Impurity Measures in Databases

Dan A. Simovici Dana Cristofor
Laurentiu Cristofor
University of Massachusetts at Boston
Department of Computer Science
Boston, Massachusetts 02125, USA
e-mail: {dsim,dana,laur}cs.umb.edu

November 16, 2001

Abstract

We introduce purity dependencies as generalizations of functional dependencies in relational databases starting from the notion of impurity measure. The impurity measure of a subset of a set relative to a partition of that set and the relative impurity of two partitions allow us to define the relative impurity of two attribute sets of a table of a relational database and to introduce purity dependencies. We discuss properties of these dependencies that generalize similar properties of functional dependencies and we highlight their relevance for approximate classifications. Finally, an algorithm that mines datasets for these dependencies is presented.

1 Introduction

Functional dependencies play an important role in the design of databases, due to their role in the normalization theory that aims to minimize redundancy and anomalies in relational databases. The identification of these dependencies satisfied by database schemas is an important topic in data mining literature (see [KM95, HKPT97]).

We propose a generalization of the notion of functional dependency starting from the notion of impurity of a subset of a set S relative to a partition of S; this notion is extended to the notion of relative impurity of two partitions. Since sets of attributes of a table of a relational database naturally generate partitions on the set of tuples (as we show in Section 3) it becomes possible to define the relative impurity of two sets of attributes. When this impurity is below a certain limit we say that the table satisfies a *purity dependency*. Purity dependencies have properties that are similar to those of functional dependencies. For example, in the presence of certain purity dependencies it is possible to limit the number of spurious tuples that occur in lossy decompositions of tables (cf. Theorem 3.5). As we show in the final section of this paper, they can be

useful in approximative classifications, that is, in classifications where certain errors are tolerable.

Unless stated otherwise, all sets considered below are finite. We begin with a few notations and definitions of terms. The set

$$\{(p_1, \ldots, p_k) \in \mathbb{R}^k \mid p_i \geq 0 \text{ and } p_1 + \cdots + p_k = 1\}.$$

will be denoted by $\mathsf{SIMPLEX}_{k-1}$ and will be referred to as the k-dimensional simplex.
A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is:

- *concave* on a set $S \subseteq \mathbb{R}$ if $f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y)$ for $\alpha \in [0,1]$ and $x, y \in S$;
- *sub-additive* on S if $f(x+y) \leq f(x) + f(y)$ for $x, y \in S$.

For example, $x - x^2$ is concave on the set $\mathbb{R}$ and $-\log x$ is sub-additive on the set $[0,1]$.

In [CHH99] a concave impurity measure is defined as a real-valued function $i : \mathsf{SIMPLEX}_{k-1} \longrightarrow \mathbb{R}$ that satisfies the following conditions:

(i) $i(\alpha\mathbf{p} + (1-\alpha)\mathbf{q}) \geq \alpha i(\mathbf{p}) + (1-\alpha)i(\mathbf{q})$ for any $\alpha \in [0,1]$ and $\mathbf{p}, \mathbf{q} \in \mathsf{SIMPLEX}_{k-1}$, with equality if and only if $\mathbf{p} = \mathbf{q}$;

(ii) if $\mathbf{p} = (p_1, \ldots, p_k)$, then $i(\mathbf{p}) = 0$ if $p_i = 1$ for some i, $1 \leq i \leq k$.

The corresponding *frequency-weighted impurity measure* is the real-valued function $I : \mathbb{N}^k \longrightarrow \mathbb{R}$ given by

$$I(n_1, \ldots, n_k) = Ni\left(\frac{n_1}{N}, \ldots, \frac{n_k}{N}\right),$$

where $N = \sum_{i=1}^{k} n_i$.

In [CHH99] it is noted that both the Gini impurity measure and the entropy can be generated using a simple one-argument function that satisfies certain conditions. In this paper we additionally require subadditivity of this function, as shown in the next definition.

Definition 1.1 A function $f : [0,1] \longrightarrow \mathbb{R}$ is a *generator* if it is concave, sub-additive, and $f(0) = f(1) = 0$.

The monogenic impurity measure induced by the generator f is the impurity measure generated by the concave impurity measure i having the form $i(p_1, \ldots, p_k) = f(p_1) + \cdots + f(p_k)$, where $(p_1, \ldots, p_k) \in \mathsf{SIMPLEX}_{k-1}$. □

It is easy to verify that such functions as $f_{\text{gini}}(p) = p - p^2$, $f_{\text{ent}}(p) = -p \log p$, $f_{\text{sq}}(p) = \sqrt{p} - p$, or

$$f_{\text{peak}}(p) = \begin{cases} p & \text{if } 0 \leq p \leq 0.5 \\ 1-p & \text{if } 0.5 < p \leq 1 \end{cases}$$

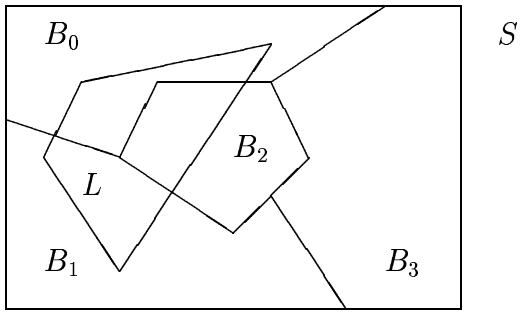

Figure 1: A 4-block partition of S and an impure subset L

are generators. (We assume that $0 \cdot \infty = 0$ in the definition of f_{ent}.) Thus, both the Gini impurity measure, induced by f_{gini}, and the entropy measure, induced by f_{ent}, are monogenic impurity measures.

Our definition of the entropy measure is an alternative approach to the axiomatic definitions of entropy, a subject much discussed in information theory (see, for Example [Khi56, Khi57, IU62, GT66, MR75]), and has the advantage of opening the possibility of some useful generalizations. For an axiomatic presentation of partition entropies see [SJ01].

