Generalized Entropy and Projection Clustering of Categorical Data

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 generalize the notion of entropy for a set of attributes of a table and we study its applications to clustering of categorical data. This new concept allows greater flexibility in identifying sets of attributes and, in a certain case, is naturally related to the average distance between the records that are the object of clustering. An algorithm that identifies clusterable sets of attributes (using several types of entropy) is also presented as well as experimental results obtained with this algorithm.

1 Introduction

In this paper we investigate clustering on categorical data, a problem that poses challenges that are quite distinct from numerical data clustering (see [AGGR98] and [CFZ99]). This type of data can be encoded numerically; however, this introduces artificial total orderings on the domains of the attributes involved which have the potential to create artificial agglomerations that are meaningless from a practical point of view. Our approach to clustering is based on identifying projections that have sufficient concentrations of values (determined by certain types of entropies associated to attribute sets). The sets of records that are formed by grouping on these sets of attributes qualify as clusters if they do not differ too much on the remaining attributes; this requirement can be enforced by limiting the value of a certain generalization of conditional entropy. Our work extends the information-theoretical approach to clustering found in [CFZ99].

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 $I(n_1, \ldots, n_k) = Ni(n_1/N, \ldots, n_k/N)$, 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 additionally require two extra conditions which are satisfied by many functions.

Definition 11 A concave function $f : [0,1] \longrightarrow \mathbb{R}$ is a *generator* if $f(0) = f(1) = 0$, f is subadditive and for $(a_1, \ldots, a_n) \in \mathsf{SIMPLEX}_{n-1}$ and $k \in [0,1]$ we have

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

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 $f_{\text{gini}}(p) = p - p^2$, $f_{\text{sq}}(p) = \sqrt{p} - p$, $f_{\text{ent}}(p) = -p \log p$, 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 $0.5 < p \leq 1$) satisfy the conditions that define generators.

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

We need to define formally the notion of database table. Let $\mathcal{U}$ be a set of symbols referred to as *attributes*. We assume that for every $A \in \mathcal{U}$ there exists a set $\mathsf{Dom}(A)$ (referred to as the domain of A) such that $|\mathsf{Dom}(A)| \geq 2$. If H is a finite subset of the set $\mathcal{U}$, an H-tuple is a function $t : H \longrightarrow \bigcup\{\mathsf{Dom}(A) \mid A \in H\}$ such that $t(A) \in \mathsf{Dom}(A)$ for every $A \in H$. A table is a triple $\tau = (T, H, \rho)$, where T, the name of the table, is a string over a given alphabet, H is a finite set of attributes, and ρ is a finite multiset of H-tuples. The set H is referred to as *the heading of the table.*

The projection of a tuple $t \in \rho$ on a set of attributes X, $X \subseteq H$, is defined as the restriction of t to X, and will be denoted by $t[X]$. The *active domain* of an attribute A relative to a table τ is the set of all values that occur under the attribute A in τ. This notion is extended to sets of attributes; namely, if X is a set of attributes of $\tau = (T, H, \rho)$, then the active domain of X, $\mathsf{aDom}_\tau(X)$, is the set of all values that appear under X in τ, $\mathsf{aDom}_\tau(X) = \{t[X] \mid t \in \rho\}$. For relational terminology and notations see, for example, [ST95].

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 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:

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

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 22 *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 23 *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 22, 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 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 $\mathsf{IMP}^f_\sigma(K) \leq \mathsf{IMP}^f_\pi(K)$ 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. ∎

3 f-Entropy

Let $\tau = (T, H, \rho)$ be a table. 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. Every block of π_X corresponds to a distinct value of $\mathsf{aDom}_\tau(X)$. Therefore, $|\pi_X| = |\mathsf{aDom}_\tau(X)|$. If U, V are two subsets of H such that $U \subseteq V$, then $\pi_V \leq \pi_U$.

Definition 31 Let $\tau = (T, H, \rho)$ be a table and let X, Y be sets of attributes, $X, Y \subseteq H$.

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

The *conditional f-entropy* of X on Y is $\mathcal{H}^f(X|Y) = \sum_{j=1}^m \frac{|C_j|}{|\rho|} \mathsf{imp}^f_{\pi_X}(C_j)$, 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 $\pi_\emptyset = \{\rho\}$, it follows that $\mathcal{H}^f(X) = \mathcal{H}^f(X|\emptyset)$ for every set of attributes X. Therefore, we have $\mathcal{H}^f(X) = \sum_{i=1}^n f\left(\frac{|B_i|}{|\rho|}\right)$, where $\pi_X = \{B_1, \ldots, B_n\}$. For the conditional entropy we have $\mathcal{H}^f(X|Y) = \frac{1}{|\rho|} \sum_{j=1}^m |C_j| \sum_{i=1}^n f\left(\frac{|B_i \cap C_j|}{|C_j|}\right)$. The concavity of the 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 $\mathcal{H}^f(X) \leq n f(\frac{1}{n})$, where n is the number of blocks of π_X, that is, the cardinality of the active domain of X.

