Rule-based classification example

(a) In rule based classification, consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules:

$R_1: A \rightarrow +$ (covers 4 positive and 1 negative examples)

$R_2: B \rightarrow +$ (-//- 30 -//- 10 -//-)

$R_3: C \rightarrow +$ (-//- 100 -//- 90 -//-)

determine which is the best and worst candidate rule according to:

(i) Rule accuracy , (ii) FOIL's information gain .

(b) Explain why rule accuracy and FOIL's information gain rank the rules differently.

ΛΥΣΗ

(a)(i) $R_1: A \rightarrow +$, antecedent $\rightarrow$ consequent

$$\text{Coverage of rule } R_1 = \frac{\#\text{of records that satisfy antecedent}}{\#\text{of total records/samples}}$$

$$\text{Accuracy of rule } R_1 = \frac{\#\text{of records that satisfy antecedent AND consequent}}{\#\text{of records that satisfy (only) antecedent}}.$$

$R_1: A \rightarrow +$: $\text{Accuracy}_{R_1} = \frac{4}{4+1} = \frac{4}{5} = 80\%$

$R_2: B \rightarrow +$: $\text{Accuracy}_{R_2} = \frac{30}{40} = \frac{3}{4} = 75\%$

$R_3: C \rightarrow +$: $\text{Accuracy}_{R_3} = \frac{100}{190} = \frac{10}{19} = 52.6\%$

$\rightarrow$ according to accuracy, the best candidate rule is R_1 and the worst is R_3.

(ii) FOIL's info gain : $\left\{ \begin{array}{l} R_0: \{\} \rightarrow \text{class} \\ R_i: \{A\} \rightarrow \text{class} \end{array} \right\}$ $\text{Gain}(R_0, R_i) = t \cdot \left(\log_2 \frac{p_i}{p_i + n_i} - \log_2 \frac{p_0}{p_0 + n_0} \right)$.

t : number of positive instances covered by both R_0 and R_1.

p_0 : number -//- by R_0

n_0 : number of negative instances covered by R_0

p_i : -//- positive -//- R_i

n_i : -//- negative -//- R_i

→ αυτό θα είναι ίσο με p_i

→ αυτό πρακτικά είναι το accuracy.

R_1 : $\text{Gain}(R_0, R_1) = 4 \cdot \left[\log_2 \left(\frac{4}{4+1} \right) - \log_2 \left(\frac{100}{100+400} \right) \right] = 4 \cdot \left(\log_2 \frac{4}{5} - \log_2 \frac{1}{5} \right) = 4 \log_2 4 = 8$.

R_2 : $\text{Gain}(R_0, R_2) = 30 \left[\log_2 \frac{30}{30+10} - \log_2 \frac{100}{500} \right] = 30 \left(\log_2 \frac{3}{4} - \log_2 \frac{1}{5} \right) \simeq 57.2$

R_3 : $\text{Gain}(R_0, R_3) = 100 \cdot \left[\log_2 \frac{100}{100+90} - \log_2 \frac{100}{500} \right] = 100 \left(\log_2 \frac{10}{19} - \log_2 \frac{1}{5} \right) \simeq 139.6$.

Therefore, based on information gain : R_3: best , R_1: worst. which is the opposite of what we found on (i).

(b) Rule accuracy only accounts for the portion / percentage of samples that satisfy the antecedent as well as the consequent out of the ones that satisfy the antecedent. However, such a rule can concern only a small portion of the total samples of the dataset and hence it might not be a particularly useful rule for splitting the data.

FOIL's information gain on the other hand has a weight (t) or a multiplier that expresses this dependence on how many samples are satisfying the rule's antecedent, and thus is a better ~~indicator~~ metric for rule ranking especially for the top-levels of a rule-based classifier.

Association rules example

(a) Perform the apriori algorithm for the given transaction database, using the minimum support 0.3 and the minimum confidence 0.77. Note that you have to show how the algorithm is performed.

Row	Transaction
1	{a b}
2	{b c}
3	{a b c}
4	{d e f}
5	{a b c}
6	{d f}
7	{c d e f}
8	{a b c d e}

(b) Assuming we have a rule $I_1 \rightarrow I_2$. Describe how to interpret the situations where the rule has:
- low support and high confidence,
- high support and low confidence.

(c) Assume that the rule {1 2} → {3 4} is in the final set of rules, and the rule {3 4} → {1 2} is not in the final set. For each of the following rules, state if the rule definitely appears in the final set, if there is a possibility that it appears in the final set, or if it definitely does not appear in the final set of rules.

