Impurity Measures and Applications to Classification and Clustering

Dan A. Simovici Dana Cristofor Laurentiu Cristofor *

Abstract

We introduce the notion of set impurity and relative impurity of two partitions of a set. These concepts allow us to define the notion of f-entropy as a generalization of the Shannon entropy. Starting from the partition defined by a set of attributes on the set of tuples of a table we define impurity dependencies in relational databases and develop a classification algorithm based on these dependencies.

The notions of impurity and conditional impurity allow us to extend and improve entropy-based clustering algorithms for collection of rows that have non-numerical entries by identifying sets of attributes (commonly referred to as subspaces) that have good clustering properties. We validate our conclusions through experiments.

1 Introduction

In this paper we use an information-theoretical approach for the approximate classification and clustering of categorical data. Starting from the notion of impurity measure of sets relative to partitions we generalize the notion of entropy of attribute sets introduced in [CFZ99]. It turns out that some of these generalizations are more efficient than classical entropy in identifying projections that have strong concentrations of values and, therefore, yield groups of rows that have good clustering properties.

Let $\mathcal{U}$ be a set whose elements are referred to as *attributes*. We assume that for every element A of $\mathcal{U}$ there is a set denoted by $\mathsf{Dom}(A)$ referred to as the *domain of* A such that $|\mathsf{Dom}(A)| \geq 2$.

A table is a function $T : R \times H \longrightarrow \bigcup D_H$, where R, a finite set, is referred to as the set of rows of T, H is a finite subset of $\mathcal{U}$ (called the *heading of* T), $D_H = \{\mathsf{Dom}(A) \mid A \in H\}$, and $T(r, A) \in \mathsf{Dom}(A)$ for every $A \in H$. The heading of T will be denoted by $\mathsf{heading}(T)$.

The set $\mathsf{aDom}_T(A)$ defined by

$$\mathsf{aDom}_T(A) = \{T(r, A) \mid r \in R\}$$

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

is *the active domain of* A *in* T.

The *projection* of the table $T : R \times H \longrightarrow \{\mathsf{Dom}(A) \mid A \in H\}$ on the set of attributes $L \subseteq \mathsf{heading}(T)$ is the table $T[L] : R \times B \longrightarrow \{\mathsf{Dom}(A) \mid A \in L\}$ given by $T[L](r, A) = T(r, A)$ for every $r \in R$ and $A \in L$.

If r is a row of the table T where $\mathsf{heading}(T) = H$, then $\{T(r, A) \mid A \in H\}$ is the content of the row r. The content of r in the table $T[L]$, denoted by $r[L]$ is the *projection of* r *on* L.

Let $T : R \times H \longrightarrow D_H$, $T' : R' \times H' \longrightarrow D_{H'}$ be two tables. The *join* of T, T' is the table $T \bowtie T' \longrightarrow P \times (H \cup H') \longrightarrow D_{H \cup H'}$, where $P \subseteq R \times R'$ consists of those pairs of rows (r, r') such that $r[A] = r'[A]$ for every $A \in H \cap H'$ and $T \bowtie T'((r, r'), A) = r[A] = r'[A]$ for $(r, r') \in P$ and $A \in H \cup H'$.

Let $\mathbb{R}$ be the set of reals. The k-dimensional simplex is the set $\mathsf{SIMPLEX}_{k-1} = \{(p_1, \ldots, p_k) \in \mathbb{R}^k \mid p_i \geq 0 \text{ and } p_1 + \cdots + p_k = 1\}$. 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 $x, y \in S$. The function f is *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 $I(n_1, \ldots, n_k) = N i(n_1/N, \ldots, n_k/N)$, where $N = \sum_{i=1}^{k} n_i$. Both the Gini impurity measure and the entropy can be generated using a simple one-argument concave function $f : [0, 1] \longrightarrow \mathbb{R}$ such that $f(0) = f(1) = 0$. In this paper we strengthen some of these conditions.

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

If, in addition, we have

$$f(ka_1) + \cdots + f(ka_n) \leq k(f(a_1) + \cdots + f(a_n)) + f(k) \tag{1}$$

for $(a_1, \ldots, a_n) \in \mathsf{SIMPLEX}_{n-1}$ and $k \in [0, 1]$, then we refer to f as a *strong generator*.

The (strong) monogenic impurity measure induced by the (strong) generator f is the impurity measure generated by a 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 $f_{\text{gini}}(p) = p - p^2$, $f_{\text{sq}}(p) = \sqrt{p} - p$, $f_{\text{ent}}(p) = -p \log p$, or f_{peak} (given by $f_{\text{peak}}(p) = p$ for $0 \leq p \leq 0.5$ and $f_{\text{peak}}(p) = 1 - p$ for $0.5 < p \leq 1$) are strong generators.