Further examples of generators are $f_{\sin}(p) = \sin \pi p$, $f_{\text{circle}} = \sqrt{p - p^2}$, $f_0(p) = 1 - e^{p^2 - p}$, where $p \in [0, 1]$, or

$$f(p) = \begin{cases} 1 & \text{if } 0 < p < 1 \\ 0 & \text{if } p = 0 \text{ or } p = 1. \end{cases}$$

For a concave function f, the inequality

$$f(p_1) + \cdots + f(p_k) \leq kf\left(\frac{1}{k}\right) \tag{1}$$

holds for every $(p_1, \ldots, p_k) \in \mathsf{SIMPLEX}_{k-1}$, and is known as *Jensen's inequality*. This implies that the largest value of the sum $f(p_1) + \cdots + f(p_k)$ is achieved if and only if $p_1 = \cdots = p_k = \frac{1}{k}$. Therefore, for the monogenic impurity measure generated by the function f we have $0 \leq i(p_1, \ldots, p_k) \leq kf(\frac{1}{k})$ for $(p_1, \ldots, p_k) \in \mathsf{SIMPLEX}_{k-1}$.

2 Impurity of Sets and Partitions

In this section we introduce the notion of impurity of a subset of a set S relative to a partition. In turn, this notion is used to define the impurity of a partition relative to another partition.

Definition 2.1 Let f be a generator, S be a set and let $\mathsf{PART}(S)$ be the set of all partitions of S. The *impurity of a subset L of S relative to a partition $\pi \in \mathsf{PART}(S)$ and generated by f* is the monogenic impurity measure induced by f:

$$\mathsf{IMP}_\pi^f(L) = |L| \left(f\left(\frac{|L \cap B_1|}{|L|}\right) + \cdots + f\left(\frac{|L \cap B_n|}{|L|}\right)\right), \tag{2}$$

where $\pi = \{B_1, \ldots, B_n\}$.

The *specific impurity of L relative to π and generated by f* is

$$\mathsf{imp}_\pi^f(L) = \frac{\mathsf{IMP}_\pi^f(L)}{|L|} = f\left(\frac{|L \cap B_1|}{|L|}\right) + \cdots + f\left(\frac{|L \cap B_n|}{|L|}\right). \tag{3}$$

When the subset L is included in one of the blocks of partition π, $\mathsf{IMP}_\pi^f(L) = 0$ and L is called a *π-pure* set; otherwise, L is called π-impure. □

Note that if L is π-impure, then $|L| \geq 2$.

In Figure 1, we show an impure set L that intersects three of the four blocks of a partition of a set S.

Note that Jensen's inequality (1) implies that the impurity of a set $L \subseteq S$ relative to a partition $\pi = \{B_1, \ldots, B_n\}$ of S as defined in (2) cannot exceed the value

$$m_{L,\pi}^f = |L| \cdot n \cdot f(\frac{1}{n}). \tag{4}$$

The next theorems present some important properties of the impurity measures.

Theorem 2.2 *Let K, L be two disjoint subsets of the set S and let $\pi \in \mathsf{PART}(S)$. Then, we have*

$$\mathsf{IMP}_\pi^f(K \cup L) \geq \mathsf{IMP}_\pi^f(K) + \mathsf{IMP}_\pi^f(L).$$

Proof. The definition of imp_π^f allows us to write

$$\begin{aligned} \mathsf{imp}_\pi^f(K \cup L) &= \sum_{\ell=1}^{n} f\left(\frac{|(K \cup L) \cap B_\ell|}{|K \cup L|}\right) \\ &= \sum_{\ell=1}^{n} f\left(\frac{|K \cap B_\ell| + |L \cap B_\ell|}{|K \cup L|}\right), \end{aligned}$$

because K and L are disjoint.

Since $\frac{|K \cap B_\ell| + |L \cap B_\ell|}{|K \cup L|}$ is a convex combination of $\frac{|K \cap B_\ell|}{|K|}$ and $\frac{|L \cap B_\ell|}{|L|}$, the concavity of f allows us to write

$$f\left(\frac{|K \cap B_\ell| + |L \cap B_\ell|}{|K \cup L|}\right) \geq \frac{|K|}{|K| + |L|} f\left(\frac{|K \cap B_\ell|}{|K|}\right) + \frac{|L|}{|K| + |L|} f\left(\frac{|L \cap B_\ell|}{|L|}\right),$$

so

$$\mathsf{imp}_{\pi}^{f}(K \cup L) \geq \frac{|K|}{|K|+|L|}\mathsf{imp}_{\pi}^{f}(K) + \frac{|L|}{|K|+|L|}\mathsf{imp}_{\pi}^{f}(L),$$

which gives immediately the desired inequality. ∎

The following corollary shows that the impurity of a set increases with the size of the set.

Corollary 2.3 *If K, L are subsets of S such that $K \subseteq L$ and $\pi \in \mathsf{PART}(S)$, then $\mathsf{IMP}_{\pi}^{f}(K) \leq \mathsf{IMP}_{\pi}^{f}(L)$.*

Proof. Let $H = L - K$. By Theorem 2.2, since K, H are disjoint, we have:

$$\mathsf{IMP}_{\pi}^{f}(L) = \mathsf{IMP}_{\pi}^{f}(K \cup H) \geq \mathsf{IMP}_{\pi}^{f}(K) + \mathsf{IMP}_{\pi}^{f}(H),$$

so $\mathsf{IMP}_{\pi}^{f}(L) \geq \mathsf{IMP}_{\pi}^{f}(K)$. ∎

Let π and σ be two partitions in $\mathsf{PART}(S)$. We write $\pi \leq \sigma$ if each block of the partition π is included in a block of the partition σ; in this case we say that *π is a finer partition than σ*. The following theorem shows that the impurity of a set increases if the partition with respect to which the impurity is computed is finer.