(i) {1 2 3} → {4}, (ii) {1} → {2 3 4}, (iii) {2 3 4} → {1} (iv) {3} → {1 2 4}

ANSWER

k=1	Support of itemset		k=2			k=3		
C_1	σ	F_1	C_2	σ	F_2	C_3	σ	F_3
{a}	4/8 = 0.5 ✓	{a}	{a b}	4/8 = 0.5 ✓	{a b}	{a b c}	3/8 = 0.375 ✓	{a b c}
{b}	5/8 = 0.625 ✓	{b}	{a c}	3/8 = 0.375 ✓	{a c}	{d e f}	2/8 = 0.25 ✗	
{c}	5/8 = 0.625 ✓	{c}	{a d}	1/8 = 0.125 ✗	{b c}			
{d}	4/8 = 0.5 ✓	{d}	{a e}	1/8 = 0.125 ✗	{d e}			
{e}	3/8 = 0.375 ✓	{e}	{a f}	0/8 = 0 ✗	{d f}			
{f}	3/8 = 0.375 ✓	{f}	{b c}	4/8 = 0.5 ✓				
			{b d}	1/8 = 0.125 ✗				
			{b e}	1/8 = 0.125 ✗				
			{b f}	0/8 = 0 ✗				
			{c d}	2/8 = 0.25 ✗				
			{c e}	2/8 = 0.25 ✗				
			{c f}	1/8 = 0.125 ✗				
			{d e}	3/8 = 0.375 ✓				
			{d f}	3/8 = 0.375 ✓				
			{e f}	2/8 = 0.25 ✗				

The support of an itemset X is: $\sigma(X) = \frac{\#\text{ of transactions containing X itemset}}{\#\text{ of all transactions} \rightarrow 8 \text{ in this example}}$

For k=1, we get all the k-itemsets (ie. 1-itemsets) in set C_1 and calculate the support of each 1-itemset. For the itemsets X where $\sigma(X) > \sigma_t = 0.3$ we say that they are frequent and keep them in set $F_k = F_1$. Here we found all 1-itemsets are frequent. Then, we move on to construct possible frequent 2-itemsets (k=2). They are $\binom{6}{2} = \frac{6!}{2!4!} = 15$ in number as shown above in the table and all 2-itemsets are kept in set C_2. Again, each itemset's support is calculated and those itemsets with $\sigma(X) > \sigma_t = 0.3$ are considered frequent and kept in frequent itemset set $F_2 = \{\{a b\}, \{a c\}, \{b c\}, \{d e\}, \{d f\}\}$. Now, to construct C_3 with possible 3-itemset we take we merge two 2-itemsets only if their $k-2 = 3-2 = 1$ first items are the same. In such a case a 3-itemset is created. Again we evaluate 3-itemsets based on their support value and if $\sigma(X) > \sigma_t = 0.3$ we include them in F_3. We find $F_3 = \{a b c\}$.

Scanned with CamScanner

Then, since we only have one 3-itemset in F_3 (set of frequent 3-itemsets), we cannot construct any 4-itemsets $\Rightarrow F_4 = \phi$. The algorithm (A priori) stops.

Finaly: 1-itemsets that are frequent: $F_1 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f\}\}$

2-itemsets that are frequent: $F_2 = \{\{a\ b\}, \{a\ c\}, \{b\ c\}, \{d\ e\}, \{d\ f\}\}$

3-itemsets that are frequent: $F_3 = \{a\ b\ c\}$.

All frequent itemsets: $FI = F_1 \cup F_2 \cup F_3$.

Next, we find possible association rules among frequent itemsets.

For each frequent k-itemset, we get $2^k - 2$ association rules.

- $\forall$ 1-itemset: $2^1 - 2 = 0$ association rules (as expected)
- $\forall$ 2-itemset: $2^2 - 2 = 2$ association rules, 5 2-itemsets $\Rightarrow 2 \cdot 5 = 10$ association rules
- $\forall$ 3-itemset: $2^3 - 2 = 6$ association rules, 1 3-itemset $\Rightarrow 6 \cdot 1 = 6$ association rules