Let f be a generator. Since f is concave it satisfies Jensen's inequality $f(p_1) + \cdots + f(p_k) \leq k f\left(\frac{1}{k}\right)$ for every $(p_1, \ldots, p_k) \in \mathsf{SIMPLEX}_{k-1}$ and 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

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* $\pi \in \mathsf{PART}(S)$ is the monogenic impurity measure induced by f:

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

where $\pi = \{B_1, \ldots, B_n\}$. The *specific impurity of L relative to π* is

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

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

Proof. The disjointness of K and L allows us to write

$$\mathsf{imp}^f_\pi(K \cup L) = \sum_{\ell=1}^{n} f\left(\frac{|(K \cup L) \cap B_\ell|}{|K \cup L|}\right) = \sum_{\ell=1}^{n} f\left(\frac{|K \cap B_\ell| + |L \cap B_\ell|}{|K \cup L|}\right).$$

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 implies

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

so $\mathsf{imp}^f_\pi(K \cup L) \geq \frac{|K|}{|K|+|L|}\mathsf{imp}^f_\pi(K) + \frac{|L|}{|K|+|L|}\mathsf{imp}^f_\pi(L)$, which gives immediately the desired inequality. ∎

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

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

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

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

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 $\mathsf{IMP}^f_\sigma(K) \leq \mathsf{IMP}^f_\pi(K)$ 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 $\{1, \ldots, n\}$. The subadditivity of f implies

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

This yields 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 $\mathsf{IMP}^f_{\pi\wedge\zeta}(K) = \mathsf{IMP}^f_\pi(K)$.*

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 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

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

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

Theorem 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 $\mathsf{IMP}^f_\pi(\zeta) \leq \mathsf{IMP}^f_\sigma(\zeta)$ 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 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 $\mathsf{IMP}^f_{\pi\wedge\zeta}(C_j \cap D) = \mathsf{IMP}^f_\pi(C_j \cap D)$, so $\mathsf{IMP}^f_{\pi\wedge\zeta}(C_j \cap D) \leq \mathsf{IMP}^f_\pi(C_j)$ due to Corollary 2.3. 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 2.8 *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 2.7, 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 $T : R \times H \longrightarrow D_H$ be a table, where $H = A_1 \cdots A_n$.

For $X \subseteq H$ define the equivalence $\equiv_X$ on R by $u \equiv_X v$ if $u[X] = v[X]$; denote the corresponding partition of the set R 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 2.7, 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 R 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 $X \rightarrow Y$ 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 suggests the following generalization of the notion of functional dependency:

Definition 3.1 Let $T : R \times H \longrightarrow D_H$ be a table and let $X, Y \subseteq H$ be two sets of attributes. T satisfies the purity dependency $X \stackrel{f,\alpha}{\longrightarrow} Y$ if $\mathsf{IMP}^f_{\pi_Y}(\pi_X) \leq \alpha$.

Let $T : R \times H \longrightarrow D_H$ be a table and let U, V be subsets of the heading H of T such that $U \cup V = H$. The number of spurious tuples of the decomposition (U, V) of the table T is $n_{UV} = |R| - |R[U] \bowtie R[V]|$. 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 3.2 *Let $T : R \times H \longrightarrow D_H$ be a table and let U, V be subsets of the heading H of T such that $U \cup V = H$. If $U \cap V \stackrel{f_0,\alpha}{\longrightarrow} U$ and $U \cap V \stackrel{f_0,\beta}{\longrightarrow} V$, where*

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

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

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

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

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

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

which gives the desired inequality. ∎

Theorem 3.3 *Let $T : R \times H \longrightarrow D_H$ be a table and let X, Y, Z be subsets of H. The following statements hold:*

1. *if T satisfies $X \to Y$ and $Y \xrightarrow{f,\alpha} Z$, then T satisfies $X \xrightarrow{f,\alpha} Z$ (left transitivity property);*
2. *if T satisfies $X \xrightarrow{f,\alpha} Y$ and $Y \to Z$, then T satisfies $X \xrightarrow{f,\alpha} Z$ (right transitivity property);*
3. *if T satisfies $X \xrightarrow{f,\alpha} Y$, then T also satisfies $X \cup Z \xrightarrow{f,\alpha} Y \cup Z$ (the augmentation property).*

Proof. Suppose that T 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 T 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 T 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 T. 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 T satisfies the purity dependency $X \cup Z \xrightarrow{f,\alpha} Y \cup Z$. ∎

Theorem 3.4 *If the table T satisfies the dependency $X \xrightarrow{f,\alpha} Y$, $X \subseteq X'$, and $Y' \subseteq Y$, then T satisfies both $X' \xrightarrow{f,\alpha} Y$ and $X \xrightarrow{f,\alpha} Y'$.*

Proof. Since T 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 3.5 Suppose that A is a binary attribute of the table T, that is, an attribute such that $|\mathsf{aDom}_T(A)| = 2$. 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 T satisfies the partition dependency $X \xrightarrow{f_{\mathrm{gini}},\alpha} A$, 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 T satisfies the purity dependency $X \xrightarrow{f_{\mathrm{peak}},\alpha}$, 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 $T : R \times H \longrightarrow D_H$ and let A be an attribute of the same table. Suppose that the tuples of T 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 T 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 3.4 shows that if T satisfies the purity dependency $X \xrightarrow{f,\alpha} A$, then T 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`.

Suppose that 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 lexicographic 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 3.5 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 sets: {*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.

5 A Generalization of Entropy

Let L be a subset of a set S and let $\pi = \{B_1, \ldots, B_n\}$ be a partition of S. We start from the observation that

$$\mathsf{imp}_{\pi}^{f_{\text{ent}}}(L) = \sum_{i=1}^{n} f_{\text{ent}}\left(\frac{|L \cap B_i|}{|L|}\right).$$

Therefore, we have

$$\begin{aligned}\mathsf{imp}_\pi^{f_{\mathrm{ent}}}(S) &= \sum_{i=1}^{n} f_{\mathrm{ent}}\left(\frac{|B_i|}{|S|}\right)\\ &= -\sum_{i=1}^{n}\frac{|B_i|}{|S|}\log\frac{|B_i|}{|S|}.\end{aligned}$$

This coincides with the entropy of a discrete random variable ξ whose range is π such that $P(\xi = B_i) = \frac{|B_i|}{|S|}$ for $1 \leq i \leq n$.