Theorem 2.4 *Let $\pi = \{B_1, \ldots, B_n\}$ and $\sigma = \{C_1, \ldots, C_m\}$ be two partitions of a set S. If $\pi \leq \sigma$, then $\mathsf{IMP}_{\sigma}^{f}(K) \leq \mathsf{IMP}_{\pi}^{f}(K)$ for every subset K of S.*

Proof. Since $\pi \leq \sigma$ every block C_j of σ is the union of some blocks of the partition π, $C_j = \bigcup\{B_h \mid h \in H_j\}$, where $H_1, \ldots, H_m$ is a partition of the set $\{1, \ldots, n\}$. Therefore,

$$\mathsf{imp}_{\sigma}^{f}(K) = \sum_{j=1}^{m} f\left(\frac{|K \cap C_j|}{|K|}\right) \leq \sum_{j=1}^{m} \sum_{h \in H_j} f\left(\frac{|K \cap B_h|}{|K|}\right) = \mathsf{imp}_{\pi}^{f}(K),$$

due to the subadditivity of f. This implies immediately the inequality of the theorem. ∎

Lemma 2.5 *Let $\pi = \{B_1, \ldots, B_n\}$, $\zeta = \{D_1, \ldots, D_p\}$ be two partitions of a set S. If K is a subset of S such that $K \subseteq D_k$ for some block $D_k \in \zeta$, then $\mathsf{IMP}_{\pi \wedge \zeta}^{f}(K) = \mathsf{IMP}_{\pi}^{f}(K)$.*

Proof. Since the blocks of the partition $\pi \cap \zeta$ have the form $B_h \cap D_j$ we have

$$\mathsf{imp}_{\pi \wedge \zeta}^{f}(K) = \sum_{h=1}^{n} \sum_{j=1}^{p} f\left(\frac{|K \cap B_h \cap D_j|}{|K|}\right).$$

Each sum $\sum_{j=1}^{p} f\left(\frac{|K \cap B_h \cap D_j|}{|K|}\right)$ contains at most one non-null term, namely

$$f\left(\frac{|K \cap B_h \cap D_k|}{|K|}\right) = f\left(\frac{|K \cap B_h|}{|K|}\right),$$

if $K \cap B_h \neq \emptyset$, which implies the desired equality. ∎

Definition 2.6 Let S be a finite set and let $\pi, \sigma \in \mathsf{PART}(S)$ be two partitions, where $\pi = \{B_1, \ldots, B_n\}$ and $\sigma = \{C_1, \ldots, C_m\}$. The impurity of σ with respect to π generated by f is given by

$$\mathsf{IMP}_{\pi}^{f}(\sigma) = \max_{C \in \sigma} \mathsf{IMP}_{\pi}^{f}(C). \tag{5}$$

A partition σ is *α-impure with respect to partition π* if $\mathsf{IMP}_{\pi}^{f}(\sigma) \leq \alpha$. □

There are other ways of defining a measure of impurity between partitions, for example we could have used the following formula:

$$\mathsf{IMP}_{\pi}^{f}(\sigma) = \sum_{j=1}^{m} \frac{|C_j|}{|S|} \mathsf{IMP}_{\pi}^{f}(C_j) \tag{6}$$

which takes into consideration the impurities of all the blocks of σ with respect to π. Both definitions lead to measures that have similar properties. We prefer to use (5) rather than (6) since, as we will show in Section 3, its values are easier to interpret.

Informally, the impurity of a partition σ with respect to a partition π gives us a measure of how well do the blocks of σ fit inside the blocks of π. When $\sigma \leq \pi$ this impurity will be 0 as proved by the following Corollary 2.8.

Theorem 2.7 *Let S be a finite set and let $\pi, \sigma, \zeta \in \mathsf{PART}(S)$ be partitions of S. We have:*

(i) *if $\sigma \leq \pi$, then $\mathsf{IMP}_{\pi}^{f}(\zeta) \leq \mathsf{IMP}_{\sigma}^{f}(\zeta)$ and $\mathsf{IMP}_{\zeta}^{f}(\sigma) \leq \mathsf{IMP}_{\zeta}^{f}(\pi)$;*

(ii) $\mathsf{IMP}_{\pi \wedge \zeta}^{f}(\sigma \wedge \zeta) \leq \mathsf{IMP}_{\pi}^{f}(\sigma)$.

Proof. The first part of the Theorem is an immediate consequence of Theorem 2.4.

To prove the second part assume that D is a block of ζ, $\pi = \{B_1, \ldots, B_n\}$, and $\sigma = \{C_1, \ldots, C_m\}$. By Lemma 2.5 we have $\mathsf{IMP}_{\pi \wedge \zeta}^{f}(C_j \cap D) = \mathsf{IMP}_{\pi}^{f}(C_j \cap D)$, so $\mathsf{IMP}_{\pi \wedge \zeta}^{f}(C_j \cap D) \leq \mathsf{IMP}_{\pi}^{f}(C_j)$ due to Corollary 2.3. Thus, $\mathsf{IMP}_{\pi \wedge \zeta}^{f}(C_j \cap D) \leq \mathsf{IMP}_{\pi}^{f}(\sigma)$ for every block $C_j \cap D$ of $\sigma \wedge \zeta$ and this implies the second inequality. ■

Corollary 2.8 *Let σ, π be two partitions of a set S. We have $\sigma \leq \pi$ if and only if $\mathsf{IMP}_{\pi}^{f}(\sigma) = 0$.*

Proof. It is easy to see that $\mathsf{IMP}_{\sigma}^{f}(\sigma) = 0$. Therefore, by the first part of Theorem 2.7, we have $\mathsf{IMP}_{\pi}^{f}(\sigma) = 0$.