Theorem 32 *Let X, X', Y, Y' be sets of attributes of the table $\tau = (T, H, \rho)$. For every 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 24, $\mathsf{IMP}^f_{\pi_X}(C) \leq \mathsf{IMP}^f_{\pi_{X'}}(C)$ for every set $C \subseteq \rho$. 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 22 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 33 *If X, X' are two sets of attributes of the table τ, 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 32. ∎

Theorem 34 *Let X, Y be two sets of attributes of a table $\tau = (T, H, \rho)$. 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

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

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

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

which, in turn, yields

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

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

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

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 35 *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 34. ∎

4 Clusters for Categorical Data

Let $\tau = (T, H, \rho)$ be a table, X be a set of attributes $X \subseteq H$ and assume that $\mathsf{aDom}_\tau(X)$ consists of N elements, $\mathsf{aDom}_\tau(X) = \{x_1, \ldots, x_N\}$. 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 $\frac{|B_i|}{|\rho|}$ for $1 \leq i \leq N$.

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

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

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 $d_{\tau,X}$ be the distance on $\mathsf{aDom}_\tau(X)$ defined by:

$$d^{0,1}_{\tau,X}(x,x') = \begin{cases} 0 & \text{if } x = x' \\ 1 & \text{otherwise} \end{cases}$$

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

Theorem 41 *The average distance $E(d^{0,1}_{\tau,X})$ equals $\mathcal{H}^{f_{gini}}(X)$.*

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

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

which gives the desired equality. ∎

Theorem 41 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 τ 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 42 Let $\tau = (T, H, \rho)$ be a table. An *(f,α)-clusterable set of attributes in τ* 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 records 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 23. 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 idenfifying those sets that can be included in `KAttrSets`.

Note that if $X \subseteq X'$, then, by Theorem 32, 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.

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

The inputs of the algorithm are the generator function f and the value α representing the limit of internal entropy.

There are `N` attributes in our dataset.

```
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;
  for each set of attributes S from Candidates do
    if H(S) <= alpha 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;
```

The external entropies of the sets of `MaxAttrSets` are also computed and we sort the maximal (f, α)-clusterable sets on their external f-entropies.

The algorithm was run on a database extracted from the "agaricus-lepiota" dataset from UCI (http://www.ics.uci.edu/AI/ML/Machine-Learning.html) by retaining the first 15 attributes of a table documenting 23 attributes for a number of 8124 types of mushrooms. We experimented using three different function generators for different values of α: f_{ent}, f_{gini}, and f_{sq}. The results are summarized below.

α	f_{gini}	
	time	sets
0.5	1	5
0.6	1	8
0.7	3	16
0.8	8	33
0.9	42	93

α	f_{gini}	
	time	sets
0.92	59	92
0.94	100	140
0.96	231	156
0.98	573	165

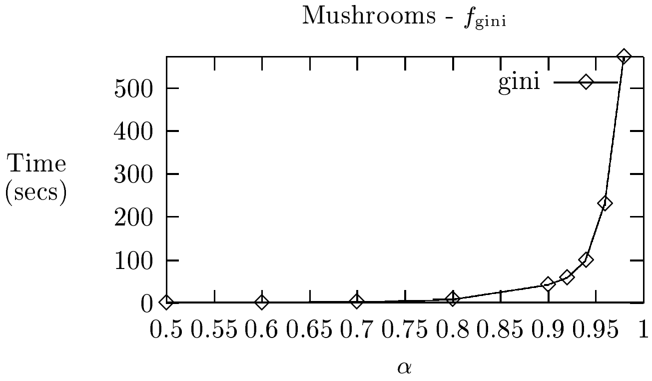
α	f_{ent}		f_{sq}	
	time	sets	time	sets
1	1	5	3	14
2	4	19	13	58
4	56	140	97	233
6	519	225	265	311
8	1541	110	527	311

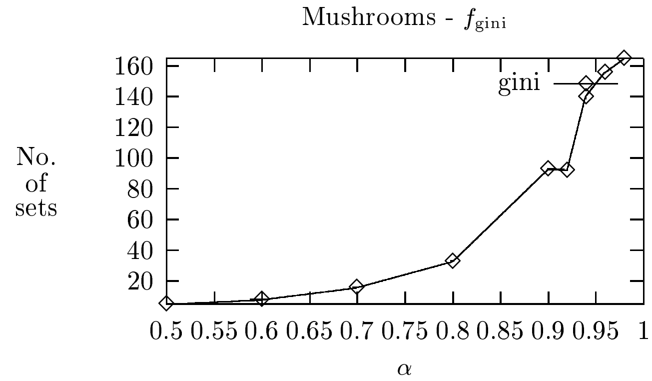
Elementary computations show that if $0 \leq p \leq 0.025$, then $f_{\mathrm{gini}}(p) \leq f_{\mathrm{ent}}(p) \leq f_{\mathrm{sq}}(p)$; for $0.025 < p \leq 0.39$ we have $f_{\mathrm{gini}}(p) \leq f_{\mathrm{sq}}(p) \leq f_{\mathrm{ent}}(p)$. Finally, if $0.39 < p \leq 1$, we have $f_{\mathrm{sq}}(p) \leq f_{\mathrm{gini}}(p) \leq f_{\mathrm{ent}}(p)$. 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 records. 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, and $\sqrt{k} - 1$ for the f_{sq}-entropy. Considering this, by specifying a limit for the entropy of an attribute, we can get an approximative idea of how many large blocks its partition will have.

