A Metric Approach to Building Decision Trees based on Goodman-Kruskal Association Index

Dan A. Simovici and Szymon Jaroszewicz

University of Massachusetts at Boston,
Department of Computer Science,
Boston, Massachusetts 02125, USA
{dsim,sj}@cs.umb.edu

Abstract. We introduce a numerical measure on sets of partitions of finite sets that is linked to the Goodman-Kruskal association index commonly used in statistics. This measure allows us to define a metric on such partions used for constructing decision trees. Experimental results suggest that by replacing the usual splitting criterion used in C4.5 by a metric criterion based on the Goodman-Kruskal coefficient it is possible, in most cases, to obtain smaller decision trees without sacrificing accuracy.
Keywords: Goodman-Kruskal association index, metric, partition, decision tree

1 Introduction

The construction of decision trees is centered around the selection algorithm of an attribute that generates a partition of the subset of the training data set that is located in the node about to be split. Over the years, several greedy techniques for choosing the splitting attribute have been proposed including the entropy gain and the gain ratio [1], the Gini index [2], the Kolmogorov-Smirnov metric [3, 4], or a metric derived from Shannon entropy [5]. In our previous work [6] we extended the metric splitting criterion introduced by L. de Mántaras by introducing metrics on the set of partitions of a finite set constructed by using generalized conditional entropy (which correspond to a generalization of entropy introduced by Daroczy [7]). This paper introduces a different type of metric on partitions of finite sets that is generated by a coefficient derived from the Goodman-Kruskal association index and shows that this metric can be applied succesfully to the construction of decision trees.
The purpose of this note is to define a metric on the set of partitions of a finite set that is derived from to the Goodman-Kruskal association index. A general framework of classification can be formulated starting with two finite random variables

$$X : \begin{pmatrix} a_1 & \cdots & a_l \\ p_1 & \cdots & p_l \end{pmatrix} \text{ and } Y : \begin{pmatrix} b_1 & \cdots & b_k \\ q_1 & \cdots & q_k \end{pmatrix}$$

We assume that we deal with a finite probability space where the elementary events are pairs of values (a_i, b_j), where a is a value of X and b_j

is a value of Y. The classification rule adopted here is that an elementary event is classified in the class that has the maximal probability. Thus, in the absence of any knowledge about X, an elementary event will be classified in the Y-class b_j if b_j corresponds to the highest value among the probabilities $P(Y = b_j)$ for $1 \leq j \leq k$. If $P(Y = b_j|X = a_i)$ is the probability of predicting the value b_j for Y when $X = a_i$, then an event that has the component $X = a_i$ will be classified in the Y-class b_j if j is the number for which $P(Y = b_j|X = a_i)$ has the largest value. The probability of misclassification committed by applying this rule is $1 - \max_{1\leq j\leq k} P(Y = b_j|X = a_i)$.
The original Goodman-Kruskal association index $\lambda_{Y|X}$ (see [8, 9]) is the relative reduction in the probability of prediction error:

$$\begin{array}{l}\lambda_{Y|X}\\ = 1 - \frac{\mathsf{GK}(X,Y)}{1-\max_{1\leq j\leq k} P(Y=b_j)}\\ = \frac{\sum_{i=1}^{l} P(X=a_i)\max_{1\leq j\leq k} P(Y=b_j|X=a_i) - \max_{1\leq j\leq k} P(Y=b_j)}{1-\max_{1\leq j\leq k} P(Y=b_j)}.\end{array}$$

In other words, $\lambda_{Y|X}$ is the proportion of the relative error in predicting the value of Y that can be eliminated by knowledge of the X-value.
The Goodman-Kruskal coefficient of X and Y that we use is defined by:

$$\begin{aligned}\mathsf{GK}(X,Y) &= \sum_{i=1}^{l} P(X = a_i)\left(1 - \max_{1\leq j\leq k} P(Y = b_j|X = a_i)\right)\\ &= 1 - \sum_{i=1}^{l} P(X = a_i) \max_{1\leq j\leq k} P(Y = b_j|X = a_i)\\ &= 1 - \sum_{i=1}^{l} \max_{1\leq j\leq k} P(Y = b_j \wedge X = a_i).\end{aligned}$$

