Metric Methods in Data Mining

Dan Simovici

University of Massachusetts at Boston,
Department of Computer Science,
Boston, Massachusetts 02125, USA

Where to find it?

At www.cs.umb.edu/∼dsim

Content

- Preliminaries
- Classification in Data Mining
- A Metric Approach to Incremental Clustering
- The Goodman-Kruskal Association Index
- A Metric Approach to Discretization
- Conclusions and Future Work

Preliminaries

Dissimilarities and Metrics

Let S be a set (patients, phenotypes, clinics, etc).
Dissimilarity: a function

$$d : S \times S \longrightarrow \mathbb{R}_+$$

such that $d(p, q) = 0$ if and only if $p = q$.

$d(p, q)$ measures the dissimilarity between two objects; if $d(p, r) < d(p, q)$ this means that r resembles more to p then q does.

If d is a dissimilarity on S such that

$$d(p,q) = d(q,p)$$

for every objects p, q, then d is a symmetric dissimilarity.

A metric on the set S is a symmetric dissimilarity that satisfies the triangular inequality:

$$d(p,q) + d(q,r) \geq d(p,r)$$

for every $p, q, r \in S$.

Why the triangular axiom?

If the triangular axiom is violated we may have the paradoxical situation of three objects u, v, w such that:

$$d(u, w) + d(v, w) < d(u, v).$$

- both u and v are similar to w, but
- u and v are very different from each other!

Examples

- Standard distance on real line:

$$d(p, q) = |p - q|$$

- Minkowski distance in $\mathbb{R}^n$:

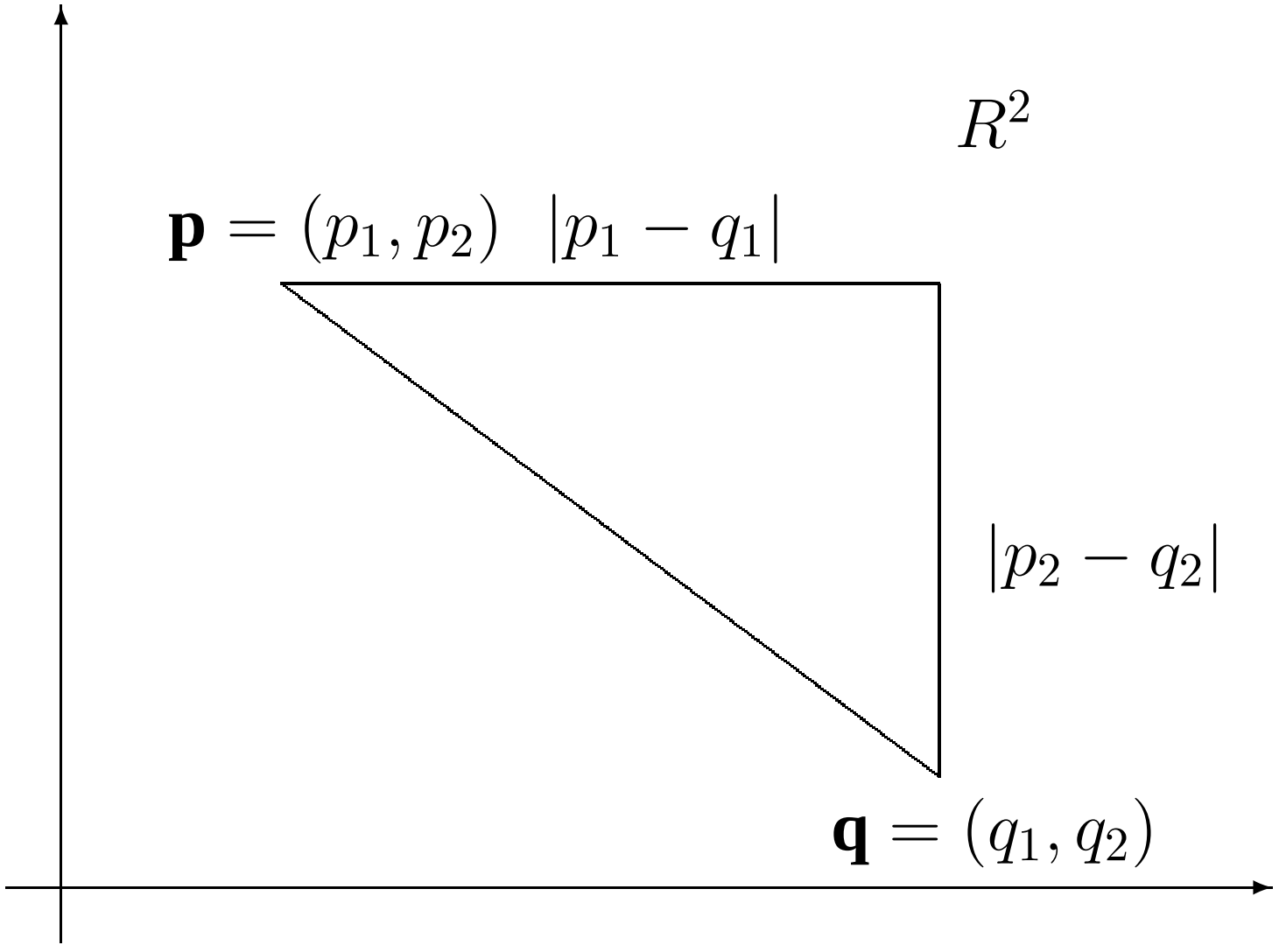
$$d_k(\mathbf{p}, \mathbf{q}) = \left(\sum_{i=1}^{n} |p_i - q_i|^k \right)^{\frac{1}{k}}$$

for $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n) \in \mathbb{R}^n$.

Examples

In $\mathbb{R}^2$:

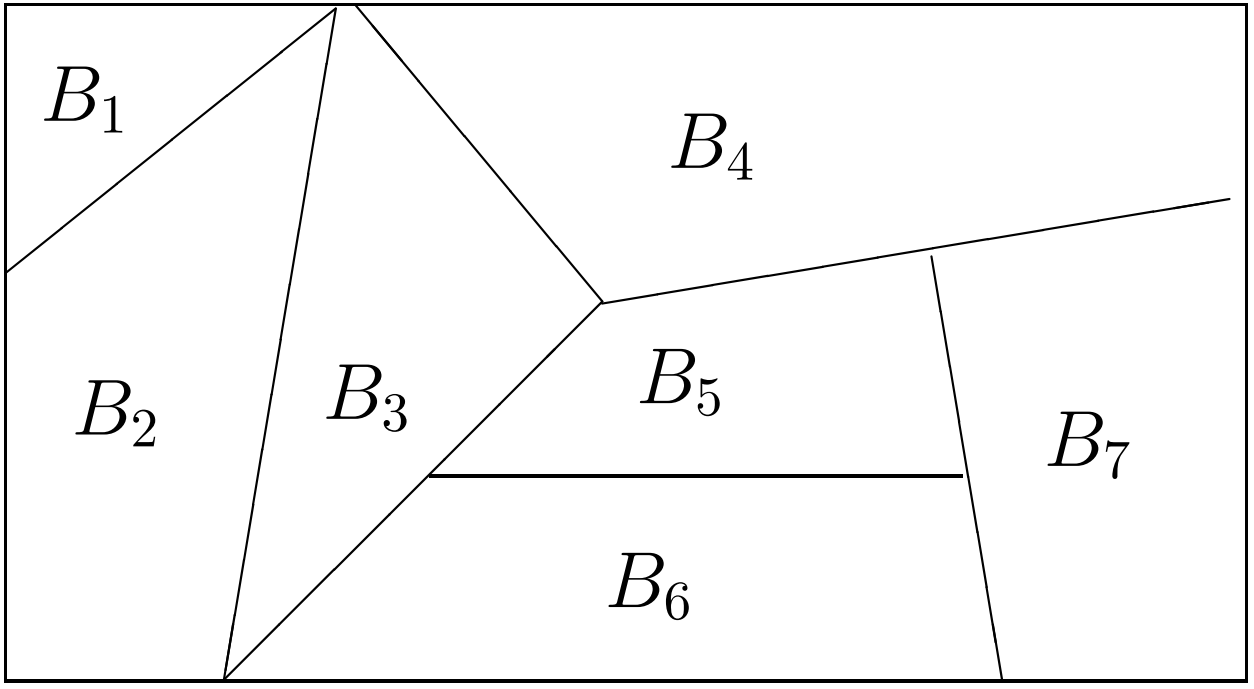
$$
\begin{aligned}
d_1(\mathbf{p},\mathbf{q}) &= |p_1 - q_1| + |p_2 - q_2| \\
&\quad \text{(Manhattan distance)} \\
d_2(\mathbf{p},\mathbf{q}) &= \sqrt{|p_1 - q_1|^2 + |p_2 - q_2|^2} \\
&\quad \text{(Euclidean distance)} \\
d_\infty(\mathbf{p},\mathbf{q}) &= \lim_{k\to\infty} d_k(\mathbf{p},\mathbf{q}) \\
&= \max\{|p_1 - q_1|, |p_2 - q_2|\} \\
&\quad \text{(Canberra distance)}
\end{aligned}
$$


R^2
$\mathbf{p} = (p_1, p_2)$
$|p_1 - q_1|$
$|p_2 - q_2|$
$\mathbf{q} = (q_1, q_2)$