Definition 5.1 Let $T : R \times H \longrightarrow D_H$ be a table, X, Y be sets of attributes, $X, Y \subseteq H$ and let f be a strong generator.

The *f-entropy* of X is $\mathcal{H}^f(X) = \mathsf{imp}^f_{\pi_X}(R)$.

The *conditional f-entropy* of X on Y is $\mathcal{H}^f(X|Y) = \sum_{j=1}^{m}\frac{|C_j|}{|R|}\mathsf{imp}^f_{\pi_X}(C_j)$, where $\pi_Y = \{C_1, \ldots, C_m\}$.

In other words, the conditional entropy $\mathcal{H}^f(X|Y)$ is the average value of the specific impurity of the blocks of the partition π_Y relative to the partition π_X.

Since $\pi_\emptyset = \{R\}$, it follows that $\mathcal{H}^f(X) = \mathcal{H}^f(X|\emptyset)$ for every set of attributes X. Therefore, we have $\mathcal{H}^f(X) = \sum_{i=1}^{n} f\left(\frac{|B_i|}{|R|}\right)$, where $\pi_X = \{B_1, \ldots, B_n\}$. For the conditional entropy we have $\mathcal{H}^f(X|Y) = \frac{1}{|R|}\sum_{j=1}^{m}|C_j|\sum_{i=1}^{n} f\left(\frac{|B_i \cap C_j|}{|C_j|}\right)$. The concavity of the strong generator f implies that the largest value of the f-entropy of a set of attributes X occurs when the blocks of the partition π_X are of equal size. Therefore, we have $\mathcal{H}^f(X) \leq nf(\frac{1}{n})$, where n is the number of blocks of π_X, that is, the cardinality of the active domain of X.

Theorem 5.2 *Let X, X', Y, Y' be sets of attributes of the table $T : R \times H \longrightarrow D_H$. For every strong generator f we have:*

1. *$X \subseteq X'$ implies $\mathcal{H}^f(X|Y) \leq \mathcal{H}^f(X'|Y)$ for every set of attributes Y;*
2. *$Y \subseteq Y'$ implies $\mathcal{H}^f(X|Y) \geq \mathcal{H}^f(X|Y')$ for every set of attributes X.*

Proof. If $X \subseteq X'$, then $\pi_{X'} \leq \pi_X$, so, by Theorem 2.4, $\mathsf{IMP}^f_{\pi_X}(C) \leq \mathsf{IMP}^f_{\pi_{X'}}(C)$ for every set $C \subseteq R$. This implies $\mathcal{H}^f(X|Y) \leq \mathcal{H}^f(X'|Y)$.

Suppose now that $Y \subseteq Y'$. Since $\pi_{Y'} \leq \pi_Y$, every block C of π_Y can be written as a union of blocks of $\pi_{Y'}$, $C = \bigcup\{C'_i \mid 1 \leq i \leq \ell\}$. The superadditivity of the impurity shown in Theorem 2.2 implies that $\mathsf{IMP}^f_\pi(C) \geq \sum_{i=1}^{\ell}\mathsf{IMP}^f_\pi(C_i)$, and this gives the second inequality. ∎

Corollary 5.3 *If X, X' are two sets of attributes of the table T, then $X \subseteq X'$ implies $\mathcal{H}^f(X) \leq \mathcal{H}^f(X')$.*

Proof. This statement follows immediately from the first inequality of Theorem 5.2. ∎

Theorem 5.4 *Let X, Y be two sets of attributes of a table $T : R \times H \longrightarrow D_H$. We have*

1. $\mathcal{H}^f(X|Y) \leq \mathcal{H}^f(X)$;
2. $\mathcal{H}^f(XY) \leq \mathcal{H}^f(X|Y) + \mathcal{H}^f(Y)$.

Proof. Suppose that $\pi_X = \{B_1, \ldots, B_n\}$ and $\pi_Y = \{C_1, \ldots, C_m\}$. To prove the first part of the theorem observe that we can write

$$\frac{|B_i|}{|R|} = \sum_{j=1}^{m} c_j \frac{|B_i \cap C_j|}{|C_j|},$$

where $c_j = \frac{|C_j|}{|R|}$ for $1 \leq j \leq m$ and $(c_1, \ldots, c_m) \in \mathsf{SIMPLEX}_{m-1}$. The concavity of f implies

$$f\left(\frac{|B_i|}{|R|}\right) \geq \sum_{j=1}^{m} c_j f\left(\frac{|B_i \cap C_j|}{|C_j|}\right),$$

which, in turn, yields

$$\mathcal{H}^f(X|Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} c_j f\left(\frac{B_i \cap C_j}{|C_j|}\right) \leq \sum_{i=1}^{n} f\left(\frac{|B_i|}{|R|}\right) = \mathcal{H}^f(X).$$

