A Decision Model for Cost Optimal Record Matching

Presenter: *Vassilios S. Verykios*

IST College / Drexel University

Affiliates Workshop on Data Quality

NISS/Telcordia - December 1st, 2000

Comparison Vector

- Given a pair of database records with partially overlapping schemata, decide whether it is a match or not.
- Compare the pairs of values stored in each common attribute/field (assume *n* common fields).
- The n comparison measurements form a *comparison vector X*.

Record Comparison

A	B	C	D	○ ○ ○	E	F

—

A	B		C	○ ○ ○	D	F

=

1. Agreement
2. Disagreement
3. Missing

1	1	3	2	○ ○ ○	2	1

Random Vector

- Even if a pair of records match, the observed value for each field comparison is different each time the observation is made.

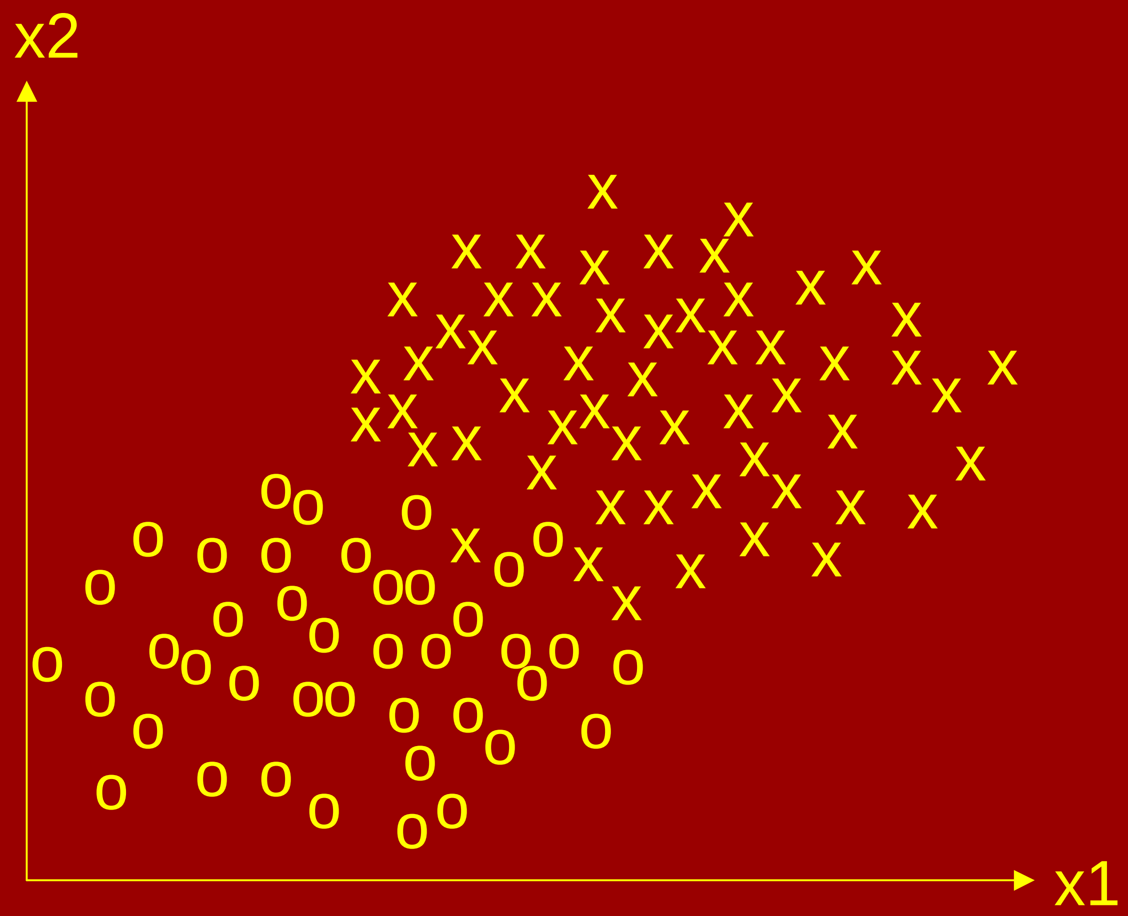
- Therefore, each *field comparison variable* is a *random variable*.

- Likewise, the *comparison vector X* is a *random vector*.

Distribution of Vectors

- Each pair of records is expressed by a comparison vector (or a sample) in an n-dimensional space.
- Many comparison vectors form a *distribution* of *X* in the n-dimensional space.
- Figure 1 shows a simple two dimensional example of two distributions corresponding to matched and unmatched pairs of records.