\+ 16 association rules in total

Association rules	confidence	
$\{a\} \rightarrow \{b\}$	$0.5/0.5 = 1$	✓
$b \rightarrow a$	$0.5/0.625 = 0.8$	✓
$a \rightarrow c$	$0.375/0.5 = 0.75$	×
$c \rightarrow a$	$0.375/0.625 = 0.6$	×
$b \rightarrow c$	$0.5/0.625 = 0.8$	✓
$c \rightarrow b$	$0.5/0.625 = 0.8$	✓
$d \rightarrow e$	$0.375/0.5 = 0.75$	×
$e \rightarrow d$	$0.375/0.375 = 1$	✓
$d \rightarrow f$	$0.375/0.5 = 0.75$	×
$f \rightarrow d$	$0.375/0.375 = 1$	✓
$ab \rightarrow c$	$0.375/0.5 = 0.75$	×
$ac \rightarrow b$	$0.375/0.375 = 1$	✓
$bc \rightarrow a$	$0.375/0.5 = 0.75$	×
$a \rightarrow bc$	$0.375/0.5 = 0.75$	×
$b \rightarrow ac$	$0.375/0.625 = 0.6$	×
$c \rightarrow ab$	$0.375/0.625 = 0.6$	×

C_t = confidence-threshold = 0.77

We keep those association rules with confidence$(X \rightarrow Y) > C_t$

Finally the association rules that are considered as strong (given C_t) are:

$\{a\} \rightarrow \{b\}$
$\{b\} \rightarrow \{a\}$
$\{b\} \rightarrow \{c\}$
$\{c\} \rightarrow \{b\}$
$\{e\} \rightarrow \{d\}$
$\{f\} \rightarrow \{d\}$
$\{ac\} \rightarrow \{b\}$.

(b) $I_1 \rightarrow I_2$.

- Low support & high confidence: $I_1 \cup I_2$ is seldom bought, but when I_1 is bought it's highly probable that I_2 is also bought. $c(I_1 \rightarrow I_2) = \sigma(I_1 \cup I_2)/\sigma(I_1)$ is high which means $\sigma(I_1)$ is small meaning that I_1 is a relatively uncommon itemset in the available transactions $\Rightarrow$ if I_2 is also uncommon: strong rule but seldom applicable. eg $\{$Ipod$\} \rightarrow \{$Special Ipod Headphones$\}$.
- High support & low confidence: $I_1 \cup I_2$ are often bought together, but I_1 is even more often bought, hence the rule is weak and someone interested in I_1 might not necessarily be interested in I_2. eg $\{$plastic bag$\} \rightarrow \{$cheap beer$\}$

(c) $\{1\ 2\} \to \{3\ 4\}$: in final set of rules $\Rightarrow c(X \to Y) > c_t$

$\{3\ 4\} \to \{1\ 2\}$: not in final set of rules $\Rightarrow c(Y \to X) < c_t$.

όπου $X = \{1\ 2\}$, $Y = \{3\ 4\}$

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{\sigma(\{1\ 2\ 3\ 4\})}{\sigma(\{1\ 2\})} > c_t$$

$$c(Y \to X) = \frac{\sigma(Y \cup X)}{\sigma(Y)} = \frac{\sigma(\{1\ 2\ 3\ 4\})}{\sigma(\{3\ 4\})} < c_t.$$

Ισχύουν τα εξής:

$\sigma(\{1\}) > \sigma(\{1\ 2\})$

$\sigma(\{2\}) > \sigma(\{1\ 2\})$

$\sigma(\{3\}) > \sigma(\{3\ 4\})$

$\sigma(\{4\}) > \sigma(\{3\ 4\})$

$\sigma(\{1\ 2\ 3\}) < \sigma(\{1\ 2\})$.

- $\{1\ 2\ 3\} \to \{4\}$: $c(\{1\ 2\ 3\} \to \{4\}) = \dfrac{\sigma(\{1\ 2\ 3\ 4\})}{\sigma(\{1\ 2\ 3\})} > c(X \to Y) > c_t$.

  άρα σίγουρα στο final set

- $\{1\} \to \{2\ 3\ 4\}$: $c(\{1\} \to \{2\ 3\ 4\}) = \dfrac{\sigma(\{1\ 2\ 3\ 4\})}{\sigma(\{1\})} < c(X \to Y)$

  άρα ίσως στο final set

  Εξαρτάται από ακριβής τιμή $\sigma(\{1\})$

  π.χ. αν $c_t = 0.8$ και βγει $c(X \to Y) = 0.9$ τότε $c(\{1\} \to \{2\ 3\ 4\}) < c(X \to Y) = 0.9$

  και π.χ. αν πάρει τιμή 0.85 τότε στο final set, αλλά αν πάρει τιμή 0.7 τότε δεν είναι στο final set.

- Ομοίως $\{2\ 3\ 4\} \to \{1\}$: ίσως

- $\{3\} \to \{1\ 2\ 4\}$: σίγουρα όχι στο final set.