On a Minimization of Variables to Represent Sparse Multi-Valued Input Decision Functions

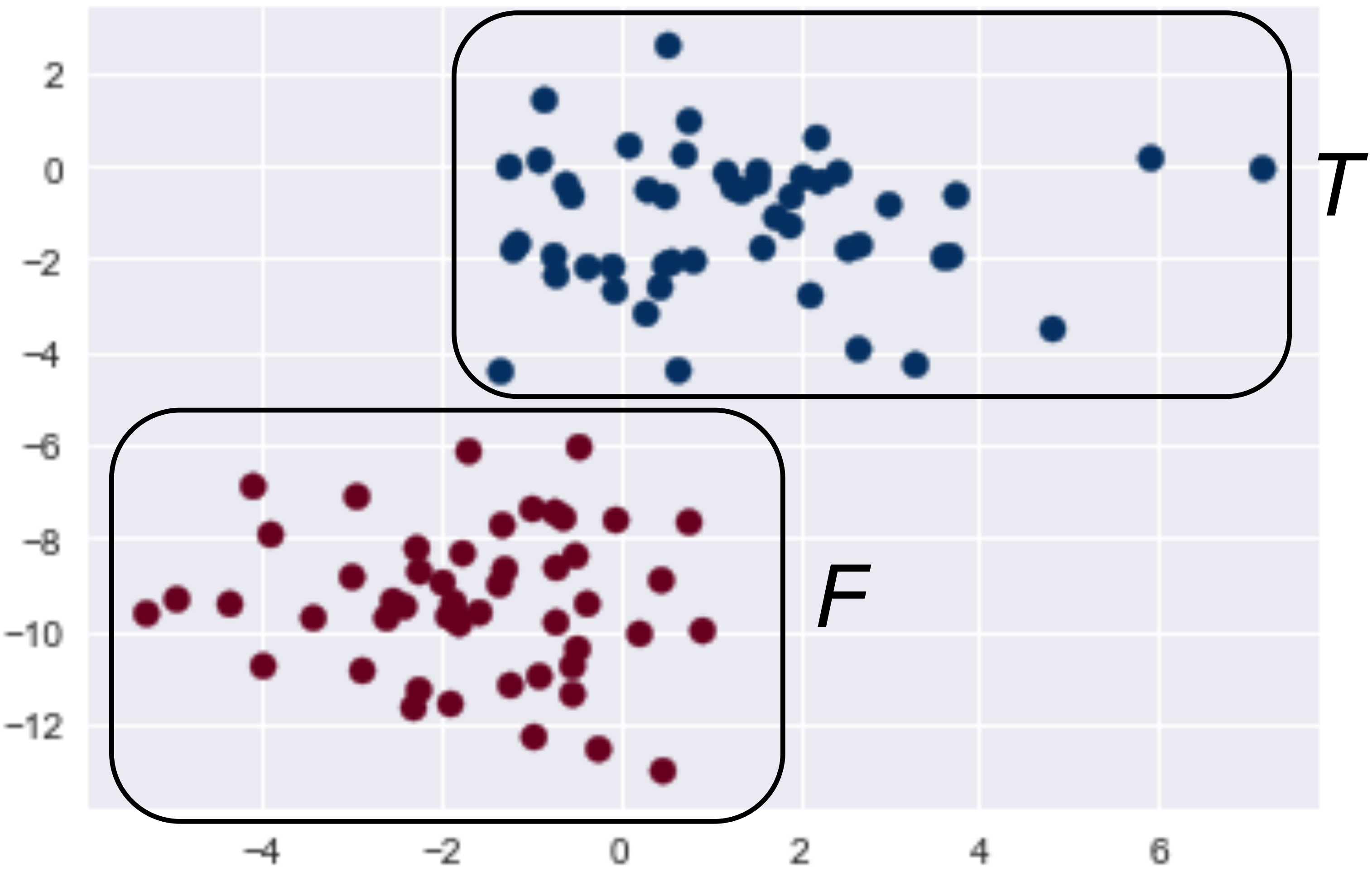
Tsutomu Sasao

Meiji University, Kanagawa, Japan

Outline of the Talk

- Introduction
- Definitions
- Minimization of Variables
- Removal of Inconsistent Instances
- Application 1: Mushrooms
- Application 2: Hepatitis
- Application 3: Breast Cancer
- Conclusions

Given two sample sets T and F, find a simple rule to distinguish them.