Thus, $\mathsf{GK}(X, Y)$ is the expected value of the probability of misclassification. This coefficient is related to $\lambda_{Y|X}$ by:

$$G(X,Y) = (1 - \lambda_{Y|X})\left(1 - \max_{1\leq j\leq k} P(Y = b_j)\right)$$

Next, we formulate a definition of the Goodman-Kruskal coefficient GK within an algebraic setting, using partitions of finite sets. The advantage of this formulation is the possibility of using lattices of partitions of finite sets and various operations on partitions.
A partition of a set S is a collection of nonempty subsets of S, $\pi = \{B_i \mid i \in I\}$ such that $B_i \cap B_j = \emptyset$ for every $i, j \in I$ such that $i \neq j$ and $\bigcup_{i\in I} B_i = S$. Note that a partition $\pi = \{B_1, \ldots, B_l\}$ of a finite set S generates a finite random variable:

$$X : \begin{pmatrix} 1 & \cdots & l \\ p_1 & \cdots & p_l \end{pmatrix},$$

where $p_i = \frac{|B_i|}{|S|}$ for $1 \leq i \leq l$, and thus, the Goodman-Kruskal coefficient can be formulated in terms of partitions of finite sets.

If $\pi, \sigma \in \mathsf{PART}(S)$ we write $\pi \leq \sigma$ if each block of π is a subset of a block of σ. This is equivalent to saying that every block of σ is a union of blocks of π. We obtain a partial ordered set $(\mathsf{PART}(S), \leq)$. The least partition of S is the unit partition $\iota_S = \{\{a\} \mid a \in S\}$; the largest partition is the one-block partition $\omega_S = \{S\}$. The partial ordered set $(\mathsf{PART}(S), \leq)$ is a semi-modular lattice (see [10]), where $\inf\{\pi, \sigma\}$ is the partition $\pi \wedge \sigma$ whose blocks consist of intersections of blocks $B \cap C$, where $B \in \pi$ and $C \in \sigma$. Note that π is covered by σ (that is, $\pi < \sigma$ and there is no $\theta \in \mathsf{PART}(S)$ such that $\pi < \theta < \sigma$) if and only if σ is obtained from π by fusing together two blocks of π.
The *trace of a partition* $\pi = \{B_1, \ldots, B_k\}$ from $\mathsf{PART}(S)$ on a subset R of S is the partition $\pi_R \in \mathsf{PART}(R)$ given by $\pi_R = \{B_1 \cap R, \ldots, B_l \cap R\}$.
If S, T are two disjoint sets and $\pi \in \mathsf{PART}(S), \sigma \in \mathsf{PART}(T)$, then we denote by $\pi + \sigma$ the partition of $S \cup T$ that consists of all the blocks of π and σ. It is easy to see that "+" is an associative partial operation.

Definition 1. *Let $\pi = \{B_1, \ldots, B_k\}$ and $\sigma = \{C_1, \ldots, C_l\}$ be two partitions of a set S. The Goodman-Kruskal coefficient of π and σ is the number:*

$$\mathsf{GK}(\pi, \sigma) = 1 - \frac{1}{|S|} \sum_{i=1}^{k} \max_{1 \leq j \leq l} |B_i \cap C_j|.$$

