \subseteq U$ was identified in [14] as the *absolute boundary* region of the set X. In the absolute boundary region, every elementary set E is probabilistically independent from the set X, i.e. $P(X \cap E) = P(X)P(E)$. The boundary area can be expressed by using of the absolute dependency gain function as the criterion: $BND^*(X, \neg X) = \cup\{E : dabs(X|E) = 0\}$.

The area of the universe characterized by at least some probabilistic connection with the target set X is called the *absolute positive region* of the classification $(X, \neg X)$. It can be expressed as $POS^*(X, \neg X) = \cup\{E : dabs(X|E) > 0\}$. Because $dabs(X|E) > 0$ is equivalent to $P(X|E) > P(X)$ or $P(X|E) < P(X)$, the *absolute positive region of the classification* $(X, \neg X)$ can be broken down into *the*

absolute positive region of the set X, $POS^*(X) = \cup\{E : P(X|E) > P(X)\}$ and *the absolute negative region of the set* X, $NEG^*(X) = \cup\{E : P(X|E) < P(X)\}$.

The absolute approximation regions form the basis of the Bayesian Rough Set Model investigated in [7]. They are also useful in the analysis of probabilistic dependencies between attributes, as demonstrated in the following sections.

6 Elementary, Composed and Binary Attributes

In many applications, the information about objects is expressed in terms of values of observations or measurements referred to as *features*. For the purpose of rough set-based analysis, the feature values are typically mapped into finite-valued numeric or symbolic domains to form composite mappings referred to as *attributes*. A common kind of mapping is dividing the range of values of a feature into a number of suitably chosen subranges via a discretisation procedure. Formally, an attribute a is a function $a : U \rightarrow a(U) \subseteq V_a$, where V_a is a finite set of values called the *domain* of the attribute a. The size of the domain of an attribute a, denoted as $com(a) = card(V_a)$, will be called a *theoretical complexity* of the attribute. The theoretical complexity reflects the maximum number of values an attribute can take. Each attribute defines a classifications of the universe U into classes corresponding to different values of the attribute. That is, each attribute value $v \in a(U)$, corresponds the set of objects $E_v^a = a^{-1}(v) = \{e \in U : a(e) = v\}$. The classes E_v^a, referred to as *a-elementary sets*, form a partition of U. The equivalence relation corresponding to this partition will be denoted as IND_a. We will divide the attributes into two categories:

- The initial, given collection of attributes A, elements of which $a \in A$ are referred to as *elementary attributes*;
- The *composed attributes*, which are formed by taking combinations of some elementary attributes.

The values of a composed attribute are combinations of values of component elementary attributes. Each composed attribute is a subset of A. For proper reference between an elementary attribute and its value, we will assume that composed attributes are ordered. For the sake of consistency, we will also treat elementary attributes a as single-element subsets of A, $\{a\} \subseteq A$, and the empty subset of A, $\{\}$ will be interpreted as a *trivial attribute*, i.e. with only one value corresponding to the whole universe U. In the context of this assumption, both elementary and composed attributes C will be perceived in two ways: as subsets $C \subseteq A$ and also as mappings $C : U \rightarrow C(U) \subseteq \otimes_{a \in C} V_a$, where $\otimes$ denotes Cartesian product operator of all domains of attributes in C, the domain of C. The theoretical complexity of a composed attribute is a product of theoretical complexities of all its elementary attribute domains, $com(C) = \prod_{a \in C} com(a)$. The theoretical complexity of a trivial attribute is one. In practical applications, the theoretical complexity estimates our ability to learn from example observations, or the *learnability* of a classification represented by an attribute. High theoretical complexity attributes lead to non-learnable classifications.

The lowest complexity, non-trivial attributes are binary-valued attributes. Every non-trivial and non-binary attribute can be replaced equivalently by a collection of binary attributes. The binary attributes are defined for each value v of the attribute a, by creating a new attribute a_v such that $a_v(e) = 1$ if $a(e) = v$ and $a_v(e) = 0$ if $a(v) \neq v$.

The composed attribute B_a consisting of the binary attributes is equivalent to the attribute a because it generates the same classification of U as the attribute a, that is, $IND_{B_a} = IND_a$. Using binary elementary attributes has a number of advantages, including the consistency of representation, ease of implementation and increased generality of minimal length rules computed by applying the idea of rough set theory value reduct [5]. Consequently, from now on in this article, we will assume that all elementary attributes are binary. The composed attributes are vectors of binary attributes. The theoretical complexity of a composed attribute containing n binary attributes can be simply calculated as 2^n. Therefore, the number of bits n can be used as an alternative complexity measure.