For a fixed j let $a_i = \frac{|B_i \cap C_j|}{|C_j|}$ and $k = \frac{|C_j|}{|R|}$. Since $(a_1, \ldots, a_n) \in \mathsf{SIMPLEX}_{n-1}$ we have

$$\sum_{i=1}^{n} f\left(\frac{|B_i \cap C_j|}{|R|}\right) \leq \frac{|C_j|}{|R|} \sum_{i=1}^{n} f\left(\frac{|B_i \cap C_j|}{|C_j|}\right) + f\left(\frac{|C_j|}{|R|}\right).$$

Summing these inequalities for $1 \leq j \leq m$ and observing that the blocks of the partition π_{XY} have the form $B_i \cap C_j$ we obtain the second inequality of the theorem. ∎

Corollary 5.5 *For every sets of attributes X, Y we have $\mathcal{H}^f(XY) \leq \mathcal{H}^f(X) + \mathcal{H}^f(Y)$.*

Proof. The inequality follows immediately from Theorem 5.4. ∎

6 Clusters for Categorical Data

Let $T : R \times H \longrightarrow D_H$ be a table, X be a set of attributes $X \subseteq H$ and assume that $\mathsf{aDom}_T(X)$ consists of N elements, $\mathsf{aDom}_T(X) = \{x_1, \ldots, x_N\}$. Each block B_i of π_X corresponds to a value x_i of the active domain of X for $1 \leq i \leq N$, and x_i appears under X with the frequency $\frac{|B_i|}{|R|}$ for $1 \leq i \leq N$.

Let $d_{T,X}$ be a distance defined on $\mathsf{aDom}_T(X)$. The average distance between the elements of $\mathsf{aDom}_T(X)$ is

$$E(d_T(X)) = \frac{\sum_{i=1}^N \sum_{j=1}^N \{d_{T,X}(x_i, x_j)|B_i| \cdot |B_j|\}}{|R|^2}.$$

The next result connects one of the generalizations of entropy with the average distance and serves as a foundation for the applications of information-theoretical methods to clustering.

Let $d_{T,X}$ be the distance on $\mathsf{aDom}_T(X)$ defined by:

$$d^{0,1}_{T,X}(x, x') = \begin{cases} 0 & \text{if } x = x' \\ 1 & \text{otherwise} \end{cases}$$

for $x, x' \in \mathsf{aDom}_T(X)$.

Theorem 6.1 *The average distance* $E(d^{0,1}_{T,X})$ *equals* $\mathcal{H}^{f_{gini}}(X)$.

Proof. The definition of the distance $d^{0,1}_{T,X}$ implies that the average distance $E(d^{0,1}_{T,X})$ can be written as

$$\begin{aligned} E(d^{0,1}_{T,X}) &= \frac{\sum_{i=1}^N \sum_{j=1}^N \{|B_i| \cdot |B_j| \mid 1 \leq i, j \leq N, i \neq j\}}{|R|^2} \\ &= 2 \cdot \frac{\sum_{i=1}^N \sum_{j=1}^N \{|B_i| \cdot |B_j| \mid 1 \leq i, j \leq N, i < j\}}{|R|^2} \\ &= \frac{|R|^2 - \sum_{i=1}^N \{|B_i|^2 \mid 1 \leq i \leq N\}}{|R|^2} = \mathcal{H}^{f_{\text{gini}}}(X), \end{aligned}$$

which gives the desired equality. ∎

Theorem 6.1 suggests that by limiting the entropy of a set of attributes X we limit the average distance between the projections of the tuples of the table on X and, therefore, increase our chances to find large clusters of tuples based on their similarity on X.

A projection of a table T on a set of attributes X presents an interest for us if it satisfies two criteria:

1. The f-entropy $\mathcal{H}^f(X)$, referred to as the *internal entropy* of X is limited. This insures that the values X-projections congregate towards a small number of clusters.

2. The impurity of each of the blocks of the partition π_X relative to the partition π_{H-X} is small. This insures that the tuples created by clustering the X-projection do not diverge excessively on the remaining attributes of the table. The definition of conditional f-entropy shows that such sets of attributes can be identified by limiting the value of the conditional entropy $\mathcal{H}^f(H - X|X)$, referred to as the *external* entropy of X is limited.

Definition 6.2 Let $T : R \times H \longrightarrow D_H$ be a table. An (f, α)-*clusterable set of attributes in* T is a set $X \subseteq H$ such that $\mathcal{H}^f(X) \leq \alpha$ and X is maximal with this property, that is, $X \subset Y$ implies $\mathcal{H}^f(Y) > \alpha$.

The *clusters determined by the* (f, α)-*clusterable set* X are the groups of rows that have equal values on X.

Next, we present an algorithm that can be used to identify clusterable sets of attributes. The algorithm is based on the monotonicity of the f-entropy shown in Corollary 2.3. Namely, if U is a set of attributes such that $\mathcal{H}^f(U) > \alpha$, then for every set V such that $U \subset V$ we have $\mathcal{H}^f(V) > \alpha$. Thus, once a set U is disqualified from being (f, α)-clusterable, no superset of U can be (f, α)-clusterable. Of course, if U, U' are (f, α)-clusterable, the set UU' is a candidate for the collection `KAttrSets` and we apply the apriori-gen technique from [AMS+96] for identifying those sets that can be included in `KAttrSets`.

Note that if $X \subseteq X'$, then, by Theorem 5.2, we have $\mathcal{H}^f(H - X|X) \geq \mathcal{H}(H - X'|X')$. Thus, in order to limit the external entropy, we focus on the maximal (f, α)-clusterable sets of attributes.