Conversely, if $\mathsf{IMP}_{\pi}^{f}(\sigma) = 0$, then $\mathsf{IMP}_{\pi}^{f}(D) = 0$ for every block D of σ. This implies $f\left(\frac{|B \cap D|}{|D|}\right) = 0$ for every block B of π, so we have either $B \cap D = \emptyset$, or $B \cap D = D$ (that is, $D \subseteq B$). This implies $\sigma \leq \pi$. ■

3 Purity Dependencies

In this section we introduce the notion of purity dependencies which are a generalization of functional dependencies based on the notion of impurity measure between two sets. We begin by introducing the notation that we will use. Let $\tau = (T, H, \rho)$ be a table, where T is the name of the table, $H = A_1 \cdots A_n$ is the heading of the table, and $\rho \subseteq \mathsf{Dom}(A_1) \times \cdots \times \mathsf{Dom}(A_n)$. Here $\mathsf{Dom}(A_i)$ is the domain of the attribute A_i for $1 \leq i \leq n$. The projection of a tuple $t \in \rho$ on a set of attributes X will be denoted by $t[X]$.

The notion of the active domain of an attribute of a table is extended to sets of attributes as follows. The *active domain* of the set of attributes X of the table τ is the set of all values that appear under X in τ, that is, $\mathsf{aDom}_\tau(X) = \{t[X] \mid t \in \rho\}$. For relational terminology and notations see, for example [Mai83, ST95].

For $X \subseteq H$ we define the equivalence $\equiv_X$ on the set of tuples ρ by $u \equiv_X v$ if $u[X] = v[X]$; the corresponding partition of the set of tuples ρ is denoted by π_X. Clearly, π_X is the partition of the tuples of ρ that would be obtained using a `group by` X clause in SQL.

Note that if U, V are two subsets of H such that $U \subseteq V$, then $\pi_V \leq \pi_U$.

Example 3.1 Our running example is the mushroom database from University of California - Irvine (see [BM98]). This dataset describes 23 attributes of 8124 different types of North American mushrooms. We adopted this dataset for several reasons: it is well-known and well-documented, its attributes are easy to understand, it has a large enough number of tuples, and it also has more nominal attributes than other UCI datasets. Thus, there was a good chance of finding interesting patterns embedded in the data.

The *class* attribute specifies whether a mushroom is edible or poisonous, so $|\mathsf{aDom}(class)| = 2$. Similarly, an attribute like *odor* has 7 values: almond, anise, creosote, fishy, foul, musty, none, pungent, and spicy, so the corresponding partition π_{odor} has seven blocks. □

The specific impurity of the set of tuples ρ relative to a partition π_X represents the entropy of the attribute set X as defined in [SJ00].

It is easy to see that a table satisfies a functional dependency $X \to Y$ if and only if $\pi_X \leq \pi_Y$, or equivalently, if and only if $\mathsf{IMP}^f_{\pi_Y}(\pi_X) = 0$, according to Corollary 2.8. This amounts to requiring that every block B of the partition π_X is a π_Y-pure set. Thus, functional dependencies could be generalized by imposing an upper bound α on the impurity of the blocks of the partition π_X relative to the partition π_Y. This suggests the following definition:

Definition 3.2 Let $\tau = (T, H, \rho)$ be a table and let $X, Y \subseteq H$ be two sets of attributes. τ satisfies the purity dependency $X \xrightarrow{f,\alpha} Y$ if $\mathsf{IMP}^f_{\pi_Y}(\pi_X) \leq \alpha$. □

In other words, the table τ satisfies the purity dependency $X \xrightarrow{f,\alpha} Y$ the largest f-impurity of a block B of π_X relative to the partition π_Y does not exceed α.

Example 3.3 We are interested in finding purity dependencies of the form $X \xrightarrow{f,\alpha} \textit{class}$ in the mushroom database mentioned in Example 3.1; in other words, we were interested in finding sets of attributes X such that $\mathsf{IMP}^{f}_{\pi_{class}}(\pi_X) \leq \alpha$. Thus, by examining the values of the attributes X we could predict with a certain error margin, the value of the attribute *class*. Of course, given the nature of this dataset, there would be little use in predicting whether a mushroom is edible or poisonous with less than 100% accuracy. However, as we will show in Section 4, even though we searched for purity dependencies with an α value different than 0, some of the facts that we discovered presented 100% accuracy.

In a rule of the form $X \xrightarrow{f,\alpha} A$ with $|\mathsf{aDom}_\tau(A)| = 2$ the properties of the blocks of π_X depend on the choice of the function f.

For example if the mushroom dataset contained a purity dependency of the form $\textit{odor} \xrightarrow{f_{\text{gini}},\alpha} \textit{class}$, then for every set of mushrooms M having the same *odor* attribute value, if $M_{ed} \subseteq M$ represents the set of edible mushrooms and $M_{po} \subseteq M$ represents the set of poisonous mushrooms, we have:

$$\frac{2|M_{ed}| \cdot |M_{po}|}{|M_{ed}| + |M_{po}|} \leq \alpha.$$

In other words, the harmonic average of the sizes of M_{ed} and M_{po} does not exceed α. Thus, one of the sets has fewer than α elements and the other has at least $|U| - \alpha$ elements.

If the mushroom database satisfies the purity dependency $\textit{odor} \xrightarrow{f_{\text{peak}},\alpha} \textit{class}$, then at least one of the sets M_{ed}, M_{po} will have size less than $\alpha/2$, so the other set will have at least $|U| - \frac{\alpha}{2}$ elements. ∎

If $\tau = (T, H, \rho)$ is a table and U, V are subsets of the heading H of τ such that $U \cup V = H$, then we have $\rho \subseteq \rho[U] \bowtie \rho[V]$. The tuples in $(\rho[U] \bowtie \rho[V]) - \rho$ are referred to as (U,V)-*spurious tuples* or simply as *spurious tuples* when the pair (U,V) is clear from context. The number of spurious tuples of the decomposition (U,V) of the table τ is denoted by n_{UV}.

The (U,V)-*pure* part of the relation ρ is the set of tuples

$$\rho^{U,V}_{pure} = \bigcup\{B \in \pi_{U\cap V} | B \text{ is either } \pi_U\text{-pure or } \pi_V\text{-pure}\}.$$

Correspondingly, the (U,V)-*impure* part of ρ is $\rho^{U,V}_{imp} = \rho - \rho^{U,V}_{pure}$. A table $\tau = (T, H, \rho)$ is (U,V)-*impure* if its content ρ coincides with its (U,V)-impure part.