7 Probabilistic Dependencies Between Attributes

The presence of non-trivial classification of the universe may improve the degree of the decision certainty. We will assume in this section that the classification IND^*_C corresponds to a composed, in general, attribute $C \subseteq A$. The degree of improvement can be quantified using the expected value $egabs(X|C)$ of the absolute gain functions assigned elementary rules $r_{X|E}$, $E \in IND^*_C$:

$$egabs(X|C) = \sum_{E \in IND^*_C} P(E) gabs(r_{X|E}) \tag{5}$$

The *expected gain function* defined by (5) measures the average degree of increase of the occurrence probability of X or $\neg X$, relative to its prior probability $P(X)$, as a result of presence of the classification knowledge, as represented by equivalence classes of the indiscernibility relation IND^*_C and the associated probabilities. The notion of the expected gain function stems from the idea of the *relative gain* function reported in [14].

The expected gain function *egabs* can also be seen as the measure of the degree of probabilistic dependency between classification represented by the relation IND and the partition of the universe corresponding to the sets X and $\neg X$. This follows from the following proposition:

Proposition 4. *The expected gain function can be expressed as*

$$egabs(X|C) = \sum_{E \in IND^*_C} |P(X \cap E) - P(X)P(E)| = \sum_{E \in IND^*_C} dabs(X, E) \tag{6}$$

The measure can be also expressed in the form:

$$egabs(X|C) = P(X) \sum_{E \in IND^*_C} gabs(E|X). \tag{7}$$

For the purpose of normalization of the expected gain function, the following Proposition 5 is useful.

Proposition 5. *The expected gain falls in the range* $0 \leq egabs(X|C) \leq 0.5$.

The target set X and the attribute C are *independent* if $egabs(X|C) = 0$. The independence can occur only if $P(X \cap E) = P(X)P(E)$, for all elementary sets $E \in IND^*_C$. That is, for the independence between X, or $\neg X$, and the partition IND^*_C to hold, the set X, or $\neg X$, must be independent with each element of the partition IND^*_C. Conversely, the strongest dependency occurs when X is definable and when $P(X) = 0.5$. This would suggest the use of the λ-dependency function $0 \leq \lambda(X|C) \leq 1$, defined by:

$$\lambda(X|C) = \frac{egabs(X|C)}{2P(X)(1 - P(X))}, \tag{8}$$

as a normalized measure of dependency between attribute C and the target classification $(X, \neg X)$. The function $\lambda(X|C) = 1$ only if X is definable in the approximation space (U, IND_C), that is if the dependency is deterministic (functional). In line with our initial assumption of $0 < P(X) < 1$, $\lambda(X|C)$ is undefined for $X = \phi$ and for $X = U$.

Finally, because elementary attributes are binary, the λ-dependency function can be used to evaluate the degree of probabilistic dependency between any composed attribute $C \subseteq A$ and an elementary attribute $a \in A$. Consequently, the dependency between elementary attribute a and composed attribute C will be denoted as $\lambda(a|C)$. To be consistent with this notation, we will use symbol d to denote the *decision attribute* representing the target classification $(X, \neg X)$.

8 Optimization and Evaluation of Attributes

One of the main advantages of rough set methodology is the ability to perform reduction of features or attributes used to represent objects. The application idea of *reduct*, introduced by Pawlak [5] allows for optimization of representation of classification knowledge by providing a systematic technique for removal of redundant attributes. It turns out that the idea of reduct is also applicable to the optimization of probabilistic knowledge representation, in particular with respect to the representation of the probabilistic dependency between a composed attribute and a binary attribute. The following theorem, based on [7], demonstrates that the probabilistic dependency measure between attributes is *monotonic*, which means that expanding a composed attribute $C \subset A$ by extra bits would never result in the decrease of dependency $\lambda(d|C)$ with the decision attribute d corresponding to the partition $(X, \neg X)$ of the universe U.

Theorem 1. *λ-dependency is monotonic, that is, for any composed attribute $C \subset A$ and an elementary attribute $a \in A$ the relation $\lambda(d|C) \leq \lambda(d|C \cup \{a\})$ holds.*

As a consequence of the Theorem 1, the notion of the *probabilistic reduct* of attributes $RED \subseteq C$ can be defined as a minimal subset of attributes preserving the dependency with the decision attribute d. That is, the reduct satisfies the following two properties:

- $\lambda(d|RED) = \lambda(d|C)$;
- for any attribute $a \in RED$: $\lambda(d|RED - \{a\}) < \lambda(d|RED)$.

The probabilistic reducts can be computed using any methods available for reduct computation in the framework of the original rough set approach. The reduct provides a method for computing fundamental factors in a probabilistic relationship.