Multi-Valued Decision Function

$$f : P^n \to B,$$

$$P = \{0, 1, \ldots, p-1\},$$

$$B = \{0, 1\}.$$

$T \cap F = \phi,\ T \subseteq P^n,\ F \subseteq P^n.$

(T, F): Training Data

(T,F) is totally defined if $T \cup F = P^n$.

(T,F) is partially defined if $T \cup F \subset P^n$.

$$\frac{|T| + |F|}{p^n} < 10^{-6}$$

$T(f) = \{\vec{a} \in P^n \mid f(\vec{a})=1\}$, and **ON-set**

$F(f) = \{\vec{b} \in P^n \mid f(\vec{b})=0\}$. **OFF-set**

Given a partially defined function (T, F),

f is an extension of (T, F), if

$T(f) \supseteq T$, and $F(f) \supseteq F$.

X1:Physics, X2:Math, X3：English, X4:Art
2：Excellent, 1:Fair, 0：Poor

Let

$T = \{(2,2,0,1),(2,1,1,2)\}$ and

$F = \{(1,1,0,1),(0,1,1,2)\}$.

Then, the function f, where

$T(f) = \{(2,*,*,*)\}$ and

$F(f) = \{(1,*,*,*),(0,*,*,*)\}$

is an extension of (T,F).

For subsets $U \subseteq P^n$,

and $S \subseteq \{1, 2, \ldots, n\}$.

The **projection** of U to S is

the set $U|_S = \{\vec{a}\,|_S \mid \vec{a} \in U\}$.

Restriction

X1:Physics, X2:Math, X3：English, X4:Art
2：Excellent, 1:Fair, 0：Poor

Let $P=\{0,1,2\}$ and $n=4$. Let
$U=\{(1,2,0,1),(0,1,1,2),(2,0,1,2)\}$
and $S=\{2,3\}$.
Then, $U|_S=\{(*,2,0,*),(*,1,1,*),(*,0,1,*)\}$.

For a partially defined function (T,F),
a subset $S \subseteq \{1,2,...,n\}$ is a support set
if $T|_S$ and $F|_S$ are disjoint.

(T,F) can be represented by the variables
for S.

X1:Physics, X2:Math, X3：English, X4:Art
2：Excellent, 1:Fair, 0：Poor

Let (T, F) be a function, where
$T = \{(0,1,1,2), (1,2,0,1)\}$ and
$F = \{(1,1,0,1), (0,1,2,2)\}$.
Then, S={2,3} is a support set , since
$T|_S$ and $F|_S$ are disjoint, where
$T|_S = \{(*, 1,1,*), (*, 2,0,*)\}$, and
$F|_S = \{(*, 1,0,*), (*, 1,2,*)\}$.
The function is represented by two variables:

$$f = {X_2}^{\{2\}}{X_3}^{\{0\}} \vee {X_2}^{\{1\}}{X_3}^{\{1\}}.$$

Minimization of Variables

Algorithm

1) For each pair $(\vec{a},\vec{b})$, in $\vec{a} \in T$ and $\vec{b} \in F$,

make a clause $C(\vec{a},\vec{b}) = z_1 \vee z_2 \vee \cdots \vee z_n$,

$z_j = 0$ $(if\ \ a_j = b_j)$, $z_j = y_j$ $(if\ \ a_j \neq b_j)$.

2) For all the pairs $(\vec{a},\vec{b})$, in $\vec{a} \in T$ and $\vec{b} \in F$,

make the product of the clauses: $R = \bigwedge C(\vec{a},\vec{b})$.

3) Convert R into sum-of-products and simplify it.

4) A product with the fewest literals corresponds to a minimum support set.

Example

		X_1	X_2	X_3	X_4
T	a_1	1	2	0	1
	a_2	0	1	1	2
F	b_1	1	1	0	1
	b_2	0	1	2	2

$T|_S = \{(*,2,0,*),(*,1,1,*)\}.$

$F|_S = \{(*,1,0,*),(*,1,2,*)\}.$

$C(\vec{a}_1,\vec{b}_1) = y_2,$

$C(\vec{a}_1,\vec{b}_2) = y_1 \vee y_2 \vee y_3 \vee y_4,$

$C(\vec{a}_2,\vec{b}_1) = y_1 \vee y_3 \vee y_4,$

$C(\vec{a}_2,\vec{b}_2) = y_3,$

$R = y_2(y_1 \vee y_2 \vee y_3 \vee y_4)(y_1 \vee y_3 \vee y_4)y_3$

$= y_2 y_3$

The minimum support set is S={2,3}.