Decision trees are built from data that has a tabular structure common in relational databases. As it is common in the relational terminology (see [11], for example), we regard a table as a triple $\tau = (T, H, \rho)$, where T is a string that gives the name of the table, $H = \{A_1, \ldots, A_n\}$ is a finite set of symbols (called the attributes of τ), and ρ is a relation, $\rho \subseteq \mathrm{Dom}(A_1) \times \cdots \times \mathrm{Dom}(A_n)$. Here $\mathrm{Dom}(A_i)$ is the domain of the attribute A_i for $1 \leq i \leq n$.
A set of attributes $L \subseteq H$ determines a partion π^L on the relation ρ, that is, on the set of tuples of the table τ, where two tuples belong to the same block if they have equal projections on L. It is easy to see that if L, K are two sets of attributes, then $\pi^{LK} = \pi^L \wedge \pi^K$.
The classical technique for building decision trees is using the entropy gain ratio as a criterion for choosing for every internal node of the tree the splitting attribute that maximizes this ratio (see [1]). The construction has an inductive character. If $\tau = (T, H, \rho)$ is the data set used to build the decision tree $\mathcal{T}$, let v be a node of $\mathcal{T}$ that is about to be split and let ρ_v be the set of tuples that corresponds to v. Suppose that the target partition of the data set ρ is θ. Then, the trace of this partition on ρ_v is θ_{ρ_v}.
Choosing the splitting attribute for a node v of a decision tree $\mathcal{T}$ for τ based on the minimal value of $\mathsf{GK}(\pi_v^A, \theta_{\rho_v})$ alone does not yield decision trees with good accuracy. A lucid discussion of those issues can be found in [12, 4]. However, we will show that the GK coefficient can be used to define a metric on the set of partitions of a finite set that can be succesfully used for choosing splitting attributes. The decision trees that result are smaller, have fewer leaves (and therefore, less fragmentation) compared with trees built by using the standard gain ratio criterion; also, they have comparable accuracy.

2 The Goodman-Kruskal metric space

The main result of this section is a construction of a metric d_{GK} on the set of partitions of a finite set that is related to the Goodman-Kruskal coefficient and can be used for constructing decision trees. To introduce this metric we need to establish several properties of GK. Unless we state otherwise, all sets considered here are finite.

Theorem 1. *Let S be a set and let $\pi, \sigma \in \mathsf{PART}(S)$. We have $\mathsf{GK}(\pi, \sigma) = 0$ if and only if $\pi \leq \sigma$.*

Proof. It is immediate that $\pi \leq \sigma$ implies $\mathsf{GK}(\pi, \sigma) = 0$. Conversely, if $\mathsf{GK}(\pi, \sigma) = 0$, then $\sum_{i=1}^{k} \max_{1 \leq j \leq l} |B_i \cap C_j| = |S|$, which means that for each block B_i of π, there is a block C_j such that $|B_i \cap C_j| = |B_i|$. This is possible only if $B_i \subseteq C_j$, that is, if $\pi \leq \sigma$, which gives the desired conclusion.

Theorem 2. *The function GK is monotonic in the first argument and dually monotonic in the second argument.*

Proof. To prove the first part of the statement let $\pi = \{B_1, \ldots, B_k\}$, $\pi' = \{B'_1, \ldots, B'_m\}$, and $\sigma = \{C_1, \ldots, C_l\}$ be three partitions of S such that $\pi \leq \pi'$. Then, for every block B'_r of π' there is a collection of blocks of π: $B_{i_1}, \ldots, B_{i_s}$ such that $B'_r = B_{i_1} \cup \cdots \cup B_{i_s}$. Consequently for every m, $1 \leq m \leq l$ we can write:

$$\begin{aligned} |B'_r \cap C_m| &= |B_{i_1} \cap C_m| + \cdots + |B_{i_s} \cap C_m| \\ &\leq \max_{1 \leq j \leq l} |B_{i_1} \cap C_j| + \cdots + \max_{1 \leq j \leq l} |B_{i_s} \cap C_j|. \end{aligned}$$