An important question is to characterize attributes that are *neutral* with respect to the relation between attribute C and d. Such attributes will have no effect on dependency with the decision attribute and will be always eliminated from any reduct. The following Theorem 2 provides the answer to this question.

Theorem 2. *If an attribute a is independent with $C \cup \{d\}$ i.e. if $\lambda(a|C \cup \{d\}) = 0$, then $\lambda(d|C \cup \{a\}) = \lambda(d|C)$.*

The above theorem suggests that for a new attribute to possibly contribute to the increase of dependency $\lambda(C|d)$, it should be correlated either with d or C. We also note that the independence of the attribute a with $C \cup \{d\}$ is a two-way property, that is, $\lambda(C \cup \{d\}|a) = 0$ if and only if $\lambda(a|C \cup \{d\}) = 0$.

Elementary and composed attributes appearing in a reduct can be evaluated with respect to their contribution to the dependency with the target attribute by adopting the notion of a *significance factor*. The significance factor $sig_{RED}(B)$ of an attribute $B \subseteq A$ represents the relative decrease of the dependency $\lambda(d|RED)$ due to removal of B from the reduct:

$$sig_{RED}(B) = \frac{\lambda(d|RED) - \lambda(d|RED - B)}{\lambda(d|RED)} \tag{9}$$

Finally, as in the original rough set approach, one can define the *core* set of elementary attributes as the ones which form the intersection of all reducts of C, if the intersection is not empty. After [5], any core attribute a satisfies the inequality $\lambda(d|C) > \lambda(d|C - \{a\})$, which leads to a simple method of core computation.

9 Conclusion

The article is an attempt to introduce a comprehensive probabilistic version of rough set theory by integrating ideas from Pawlak's classical rough set model, elements of probability theory with its notion of probabilistic independence, the variable precision model of rough sets and the Bayesian model. The novel aspects of the approach include the introduction of measures of inter-set dependencies, based on the notion of absolute certainty gain and probabilistic dependence, the

adaptation of the absolute certainty gain to probabilistic rule evaluation, the introduction of the notion of a composed attribute and of the attribute dependency measure based on the idea of expected gain function and its application to attribute optimization and evaluation. The presented ideas seem to connect well with the general methodology of rough sets, hopefully leading to new applications and better understanding of fundamental issues of learning from data.

References

1. Beynon, M. The elucidation of an iterative procedure to β-reduct selection in the variable precision rough set model. Proc. RSCTC'2004, LNAI 1711, 412-417.
2. Grzymala-Busse, J. LERS-A System for learning from examples based on rough sets. Intelligent Decision Support, Kluwer, 1991, 3-18.
3. Greco, S. Matarazzo, B. Slowinski, R. Stefanowski, J. Variable consistency model of dominance-based rough set approach. Proc. RSCTC'2000, LNAI 2005, 170-179.
4. Murai, T. Sanada, M. Kudo, M. A note on Ziarko's variable precision rough set model in non-monotonic reasoning. Proc. RSCTC'2004, LNAI 1711, 103-108.
5. Pawlak, Z. Rough sets - Theoretical Aspects of Reasoning About Data. Kluwer, 1991.
6. Mieszkowicz, A. Rolka, L. Remarks on approximation quality in variable precision rough set model. Proc. RSCTC'2004, LNAI 1711, 402-411.
7. Slezak, D., Ziarko, W. Investigation of the Bayesian rough set model. Intl. Journal of Approximate Reasoning, vol. 40(1-2), 2005, 81-91.
8. Skowron, A. Rauszer C. The discernibility matrices and functions in information systems. ICS Report 1/91, Warsaw University of Technology, 1991.
9. Wong, M. Ziarko, W. Comparison of the probabilistic approximate classification and the fuzzy set model. Intl. Journal for Fuzzy Sets and Systems, vol. 21, 1986, 357-362.
10. Yao, Y., Wong, M. A decision theoretic framework for approximating concepts. Intl. Journal of Man-Machine Studies, 37, 1992, 793-809.
11. Yao, Y. Probabilistic approaches to rough sets. Expert Systems, vol. 20(5), 2003, 287-291.
12. Yao, Y. Zhong, N. An analysis of quantitative measures associated with rules. Proc. PAKDD'99, LNAI 1574, 479-488.
13. Ziarko, W. Variable precision rough sets model. Journal of Computer and Systems Sciences, vol. 46(1), 1993, 39-59.
14. Ziarko, W. Set approximation quality measures in the variable precision rough set model. Soft Computing Systems, Management and Applications, IOS Press, 2001, 442-452.
15. Ziarko, W. Shan, N. A method for computing all maximally general rules in attribute-value systems. Computational Intelligence, vol. 12(2), 1996, 223-234.