Partitions

PART(S): set of partitions of set S

Partition $\pi = \{B_1, \ldots, B_7\}$

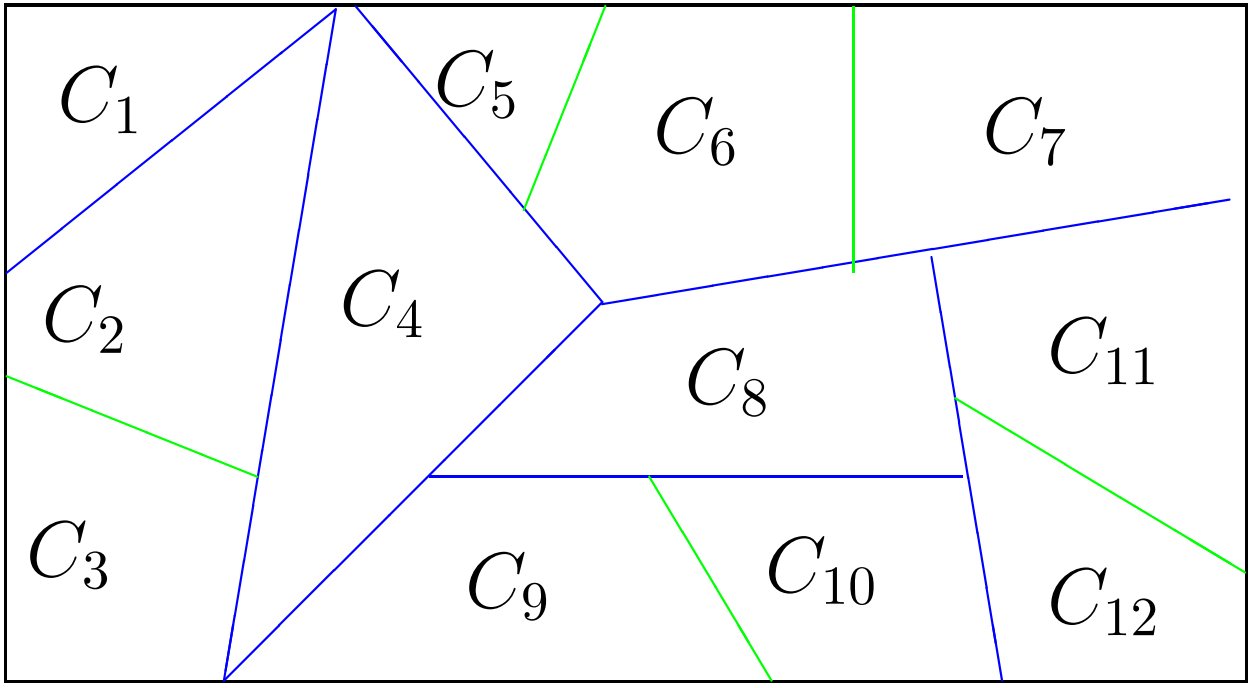


$B_1, \ldots, B_7$ are the blocks of π

Partitions Partial Order

$\sigma \leq \pi$ if each block C of σ is included in a block of π.

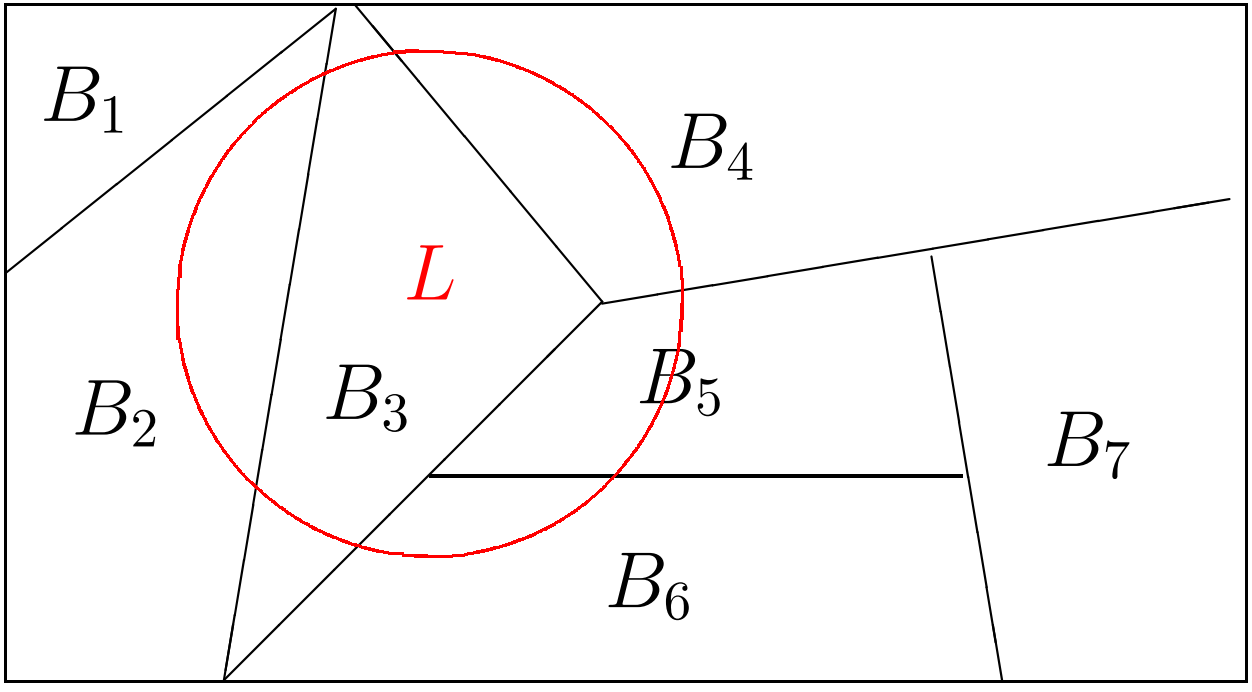
Partition $\sigma = \{C_1, \ldots, C_{12}\} \leq \pi$



Let $L \subseteq S$ and $\pi = \{B_1, \ldots, B_n\}$. The *trace of the partition* π on L is:

$$\pi_L = \{B_i \cap L \,|\, 1 \leq i \leq k \text{ and } B_i \cap L \neq \emptyset\}.$$

Trace of partition $\pi = \{B_1, \ldots, B_7\}$ on set L



Key Issue for Data Mining:

Defining Dissimilarities and Metrics for Partitions

Shannon's Entropy

For random variables...

The Shannon entropy is introduced for a random variable distribution

$$X : \begin{pmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p_n \end{pmatrix}$$

is $\mathcal{H}(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$.

Shannon entropy

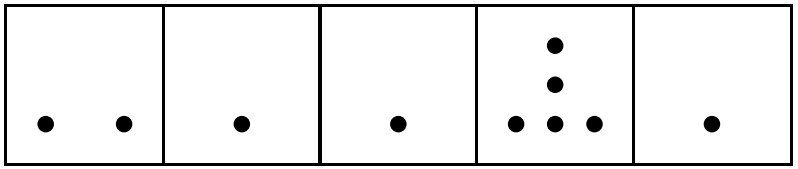
... for partitions

A partition $\pi = \{B_1, \ldots, B_m\}$ on a finite, nonempty set A generates naturally a random variable:

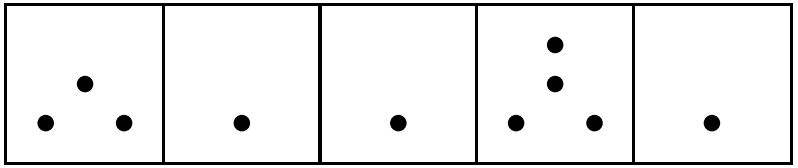
$$X_\pi : \begin{pmatrix} B_1 & \cdots & B_m \\ \frac{|B_1|}{|S|} & \cdots & \frac{|B_m|}{|S|} \end{pmatrix}$$

We define the Shannon entropy of π as the Shannon entropy of X_π.

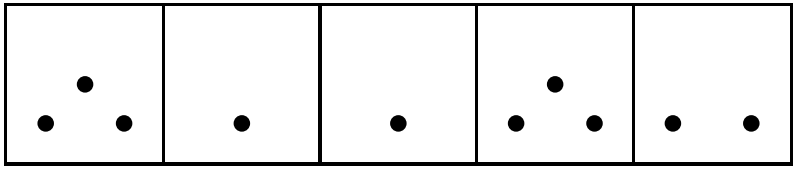
Measuring concentration of values



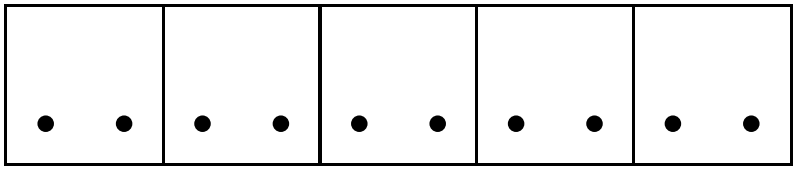
$\mathcal{H}_1(\pi_4) = 1.9609$



$\mathcal{H}_1(\pi_3) = 2.0464$



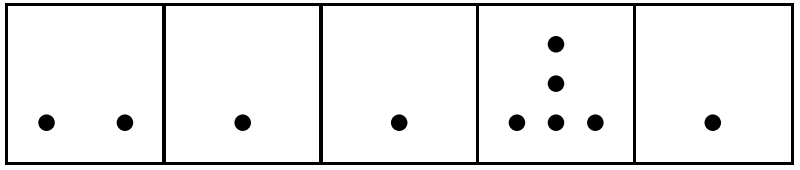
$\mathcal{H}_1(\pi_2) = 2.1709$



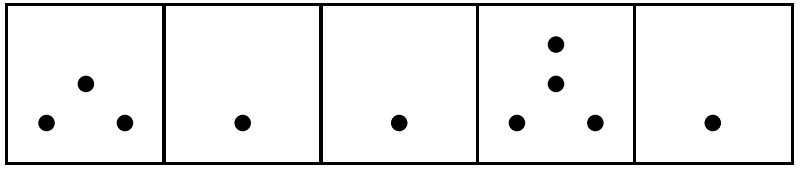
$\mathcal{H}_1(\pi_1) = 2.3219$

Gini's Index

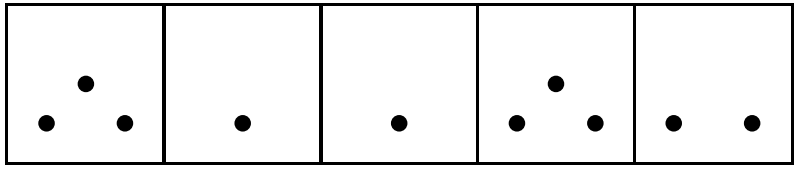
$$\mathcal{H}_2(\pi) = 1 - \sum_{i=1}^{n} p_i^2$$



