Mining for Purity Dependencies in Databases

Dan A. Simovici, Dana Cristofor, and Laurentiu Cristofor

University of Massachusetts at Boston,
Department of Mathematics and Computer Science,
Boston, Massachusetts 02446, USA,
{dsim,dana,laur}@cs.umb.edu

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 knowledge embedded in these functional dependencies makes them significant for data mining.

In the design process of relational databases functional dependencies are identified in the analysis phase of the design and they reflect the semantics of the database schema in the context of the application for which the database is developed. Data mining, and machine learning aim to discover functional dependencies from warehoused data and this can help in identifying knowledge embedded in data and in refining the design of database schemas.

In this paper 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 32). 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.

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* (*supra-additive*) on S if $f(x+y) \leq f(x) + f(y)$ ($f(x+y) \geq f(x) + f(y)$) for $x, y \in S$.

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

The monogenic impurity measure induced by the generator f is the impurity measure generated by concave impurity measure i of 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 $f_{\text{gini}}(p) = p - p^2$, $f_{\text{ent}}(p) = -p \log p$, or

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

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.

Further examples of generators are $f_{\sin}(p) = \sin \pi p$, $f_{\text{circle}} = \sqrt{p - p^2}$, $f_1(p) = 1 - e^{p^2 - p}$, where $p \in [0, 1]$, or

$$f_0(p) = \begin{cases} 1 & \text{if } 0 < p < 1 \\ 0 & \text{if } p = 0 \text{ or } p = 1. \end{cases}$$

For a concave function f we have the inequality

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

for every $(p_1, \ldots, p_k) \in \mathsf{SIMPLEX}_{k-1}$, also 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 21 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* $\pi \in \mathsf{PART}(S)$ is the monogenic impurity measure induced by f:

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

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

The *specific impurity of L relative to π* is

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

□

Note that Jensen's inequality implies that the largest value of the impurity of a set $L \subseteq S$ relative to a partition $\pi = \{B_1, \ldots, B_n\}$ of S is $|L| \cdot n \cdot f(\frac{1}{n})$.

Theorem 22 *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. ∎

Corollary 23 *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 22, 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 $\mathsf{IMP}_\pi^f(L) \geq \mathsf{IMP}_\pi^f(K)$. ∎

Theorem 24 *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 25 *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

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

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 26 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

$$\mathsf{IMP}^f_\pi(\sigma) = \max_{C\in\sigma} \mathsf{IMP}^f_\pi(C).$$

A partition σ is *α-pure with respect to partition* π if $\mathsf{IMP}^f_\pi(\sigma) \leq 1 - \alpha$. □

Theorem 27 *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 $\mathsf{IMP}^f_\pi(\zeta) \leq \mathsf{IMP}^f_\sigma(\zeta)$ and $\mathsf{IMP}^f_\zeta(\sigma) \leq \mathsf{IMP}^f_\zeta(\pi)$;*
(ii) $\mathsf{IMP}^f_{\pi\wedge\zeta}(\sigma \wedge \zeta) \leq \mathsf{IMP}^f_\pi(\sigma)$.

Proof. The first part of the Theorem is an immediate consequence of Theorem 24.

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 25 we have $\mathsf{IMP}^f_{\pi\wedge\zeta}(C_j \cap D) = \mathsf{IMP}^f_\pi(C_j \cap D)$, so $\mathsf{IMP}^f_{\pi\wedge\zeta}(C_j \cap D) \leq \mathsf{IMP}^f_\pi(C_j)$ due to Corollary 23. Thus, $\mathsf{IMP}^f_{\pi\wedge\zeta}(C_j \cap D) \leq \mathsf{IMP}^f_\pi(\sigma)$ for every block $C_j \cap D$ of $\sigma \wedge \zeta$ and this implies the second inequality. ∎

Corollary 28 *Let σ, π be two partitions of a set S. We have $\sigma \leq \pi$ if and only if $\mathsf{IMP}^f_\pi(\sigma) = 0$.*

Proof. It is easy to see that $\mathsf{IMP}^f_\sigma(\sigma) = 0$. Therefore, by the first part of Theorem 27, we have $\mathsf{IMP}^f_\pi(\sigma) = 0$.

Conversely, if $\mathsf{IMP}^f_\pi(\sigma) = 0$, then $\mathsf{IMP}^f_\pi(D) = 0$ 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

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 *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 [ST95].

For $X \subseteq H$ define the equivalence $\equiv_X$ by $u \equiv_X v$ if $u[X] = v[X]$; denote the corresponding partition of the set ρ of tuples by π_X.

If U, V are two subsets of H such that $U \subseteq V$, then $\pi_V \leq \pi_U$. Therefore, by Theorem 27, we have $\mathsf{IMP}^f_{\pi_U}(\pi_Z) \leq \mathsf{IMP}^f_{\pi_V}(\pi_Z)$ and $\mathsf{IMP}^f_{\pi_Z}(\pi_V) \leq \mathsf{IMP}^f_{\pi_Z}(\pi_U)$ for every $U, V, Z \subseteq H$ such that $U \subseteq V$.