Splitting the content of the table $\tau = (T, H, \rho)$ in its (U,V)-pure part $\rho^{U,V}_{pure}$ and its impure part $\rho^{U,V}_{imp}$ is significant for the decomposition (U,V) because only the impure part generates spurious tuples as we show in the next example.

Example 3.4 Let $\tau = (T, ABC, \rho)$ be the table:

T

	A	B	C
t_1	a_1	b_1	c_1
t_2	a_2	b_1	c_2
t_3	a_1	b_2	c_2
t_4	a_2	b_2	c_1
t_5	a_1	b_3	c_3
t_6	a_1	b_3	c_4

and let $U = AB$ and $V = BC$. The partition $\pi_{U \cap V} = \pi_B$ has the blocks $B_1 = \{t_1, t_2\}$, $B_2 = \{t_3, t_4\}$, and $B_3 = \{t_5, t_6\}$. Note that B_1 and B_2 are neither π_{AB}-pure, nor π_{BC}-pure; however, B_3 is π_{AB}-pure. Thus, $\rho_{pure}^{AB,BC} = \{t_5, t_6\}$, while $\rho_{imp}^{AB,BC} = \{t_1, t_2, t_3, t_4\}$.

The relation $\rho[AB] \bowtie \rho[BC]$ is given by

$T[AB] \bowtie T[BC]$

	A	B	C	
$t_1[AB] \bowtie t_1[BC]$	a_1	b_1	c_1	
$t_1[AB] \bowtie t_2[BC]$	a_1	b_1	c_2	$\surd$
$t_2[AB] \bowtie t_1[BC]$	a_2	b_1	c_1	$\surd$
$t_2[AB] \bowtie t_2[BC]$	a_2	b_1	c_2	
$t_3[AB] \bowtie t_3[BC]$	a_1	b_2	c_2	
$t_3[AB] \bowtie t_4[BC]$	a_1	b_2	c_1	$\surd$
$t_4[AB] \bowtie t_3[BC]$	a_2	b_2	c_2	$\surd$
$t_4[AB] \bowtie t_4[BC]$	a_2	b_2	c_1	
$t_5[AB] \bowtie t_5[BC]$	a_1	b_3	c_3	
$t_6[AB] \bowtie t_6[BC]$	a_1	b_3	c_4	

The spurious tuples are marked by $\surd$; note that they all result by joining tuples from the impure part of ρ. □

Purity dependencies induce limits on the number of spurious tuples that can be generated by a decomposition of a table, as shown in the next theorem.

Next, we extend a well-known property of functional dependencies.

Theorem 3.5 *Let $\tau = (T, H, \rho)$ be a table and let U, V be subsets of the heading H of τ such that $U \cup V = H$ and $U \cap V \neq \emptyset$. If τ is (U, V)-impure and it satisfies both $U \cap V \xrightarrow{f,\alpha} U$ and $U \cap V \xrightarrow{f,\beta} V$, where*

$$f(p) = \begin{cases} 1 & \text{if } 0 < p < 1 \\ 0 & \text{if } p = 0 \text{ or } p = 1, \end{cases}$$

then $n_{UV} \leq \frac{\alpha\beta}{4} |\mathsf{aDom}(U \cap V)| - |\rho|$.

Proof. Let B be a block of the partition $\pi_{U\cap V}$. Since if B is not a π_U-pure set, we have $|B| \geq 2$ and $\mathsf{IMP}^f_{\pi_U}(B) = |B|r_B$, where r_B is the number of blocks C of the partition π_U such that $B \cap C \neq \emptyset$; similarly, let q_B be the number of blocks D of π_V such that $B \cap D \neq \emptyset$.

The number of spurious tuples n_{UV} is given by:

$$n_{UV} = \sum \{r_B q_B \mid B \in \pi_{U\cap V}\} - |\rho|$$

Since $\mathsf{IMP}^f_{\pi_U}(\pi_{U\cap V}) \leq \alpha$ and $\mathsf{IMP}^f_{\pi_V}(\pi_{U\cap V}) \leq \beta$, if B is a block of $\pi_{U\cap V}$, then $r_B \leq \frac{\alpha}{|B|}$ and $q_B \leq \frac{\beta}{|B|}$, so

$$\begin{aligned} n_{UV} &\leq \sum \left\{ \frac{\alpha\beta}{|B|^2} \mid B \in \pi_{U\cap V} \right\} - |\rho| \\ &\leq \alpha\beta \sum \left\{ \frac{1}{|B|^2} \mid B \in \pi_{U\cap V} \right\} - |\rho| \\ &\leq \alpha\beta \frac{|\pi_{U\cap V}|}{4} - |\rho| \\ &\leq \frac{\alpha\beta}{4} |\mathsf{aDom}(U \cap V)| - |\rho|, \end{aligned}$$

which gives the desired inequality. We have used here the fact that all blocks of $\pi_{U\cap V}$ are π_U-impure and π_V-impure, which implies $|B| \geq 2$. ∎

Example 3.6 Let $\tau' = (T', ABC, \rho')$ be the table:

T'

	A	B	C
t_1	a_1	b_1	c_1
t_2	a_2	b_1	c_2
t_3	a_1	b_2	c_2
t_4	a_2	b_2	c_1

Clearly, τ' is an impure table since it consists of the "impure part" of the table τ introduced in Example 3.4. Note that τ' satisfies the impurity dependencies $B \xrightarrow{f,4} AB$ and $B \xrightarrow{f,4} AC$, where f is the function defined in Theorem 3.5 and each block of π_B consists of two rows and intersects two blocks of π_{AB} and of π_{BC}. Thus, the number of spurious tuples is no more than $4 \cdot 4 \cdot \frac{2}{4} - 4 = 4$, and, indeed, it is easy to see that the decomposition (U, V) yields exactly four spurious tuples. □

Next, we generalize the Armstrong inference rules for functional dependencies (see [Mai83, ST95]) to purity dependencies.