$f = X_2^{\{2\}} X_3^{\{0\}} \vee X_2^{\{1\}} X_3^{\{1\}}$

Minimization of Monotone Increasing Functions

$f : P^n \rightarrow B$ is monotone increasing

if $f(\vec{a}) \geq f(\vec{b})$ for any $\vec{a},\ \vec{b} \in P^n$

such that $\vec{a} \geq \vec{b}$.

Example：Entrance Examination

	X1	X2	X3
T	2	1	1
	1	2	1
	1	1	2
F	1	0	0
	0	1	0
	0	0	1

X1:Physics, X2:Math, X3：English
2：Excellent, 1:Fair, 0：Poor

$$f_1 = X_1^{\{2\}} X_2^{\{1\}} X_3^{\{1\}} \vee X_1^{\{1\}} X_2^{\{2\}} X_3^{\{1\}} \vee X_1^{\{1\}} X_2^{\{1\}} X_3^{\{2\}}$$

Example: Entrance Examination

	X1	X2	X3
T	2	1	1
	1	2	1
	1	1	2
F	1	0	0
	0	1	0
	0	0	1

X1:Physics, X2:Math, X3：English
2：Excellent, 1:Fair, 0：Poor

$$f_2 = X_1^{\{1\}} X_2^{\{0\}} X_3^{\{0\}} \vee X_1^{\{0\}} X_2^{\{1\}} X_3^{\{0\}} \vee X_1^{\{0\}} X_2^{\{0\}} X_3^{\{1\}}$$

Example: Entrance Examination

	X1	X2	X3
T	2	1	1
	1	2	1
	1	1	2
F	1	0	0
	0	1	0
	0	0	1

X1:Physics, X2:Math, X3：English
2：Excellent, 1:Fair, 0：Poor

$$f_3 = X_1^{\{2\}} X_1^{\{1,2\}} X_1^{\{1,2\}}$$
$$\vee X_1^{\{1,2\}} X_1^{\{2\}} X_1^{\{1,2\}}$$
$$\vee X_1^{\{1,2\}} X_1^{\{1,2\}} X_1^{\{2\}}$$
$$f_3(2,2,1) = 1$$

Decision of Entrance must be Monotone Increasing

Example: Entrance Examination

	X1	X2	X3
T	2	1	1
	1	2	1
	1	1	2
F	1	0	0
	0	1	0
	0	0	1

X1:Physics, X2:Math, X3：English
2：Excellent, 1:Fair, 0：Poor

$$f_4 = X_1^{\{1,2\}} X_2^{\{1,2\}} X_3^{\{1,2\}}$$

$$f_4(1,1,1) = 1$$

Example: Entrance Examination

	X1	X2	X3
T	2	1	1
	1	2	1
	1	1	2
F	1	0	0
	0	1	0
	0	0	1

X1:Physics, X2:Math, X3：English
2：Excellent, 1:Fair, 0：Poor

$$f_5 = X_1^{\{1,2\}} X_2^{\{1,2\}} \vee X_2^{\{1,2\}} X_2^{\{1,2\}} \vee X_1^{\{1,2\}} X_3^{\{1,2\}}$$

$$f_5(1,1,0) = 1$$

Example: Entrance Examination

	X1	X2	X3
T	2	1	1
	1	2	1
	1	1	2
F	1	0	0
	0	1	0
	0	0	1

X1:Physics, X2:Math, X3：English
2：Excellent, 1:Fair, 0：Poor

$$f_6 = X_1^{\{1,2\}} X_2^{\{1,2\}}$$

$$f_6(1,1,*) = 1$$

Theorem

Let $T \bigcap F = \phi$, where $T, F \subseteq P^n$.

Then, (T, F) has a monotone increasing extension iff there is no pair $(\vec{a}, \vec{b})$, such that $\vec{a} \in T$, and $\vec{b} \in F$ and $\vec{a} < \vec{b}$.