It is easy to see that if an attribute set partitions a set of tuples in k blocks, then its maximum entropy would be $1 - \frac{1}{k}$ for the f_{gini}-entropy, $\log k$ for the f_{ent}-entropy, $\sqrt{k} - 1$ for the f_{sq}-entropy, and in general $nf(1/n)$ for an f-entropy. By specifying a value k and computing the maximum f-entropy and then looking for all maximal sets of attributes that have entropy less than the maximum value we computed, we obtain sets of attributes that can be clustered in approximatively k clusters.

Our algorithm has three input parameters: the number of blocks k in which our sets of attributes should partition the rows of the database, a percentage of the attributes p that we want to use in our clustering process, and the type of strong generator function f that we want to use.

The output of the algorithm consists of all sets of attributes found with internal entropy less than $kf(1/k)$ from the first pN attributes ordered by the values of their internal entropy, where N is the total number of attributes. The parameter p allows us to eliminate attributes with larger entropy.

We use three collections of sets of attributes: `Candidates`, `KAttrSets` and `MaxAttrSets`. When the algorithm ends, the collection `MaxAttrSets` will contain the $(f, kf(1/k))$-clusterable sets.

```
Candidates = empty; KAttrSets = empty; MaxAttrSets = empty;
add to Candidates all one-attribute sets;
for step = 1 to N do
  scan dataset and compute for each set S from Candidates
     the internal entropy of S;

  retain from the set of attributes only pN attributes
  with the smallest internal entropies and discard the rest;

  for each set of attributes S from Candidates do
```

```
    if H(S) <= kf(1/k) do
      add S to KAttrSets;
      add S to MaxAttrSets;
    endif;
  apriori_gen(Candidates, KAttrSets);
  if Candidates is empty
    break;
endfor;
remove non maximal sets from MaxAttrSets;
scan dataset and compute for each set S from MaxAttrSets
  the external entropy of S
```

As in Section 4, the algorithm was run on the "agaricus-lepiota" dataset. We experimented using three different strong generators for different values of k: f_{ent}, f_{gini}, and f_{sq}.

The results for $p = 0.5$ and $p = 0.7$ are summarized below.

$p = 0.5$

k	f_{gini}		f_{ent}		f_{sq}	
	time	sets	time	sets	time	sets
2	1	5	1	4	1	5
3	2	10	1	7	1	6
4	6	26	2	13	2	13
5	8	27	4	21	3	16
10	48	41	15	49	12	52
20	96	36	54	71	40	97
30	134	15	89	67	71	80
40	136	1	113	45	100	63
50	136	1	129	17	116	42
60	136	1	138	17	129	32
70	136	1	141	6	142	25
80	136	1	147	11	141	5
90	136	1	147	1	143	8
100	136	1	150	1	150	1

$p = 0.7$

k	f_{gini}		f_{ent}		f_{sq}	
	time	sets	time	sets	time	sets
2	1	5	1	6	1	6
3	6	22	3	12	1	6
4	13	42	8	33	4	20
5	23	65	13	47	5	21
10	233	178	88	200	33	121
20	1041	368	569	419	217	374
30	2156	374	1271	596	603	512
40	2889	473	2068	612	1159	628
50	4058	274	2820	453	1659	662
60	4775	233	3627	358	2224	696
70	5169	210	4061	438	2827	598
80	5413	215	4786	281	3262	491
90	5908	184	5034	274	4145	608
100	7271	171	6505	425	5149	396

Elementary computations show that if $0 \leq p \leq 0.025$, then $f_{\mathrm{gini}}(p) \leq f_{\mathrm{ent}}(p) \leq f_{\mathrm{sq}}(p)$; for $0.025 < p \leq 0.39$ we have $f_{\mathrm{gini}}(p) \leq f_{\mathrm{sq}}(p) \leq f_{\mathrm{ent}}(p)$. Finally, if $0.39 < p \leq 1$, we have $f_{\mathrm{sq}}(p) \leq f_{\mathrm{gini}}(p) \leq f_{\mathrm{ent}}(p)$. Thus, f_{sq} "penalizes" the sets of attributes whose partitions contain small blocks (containing fewer than 2.5% of the set of all tuples); such sets are not good candidates for defining clusters of rows.

As an example of clusters determined by an $(f, kf(1/k))$-clusterable set of attributes, consider the attribute set {*bruises, gill-spacing, gill-size, stalk-shape, veil-type*} obtained when we used f_{sq} with $k = 10$. The internal entropy of this cluster was 1.87871 and the external entropy was 1.39457. There are 11 clusters formed by grouping on these attributes, and they contain 1540, 272, 1728, 288, 768, 152, 1000, 2016, 256, 8 and 96 rows, respectively, as shown in the next table. The external impurity of a block B, $\mathsf{imp}^{f_{\mathrm{sq}}}_{\pi H-X}(B)$ is denoted by $E(B)$ in Figure 1.

The attributes that were considered in the computation of external entropy were class, cap-surface, gill-attachment, stalk-surface-above-ring, stalk-surface-below-ring, veil-color, ring-number.

The graph of the number of sets with good clustering properties as a function of the anticipated number of blocks (k) is unimodal and reaches its maximum for a number of anticipated blocks that depends on p, the fraction of the number of attributes retained from the table, and on the generator function involved in the experiment. After this maximum is reached, the number of sets decreases because the algorithm accepts sets with a larger number of attributes until it will eventually accept the set of all attributes, which happens, when $p = 0.5$, for f_{gini} for $k = 40$ and for f_{ent} and f_{sq} for values of k of 90 and 100 respectively. The graph of the time as a function of k shows that the time used by the algorithm grows until the moment the algorithm reaches its saturation point by finding the set of all attributes. This effect is more apparent for relatively small values of p (see, for example, Figures 2 to 5).