Theorem 3.7 *Let $\tau = (T, H, \rho)$ be a table and let X, Y, Z be subsets of H. The following statements hold:*

1. *if τ satisfies $X \to Y$ and $Y \xrightarrow{f,\alpha} Z$, then τ satisfies $X \xrightarrow{f,\alpha} Z$ (left transitivity property);*

2. *if τ satisfies $X \xrightarrow{f,\alpha} Y$ and $Y \to Z$, then τ satisfies $X \xrightarrow{f,\alpha} Z$ (right transitivity property);*

3. *if τ satisfies $X \xrightarrow{f,\alpha} Y$, then $X \cup Z \xrightarrow{f,\alpha} Y \cup Z$ (the augmentation property).*

Proof. Suppose that τ satisfies $X \to Y$ and $Y \xrightarrow{f,\alpha} Z$. Then, we have $\pi_X \leq \pi_Y$ and $\mathsf{IMP}^f_{\pi_Z}(\pi_Y) \leq \alpha$. By Theorem 2.7, we have $\mathsf{IMP}^f_{\pi_Z}(\pi_X) \leq \mathsf{IMP}^f_{\pi_Z}(\pi_Y)$, so $\mathsf{IMP}^f_{\pi_Z}(\pi_X) \leq \alpha$, which justifies the first part of the theorem.

Now, suppose that τ satisfies $X \xrightarrow{f,\alpha} Y$ and $Y \to Z$, so $\mathsf{IMP}^f_{\pi_Y}(\pi_X) \leq \alpha$ and $\pi_Y \leq \pi_Z$. Again, by Theorem 2.7, we have $\mathsf{IMP}^f_{\pi_Z}(\pi_X) \leq \mathsf{IMP}^f_{\pi_Y}(\pi_X)$, which implies that τ satisfies $X \xrightarrow{f,\alpha} Z$.

To prove the augmentation property observe that $\pi_{U \cup V} = \pi_U \wedge \pi_V$ for all sets of attributes U, V of τ. The second part of Theorem 2.7 implies that

$$\mathsf{IMP}^f_{\pi_{Y \cup Z}}(\pi_{X \cup Z}) = \mathsf{IMP}^f_{\pi_Y \wedge \pi_Z}(\pi_X \wedge \pi_Z) \leq \mathsf{IMP}^f_{\pi_Y}(\pi_X) \leq \alpha,$$

which means that τ satisfies the purity dependency $X \cup Z \xrightarrow{f,\alpha} Y \cup Z$. ∎

Theorem 3.8 *If the table τ satisfies the dependency $X \xrightarrow{f,\alpha} Y$, $X \subseteq X'$, and $Y' \subseteq Y$, then τ satisfies both $X' \xrightarrow{f,\alpha} Y$ and $X \xrightarrow{f,\alpha} Y'$.*

Proof. Since τ satisfies $X \xrightarrow{f,\alpha} Y$ we have $\mathsf{IMP}^f_{\pi_Y}(\pi_X) \leq \alpha$. As we noticed above, $X \subseteq X'$ and $Y' \subseteq Y$ imply $\mathsf{IMP}^f_{\pi_Y}(\pi_{X'}) \leq \mathsf{IMP}^f_{\pi_Y}(\pi_X) \leq \alpha$ and $\mathsf{IMP}^f_{\pi_{Y'}}(\pi_X) \leq \mathsf{IMP}^f_{\pi_Y}(\pi_X) \leq \alpha$, which give the desired conclusions.

Alternatively, the statement follows from the transitivity properties. ∎

The purity dependencies that we introduced are essentially based on the generalization of the notion of entropy formulated in terms of partitions. In a different but related direction, Kivinen and Manilla ([KM95]) studied functional dependencies that are approximatively satisfied by tables. They introduced three pairs of measures (denoted by G_i, g_i for $1 \leq i \leq 3$) that evaluate the extent to which a table violates a functional dependency. Specifically, $G_1(X \to Y, \tau)$ equals the number of pairs of tuples in ρ that violate $X \to Y$, $G_2(X \to Y, \tau)$ gives the number of tuples that participate in such a violation, and $G_3(X \to Y, \tau)$ is the minimum number of tuples that must be removed from ρ to obtain a relation that satisfies $X \to Y$. The g_is are given by $g_1(X \to Y, \tau) = \frac{G_1(X \to Y, \tau)}{|\rho|^2}$ and $g_i(X \to Y, \tau) = \frac{G_i(X \to Y, \tau)}{|\rho|}$ for $i = 2, 3$, respectively. The measures G_1 and G_2 can be expressed using the partitions generated by attribute sets; however, they are not impurity measures in the sense of this paper. Indeed, let X, Y be two sets of attributes of the table $\tau = (T, H, \rho)$, and let $\pi_X = \{B_1, \ldots, B_n\}$,

and $\pi_Y = \{C_1, \ldots, C_m\}$. We can write:

$$G_1(X \to Y, \tau) = \sum_{i=1}^{n} \sum_{\substack{j,l=1 \\ j \neq l}}^{m} |B_i \cap C_j| \cdot |B_i \cap C_l|.$$

Similarly, $G_2(X \to Y, \tau) = \sum\{|B_i| \mid B_i \text{ is } \pi_Y - \text{impure}\}$.

4 Purity Dependencies and Approximate Classifications

Let X be a set of attributes of a table $\tau = (T, H, \rho)$ and let A be an attribute of the same table. Suppose that the tuples of τ are classified in groups based on the values of A and that we must determine the groups where the tuples belong based on tests on the remaining attributes. Of course, we are interested in using a minimal number of tests in the classification process. This task is frequently encountered in areas such as biology, medicine, social sciences, etc.

We mention here an important difference between this problem and the problem of finding a decision tree: this is the fact that whereas in a decision tree we can use a large number of attributes on different paths of the decision tree, here we are looking for a fixed set of attributes, whose examination will allow us to classify a tuple. We will look for such sets of attributes that are *minimal* in the sense that any of their subsets will not allow us to perform the classification with the desired precision.