Removal of Inconsistent Instances

Example：Inconsistent Pairs

		X1	X2	X3	X4
T	a1	0	0	2	1
	a2	0	0	1	2
	a3	1	1	0	0
	a4	2	2	2	2
F	b1	2	1	0	0
	b2	1	2	0	0
	b3	0	0	2	2
	b4	0	0	0	0

X1:Physics, X2:Math, X3：English, X4:Art
2：Excellent, 1:Fair, 0：Poor

$$\vec{a}_3 < \vec{b}_1$$

$$\vec{a}_3 < \vec{b}_2$$

$$\vec{a}_1 < \vec{b}_3$$

$$\vec{a}_2 < \vec{b}_3$$

If there is any Inconsistency, then Professors may be sued.

Example：Inconsistent Pairs

		X1	X2	X3	X4
T	a1	0	0	2	1
	a2	0	0	1	2
	a3	1	1	0	0
	a4	2	2	2	2
F	b1	2	1	0	0
	b2	1	2	0	0
	b3	0	0	2	2
	b4	0	0	0	0

X1:Physics, X2:Math, X3：English, X4:Art
2：Excellent, 1:Fair, 0：Poor

$$\vec{a}_3 < \vec{b}_1$$

$$\vec{a}_3 < \vec{b}_2$$

$$\vec{a}_1 < \vec{b}_3$$

$$\vec{a}_2 < \vec{b}_3$$

Example：Consistent Pairs

		X1	X2	X3	X4
T	a1	0	0	2	1
	a2	0	0	1	2
	a4	2	2	2	2
F	b1	2	1	0	0
	b2	1	2	0	0
	b4	0	0	0	0

X1:Physics, X2:Math, X3：English, X4:Art
2：Excellent, 1:Fair, 0：Poor

There is no pair $(\vec{a}_i, \vec{b}_j)$ such that $\vec{a}_i < \vec{b}_j$.

$$f = X_3^{\{2\}} X_4^{\{1,2\}} \vee X_3^{\{1,2\}} X_4^{\{2\}}$$

Application 1: Mushrooms

Training Data

- # of instances: 5644
  - Poisonous:2156
  - Edible:3488
- # of variables: 22
- The function is

$$f: P_1 \times P_2 \times \cdots \times P_{22} \rightarrow B.$$

$$P_i = \{0, 1, \ldots, p_{\mathrm{i}} - 1\}.$$

$$p_1 = 6, p_2 = 4, p_3 = 10, p_4 = 2,$$

$$p_5 = 9, p_6 = 4, p_7 = 3, p_8 = 2,$$

$$p_9 = 12, p_{10} = 2, p_{11} = 6, p_{12} = 4,$$

$$p_{13} = 4, p_{14} = 4, p_{15} = 9, p_{16} = 2,$$

$$p_{17} = 4, p_{18} = 3, p_{19} = 8, p_{20} = 9,$$

$$p_{21} = 6, p_{22} = 7.$$

Multi-Valued Approach

- Found a 3-variable solution to represent poisonous mushrooms.

$$X_5^{\{2,3,4,5,7\}} \vee X_{21}^{\{0,1,2,3,5\}} X_{22}^{\{1,5\}} \vee X_{21}^{\{1,4\}} X_{22}^{\{0,2,5\}}$$

- X_5: denotes odor (9-valued)
- X_{21}: denotes population (6-valued),
- X_{22}: denotes habitat (7-valued).

Two-Valued Approach

- Prof. Boros' group in Rutgers University converted 22 multi-valued variables into 125 two-valued variables.
- They found a solution with 6 two-valued variables to represent poisonous mushrooms.
- Our multi-valued approach required only 3 variables.

Application 2: Hepatitis

Data Set

- # of instances: 80
  - Died: 13
  - Survived: 67
- # of variables: 19
  - Two-valued:13
  - Real-valued: 6

Multi-Valued Approach

- Minimum support set: $\{X_1, X_{15}, X_{17}, X_{18}\}$
- X_1: Age (9-valued)
- X_{15}: Alkaline phosphatase (7-valued)
- X_{17}: Albumin (7-valued)
- X_{18}: Prothrombin time (10-valued)