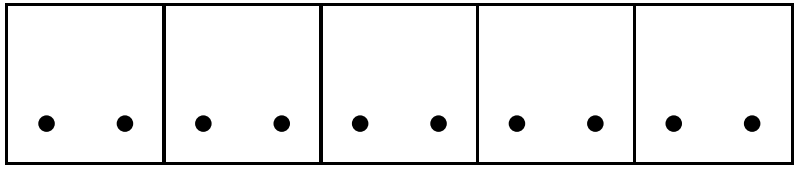
$$\mathcal{H}_1(\pi_4) = 0.68$$



$$\mathcal{H}_1(\pi_3) = 0.72$$



$$\mathcal{H}_1(\pi_2) = 0.79$$



$$\mathcal{H}_1(\pi_1) = 0.80$$

Generalized Entropy of Partitions

Daróczy's β-generalized entropy of $\pi = \{B_1, \ldots, B_n\}$:

$$\mathcal{H}_\beta(\pi) = \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_{i=1}^{n} \left(\frac{|B_i|}{|S|} \right)^\beta \right).$$

For $\beta = 2$ we obtain the Gini index. Also, $\lim_{\beta \to 1} \mathcal{H}_\beta(\pi)$ is Shannon's entropy

$$\mathcal{H}(\pi) = -\sum_{i=1}^{n} \frac{|B_i|}{|S|} \log_2 \frac{|B_i|}{|S|}$$

Set Purity and Entropy

$\mathcal{H}(\pi_L)$ measures the impurity of the set L relative to the partition π: the larger the entropy, the more L is scattered among the blocks of π.

If $\pi, \sigma \in \mathsf{PART}(S)$, the average impurity of the blocks of σ relative to π is the *conditional entropy of π relative to σ*:

$$\mathcal{H}(\pi|\sigma) = \sum_{j=1}^{m} \frac{|Q_j|}{|S|} \mathcal{H}(\pi_{Q_j}),$$

where $\sigma = \{Q_1, \ldots, Q_m\}$.

Generalized Conditional Entropy

For $\pi, \sigma \in \mathsf{PART}(S)$ such that

$$\pi = \{P_1, \ldots, P_k\}$$
$$\sigma = \{Q_1, \ldots, Q_m\}$$

the conditional β-entropy $\mathcal{H}_\beta(\pi|\sigma)$ is:

$$\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{m} \left(\frac{|Q_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{Q_j})$$
$$= \frac{1}{(2^{1-\beta}-1)|S|^\beta} \left(\sum_{i=1}^{k}\sum_{j=1}^{m} |P_i \cap Q_j|^\beta - \sum_{j=1}^{m} |Q_j|^\beta\right)$$

Properties of Conditional Entropy

- $\mathcal{H}_\beta(\pi|\sigma) = 0$ if and only if $\sigma \leq \pi$;
- $\mathcal{H}_\beta(\pi \wedge \sigma) = \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma) = \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\pi)$;
- $\mathcal{H}_\beta(\pi|\sigma) = 0$ is dually monotonic with respect to π and is monotonic with respect to σ.

Metrics on Partition Sets

López de Mántaras:

$$d(\pi,\sigma) = \mathcal{H}(\pi|\sigma) + \mathcal{H}(\sigma|\pi)$$

Simovici and Jaroszewicz:

$$\begin{aligned} d_\beta(\pi,\sigma) &= \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma|\pi) \\ &= \frac{1}{(2^{1-\beta}-1)|S|^\beta}\left(2\cdot\sum_{i=1}^{k}\sum_{j=1}^{m}|P_i\cap Q_j|^\beta \right. \\ &\quad \left. -\sum_{i=1}^{n}|P_i|^\beta - \sum_{j=1}^{m}|Q_j|^\beta\right). \end{aligned}$$

$$\lim_{\beta\to 1} d_\beta(\pi,\sigma) = d(\pi,\sigma)$$

Tables

A database table $\tau = (T, H, \rho)$, where T is the name, $H = A_1 \cdots A_n$ is the header, $\mathrm{Dom}(A_i)$ is domain of A_i and $\rho = \{t_1, \ldots, t_m\}$, $\rho \subseteq \mathrm{Dom}(A_1) \times \cdots \times \mathrm{Dom}(A_n)$ is its content:

T

	A_1	A_2	$\cdots$	A_n
t_1	a_{11}	a_{12}	$\cdots$	a_{1n}
t_2	a_{21}	a_{22}	$\cdots$	a_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_m	a_{m1}	a_{m2}	$\cdots$	a_{mn}

Partitions induced by Attribute Sets

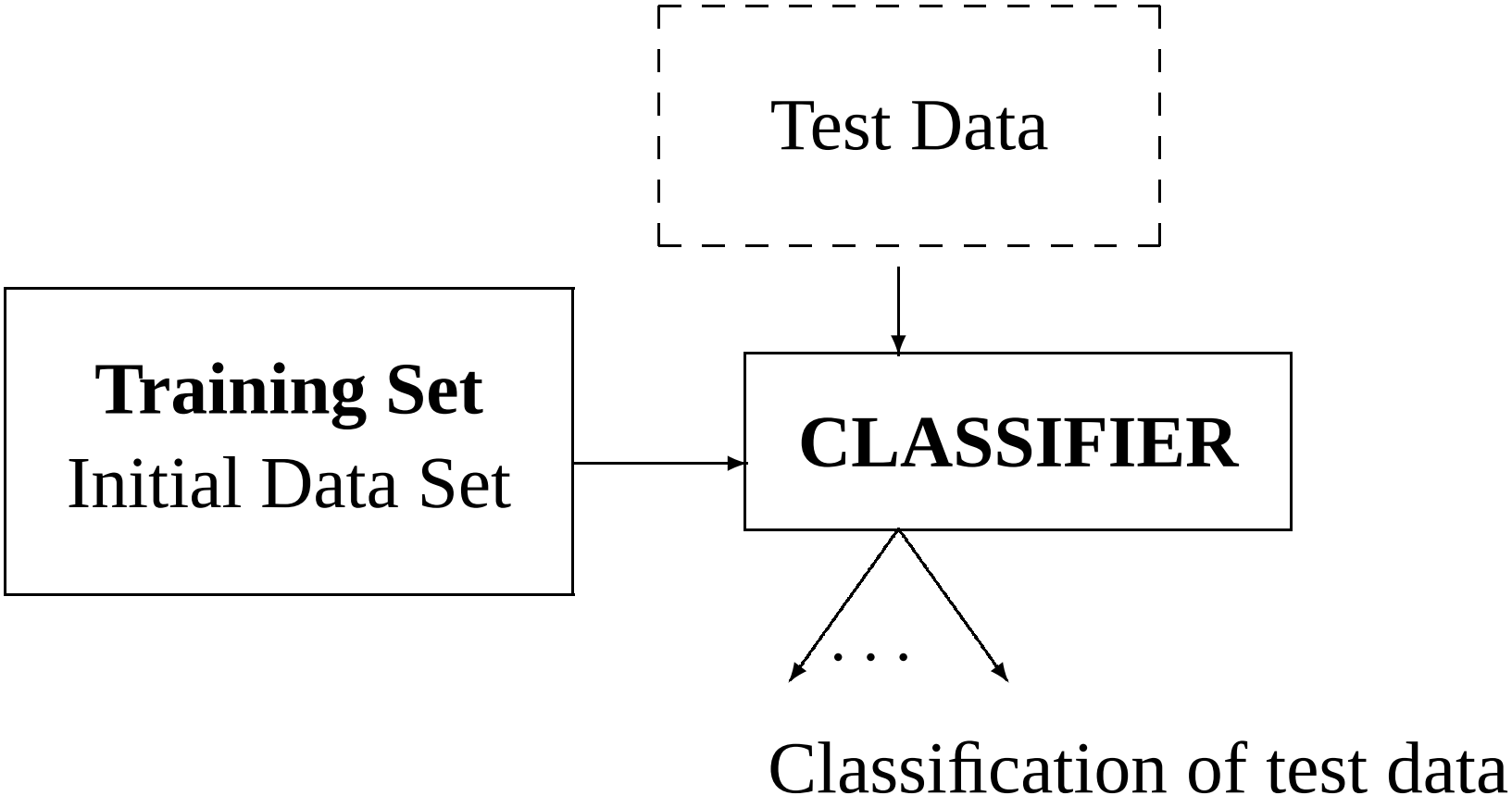
Every attribute set $K \subseteq H$ induces a partition π_K:

T

	$\cdots$	$\longleftarrow K \longrightarrow$	$\cdots$
t_1	$\cdots$	k_1	$\cdots$
t_2	$\cdots$	k_1	$\cdots$
t_3	$\cdots$	k_1	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_l	$\cdots$	k_p	$\cdots$
t_{l+1}	$\cdots$	k_p	$\cdots$
t_{l+2}	$\cdots$	k_p	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{n-1}	$\cdots$	k_r	$\cdots$
t_n	$\cdots$	k_r	$\cdots$