Set of attributes					Block B		
bruises	gill-spacing	gill-size	stalk-shape	veil-type	count	groups $\times$ tuples	$E(B)$
no	*close*	*broad*	*enlarging*	*partial*	1540	2×648 2×96 2×18 2×8	1.157
no	*close*	*narrow*	*enlarging*	*partial*	272	1×96 3×48 2×16	1.339
no	*close*	*narrow*	*tapering*	*partial*	1728	8×216	1.828
no	*crowded*	*broad*	*enlarging*	*partial*	288	8×36	1.828
no	*crowded*	*broad*	*tapering*	*partial*	768	8×96	1.828
no	*crowded*	*narrow*	*enlarging*	*partial*	152	2×48 1×8 8×6	1.942
yes	*close*	*broad*	*enlarging*	*partial*	1000	2×256 1×192 2×112 2×36	1.488
yes	*close*	*broad*	*tapering*	*partial*	2016	2×864 4×72	1.065
yes	*close*	*narrow*	*enlarging*	*partial*	256	2×128	0.414
yes	*crowded*	*narrow*	*enlarging*	*partial*	8	2×4	0.414
yes	*crowded*	*narrow*	*tapering*	*partial*	96	2×48	0.414

Figure 1: Blocks Determined by a Set of Attributes

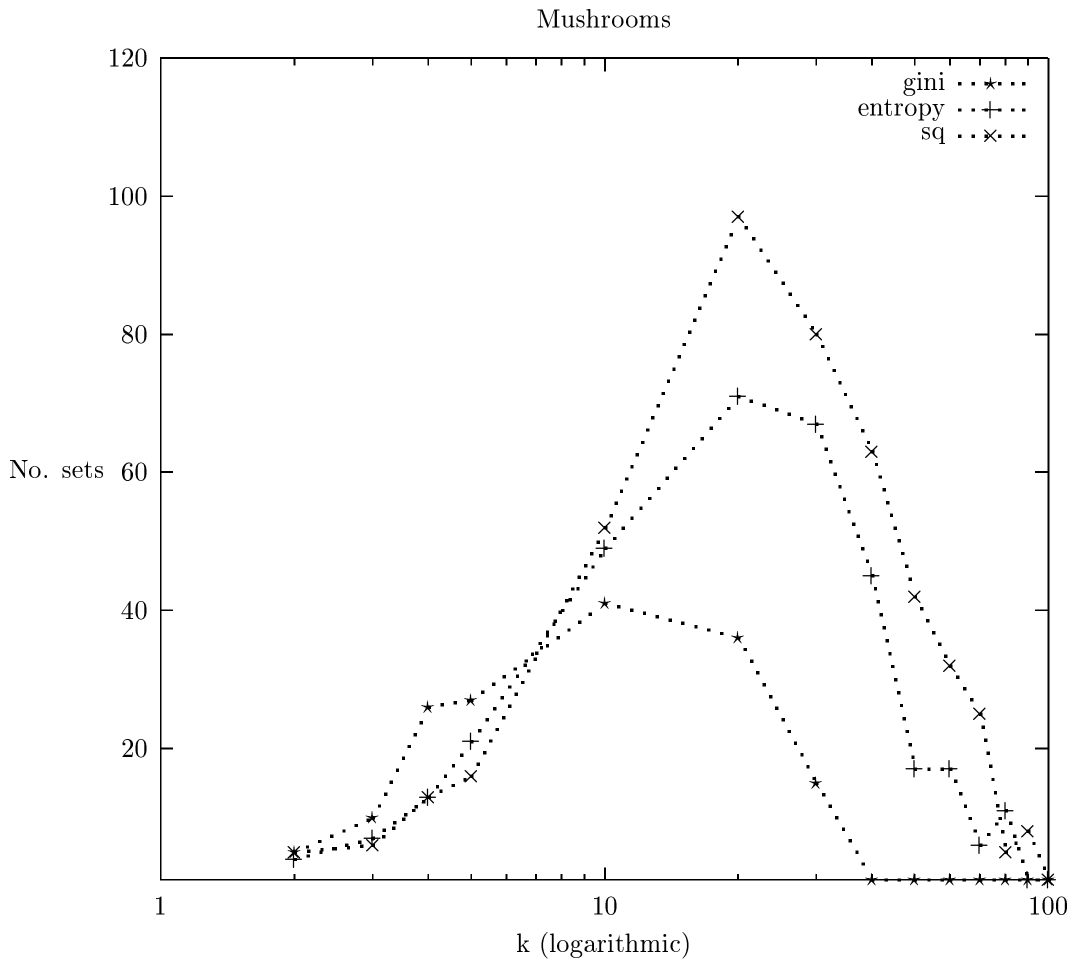

Figure 2: Number of Attribute Sets for $p = 0.5$

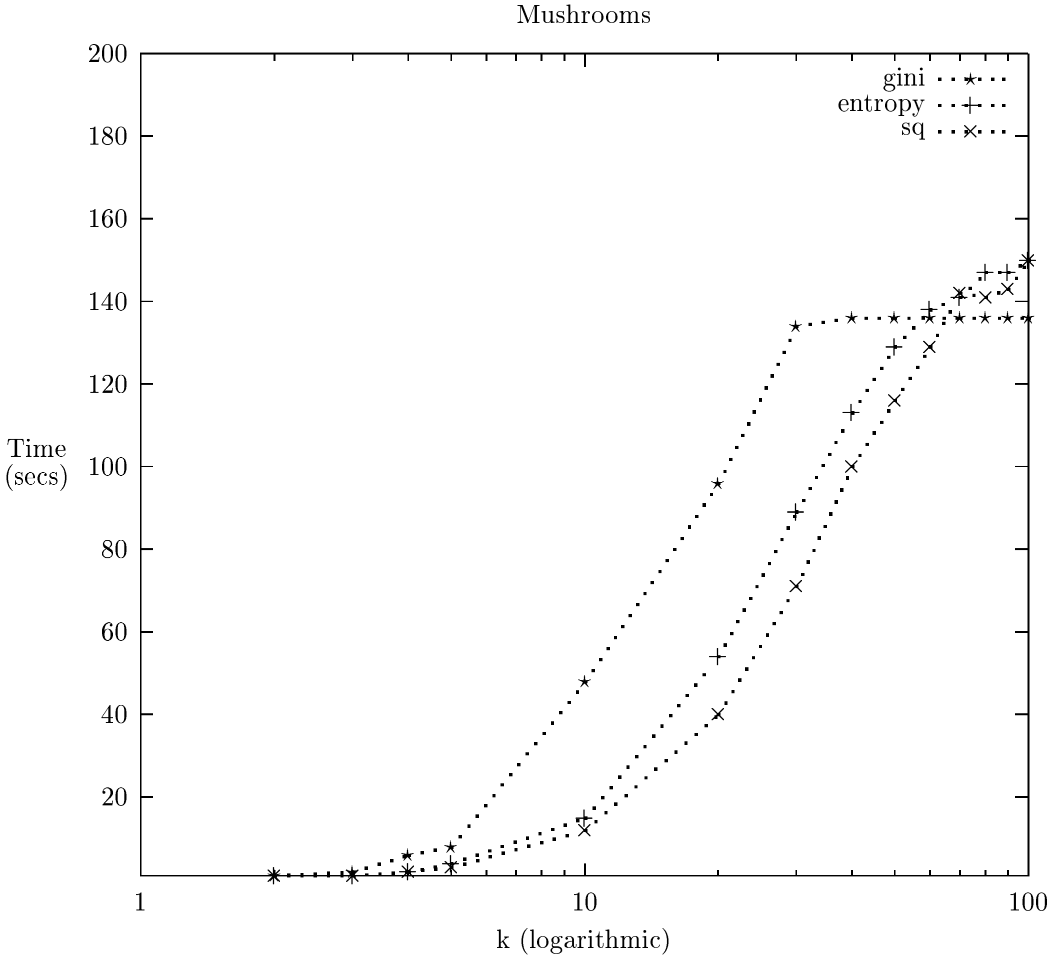

Figure 3: Time for $p = 0.5$

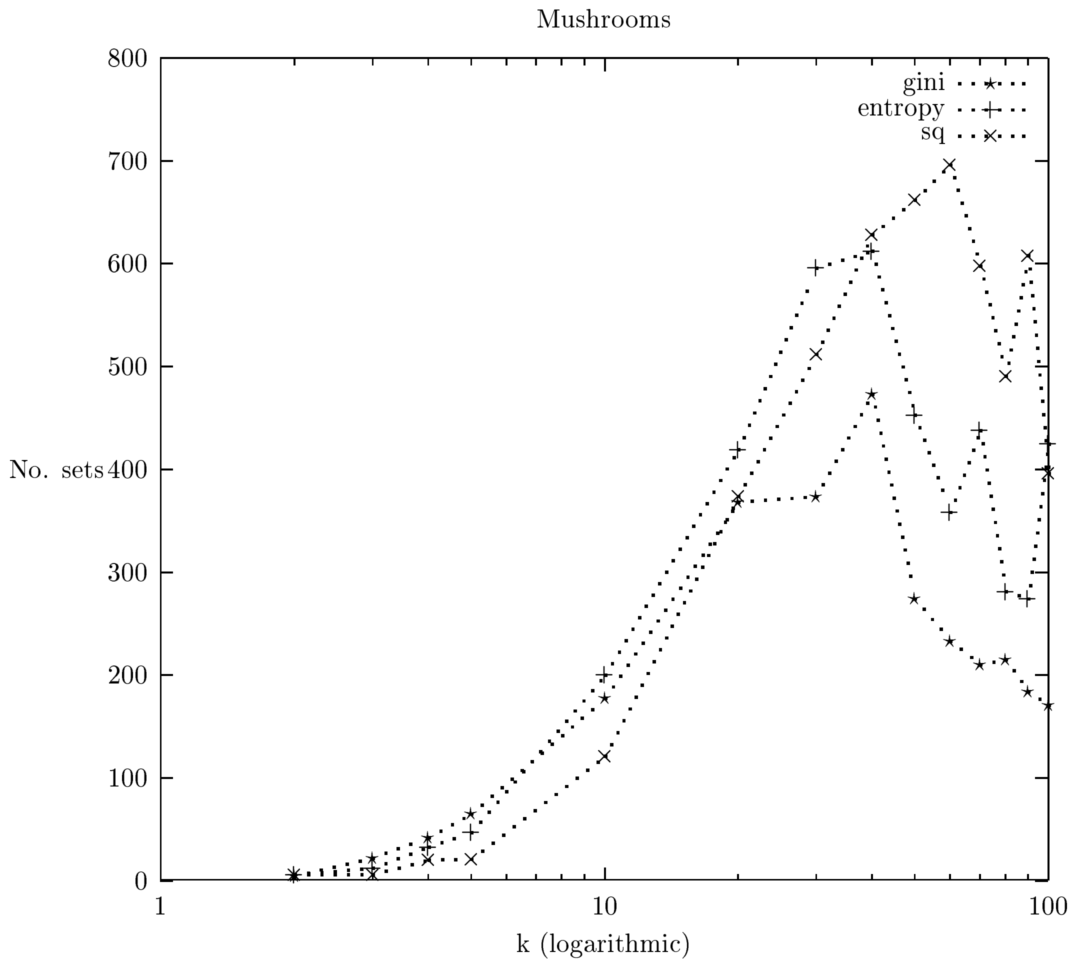

Figure 4: Number of Attribute Sets for $p = 0.7$

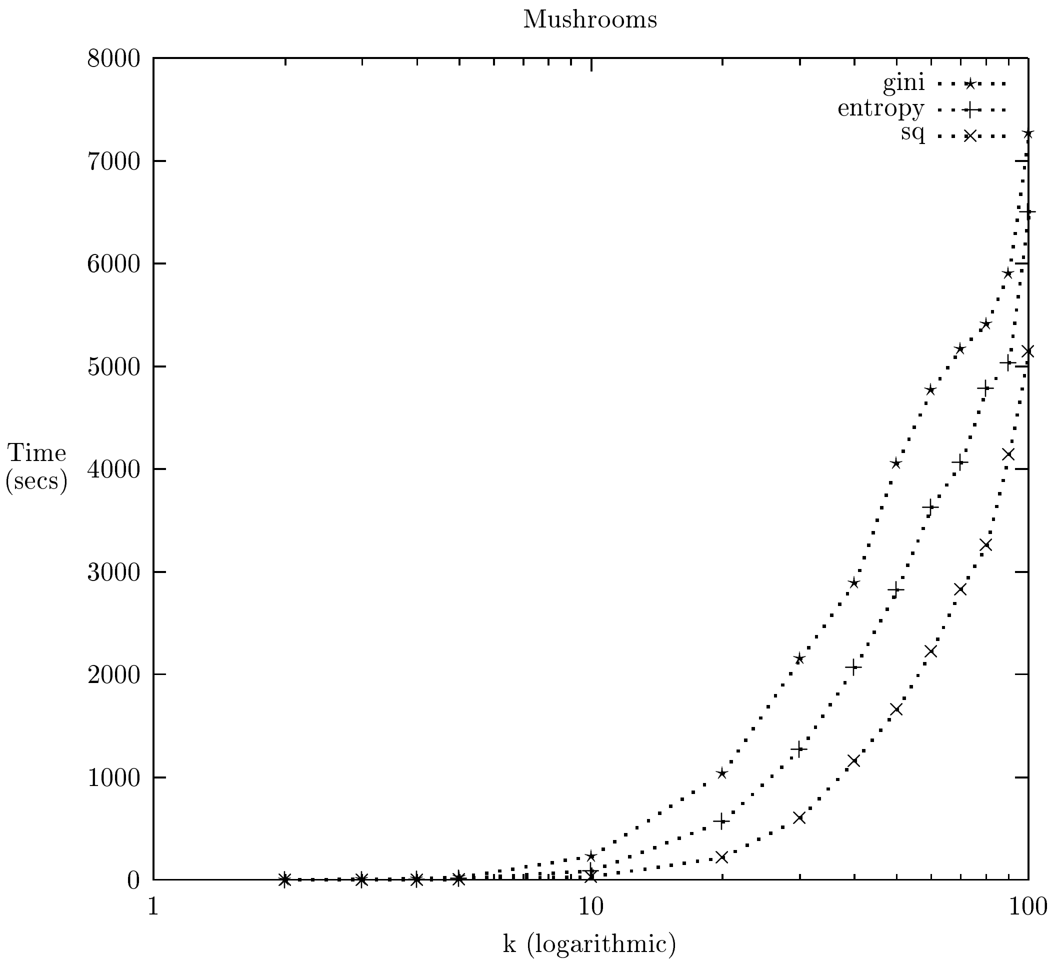

Figure 5: Time for $p = 0.7$

7 Open Problems

An investigation of classes of generator functions is necessary to determine the types of clusterable sets that are "favored" by these classes of generators.

The definition of clusters as groups of records determined by certain sets of attributes that have limited external impurities must be further refined and techniques that allow the user to choose a covering of the set of records by (f, α)-clusters must be developed.

Also, an important open question is the axiomatization of the notion of f-entropy of an attribute set. Such an axiomatization should relax known axiomatizations of the notion of entropy (see [JS99]) to accommodate the generalizations discussed in this paper.

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.

[CFZ99] C. H. Cheng, A. W. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 84–93, 1999.

[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.

[JS99] S. Jaroszewicz and D. A. Simovici. On axiomatization of conditional entropy. In *Proceedings of the 29th International Symposium for Multiple-Valued Logic*, pages 24–31, Freiburg, Germany, 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.