Thus, we obtain:

$$\max_{1 \leq j \leq l} |B'_r \cap C_j| \leq \max_{1 \leq j \leq l} |B_{i_1} \cap C_j| + \cdots + \max_{1 \leq j \leq l} |B_{i_s} \cap C_j|,$$

which implies $\mathsf{GK}(\pi, \sigma) \geq \mathsf{GK}(\pi', \sigma)$.
To prove the second part, let σ, σ' be two partitions such that $\sigma \leq \sigma'$. We show that $\mathsf{GK}(\pi, \sigma) \geq \mathsf{GK}(\pi, \sigma')$. It suffices to show that σ' covers σ, that is, $\sigma = \{C_1, \ldots, C_{l-2}, C_{l-1}, C_l\}$ and $\sigma' = \{C_1, \ldots, C_{l-2}, C_{l-1} \cup C_l\}$. In other words, the blocks of σ' coincide with the blocks of σ with the exception of one block that is obtained by fusing two blocks of σ. Note that for a given block B_i of π we have:

$$\max_{1 \leq j \leq l} |B_i \cap C_j| \leq \max\{\max_{1 \leq j \leq l-2} |B_i \cap C_j|, |B_i \cap (C_{l-1} \cup C_l)|\},$$

which implies $\mathsf{GK}(\pi, \sigma) \geq \mathsf{GK}(\pi, \sigma')$. □
The next result has a technical character:

Theorem 3. *For every three partitions θ, π, σ of a finite set S we have:*

$$\mathsf{GK}(\pi \wedge \theta, \sigma) + \mathsf{GK}(\theta, \pi) \geq \mathsf{GK}(\theta, \pi \wedge \sigma).$$

Proof. See Appendix A.

Theorem 4. *Let θ, π, σ be partitions of a set S. We have*

$$GK(\theta, \pi) + GK(\pi, \sigma) \geq GK(\theta, \sigma).$$

Proof. Note that

$$\mathsf{GK}(\theta, \pi) + \mathsf{GK}(\pi, \sigma) \geq \mathsf{GK}(\theta, \pi) + \mathsf{GK}(\pi \wedge \theta, \sigma)$$

due to the monotonicity of GK in its first argument. By Theorem 3

$$\mathsf{GK}(\theta, \pi) + \mathsf{GK}(\pi, \sigma) \geq \mathsf{GK}(\theta, \pi \wedge \sigma) \geq \mathsf{GK}(\theta, \sigma),$$

because of the dual monotonicity of GK in its second argument. □

Corollary 1. *The mapping $d_{GK} : \mathsf{PART}(S) \times \mathsf{PART}(S) \longrightarrow \mathbb{R}$ given by:*

$$d_{GK}(\pi, \sigma) = GK(\pi, \sigma) + GK(\sigma, \pi)$$

for $\pi, \sigma \in \mathsf{PART}(S)$, is a metric on the set $\mathsf{PART}(S)$.

Proof. By Theorem 1 we have $d_{GK}(\pi, \sigma) = 0$ if and only if $\pi = \sigma$. Also, the definition of d_{GK} implies $d_{GK}(\pi, \sigma) = d_{GK}(\sigma, \pi)$ for every $\pi, \sigma \in \mathsf{PART}(S)$.
Finally, the triangular inequality $d_{GK}(\pi, \sigma) + d_{GK}(\sigma, \theta) \geq d_{GK}(\pi, \theta)$ for $\pi, \sigma, \theta \in \mathsf{PART}(S)$ follows immediately from Theorem 4. □

3 The Goodman-Kruskal Splitting Criterion for Decision Trees

Let $\tau = (T, H, \rho)$ be the table that contains the training data set that is used to build a decision tree $\mathcal{T}$. Assume that we are about to expand the node v of the tree $\mathcal{T}$. Using the notations introduced in Section 1, we choose to split the note v using an attribute A_i that minimizes the distance $d_{GK}(\pi_{\rho_v}^{A_i}, \theta_{\rho_v})$.
The d_{GK} metric does not favor attributes with large domains as splitting attributes, an issue that is important for building decision trees.