Classification in Data Mining

General Classification Model

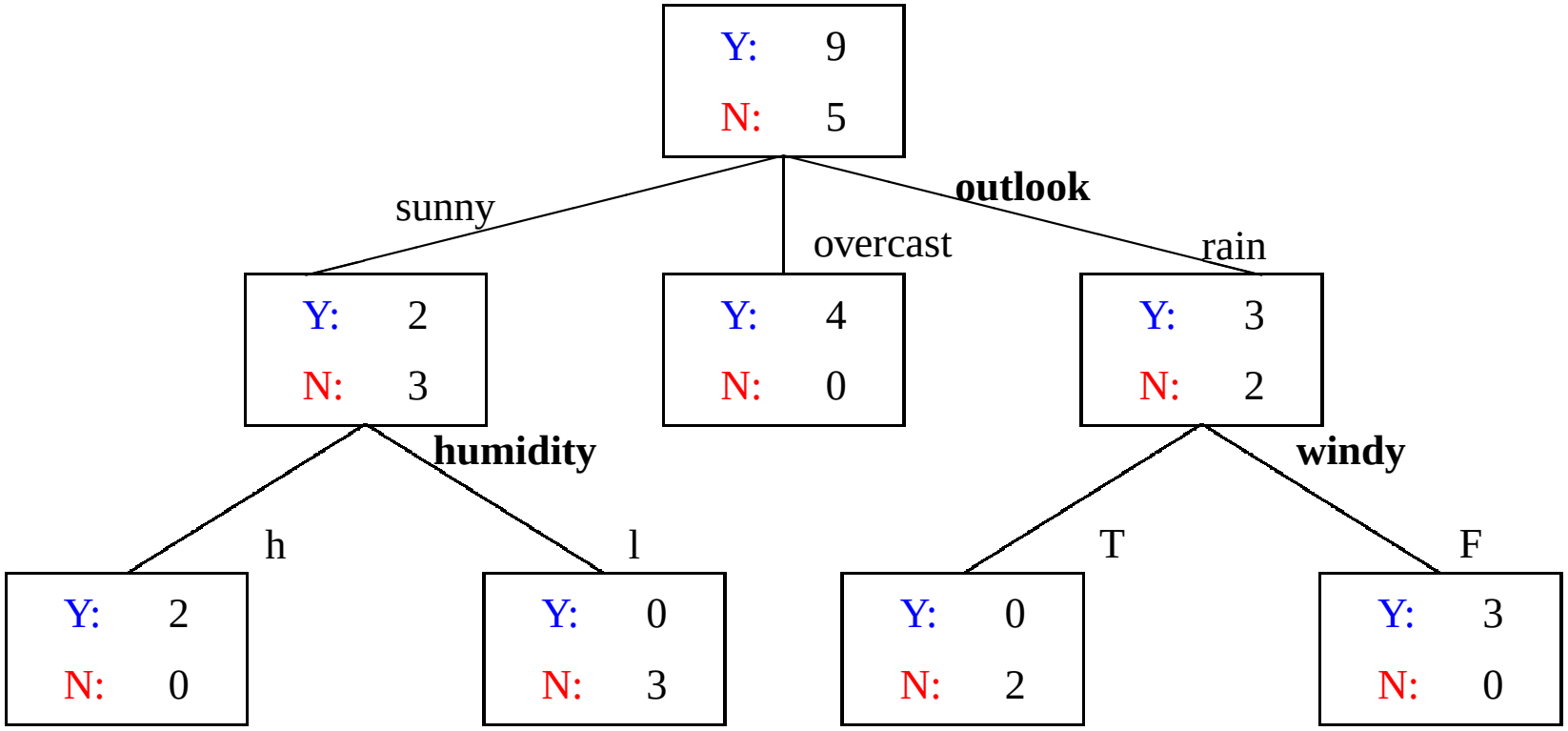


Decision Trees as Classifiers

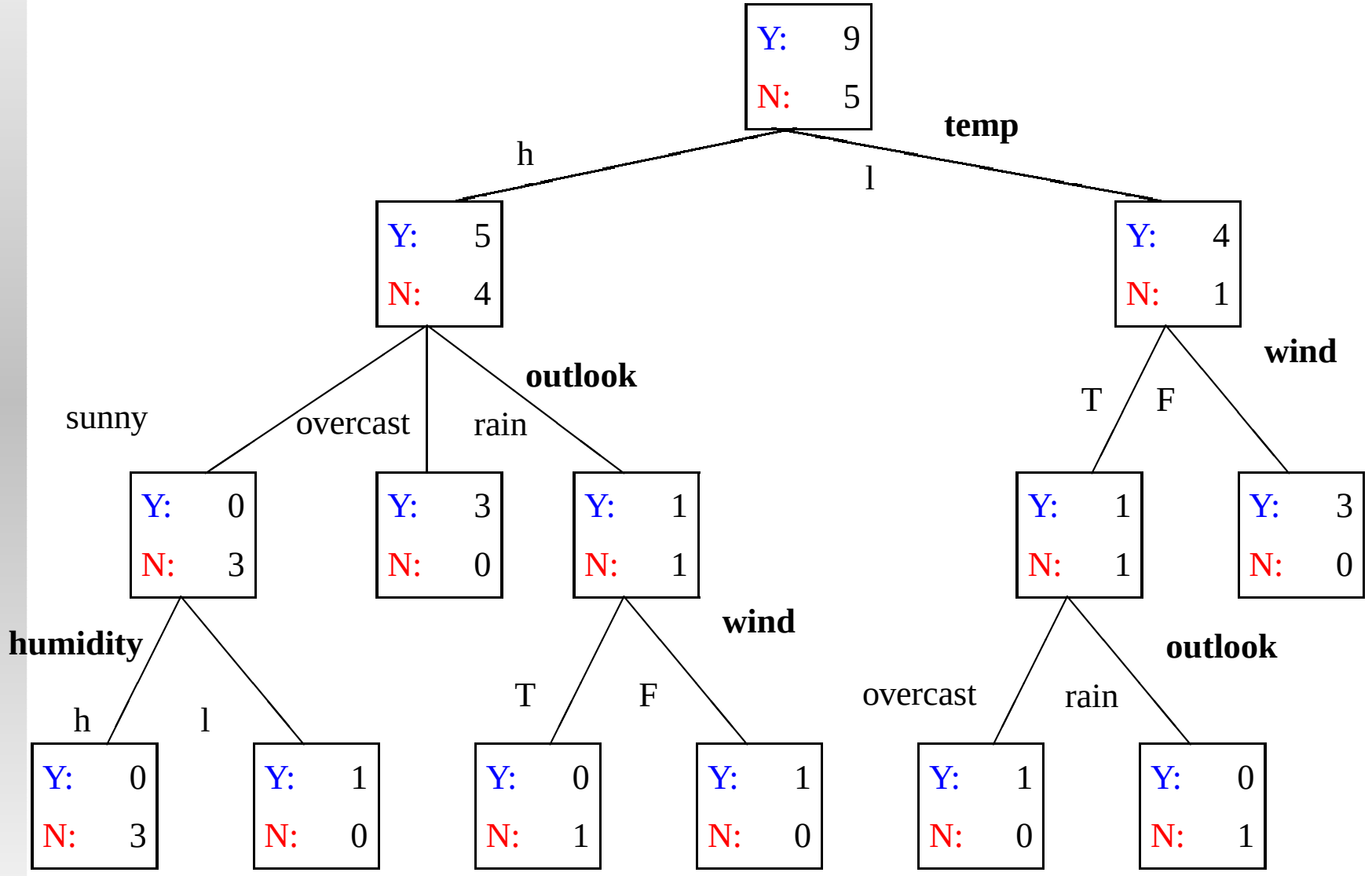
Data set for predicting weather for tennis

	outlook	temp	hum	windy	play
1	sunny	h	h	F	no
2	sunny	h	h	T	no
3	overcast	h	h	F	yes
4	rain	l	h	F	yes
5	rain	l	h	F	yes
6	rain	l	l	T	no
7	overcast	l	l	T	yes
8	sunny	h	h	F	no
9	sunny	l	l	F	yes
10	rain	h	h	F	yes
11	sunny	h	l	T	yes
12	overcast	h	h	T	yes
13	overcast	h	h	F	yes
14	rain	h	h	T	no

A Decision Tree



Another Decision Tree



DT as predictive models

- a DT is a mapping of observations about an item to conclusions about class of the item;
- each interior node corresponds to a variable;
- an arc to a child represents a possible value of that variable;
- a leaf represents the predicted value of the class given the values of the variables represented by the path from the root; thus, a leaft must be as pure as possible.

The Yield of a DT:

- a set of rules allowing new data to be classified

Thus, we prefer trees which:

- have few leaves (less fragmentation);
- have relatively small depth (generate simple rules).

The choice of the splitting attribute for DTs

- The leaves of a DT should contain C-pure or almost C-pure sets of objects $\Rightarrow$ for each leaf we must have $\mathcal{H}_\beta(\pi^C_{S_w})$ as close to 0 as possible.
- To take into account the size of the leaves note that the collection of sets of objects assigned to the current leafs is a partition κ of S and that we need to minimize:

$$\sum_w \left(\frac{|S_w|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi^C_{S_w}),$$

  which is $\mathcal{H}(\pi^C|\kappa)$.
- $\mathcal{H}(\pi^C|\kappa) = 0$ iff if $\kappa \leq \pi^C$.

The Information Gain Criterion

$$\mathcal{H}_\beta(\pi^C_{S_w}) - \mathcal{H}_\beta(\pi^C_{S_w}|\pi^A_{S_w}).$$

- When $\beta \to 1$ we obtain the information gain linked to Shannon entropy. When $\beta = 2$ one obtains the selection criteria for the Gini index using the CART algorithm.
- This criterion favors attributes with large domains, which in turn, generate bushy trees.

The Information Gain Ratio

$$\frac{\mathcal{H}_\beta(\pi^C_{S_w}) - \mathcal{H}_\beta(\pi^C_{S_w}|\pi^A_{S_w})}{\mathcal{H}_\beta(\pi^A_{S_w})},$$

This introduces the compensating divisor $\mathcal{H}_\beta(\pi^A_{S_w})$.

A New Crierion: The Metric Choice

Minimizes the distance

$$d_\beta(\pi^C_{S_w}, \pi^A_{S_w}) = \mathcal{H}_\beta(\pi^C_{S_w}|\pi^A_{S_w}) + \mathcal{H}_\beta(\pi^A_{S_w}|\pi^C_{S_w}).$$

Advantages:

- limits both conditional entropies $\mathcal{H}_\beta(\pi^C_{S_w}|\pi^A_{S_w})$ and $\mathcal{H}_\beta(\pi^A_{S_w}|\pi^C_{S_w})$;
- first limitation provides a high information gain;
- the second limitation insures that attributes with large domains are not favored over attributes with smaller domains.