Two-Valued Approach

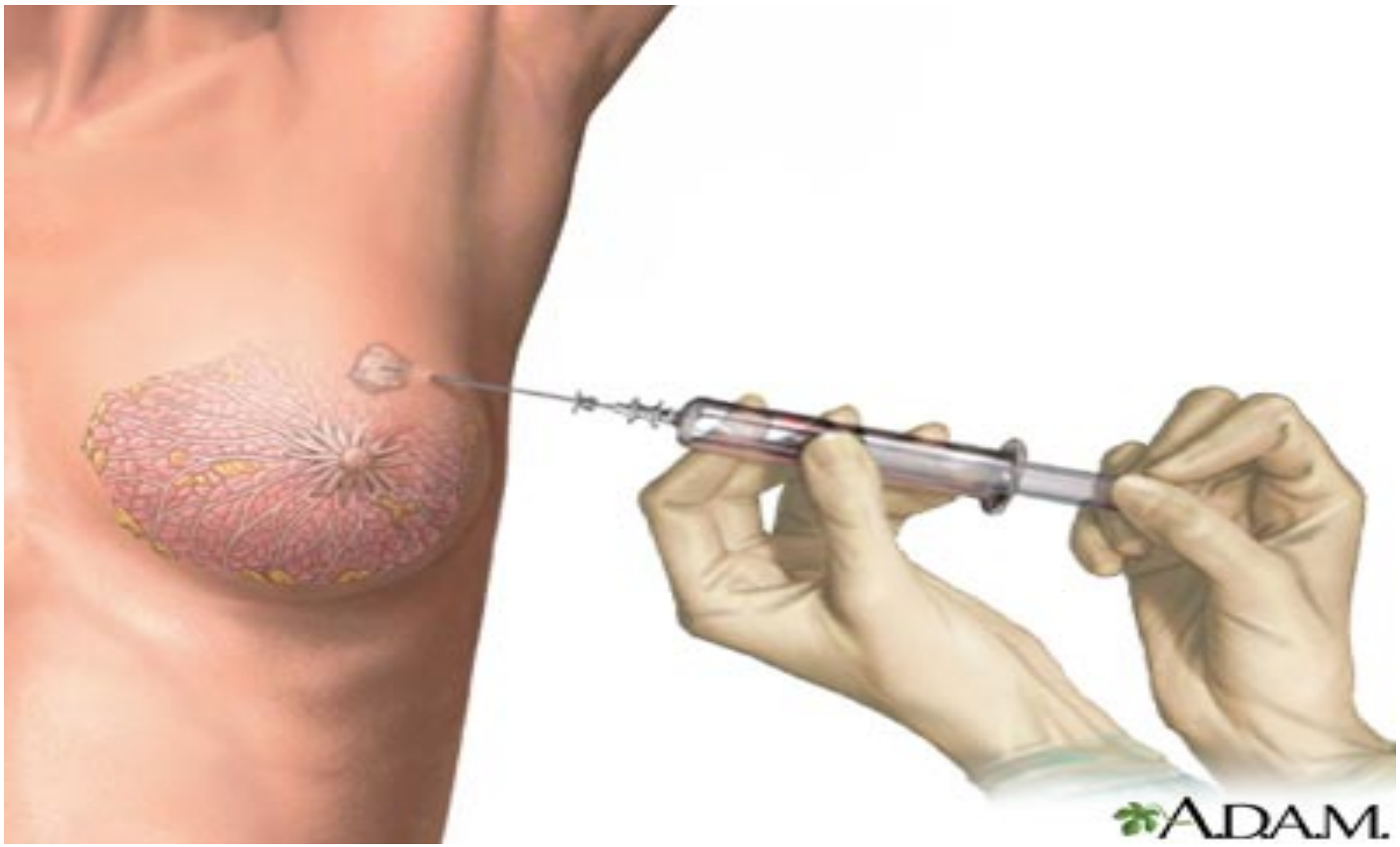
- Prof. Boros' group of Rutgers University used 46 two-valued variables to represent the function.
- They found a solution with 7 two-valued variables.
- Our multi-valued approach required only 4 variables.

Two-Valued Approach

- X4: Antivirals
- X11: Spider Angioma (spider nevus)
- X13: Varices
- X15>120 :Alkaline phosphatase
- X15> 200
- X18>50 :Prothrombin time
- X19: Histology

Application 3: Breast Cancer (Monotone Increasing Function)

Fine Needle Aspiration (FNA)