Theorem 5. *Let S be a finite set and let $\pi, \pi', \sigma \in \mathsf{PART}(S)$ be such that $\pi' \leq \pi$. If there exists a block C of σ and a block B of π such that $B \subseteq C$, then $d_{GK}(\pi, \sigma) \leq d_{GK}(\pi', \sigma)$.*

Proof. We can assume, without restricting generality, that π' is covered by π, that is, $\pi = \{B_1, \ldots, B_k\}$, $B = B_k$, $\pi' = \{B_1, \ldots, B'_k, B''_k\}$, where $B_k = B'_k \cup B''_k$. Also, let $\sigma = \{C_1, \ldots, C_l\}$, where $C_l = C$.
Theorem 2 implies that $\mathsf{GK}(\sigma, \pi') \leq GK(\sigma, \pi)$ (due to the dual monotonicity in the second argument of GK). We prove that, under the assumptions made in the theorem, we have $\mathsf{GK}(\pi', \sigma) = \mathsf{GK}(\pi, \sigma)$, which implies the desired inequality. Indeed, note that:

$$\begin{aligned}
&\mathsf{GK}(\pi', \sigma) \\
&= 1 - \frac{1}{|S|}\left(\sum_{i=1}^{k-1} \max_{1 \leq j \leq l} |B_i \cap C_j| + \max_{1 \leq j \leq l} |B'_k \cap C_j| + \max_{1 \leq j \leq l} |B''_k \cap C_j|\right) \\
&= 1 - \frac{1}{|S|}\left(\sum_{i=1}^{k-1} \max_{1 \leq j \leq l} |B_i \cap C_j| + |B'_k| + |B''_k|\right) \\
&= 1 - \frac{1}{|S|}\sum_{i=1}^{k} \max_{1 \leq j \leq l} |B_i \cap C_j| = \mathsf{GK}(\pi, \sigma)
\end{aligned}$$

because $B'_k, B''_k \subseteq B_k \subseteq C$. □

We note that the Theorem 5 is similar to the property of the metric generated by the Shannon entropy obtained by L. de Mántaras in [5] and generalized by us in [6].

Next, we compare parameters of decision trees constructed on UCI machine learning datasets [13] by using Entropy Gain Ratio and the Goodman-Kruskal distance d_{GK}. The experiments have been conducted using the J48 (a variant of C4.5) algorithm from the Weka Package [14], modified to use different splitting criteria. The pruning steps of decision tree construction are left unchanged. To verify the accuracy, we used 5-fold cross-validation. For each splitting criterion we present three characteristics of the generated trees: accuracy (percentage of correctly predicted cases), size of the tree (total number of nodes) and the number of leaves in the tree. All are averaged over the 5-fold of cross-validation.

Overall d_{GK} produced smaller trees for 24 out of 33 datasets considered. In 4 cases (`anneal.ORIG, clev.-14-heart-disease, sick, vote`) over 50% reduction was achieved. In one case (`pima-diabetes`) a sharp increase was observed. On average the trees obtained were 10% smaller.

The accuracy of trees constructed using d_{GK} was on average 1.67% worse than that of trees constructed using standard Weka version. In one case (`autos`) the decrease was significant but for all other cases it was rather moderate, and in a few cases d_{GK} produced more accurate trees. Small tree size is an advantage since, in general, small trees are much easier to understand. The total number of nodes and the number of leaves in the tress were highly correlated so we can talk simply about size of the tree.