Experimental Results - I

Audiology

β	accuracy	size	leaves
2.50	53.54	53	36
2.25	54.42	53	36
2.00	54.87	54	37
1.75	53.10	47	32
1.50	76.99	29	19
1.25	78.32	29	19
1.00	76.99	29	19
0.75	76.99	29	19
0.50	76.99	29	19
0.25	78.76	33	21
J48	77.88	54	32

Hepatitis

β	accuracy	size	leaves
2.50	81.94	15	8
2.25	81.94	9	5
2.00	81.94	9	5
1.75	83.23	9	5
1.50	84.52	9	5
1.25	84.52	11	6
1.00	85.16	11	6
0.75	85.81	9	5
0.50	83.23	5	3
0.25	82.58	5	3
J48	83.87	21	11

Primary-tumor

β	accuracy	size	leaves
2.50	34.81	50	28
2.25	35.99	31	17
2.00	37.76	33	18
1.75	36.28	29	16
1.50	41.89	40	22
1.25	42.18	38	21
1.00	42.48	81	45
0.75	41.30	48	27
0.50	43.36	62	35
0.25	44.25	56	32
J48	39.82	88	47

Vote

β	accuracy	size	leaves
2.50	94.94	7	4
2.25	94.94	7	4
2.00	94.94	7	4
1.75	94.94	7	4
1.50	95.17	7	4
1.25	95.17	7	4
1.00	95.17	7	4
0.75	94.94	7	4
0.50	95.17	9	5
0.25	95.17	9	5
J48	94.94	7	4

A Metric to Incremental Clustering

What is Clustering?

Clustering is an unsupervised learning process that partitions data such that similar data items are grouped together in sets referred to as clusters. Applications of clustering:

- condensing and identifying patterns in data;
- classifying data.

Clustering is hard

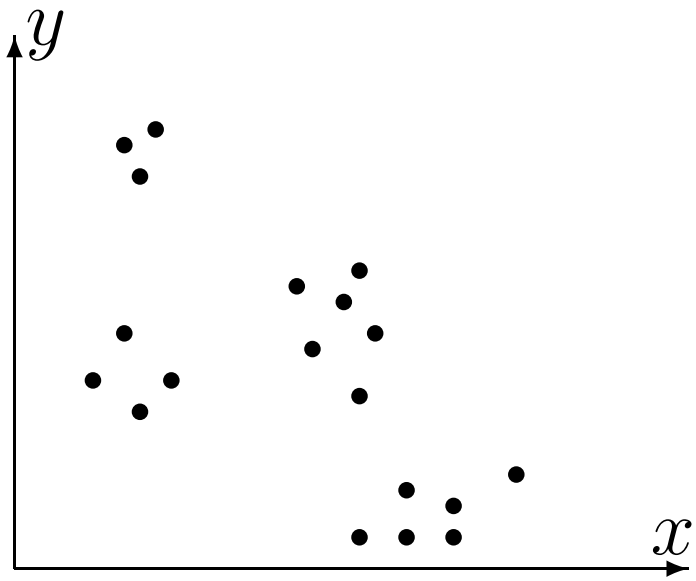
- “there is no clustering technique that is universally applicable in uncovering the variety of structures present in multidimensional data sets”
- what exactly does it mean that objects that belong to the same cluster are similar? how similar?
-
- what exactly does it mean that objects that belong to different cluster are dissimilar? how dissimilar?

Points given analytically

	x	y
p_1	6	13
p_2	8	16
p_3	9	11
p_4	11	13
p_5	19	19
p_6	20	15
p_7	22	18
p_8	23	20
p_9	24	16
p_{10}	23	12

	x	y
p_{11}	23	3
p_{12}	26	3
p_{13}	29	3
p_{14}	26	6
p_{15}	29	5
p_{16}	33	7
p_{17}	9	26
p_{18}	8	28
p_{19}	10	29

“Natural Clusters”



4 visible clusters! Eyes + brain solve problem instantly!

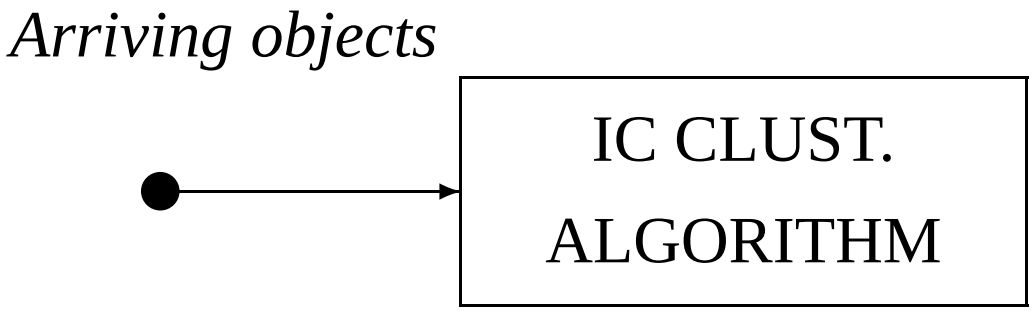
Computationally feasible but expansive

For n objects we need to:

- compute $\frac{n(n-1)}{2}$ inter-object distances;
- keep the distance matrix in memory;
- manage object groups.

For 1,000,000 objects we need to manage about 500 billion numbers!

Incremental clustering of nominal data



- main memory usage is minimal since there is no need to keep in memory the inter-object distances
- algorithms are scalable with respect to the size of the set of objects and the number of attributes.

We seek a clustering $\kappa = \{C_1, \ldots, C_n\} \in \mathsf{PART}(S)$ such that the total distance from κ to the partitions of the attributes:

$$D(\kappa) = \sum_{i=1}^{n} d_2(\kappa, \pi^{A_i})$$

is minimal.
The definition of d_2 allows us to write:

$$d_2(\kappa, \pi^A) = \sum_{i=1}^{n} |C_i|^2 + \sum_{j=1}^{m_A} |B_{a_j}^A|^2 - 2 \sum_{i=1}^{n} \sum_{j=1}^{m_A} |C_i \cap B_{a_j}^A|^2$$

Suppose now that t is a new object. The following cases may occur:

1. the object t is added to an existing cluster C_k;
2. a new cluster, C_{n+1} is created that consists only of t.

Also, from the point of view of partition π^A, t is added to the block $B^A_{t[A]}$, which corresponds to the value $t[A]$ of the A-component of t.
Thus, if $\min_k \sum_A |C_k \oplus B^A_{t[A]}| < \sum_A |B^A_{t[A]}|$ we add t to a cluster C_k for which $\sum_A |C_k \oplus B^A_{t[A]}|$ is minimal; otherwise, we create a new one-object cluster.

Experimental Results

Initial Run		Random Permutation		
Cluster	Size	Cluster	Size	Distribution (Original cluster)
1	1548	1	1692	1692 (2)
2	1693	2	1552	1548 (1), 3 (3), 1 (2)
3	1655	3	1672	1672 (5)
4	1711	4	1711	1711 (4)
5	1672	5	1652	1652 (3)
6	1616	6	1616	1616 (6)
7	1	7	85	85 (8)
8	85	8	10	10 (9)
9	10	9	8	8 (10)
10	8	10	1	1 (11)
11	1	11	1	1 (7)

Clustering mushrooms

Cl. num.	Poisonous/Edible	Total	Percentage of dominant group
1	825/2752	3577	76.9%
2	8/1050	1058	99.2%
3	1304/0	1304	100%
4	0/163	163	100%
5	1735/28	1763	98.4%
6	0/7	7	100%
7	0/192	192	100%
8	36/16	52	69%
9	8/0	8	100%

Goodman-Kruskal Association Index and Metric

The Goodman-Kruskal Coefficient

Let X, Y be two discrete random variables. The *Goodman-Kruskal coefficient* of X and Y is defined by

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

Classification rule: an elementary event is classified in the class that has the maximal probability.

GK Classification Rule

- $P(Y = b_j|X = a_i)$: the probability of predicting the value b_j for Y when $X = a_i$
  An event that has the component $X = a_i$ is 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:

$$1 - \max_{1 \leq j \leq k} P(Y = b_j|X = a_i).$$

$\mathsf{GK}(X, Y)$ is the expected probability that in a randomly chosen case the value of Y will be incorrectly predicted from X.
$\lambda_{Y|X}$ is the relative reduction in the probability of prediction error:

$$\lambda_{Y|X} = 1 - \frac{\mathsf{GK}(X, Y)}{1 - \max_{1 \leq j \leq k} P(Y = b_j)}$$

$\lambda_{Y|X}$ 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 for Partitions

Consider two partitions

$$\pi = \{B_1, \ldots, B_l\} \text{ and } \sigma = \{C_1, \ldots, C_k\}.$$

Define the *Goodman-Kruskal coefficient* of these partitions $\mathsf{GK}(\pi, \sigma)$ as the number:

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

Partitions and Random Variables

The partitions π, σ define two random variables

$$X : \begin{pmatrix} 1 & \cdots & l \\ \frac{|B_1|}{|S|} & \cdots & \frac{|B_l|}{|S|} \end{pmatrix} \text{ and } Y : \begin{pmatrix} 1 & \cdots & k \\ \frac{|C_1|}{|S|} & \cdots & \frac{|C_k|}{|S|} \end{pmatrix}.$$

such that conditional probability $P(Y = j|X = i)$ is given by:

$$P(Y = j|X = i) = \frac{P(Y = j \wedge X = i)}{P(X = i)} = \frac{|C_j \cap B_i|}{|B_i|}.$$

Interpretation of GK

For a fixed i, the largest error in predicting Y is:

$$1 - \max_{1 \le j \le k} P(Y = j | X = i) = 1 - \max_{1 \le j \le k} \frac{|C_j \cap B_i|}{|B_i|}.$$

The expected value of the largest error in predicting Y is:

$$\sum_{i=1}^{l} \frac{|B_i|}{|S|} \cdot \left(1 - \max_{1 \le j \le k} \frac{|C_j \cap B_i|}{|B_i|} \right)$$

$$= \quad 1 - \sum_{i=1}^{l} \max_{1 \le j \le k} \frac{|C_j \cap B_i|}{|S|},$$

which is exactly $\mathsf{GK}(X, Y)$.