Note that the specific impurity of the set of tuples ρ relative to a partition π_X is 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 28. This suggests the following generalization of the notion of functional dependency:

Definition 31 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 $\mathsf{IMP}^f_{\pi_Y}(\pi_X) \leq \alpha$. □

Let $\tau = (T, H, \rho)$ be a table and let U, V be subsets of the heading H of τ such that $U \cup V = H$. The number of spurious tuples of the decomposition (U, V) of the table τ is $n_{UV} = |\rho| - |\rho[U] \bowtie \rho[V]|$. 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.

The next theorem extends a well-known property of functional dependencies.

Theorem 32 *Let $\tau = (T, H, \rho)$ be a table and let U, V be subsets of the heading H of τ such that $U \cup V = H$. If $U \cap V \xrightarrow{f_0,\alpha} U$ and $U \cap V \xrightarrow{f_0,\beta} V$, where*

$$f_0(p) = \begin{cases} 1 & \text{if } 0 < p < 1 \\ 0 & \text{if } p = 0 \text{ or } p = 1, \end{cases}$$

then $n_{UV} \leq |\rho| \cdot (\alpha\beta - 1)$.

Proof. Let B be a block of the partition $\pi_{U \cap V}$. Note that if B is not a π_U-pure set, then $\mathsf{IMP}^{f_0}_{\pi_U}(B) = |B| p_B$, where p_B is the number of blocks C of the partition π_U such that $B \cap C$ is a proper subset of B; similarly, let q_B be the number of blocks D of π_V such that $B \cap D$ is a proper subset of B.

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

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

because any block B that is π_U- or π_V-pure does not contribute to the number of spurious tuples.

Since $\mathsf{IMP}^{f_0}_{\pi_U}(\pi_{U\cap V}) \leq \alpha$ and $\mathsf{IMP}^{f_0}_{\pi_V}(\pi_{U\cap V}) \leq \beta$, if B is a block of $\pi_{U\cap V}$, then $p_B \leq \frac{\alpha}{|B|}$ and $q_B \leq \frac{\beta}{|B|}$, so

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

which gives the desired inequality. ∎

Theorem 33 *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 27, 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 27, 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 27 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 34 *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. ∎

Example 35 Suppose that A is a binary attribute of the table $\tau = (T, H, \rho)$, that is, an attribute such that $|\mathsf{aDom}_\tau(A)| = 2$. We are interested in purity dependencies of the form $X \xrightarrow{f,\alpha} A$; in other words we examine sets of attributes

X such that $\mathsf{IMP}^f_{\pi_A}(\pi_X) \leq \alpha$. The properties of the blocks of π_X depend on the choice of the function f.

If the table τ satisfies the partition dependency $X \stackrel{f_{\text{gini}},\alpha}{\longrightarrow} A$, where A is a binary attribute, then for every block of tuples U of π_X that is split into the sets U' and U'' by the partition π_A we have

$$\frac{2|U'| \cdot |U''|}{|U'| + |U''|} \leq \alpha.$$

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

If τ satisfies the purity dependency $X \stackrel{f_{\text{peak}},\alpha}{\longrightarrow}$, then at least one of the sets U', U'' have size less than $\alpha/2$, so the other set has at least $|U| - \frac{\alpha}{2}$ elements. □

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.

Note that if τ satisfies the purity dependency $X \stackrel{f,\alpha}{\longrightarrow} A$, 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 34 shows that if τ satisfies the purity dependency $X \stackrel{f,\alpha}{\longrightarrow} A$, then τ also satisfies $X' \stackrel{f,\alpha}{\longrightarrow} A$ 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 $X \stackrel{f,\alpha}{\longrightarrow} A$. 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.

We implemented the above algorithm and we ran it on the "agaricus-lepiota" dataset from UCI

`http://www.ics.uci.edu/AI/ML/Machine-Learning.html`.

This is a mushroom database documenting 23 attributes for a number of 8124 types of mushrooms. One of these attributes indicates whether a mushroom is edible or poisonous. There are 4208 edible types of mushrooms, the rest of 3916 being poisonous mushrooms.

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_{circle}	f_{gini}	f_{ent}	f_{peak}	f_{circle}
2000	46	120	113	154	153	1367	834	1615
1750	53	177	109	195	278	2110	1128	2486
1500	86	238	121	252	451	3926	1750	4029
1250	95	310	142	412	770	5275	2365	6501
1000	128	433	264	466	1217	9062	4101	10787
750	170	530	332	583	2905	16897	8295	20818
500	376	914	434	1074	7559	39659	20017	48191
250	796	1434	664	1725	33189	126141	76257	148874

Our algorithm presents the user with the minimal sets sorted in lexiocographic order beginning with the minimal 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} 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.

This situation conforms to Example 35 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 however for 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.

Among the minimal sets returned by the algorithm using the gini function and an α value of 250 are the following combinations: {gill_size, spore_print_color} and {spore-print-color, habitat}.

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

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.

[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.

[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.

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