Experimental Results: Entropy Gain Ratio vs. d_{GK}

	Entropy Gain Ratio			d_{GK}		
Dataset	acc	tree size	no. of leafs	acc	tree size	no. of leafs
anneal	98.55	46.4	36.2	98.55	37.2	26.4
anneal.ORIG	90.20	63	42.6	86.30	28.2	17.8
audiology	78.76	46	29	77.41	37.4	24
autos	80	64.6	48.2	67.80	49.6	27.6
balance-scale	78.4	73.8	37.4	77.76	57	29
breast-cancer	73.09	21.2	17.2	73.78	18	13.4
wisc-breast-cancer	94.12	17.4	7.2	94.85	17	9
horse-colic	85.85	8.4	5.8	81.78	7.6	4.4
credit-rating	86.23	29.2	20.8	83.91	20.4	11.6
german-credit	72.9	108	77.6	69.5	63.4	36.8
pima-diabetes	75.65	42.6	21.8	70.96	88.6	44.8
Glass	67.26	39.4	20.2	70.09	33.4	17.2
clev.-14-heart-disease	77.53	41.4	10.4	75.89	16.4	9
hung.-14-heart-disease	78.57	9.8	6.4	80.28	10	6.2
heart-statlog	75.55	26.6	13.8	71.85	17.4	9.2
hepatitis	78.06	13.4	7.2	82.58	9	5
hypothyroid	99.46	25.8	13.4	99.39	21	11
ionosphere	89.73	25.8	13.4	88.89	16.2	8.6
iris	95.33	8.2	4.6	95.33	6.6	3.8
kr-vs-kp	99.15	51.8	27.4	98.46	76.4	39.8
labor	78.63	6.8	4	84.09	3	2
lymphography	80.41	24.4	14.8	79.01	14.8	8.8
mushroom	100	29.4	24.4	100	31.8	25
primary-tumor	40.99	77	41.2	43.64	38.8	21.4
segment	97.09	81.8	41.4	94.02	67	34
sick	98.75	42.6	23.6	98.35	18.2	10.8
sonar	74.03	23.8	12.4	69.16	29.4	15.2
soybean	91.21	89.4	58.4	90.19	105.2	71.2
splice	94.04	199.6	160.8	93.51	194.4	156.6
vehicle	72.10	117.8	59.4	65.60	128.2	64.6
vote	96.55	11	6	94.71	4.6	2.8
vowel	78.18	200.4	120.2	63.43	235	125.8
zoo	93.09	14.6	7.8	93.09	14.6	7.8
average	83.92	50.95	31.84	82.25	45.93	27.29