Properties of GK

- We have $\mathsf{GK}(\pi, \sigma) = 0$ if and only if $\pi \leq \sigma$.
- The function GK is monotonic in its first argument and dually monotonic in its second:
  - If π, π', σ are three partitions of the set S such that $\pi \leq \pi'$, then $\mathsf{GK}(\pi, \sigma) \leq \mathsf{GK}(\pi', \sigma)$.
  - If π, σ', σ are three partitions of the set S such that $\sigma \leq \sigma'$, then $\mathsf{GK}(\pi, \sigma) \geq \mathsf{GK}(\pi, \sigma')$.
- GK satisfies a triangular inequality: $\mathsf{GK}(\pi, \sigma) \leq \mathsf{GK}(\pi, \tau) + \mathsf{GK}(\tau, \sigma)$.

Metric Associated to GK

The Goodman-Kruskal coefficient allows us to define a metric on $\mathsf{PART}(S)$.

Let $d_{GK} : \mathsf{PART}(S) \times \mathsf{PART}(S) \longrightarrow \mathbb{R}$ be

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

for $\pi, \sigma \in \mathsf{PART}(S)$.

The function d_{GK} is a metric on the set $\mathsf{PART}(S)$.

Goodman-Kruskal Coefficient for Attribute Sets

Let K, L be two sets of attributes of a table.
Define $\mathsf{GK}(K, L) = \mathsf{GK}(\pi_K, \pi_L)$: the expected error that occurs when we try to predict the value of $t[L]$ from the value of $t[K]$.

- If $K_1 \subseteq K_2$, then $\pi_{K_2} \leq \pi_{K_1}$, so $\mathsf{GK}(K_2, L) \leq \mathsf{GK}(K_1, L)$.
- If $L_1 \subseteq L_2$, then $\mathsf{GK}(K, L_2) \leq \mathsf{GK}(K, L_1)$.

Goodman-Kruskal Metric on Attribute Sets

Define $d_{GK}(K, L) = d_{GK}(\pi_K, \pi_L)$ for any two sets of attributes K, L.

The new metric can be used for:

- constructing classifiers;
- discretization of continuous attributes;
- attribute clustering, feature selection and data compression.

ϵ-predictors

An ϵ-*predictor* for a set of attributes Y is a set of attributes K such that $\mathsf{GK}(K, Y) \leq \epsilon$.

- If K is an ϵ-predictor for Y, then any superset K' of K is also a ϵ-predictor for Y.
- An ϵ-predictor such that no of its proper subsets is an ϵ-predictor is called *minimal*.

An Apriori-like Algorithm for ϵ-predictors

Input: A set of attributes H, a target attribute Y, $Y \notin H$

Output: Set P of all minimal ϵ-predictors from H.

(1) $\mathsf{Cand} = \{\{A\} : A \in H\}$;

(2) $P = \emptyset$;

(3) $\mathsf{P} = \mathsf{P} \cup \{K \in \mathsf{Cand} : \mathsf{GK}(K, Y) \leq \epsilon\}$;

(4) $\mathsf{Cand} = \mathsf{Cand} \setminus \mathsf{P}$;

(5) $\mathsf{Cand} = \{L \subseteq H :$ for all $K \subset L$, $|K| = |L| - 1$ we have $K \in \mathsf{Cand}\}$;

(6) goto (3);

- If a set is a nonminimal predictor, so are all of its supersets, which can thus be skipped.
- Initialize candidate set of predictors Cand to include one-set attributes.
- The set of minimal predictors P is constructed starting from Cand.

- Initialize P to include all singleton predictors whose error is below the threshold ϵ. Remove those from C and the search for minimal two-attribute predictors makes use of the remaining candidate attributes, etc.
- The stopping condition could be exceeding the maximum predictor size or finding a predictor with desired prediction error.

Experimental Results – KHAN

J. Khan et.al.: Classification and Diagnostic Prediction of Cancers using gene expression profiling and artificial neural networks, Nature Medicine, vol 7., 2001

Differential diagnosis of four small round blue cell tumors of childhood (SRBCTs) :

NB: neuroblastoma

RMS: rhabdomyosarcoma

BL: Burkitt lymphoma

EWS: Ewing family of sarcomas

Previous work:

single layer neural networks (Khan)
logistic regression model (Weber)
SVMs (Mukerjee)
combined classifiers (Yeo)

Khan Data

- 2308 genes were measured using cDNA microarrays
- Training Data: 63 cases (12 NB, 20 RMS, 8 BL, and 23 EWS)
- Test Data: 25 cases (6 EWS, 5 RMS, 6 NB, 3 BL, and 5 non-SRBCTs)
- The test cases include 5 cases which do not belong to any of the predicted SRBCT types. Such cases are not present in the training set.

Preprocessing

Replace each class attribute with 4 binary attributes, one for each cancer type.

original attribute	computed attributes			
Cancer type	NB	RMS	BL	EWS
NB	1	0	0	0
EWS	0	0	0	1
RMS	0	1	0	0
other	0	0	0	0

- A separate predictor is built for each binary attribute to allow for handling of cases of type ‘other’ present in the test set, but absent in the training set.
- We expect that for ‘other’ cancer type all of the predictors will give the value of 0 thus indicated that none of the 4 cancer types is present.

- Predictors may contradict each other (infrequently, because low error rate of individual classifiers).
- If presence of more than one cancer type is predicted consider it misclassified.
- Small predictors decrease the risk of overfitting (small number of training cases!)

Discretization

- Every gene expression level X attribute is discretized into two intervals: $X \leq T$ and $X > T$.
- T is chosen such that the Shannon entropy $H(Y|X')$ of the target Y conditional on the discretized attribute X' is minimal.
- A separate discretization (using Fayyad-Irani) has been performed with respect to each cancer type.

Limitations on the Computation

- We find all predictors with 1 or 2 attributes, allowing up to one misclassified instance on the training set.
- The stopping rule: reaching the maximum prescribed size of the predictor, or obtaining an error rate less than to $\frac{1}{t}$, where t is the size of the training set.
- All but 30 most predictive attributes are discarded.
- For each cancer type the first predictor with minimal training error is manually picked at random (without looking at its test set performance to avoid bias in the choice).

Cancer type	selected predictor	image ids	mtr	mte	1GP	2GP
BL	WAS $\leq 0.69 \Rightarrow$ BL	236282	0	1	15	5
EWS	FCGRT $\leq 1.59 \Rightarrow$ EWS	770394	1	3	2	10
NB	MAP1B > 2.17	629896 - 383188	0	0	2	28
	or RCV1 $> 1.98 \Rightarrow$ NB					
RMS	TNNT2 > 0.55	298062 - 796258	0	2	0	25
	or SGCA $> 0.44 \Rightarrow$ RMS					

Legend:

mte misclassified cases in test set

mtr misclassified cases in training set

1GP number of one-gene predictors

2GP number of two-gene predictors

- A fairly large number (12–30) of very simple predictors have been found for each cancer type.
- Each of those predictors has very good classification rate on the training set: up to one misclassified case is allowed.
- The results show that there are many genes based on which a diagnosis can be made for each cancer type.
- All genes except for the one that predicts BL were reported among the 96 selected in Khan.

Bonferroni Correction

The probability that a single gene expression predicts BL perfectly on training set when there is no correlation between the gene and the tumor type:

$$2 \cdot \frac{55! \cdot 8!}{63!} = 5.16 \cdot 10^{-10}$$

much less than $0.05/2308 = 2.16 \cdot 10^{-5}$, the 5% significance level after Bonferroni correction. This shows that selected gene is with very high probability related to BL.

- If a classifier for only one type of tumor gave a positive prediction, then the instance was classified as this type of tumor.
- If none of them gave positive prediction we declared the case as ‘other tumor type’.
- If more than one classifier was active the case was considered a prediction error.
- The combined classifier used a total of 6 genes and classified correctly 19 out of 25 test cases.
- Out of the 6 misclassified cases, 2 gave classifications when the real outcome was ‘other’, 3 SRBCT cases were undetected, and there was 1 conflict.

Experimental Results - GOLUB

- Training data: 38 cases (27 acute lymphoblastic leukemia and 11 acute myelocytic leukemia) Test data: 34 cases (20 ALL and 14 AML);
- Data involves 6817 genes.
- We discretized the gene expression levels using Fayyad-Irani
- 20 genes were retained for which the Goodman-Kruskal coefficient was below 0.04.

- Five single-genes predictors and 66 two-gene predictors were identified.
- We identified two two-genes predictors (MGST1, APLP2 and CD33, CystatinA) for which the errors on the test set are 0 and 0.0294118, respectively.
- CD33 was among the 50 genes selected by Golub et al.

Distribution of the errors on the test set for the remaining set of minimal two-genes predictors:

Error Interval	Number of 2-attribute predictors
$[0.0, 0.05]$	2
$(0.05, 0.10]$	9
$(0.10, 0.15]$	10
$(0.15, 0.20]$	7
$(0.20, 0.25]$	13
$(0.25, 0.30]$	14
$(0.30, 0.35]$	3
$(0.35, 0.40]$	4
$(0.40, 0.45]$	3

Voting Mechanism

- We retained 19 one-attribute predictors whose prediction error on the training set did not exceed 5.3% (that is, two errors out of the 38 training cases).
- A vote was taken, and the instance was classified according to the majority vote.
- We obtained 3 errors on the test set of 34 cases. Namely, the errors occurred on the 57th, 60th and 66th cases of the original Golub test set ("unclassifiable" in the original study (Golub)).

Advantages of Using GK

- The Goodman-Kruskal dissimilarity GK is a simple, but powerful measure of predictive power that can be used to produce robust classifiers.
- The small number of training cases makes reliable construction of more complex models like Bayesian networks or C4.5 trees very hard or even impossible.
- Naive Bayesian classifiers suffer from independence assumptions which may not be satisfied in the microarray setting where most genes are correlated with each other.
- The Bonferroni correction, though conservative, yields valid results.