Note that if τ satisfies the purity dependency $X \xrightarrow{f,\alpha} A$, then, the impurity of every group of tuples defined by a common X-value relative to the partition generated by A is less than α; this implies that the A blocks are approximate unions of X-groups.

Theorem 3.8 shows that if τ satisfies the purity dependency $X \xrightarrow{f,\alpha} A$, then τ also satisfies $X' \xrightarrow{f,\alpha} A$ for every superset X' of X. Thus, it is important to determine those sets X that are minimal with respect to set inclusion such that $X \xrightarrow{f,\alpha} A$. Next, we present an algorithm for finding these minimal sets.

The input of the algorithm is the value α and the attribute A. The algorithm yields the collection `Minimal` that consists of all minimal sets of attributes X such that $\mathsf{IMP}^f_{\pi_A}(\pi_X) \leq \alpha$.

The algorithm uses the collections of sets of attributes: `Candidates`, and `NewCandidates`.

There are N attributes in our dataset.

```
Minimal = empty
Candidates = empty
NewCandidates = empty
```

add to `Candidates` all sets of attributes

```
        consisting of one attribute distinct from A;
for step = 1 to N-1 do
        scan dataset and compute IMP^f_{π_A}(π_X) for each set of
              attributes X from Candidates;
        foreach set of attributes X from Candidates do
           if IMP^f_{π_A}(π_X) ≤ α do
              add X to Minimal;
              remove X from Candidates;
           endif;
        apriori_gen(Candidates, NewCandidates);
        if NewCandidates is empty
           end algorithm and return Minimal;
        Candidates = NewCandidates;
endfor;
return Minimal;
```

The *apriori_gen* procedure is the same procedure used in the Apriori algorithm [AMS+96].

This algorithm resembles Apriori; however, whereas in Apriori the frequent set property is hereditary (i.e., is inherited from a set by its subsets), we use in our algorithm the fact that the property $\{X \mid X \subseteq H, X \xrightarrow{f,\alpha} A\}$ contains all the supersets of any of its members, which means that it is dually hereditary.

This algorithm was implemented and executed on the mushroom dataset from UCI introduced in Example 3.1. The *class* attribute of the mushroom database was chosen as A.

The results of an experiment using four different types of functions for different values of α are summarized in the table below:

	Number of Minimal Sets of Attributes Found				Total number of candidate sets examined			
α	f_{gini}	f_{ent}	f_{peak}	f_{sq}	f_{gini}	f_{ent}	f_{peak}	f_{sq}
2000	46	120	113	18	153	1367	834	57
1750	53	177	109	45	278	2110	1128	160
1500	86	238	121	69	451	3926	1750	290
1250	95	310	142	93	770	5275	2365	524
1000	128	433	264	107	1217	9062	4101	952
750	170	530	332	216	2905	16897	8295	2387
500	376	914	434	425	7559	39659	20017	7742
250	796	1434	664	984	33189	126141	76257	35274

Our algorithm presents the user with the minimal sets sorted in lexicographic order beginning with the sets of the smallest cardinality.

For example, among the minimal sets returned by the algorithm using the f_{peak} generator and $\alpha = 250$ we have the sets of attributes:

$$\{odor\} \text{ and } \{cap\_color, spore\_print\_color\}.$$

Using a relational database we verified that the *odor* is indeed a very good classifier criterion.

The edible mushrooms are classified using *odor* in one of 3 sets of cardinalities 400, 400, and 3408. The poisonous mushrooms are classified using *odor* in one of 7 blocks of cardinalities 192, 2160, 36, 120, 256, 576, and 576. There is only one set characterized by *odor* that contains both edible and poisonous mushrooms and this set corresponds to the value "none" (no odor) and contains 3408 edible mushrooms and 120 poisonous mushrooms. It is interesting to note that even though we searched for rules that were impure, we discovered interesting facts that hold in 100% of the cases.

This situation conforms to Example 3.3 since we have indeed in the smaller intersection less elements than half of the value of α (which was 250 in this case). We can conclude that we can classify very well mushrooms based on *odor*, the classification needing to be refined only in the case of odorless mushrooms.

As another example, the edible mushrooms are classified using *cap_color* and *spore_print_color* in 25 sets. The poisonous mushrooms are classified in 20 sets. There are 11 sets characterized by particular values of *cap_color* and *spore_print_color* that contain both edible and poisonous mushrooms:

cap_color	*spore_print_color*	edible	poisonous
cinnamon	white	32	12
red	white	48	876
gray	black	440	32
gray	brown	440	32
brown	black	488	64
brown	brown	536	64
brown	white	96	892
white	chocolate	16	96
white	black	256	96
white	brown	280	96
white	white	144	8

As this table shows, the ambiguous classifications tend to contain predominantly either poisonous or edible mushrooms. In this case we were not able to find 100% accurate rules of classifying mushrooms based on cap color and spore color.

Note that all the above mentioned minimal sets were found both using the f_{gini} and the f_{peak} generators.

To compare results obtained with various generator functions in constructing the set of all minimal sets of attributes X such that $X \xrightarrow{f,\alpha} A$ consider the set $\mathcal{G}$ of all generator functions and define an equivalence $\sim$ on the set $\mathcal{G} \times \mathbb{R}_+$ by $(f, \alpha) \sim (f', \alpha')$ if

$$\frac{\alpha}{f\left(\frac{1}{|\mathsf{aDom}(A)|}\right)} = \frac{\alpha'}{f'\left(\frac{1}{|\mathsf{aDom}(A)|}\right)}.$$

If $(f, \alpha) \sim (f', \alpha')$, then it is easy to see that we have $m^f_{L,\pi} \leq \alpha$ if and only if $m^{f'}_{L,\pi} \leq \alpha'$, where $m^f_{L,\pi}$ was introduced by Equality (4). This gives us a base for comparing the results obtained for several generator functions. For instance, in the case of a binary attribute A the partition π_A has two blocks. Therefore, we have $(f, \alpha) \sim (f', \alpha')$ if $\frac{\alpha}{f(0.5)} = \frac{\alpha'}{f'(0.5)}$. Thus, the experimental results obtained