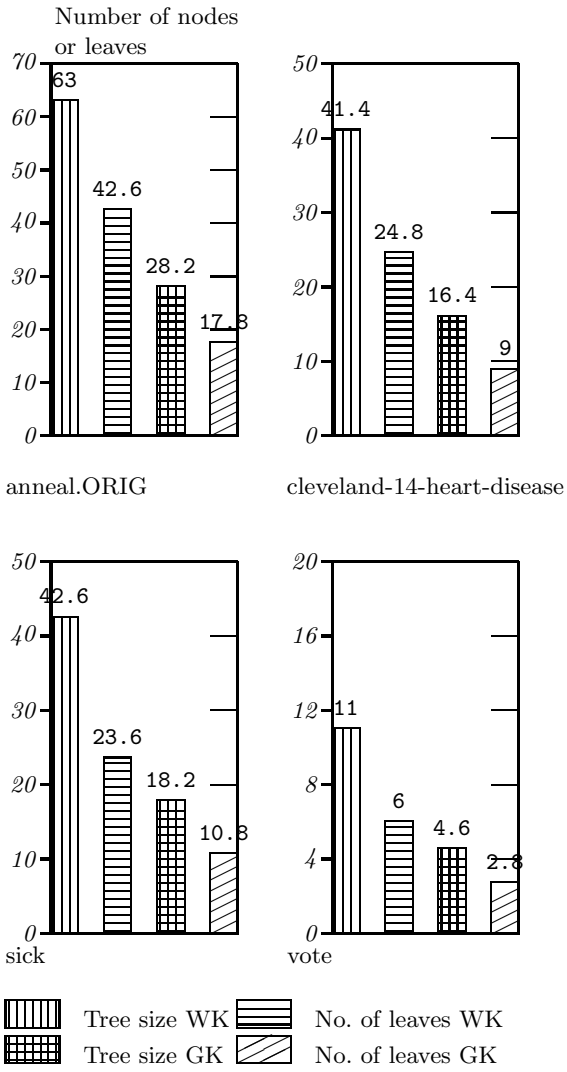

Fig. 1. Comparative Experimental Results

The best results obtained from experiments are also shown in Figure 1. Splitting nodes by using an attribute A that minimizes $\mathsf{GK}(\pi^A_{\rho_v}, \theta_{\rho_v})$ instead of $d_{GK}(\pi^A_{\rho_v}, \theta_{\rho_v})$ may result in a substantial loss of accuracy. For example, in the case of the `hungarian-14-heart-disease` dataset, the accuracy obtained using GK, under comparable conditions (averaging over 5-fold cross validation) is just 70.05% compared to 78.57% obtained by using the entropy gain ratio, or 80.28% obtained in the case of d_{GK}. This confirms the claim in the literature of the unsuitability of using $\mathsf{GK}(\pi^A_{\rho_v}, \theta_{\rho_v})$ alone as a splitting criterion.

A Proof of Theorem 3

We begin by showing that if $S_1, \ldots, S_n$ are pairwise disjoint sets, and $\pi_r, \sigma_r \in \mathsf{PART}(S_r)$ for $1 \leq r \leq n$, then

$$\mathsf{GK}(\pi_1 + \cdots + \pi_n, \sigma_1, \ldots, \sigma_n) = \sum_{r=1}^{n} \frac{|S_r|}{|S|} \mathsf{GK}(\pi_r, \sigma_r). \tag{1}$$

Let $\pi_p = \{B_1^p, \ldots, B_{l_p}^p\}$ and $\sigma_q = \{C_1^q, \ldots, C_{k_q}^q\}$ for $1 \leq p, q \leq n$. Then, we can write:

$$\begin{aligned}
&\mathsf{GK}(\pi_1 + \cdots + \pi_n, \sigma_1 + \cdots + \sigma_n) \\
&= 1 - \frac{1}{|S|} \sum_{p,i} \max_{q,j} |B_i^p \cap C_j^q| \\
&= 1 - \frac{1}{|S|} \sum_{p,i} \max_{p,j} |B_i^p \cap C_j^p| \\
&\quad \text{(because } p \neq q \text{ implies } B_i^p \cap C_j^q = \emptyset\text{)} \\
&= \sum_{p=1}^{n} \frac{|S_p|}{|S|} \left(1 - \sum_{p=1}^{n} \frac{1}{|S_p|} \sum_{i=1}^{l_p} \max_{1 \leq j \leq k_p} |B_i^p \cap C_j^p| \right) \\
&= \sum_{p=1}^{n} \frac{|S_p|}{|S|} \mathsf{GK}(\pi_p, \sigma_p),
\end{aligned}$$

which is the desired equality.
Let now $\mathcal{K}(\sigma)$ be the number:

$$\mathcal{K}(\sigma) = \mathsf{GK}(\omega_S, \sigma) = 1 - \frac{1}{|S|} \max_{1 \leq j \leq k} |C_j|.$$

We claim that if $\pi, \sigma \in \mathsf{PART}(S)$, then:

$$\mathsf{GK}(\pi, \sigma) \geq \mathcal{K}(\pi \wedge \sigma) - \mathcal{K}(\pi). \tag{2}$$

Let $\pi = \{B_1, \ldots, B_k\}$ and $\sigma = \{C_1, \ldots, C_l\}$. We can write:

$$\begin{aligned}
\mathsf{GK}(\pi, \sigma) &= |S| - \sum_{i=1}^{k} \max_{1 \le j \le l} |B_i \cap C_k| \\
&= \sum_{i=1}^{k} (|B_i| - \max_{1 \le j \le l} |B_i \cap C_j| \\
&\geq \max_{1 \le i \le k} (|B_i| - \max_{1 \le j \le l} |B_i \cap C_j|) \\
&\geq \max_{1 \le i \le k} |B_i| - \max_{1 \le i \le k, 1 \le j \le l} |B_i \cap C_j| \\
&= \mathcal{K}(\pi \wedge \sigma) - \mathcal{K}(\pi),
\end{aligned}$$