A Metric Approach to Discretization

From numerical to nominal

Previous work on discretization:

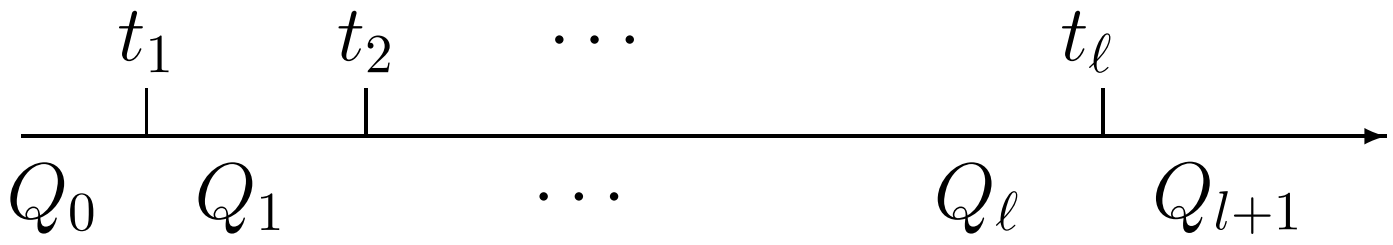
- fixed k-interval discretization (J. Dougherty, R. Kohavi, M. Sahami, 1995)
- fuzzy discretization (Kononenko 1992-1993)
- Shannon-entropy discretization (Fayyad and Irani, 1993)
- proportional k-interval discretization (Yang and Web, 2001, 2003)
- highly dependent attributes (M. Robnik and I. Kononenko, 1995)

Why Metric Discretization?

- a generalization of Fayyad-Irani discretization technique
- a geometric criterion for halting the discretization process
- better results in building
  - naive Bayes classifiers
  - decision trees

Discretization of a numeric attribute B

Set of cutpoints: $S = \{t_1, \ldots, t_\ell\}$ in $\mathsf{aDom}(B)$, where $t_1 < t_2 < \cdots < t_\ell$.



Discretization partition of $\mathsf{aDom}(B)$:

$$\pi^S = \{Q_0, \ldots, Q_\ell\}$$

Boundary Points

Recall that π^A the partition of the set of tuples of a table determined by the values of an attribute A:

```
select * from T group by A
```

$t_1, \ldots, t_n$: the list of tuples sorted on the values of an attribute B.

$\pi_{B,A}$ is the partition of $\mathsf{aDom}(B)$ that consists of the longest runs of *consecutive* B-components of the tuples in this list that belong to the *same block* K of the partition π_A.

The *boundary points* of the partition $\pi_{B,A}$ are the least and the largest elements of each of the blocks of the partition $\pi_{B,A}$.

We have $\pi_{B,A*} \leq \pi^A$ for any attribute B.

Main Results -I

Theorem: Let T be a table where the class of the tuples is determined by the attribute A and let $\beta \in (1, 2]$.
If S is a set of cutpoints such that the conditional entropy $\mathcal{H}_\beta(\pi_A|\pi_*^S)$ is minimal among the set of cutpoints with the same number of elements, then S consists of boundary points of the partition $\pi_{B,A}$ of $\mathsf{aDom}(B)$.

Main Results -II

Theorem: Let $\beta \in (1, 2]$.
If S is a set of cutpoints such that the distance $d_\beta(\pi^A, \pi^S_*)$ is minimal among the set of cutpoints with the same number of elements, then S consists of boundary points of the partition $\pi_{B,A}$ of $\mathsf{aDom}(B)$.

To discretize $\mathsf{aDom}(B)$ we seek a set of cutpoints such that

$$d_\beta(\pi^A, \pi^S_*) = \mathcal{H}_\beta(\pi^A|\pi^S_*) + \mathcal{H}_\beta(\pi^S_*|\pi^A)$$

is minimal.

Seek a set of cutpoints S such that the partition π^S_* induced on the set of rows is as close as possible to the target partition π^A.

Discretization Algorithm

Input: A table T, a class attribute A
and a real-valued attribute B.

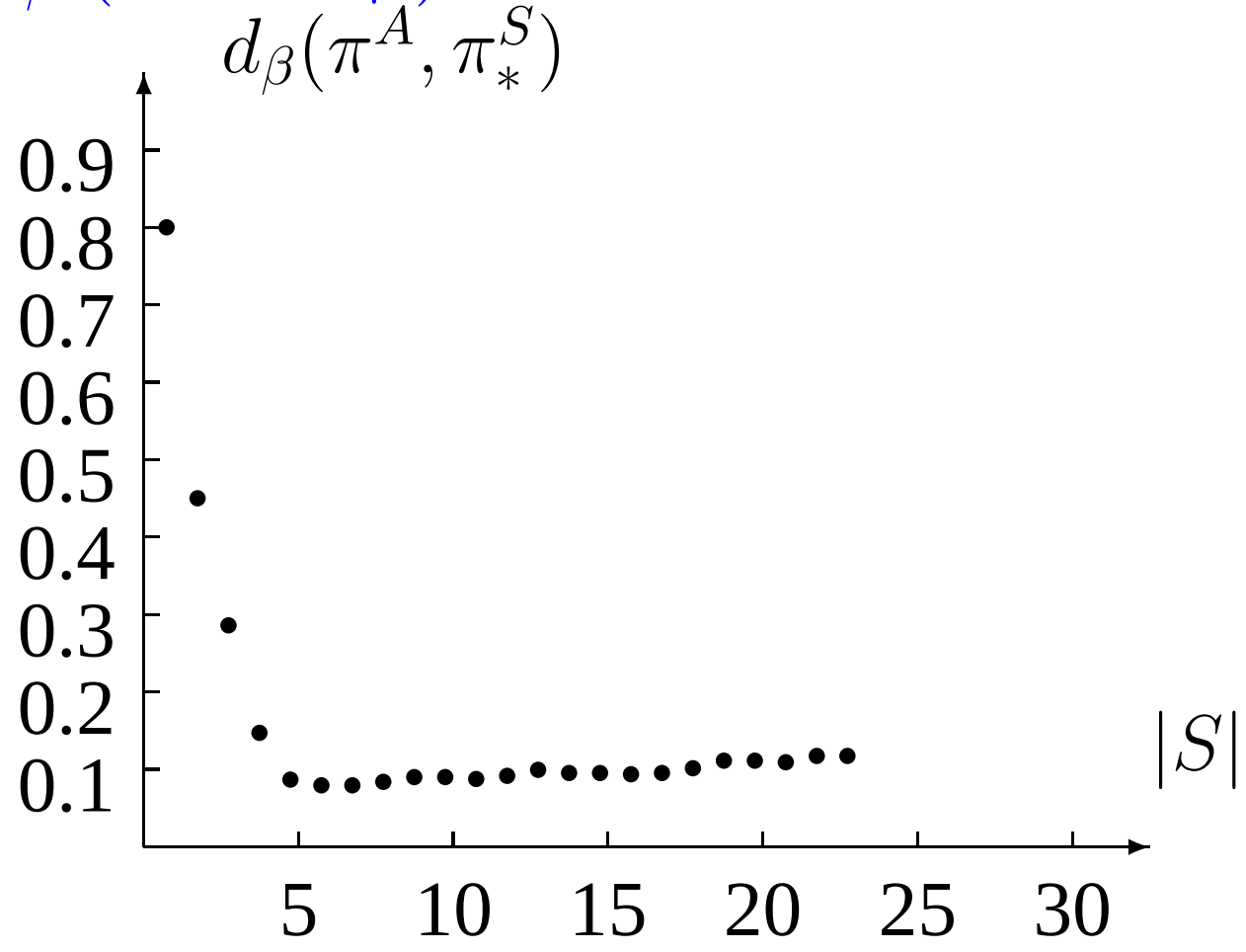
Output: A discretized attribute B.

BP is the set of boundary points of partition $\pi_{B,A*}$

Method:

sort T on B;
compute BP;
$S = \emptyset$; $d = \infty$;
while BP $\neq \emptyset$ do
    let $t =$ arg min $_{t \in \mathsf{BP}} d_\beta(\pi^A, \pi_*^{S \cup \{t\}})$;
    if $d \geq d_\beta(\pi^A, \pi_*^{S \cup \{t\}})$ then
        begin
            $S = S \cup \{t\}$; BP $=$ BP $- \{t\}$;
            $d = d_\beta(\pi^A, \pi_*^S)$
        end
    else exit while loop;
end while
for $\pi_*^S = \{Q_0, \ldots, Q_\ell\}$ replace
every value in Q_i by i for $0 \leq i \leq \ell$.

$d_\beta(\pi^A, \pi^S_*)$ as a function of $|S|$



78% of the total time is spent on decreasing the distance by the last 1%

$$d_\beta(\pi^A, \pi_*^S) = \mathcal{H}_\beta(\pi^A|\pi_*^S) + \mathcal{H}_\beta(\pi_*^S|\pi^A)$$

If $S \subseteq S'$ then $\pi^S \geq \pi^{S'}$ and

$$\begin{aligned} \mathcal{H}_\beta(\pi^A|\pi_*^S) &\geq \mathcal{H}_\beta(\pi^A|\pi_*^{S'}) \\ \mathcal{H}_\beta(\pi_*^S|\pi^A) &\leq \mathcal{H}_\beta(\pi_*^{S'}|\pi^A). \end{aligned}$$

Process starts with $S = \emptyset$, so $\pi_*^S = \omega$.
Practical halting criterion:

$$|d - d_\beta(\pi^A, \pi_*^{S\cup\{t\}})| > 0.01d.$$

Experimental Results

- Accuracy measured in stratified 10-fold cross-validation
- UCI datasets with $\beta \in \{1.5, 1.8.1.9, 2\}$

Experimental Results - I

heart-c:

Method	Size	Leaves	Accuracy
standard	51	30	79.20
$\beta = 1.5$	20	14	77.36
$\beta = 1.8$	28	18	77.36
$\beta = 1.9$	35	22	76.01
$\beta = 2.0$	54	32	76.01

glass:

standard	57	30	57.28
$\beta = 1.5$	32	24	71.02
$\beta = 1.8$	56	50	77.10
$\beta = 1.9$	64	58	67.57
$\beta = 2.0$	92	82	66.35

