Censoring and Pricing Data

Peter Buneman   Dan Suciu

July 27, 2007

1 The Censoring Problem

Alice has set of items $U = \{a_1, \ldots, a_n\}$, and wishes to publish a subset of it, $S \subseteq U$. A powerful censor, Carol, controls what Alice is allowed to publish, but Carol's rulings are not very clear. Carol gives only some positive, and some negative examples of what is publishable and what is not, and let's Alice figure out for herself what to do in all other cases. Carol says:

- Positive examples: $P_1, \ldots, P_k \subseteq U$. For any $i = 1, k$, Carol says: "it is acceptable to publish P_i or any subset thereof".
- Negative examples: $N_1, \ldots, N_m \subseteq U$. For any $j = 1, m$, Carol says: "it is forbidden to publish N_j or any superset thereof".

Now Alice has a set $S \subseteq U$ and wants a numerical function $M(S) \in [0, 1]$ that measures her risk of being punished by Carol if she publishes S. More precisely, she wants the measure to satisfy the following:

- $\forall i = 1, k$, $M(P_i) = 0$
- $\forall j = 1, m$, $M(N_j) = 1$
- If $S \subseteq S'$ then $M(S) \leq M(S')$
- Suppose we add a new positive set P_{k+1} which is identical to an existing one: $P_{k+1} = P_i$ for some $i = 1, k$. Let M' be the new function. Then we want $M' = M$. Similarly if we add a new negative set identical to an existing one.

1.1 A Solution for a Restricted Case

Let's study the problem for negative sets only: $N_1, \ldots, N_m$. We define below a function $M(S)$ that satisfies the criteria above. For that we need some notations:

$$\begin{aligned}
N_I &= (\bigcap_{i\in I})N_i \cap (\bigcap_{i\notin N_i}(U - N_i)), \forall I \subseteq \{1,\ldots,m\} \\
Cells &= \{I \mid N_I \neq \emptyset\} \\
\alpha_I &= \frac{|S \cap N_I|}{|N_I|}, \forall I \in Cells \\
i\uparrow &= \{I \mid I \in Cells, i \in I\}, \forall i \in \{1,\ldots,m\}
\end{aligned}$$

Let E_I be the statement: $N_I \subseteq S$. Denote $M(E_I) = \alpha_I$: that is $M(E_I)$ is now a continuous measure of the truthfulness of the statement $N_I \subseteq S$. The logical statement $N_i \subseteq S$ is now the conjunction $\bigwedge_{I\in i\uparrow} E_I$, and its measure is $\prod_{I\in i\uparrow} M(E_I)$, since we assume that any set of events E_I are independent, i.e. cells are independent.

We want now to define the measure $M(S)$ of the truthfulness of the logical statement:

$$\bigvee_{i=1,m} \bigwedge_{I\in i\uparrow} E_I$$

We treat now M as a probability function and use the inclusion/exclusion formula:

$$\begin{aligned}
M(S) &= M(\bigvee_{i=1,m} \bigwedge_{I\in i\uparrow} M(E_I)) \\
&= \sum_{J\subseteq\{1,\ldots,m\}, J\neq\emptyset} (-1)^{|J|-1} M(\bigwedge_{i\in J, I\in i\uparrow} E_I) \\
&= \sum_{J\subseteq\{1,\ldots,m\}, J\neq\emptyset} (-1)^{|J|-1} \prod_{I:\exists i\in J, I\in i\uparrow} \alpha_I \qquad (1)
\end{aligned}$$

Example 1.1 Let $m = 2$. Then we have two "negative" examples from Carol: N_1 and N_2, and assume $U = N_1 \cup N_2$, hence there are only three cells: $Cell = \{1, 12, 2\}$ (more formally $\{\{1\}, \{1,2\}, \{2\}\}$). Let $S \subseteq N_1 \cup N_2$ and denote:

$$\begin{aligned}
N_{12} &= N_1 \cap N_2 \\
\alpha_1 &= |S \cap N_1|/|N_1| \\
\alpha_2 &= |S \cap N_2|/|N_2| \\
\alpha_{12} &= |S \cap N_{12}|/|N_{12}|
\end{aligned}$$

We have $1\uparrow = \{\{1\}, \{1,2\}\}$, $2\uparrow = \{\{1,2\}, \{2\}\}$ and Formula (1) gives:

$$\begin{aligned}
M(S) = {} & \alpha_1 {*} \alpha_{12} \quad \text{for } J = \{1\} \\
& + \alpha_2 {*} \alpha_{12} \quad \text{for } J = \{2\} \\
& - \alpha_1 {*} \alpha_{12} {*} \alpha_2 \quad \text{for } J = \{1,2\}
\end{aligned}$$

2 The Pricing Problem

Now Alice wants to sell subsets of U, but she has decided on prices only for some of the subsets. More precisely she decides on the following prices:

$$M(N_1) = p_1, \ldots, M(N_m) = p_m$$

What is the price $M(S)$ of any set $S \subseteq U$? For a set of cells $\mathcal{I} \subseteq Cells$ define:

$$\beta_{\mathcal{I}} = \prod_{I \in \mathcal{I}} \alpha_I \times \prod_{I \notin \mathcal{I}} (1 - \alpha_I)$$

Then the following pricing function satisfies our criteria:

$$M(S) = \sum_{i \in \{1,\ldots,m\}} c_i * \beta_{i\uparrow} \qquad (2)$$

Example 2.1 Suppose we are given the following prices:

$$\begin{aligned} M(N_1) &= p_1 \\ M(N_2) &= p_2 \\ M(U) &= p_3 \end{aligned}$$

Here we have:

$$\begin{aligned} Cells &= \{13, 123, 23, 3\} \\ 1\uparrow &= \{13, 123\} \\ 2\uparrow &= \{123, 23\} \\ 3\uparrow &= \{13, 123, 23, 3\} \\ \beta_{1\uparrow} &= \alpha_{13} * \alpha_{123} * (1 - \alpha_{23}) * (1 - \alpha_3) \\ \beta_{2\uparrow} &= (1 - \alpha_{13}) * \alpha_{123} * \alpha_{23} * (1 - \alpha_3) \\ \beta_{3\uparrow} &= \alpha_{13} * \alpha_{123} * \alpha_{23} * \alpha_3 \end{aligned}$$

It follows that the price function given by Equation (2) is:

$$\begin{aligned} M(S) &= p_1 \beta_{1\uparrow} + p_2 \beta_{2\uparrow} + p_3 \beta_{3\uparrow} \\ &= p_1 * \alpha_{13} * \alpha_{123} * (1 - \alpha_{23}) * (1 - \alpha_3) \\ &\quad + p_2 * (1 - \alpha_{13}) * \alpha_{123} * \alpha_{23} * (1 - \alpha_3) \\ &\quad + p_3 * \alpha_{13} * \alpha_{123} * \alpha_{23} * \alpha_3 \end{aligned}$$

2.1 Solving the Censoring Problem

We can now solve the censoring problem in the general case, by reducing it to the pricing problem. Set the following prices:

$$\begin{aligned} M(P_1) = \ldots = M(P_k) = 0 \\ M(N_1) = \ldots = M(N_m) = M(U) = 1 \end{aligned}$$

In the particular case where $k = 0$ (i.e. no positive examples) Equation (2) becomes Equation (1).