which proves the inequality (2).
Let $\pi = \{B_1, \ldots, B_k\}$, $\theta = \{D_1, \ldots, D_m\}$ and $\sigma = \{C_1, \ldots, C_l\}$. We have:

$$\pi \wedge \theta = \pi_{D_1} + \cdots + \pi_{D_m} = \theta_{B_1} + \cdots + \theta_{B_k}.$$

Consequently, by Equality (1), we have:

$$\begin{aligned}
\mathsf{GK}(\pi \wedge \theta, \sigma) &= \mathsf{GK}(\pi_{D_1} + \cdots + \pi_{D_m}, \sigma) \\
&= \sum_{h=1}^{m} \frac{|D_h|}{|S|} \mathsf{GK}(\pi_{D_h}, \sigma_{D_h}).
\end{aligned}$$

Also, we have

$$\mathsf{GK}(\theta, \pi) = \sum_{h=1}^{m} \frac{|D_h|}{|S|} \mathcal{K}(\pi_{D_h}),$$

which implies

$$\mathsf{GK}(\pi \wedge \theta, \sigma) + \mathsf{GK}(\theta, \pi) = \sum_{h=1}^{m} \frac{|D_h|}{|S|} \left(\mathsf{GK}(\pi_{D_h}, \sigma_{D_h}) + \mathcal{K}(\pi_{D_h})\right).$$

The Inequality (2) implies:

$$\mathsf{GK}(\pi_{D_h}, \sigma_{D_h}) + \mathcal{K}(\pi_{D_h}) \geq \mathcal{K}(\pi_{D_h} \wedge \sigma_{D_h}) = \mathcal{K}((\pi \wedge \sigma)_{D_h}),$$

so we may conclude that:

$$\mathsf{GK}(\pi \wedge \theta, \sigma) + \mathsf{GK}(\theta, \pi) \geq \sum_{h=1}^{m} \frac{|D_h|}{|S|} \mathcal{K}((\pi \wedge \sigma)_{D_h}) = \mathsf{GK}(\theta, \pi \wedge \sigma).$$

□

References

1. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA (1993)
2. Breiman, L., Friedman, J.H., Ohlsen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman & Hall/CRC, Boca Raton (1984) Republished 1993.

3. Utgoff, P.E.: Decision tree induction based on efficient tree restructuring. Technical Report 95-18, University of Massachusetts, Amherst (1995)
4. Utgoff, P.E., Clouse, J.A.: A Kolmogorov-Smirnoff metric for decision tree induction. Technical Report 96-3, University of Massachusetts, Amherst (1996)
5. de Mántaras, R.L.: A distance-based attribute selection measure for decision tree induction. Machine Learning **6** (1991) 81–92
6. Simovici, D.A., Jaroszewicz, S.: Generalized conditional entropy and decision trees. In: Proceedings of EGC 2003, Lyon, France (2003) 369–380
7. Daróczy, Z.: Generalized information functions. Information and Control **16** (1970) 36–51
8. Goodman, L.A., Kruskal, W.H.: Measures of Association for Cross-Classification. Volume 1. New York, Springer-Verlag (1980)
9. Liebtrau, A.M.: Measures of Association. SAGE, Beverly Hills, CA (1983)
10. Grätzer, G.: General Lattice Theory. Second edn. Birkhäuser, Basel (1998)
11. Simovici, D.A., Tenney, R.L.: Relational Database Systems. Academic Press, New York (1995)
12. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman and Hall, Boca Raton (1998)
13. Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences, http://www.ics.uci.edu/∼mlearn/MLRepository.html (1998)
14. Witten, I.H., Frank, E.: Data Mining. Morgan-Kaufmann, San Francisco (2000)