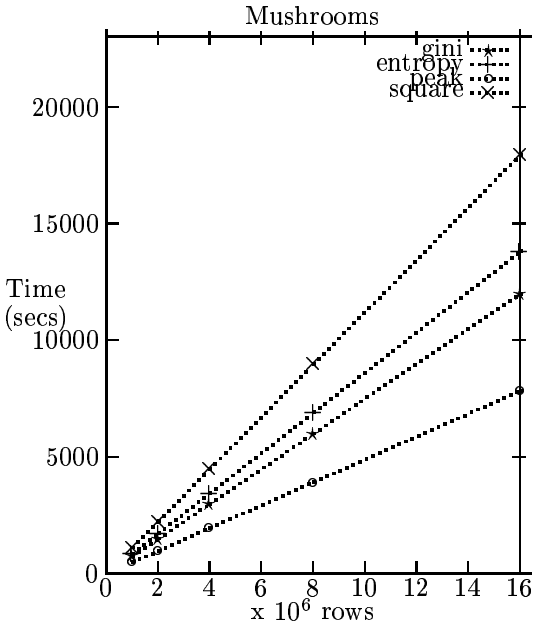

Figure 2: Dependency of time requirement on the number of tuples

in determining the purity dependencies of the form $X \xrightarrow{f,\alpha} A$ for $\alpha = 1000$ and $f = f_{\text{gini}}$ are comparable to results obtained through experiments described in the following table:

f	f_{gini}	f_{ent}	f_{peak}	f_{sq}
α	1000	2000	2000	828

Our algorithm is scalable relative to the number of tuples as shown by its time performance in an experiment whose results are shown in Figure 2. In this experiment we sought to determine the minimal sets of attributes X that satisfy a purity dependency $X \xrightarrow{f,\alpha} A$ for a fixed A by increasing the threshold α in proportion with the number of tuples. The set of purity dependencies was kept constant across these experiments by replicating the initial dataset for a sufficient number of times to achieve the desired number of tuples.

5 Conclusions and Open Problems

In this paper we introduced purity dependencies as generalizations of functional dependencies by using the notion of relative impurity of partitions. They can also be regarded as reflecting an "approximative" satisfaction of functional dependencies by tables in relational databases. As Theorem 3.5 shows, they can

be linked with table decompositions with limited information loss. Regardless of the generator function of the impurity measure used, purity dependencies have properties that are similar but not identical to Armstrong's Rules of functional dependencies (cf. Theorem 3.7).

The approach to classification that we propose in this paper generalizes well-known classification techniques in data mining such as the one proposed in the CART system which is based on the Gini impurity measure (see [BFOS84]). The algorithm described in our paper is reasonably fast and is scalable.

The role played by impurity measures in the definition and study of purity dependencies suggests the need to investigate an axiomatization of these measures. Specifically, given a finite set S and a partition $\pi = \{B_1, \ldots, B_n\}$ we are seeking necessary and sufficient conditions for a function $\mu : \mathcal{P}(S) \longrightarrow \mathbb{R}$ to respect the formula:

$$\mu(L) = |L| \sum_{i=1}^{n} f\left(\frac{|L \cap B_i|}{|L|}\right)$$

for some generator function f. The properties of impurity measures obtained in this paper constitute a good starting point in this investigation.

6 Acknowledgement

The authors would like to thank to the anonymous reviewer whose remarks have greatly contributed to the improvement of this work.

References

[AMS+96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press, Menlo Park, 1996.

[BFOS84] L. Breiman, J. H. Friedman, R. A. Ohlsen, and C. J. Stone. *Classification and Regression Trees.* Chapman & Hall/CRC, Boca Raton, 1984. Republished 1993.

[BM98] C. L. Blake and C. J. Merz. *UCI Repository of machine learning databases.* University of California, Irvine, Dept. of Information and Computer Sciences, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[CHH99] D. Coppersmith, S.J. Hong, and J.R.M. Hosking. Partitioning nominal attributes in decision trees. *Data Mining and Knowledge Discovery*, 3:197–217, 1999.

[GT66] S. Guiasu and R. Theodorescu. *Mathematical Information Theory (in Romanian).* Editura Academiei, Bucharest, 1966.

[HKPT97] Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen. Efficient discovery of functional and approximate dependencies using partitions (extended version). TR C-79, University of Helsinki, Dept. of Computer Science, Helsinki, Finland, 1997.

[IU62] R. S. Ingarden and K. Urbanik. Information without probability. *Coll. Math.*, 1:281–304, 1962.

[Khi56] A. Ia. Khinchin. On the fundamental theorem of information theory (in russian). *Usp. Mat. Nauk*, 11:17–75, 1956.

[Khi57] A. Ia. Khinchin. *Mathematical Foundations of Information Theory.* Dover Publ., New York, 1957.

[KM95] J. Kivinen and H. Mannila. Approximate dependency inference from relations. *Theoretical Computer Science*, 149(1):129–149, 1995.

[Mai83] D. Maier. *The Theory of Relational Databases.* Computer Science Press, Rockville, Maryland, 1983.

[MR75] A. M. Mathai and P. N. Rathie. *Basic Concepts in Information Theory and Statistics — Axiomatic Foundations and Applications.* Halsted Press, John Wiley & Sons, New York, 1975.

[SJ00] D. A. Simovici and S. Jaroszewicz. On information-theoretical aspects of relational databases. In C. Calude and G. Paun, editors, *Finite versus Infinite.* Springer Verlag, London, 2000.

[SJ01] D. A. Simovici and S. Jaroszewicz. An axiomatization of generalized entropy of partitions. In *Proceedings of the 29th International Symposium for Multiple-Valued Logic*, pages 259–266, Warsaw, Poland, 2001.

[ST95] D. A. Simovici and R. L. Tenney. *Relational Database Systems.* Academic Press, New York, 1995.