Experimental Results - II

ionosphere:

standard	35	18	90.88
$\beta = 1.5$	15	8	95.44
$\beta = 1.8$	19	12	88.31
$\beta = 1.9$	15	10	90.02
$\beta = 2.0$	15	10	90.02

iris:

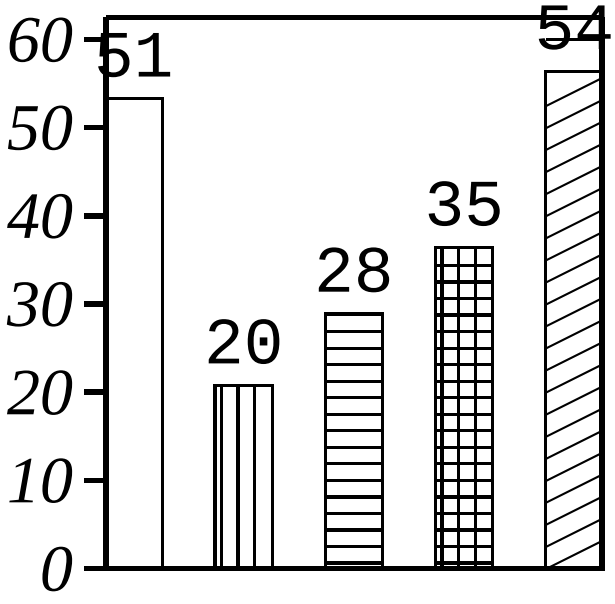
standard	9	5	95.33
$\beta = 1.5$	7	5	96
$\beta = 1.8$	7	5	96
$\beta = 1.9$	7	5	96
$\beta = 2.0$	7	5	96

Experimental Results - III

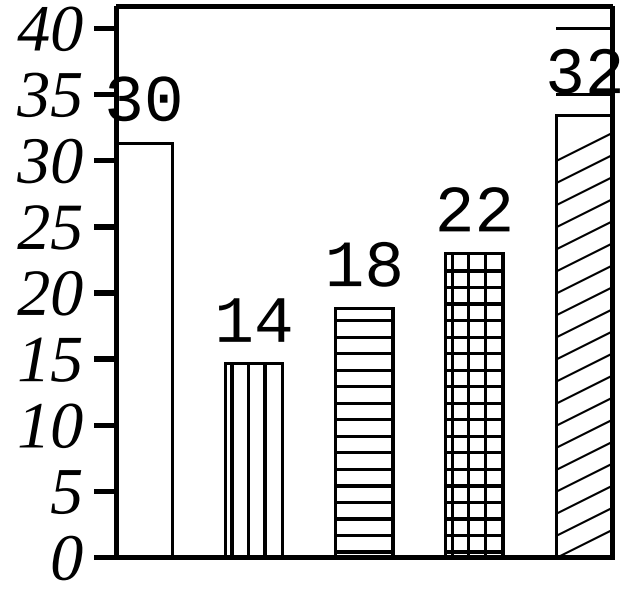
diabetes:

standard	43	22	74.08
$\beta = 1.8$	5	3	75.78
$\beta = 1.9$	7	4	75.39
$\beta = 2.0$	14	10	76.30

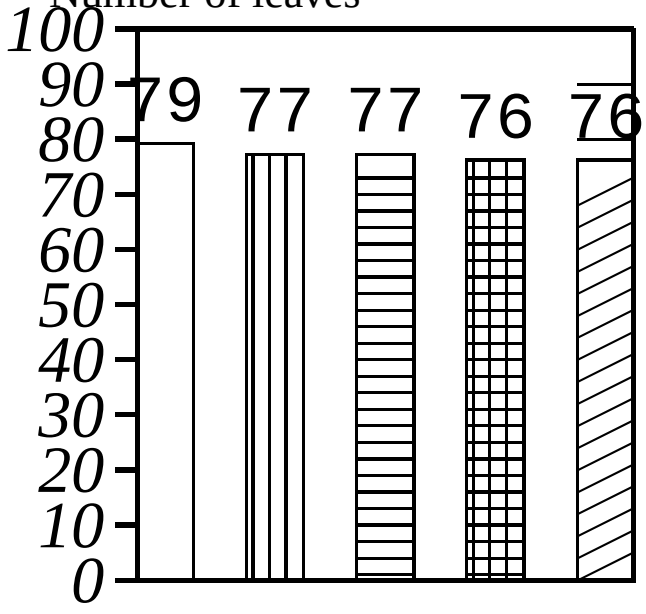
Heart-c



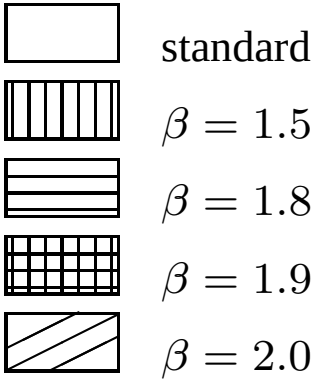
Tree size



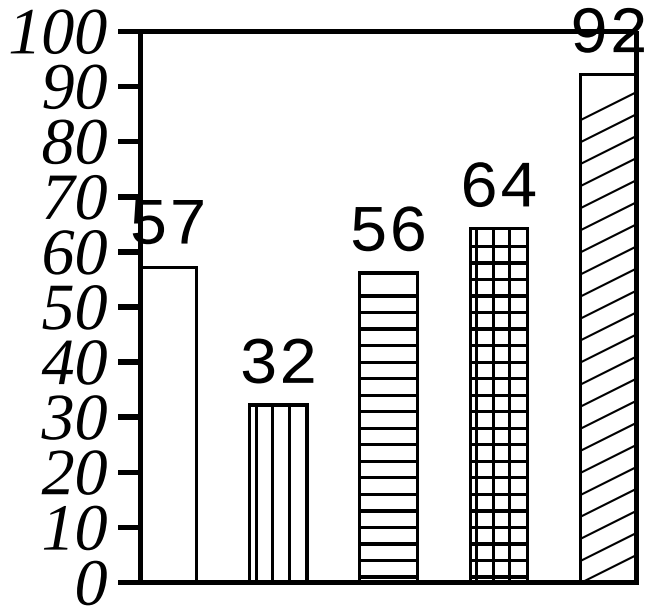
Number of leaves



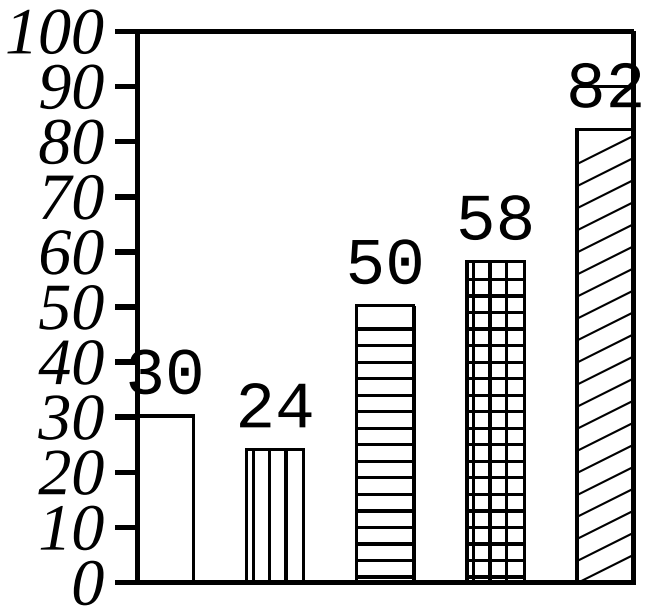
Accuracy



Glass

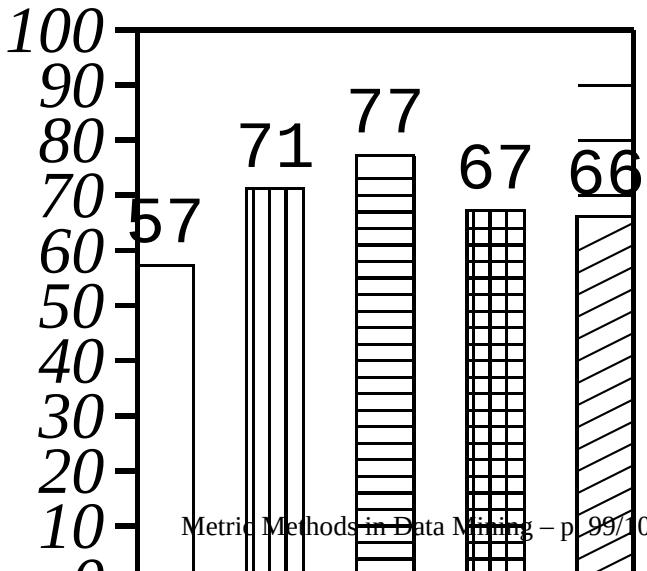


Tree size



Number of leaves

Accuracy



standard

$\beta = 1.5$

$\beta = 1.8$

$\beta = 1.9$

$\beta = 2.0$

Naive Bayes Classifiers

Error Rate

Discretization Method	Diabetes	Glass	Ionosphere	Iris
$\beta = 1.5$	34.9	25.2	4.8	2.7
$\beta = 1.8$	24.2	22.4	8.3	4
$\beta = 1.9$	24.9	23.4	8.5	4
$\beta = 2.0$	25.4	24.3	9.1	4.7
weighted prop	25.5	38.4	10.3	6.9
prop.	26.3	33.6	10.4	7.5

Conclusions and Future Work

Conclusions

An appropriate choice of β yields better classifiers and discretization methods.
Open issues:

- identifying simple parameters of data sets that inform the best choice of β;
- metric discretization for data with missing values.
- the metric space of attributes can be used to cluster attributes:
  - similar attribute are grouped in clusters, that may have a significance.
  - retaining one attribute per cluster (e.g., the medoid) allows for data compression and for simplification of decision techniques.

Thanks for Listening!