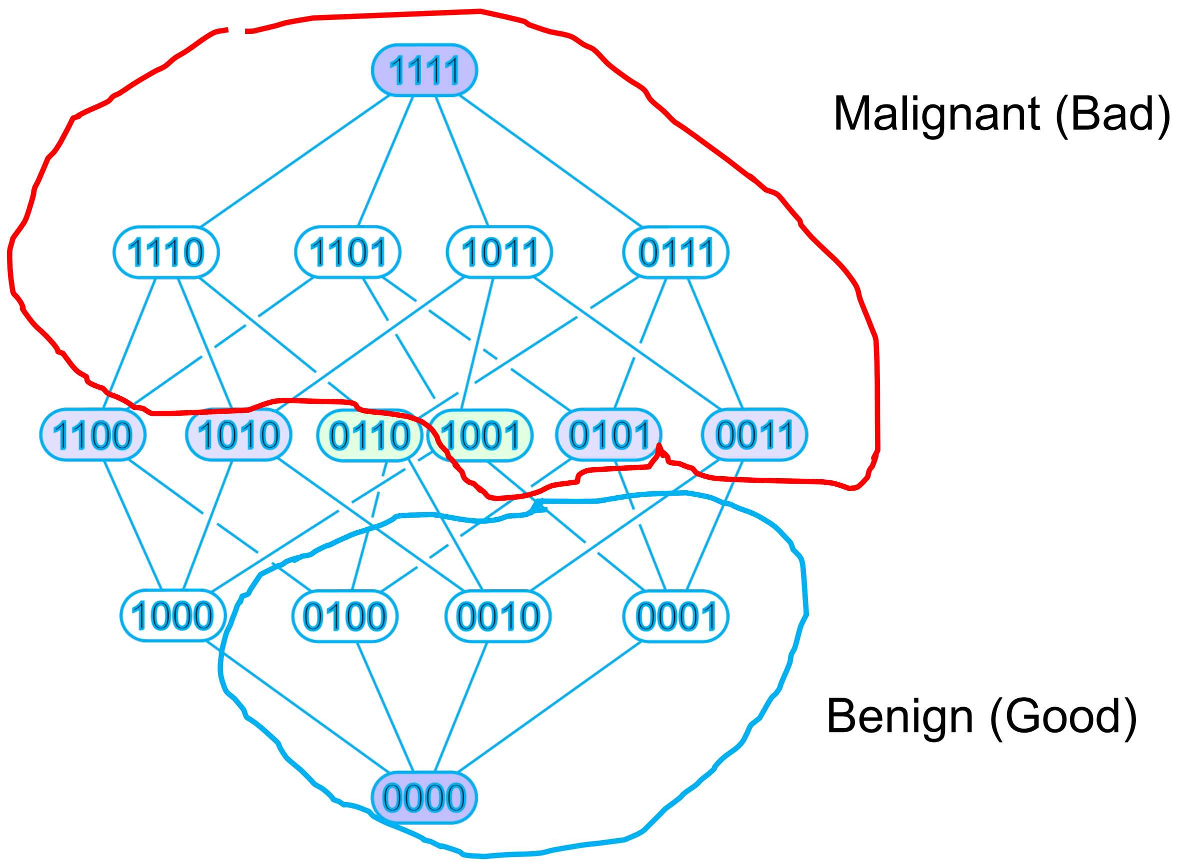
Data Set

- # of instances: 683
  - Benign: 444
  - Malignant: 239
- # of variables: 9
  - Each variable takes10 values *P*= {1,2,...,10}.
  - Larger value implies malignant tumor.
  - Smaller value implies benign tumor.
- Assume this represents a monotone increasing function.

Number of Conflicting Pairs

- 18 pairs among 444x 239=106116.
- Removed
  - 2 benign
  - 4 malignant
- Training set consists of
  - 442 benign
  - 235 malignant
- After simplification using monotone property:
  - 25 benign
  - 232 malignant.

Hasse Diagram of Monotone Increasing Function



Multi-Valued Approach

- Found a 7-variable solution:
  - $\{X_1, X_4, X_5, X_6, X_7, X_8, X_9\}$

Two-Valued Approach

- Prof. Boros' group of Rutgers University represented a 10-valued variable by 9 two-valued variables.
- They used 9 x 9=81 two-valued variables to represent the function.
- They found a solution with 11 two-valued variables.
- Our multi-valued approach required only 7 variables.

Conclusion

- Showed a method to minimize support sets for multi-valued input decision functions.
- Showed a method to make a consistent training set by removing the minimum number of instances.
- Minimized the support sets for poisonous mushrooms; hepatitis; and breast cancer.
- Showed that the multi-valued approach is direct and efficient.

Comments

- Logic minimization is useful for **data mining.**
- Especially, for medical applications, doctors must explain the reason of their decision to patients and insurance companies.
- **The rule must be simple**, so that everybody can understand.

- Detail will be shown in ISMVL-2019 proceedings, May 21-23, Fredericton, Canada.