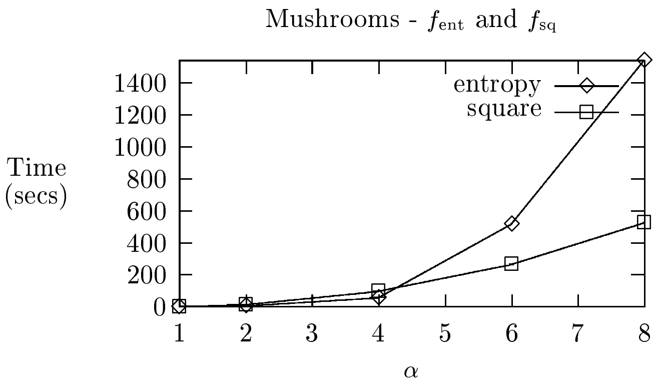
As an example of clusters determined by an (f, α)-clusterable set of attributes, consider the attribute set *bruises gill-attachment gill-spacing* obtained when we used f_{ent} with $\alpha = 2$, f_{gini} with $\alpha = 0.7$, and f_{sq} with $\alpha = 1$ and had a minimal external f_{sq}-entropy. The clusters formed by grouping on these attributes contain 210, 3330, 1208, 3272 and 104 records, respectively, as shown in the next table.

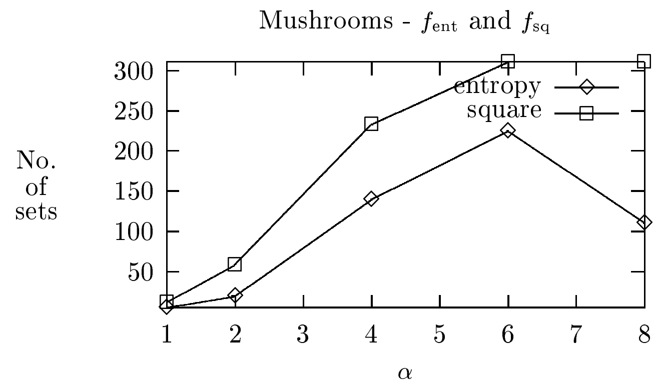
Set of attributes			Block B		
bruises	gill-attachment	gill-spacing	count	external clustering	$\mathsf{imp}^{f_{\mathrm{sq}}}_{\pi_{H-X}}(B)$
no	*attached*	*close*	210	12×16 tuples 18×1 tuple	3.554
no	*free*	*close*	3330	72×18 tuples 228×6 tuples 48×4 tuples 48×2 tuples 18×1 tuple	18.690
no	*free*	*crowded*	1208	216×4 tuples 144×2 tuples 56×1 tuple	18.899
yes	*free*	*close*	3272	144×12 tuples 120×8 tuples 72×4 tuples 148×2 tuples	19.830
yes	*free*	*crowded*	104	48×2 tuples 8×1 tuples	6.440

It is clear that the group with minimal impurity is obtained for *bruises* = 'no', *gill-attachment* = 'attached', and gill-spacing = 'closed'.


Mushrooms - f_{gini}
Time (secs)
gini
0
100
200
300
400
500
0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
α


Mushrooms - f_{gini}
No. of sets
gini
20
40
60
80
100
120
140
160
0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
α


Mushrooms - f_{ent} and f_{sq}
Time (secs)
entropy
square
0
200
400
600
800
1000
1200
1400
1 2 3 4 5 6 7 8
α


Mushrooms - f_{ent} and f_{sq}
No. of sets
entropy
square
50
100
150
200
250
300
1 2 3 4 5 6 7 8
α

5 Open Problems

Our technique for identifying clusters could be improved by performing an initial step of eliminating all attributes that have a high f-entropy. These attributes can be considered as "noise" for the purpose of clustering, so by ignoring them we would clean the data and would make the values of external entropy more meaningful.

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 accomodate the generalizations discussed in this paper.

References

[AGGR98] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering for high dimensional data for data mining applications. In *Proceedings of the 1998 ACM International Conference on Management of Data*, pages 94–105, 1998.

[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.

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