Figure 1

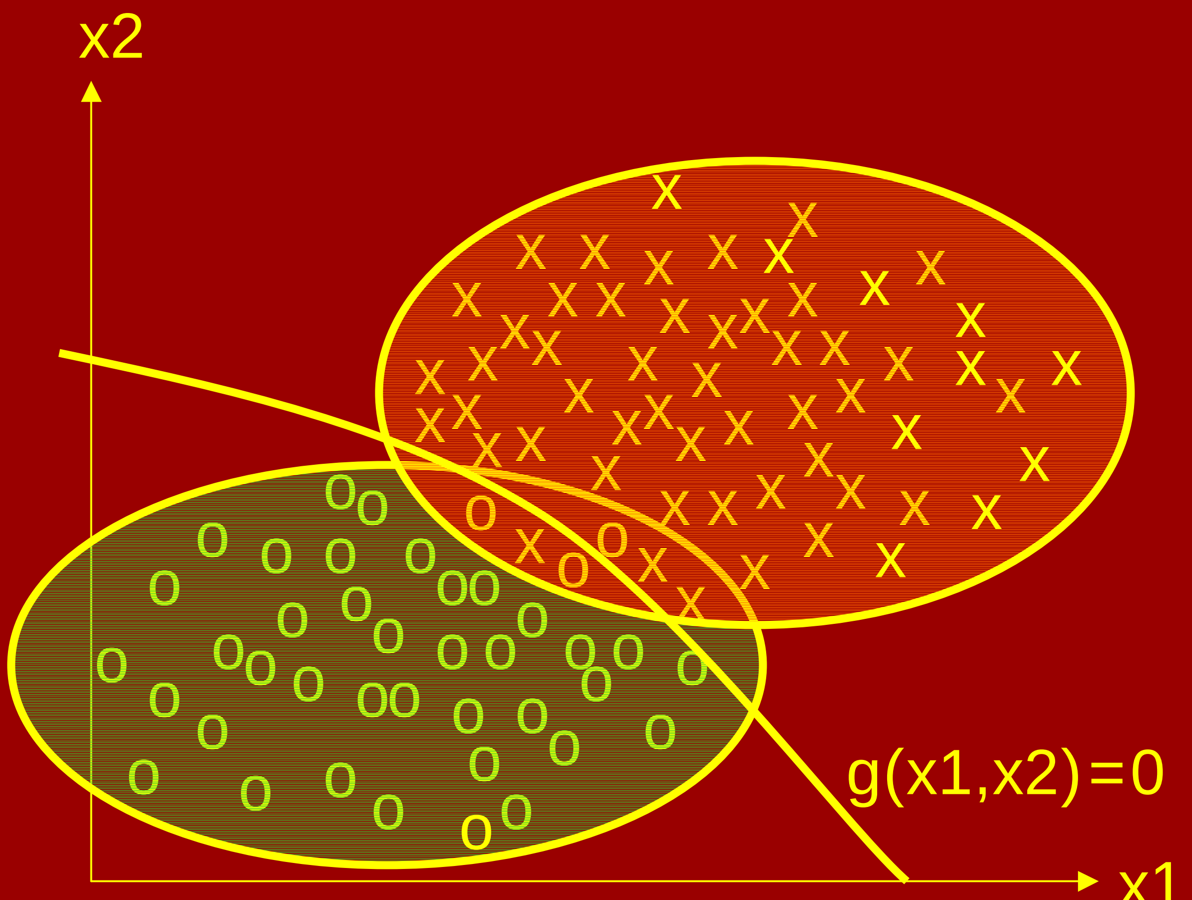


Distributions of samples from matched and unmatched record pairs.

Classifiers

- If we know these two distributions of X from past experience, we can set up a boundary between these two distributions, *g(x1,x2)=0*, which divides the two-dimensional space into two regions.
- Once the boundary is selected, we can classify a sample without a class label to a matched or unmatched, depending on the sign of *g(x1,x2).*
- We call *g(x1,x2)* a *discriminant function* and a system that detects the sign of *g(x1,x2)* a *classifier*.

Figure 2



Distributions of samples from matched and unmatched record pairs.

Learning

- In order to design a classifier, we must study the characteristics of the distribution of X for each category and find a proper discriminant function.
- This process is called *learning*.
- Samples used to design a classifier are called *learning* or *training samples*.

Statistical Hypothesis Testing

- What is the best classifier, assuming that the distributions of the random vectors are given?

- *Bayes classifier* minimizes the *probability of classification error*.

Distribution and Density Functions

- Random vector *X*
- Distribution function *P(X)*
- Density function *p(X)*
- Class *i* density or conditional density of class *i* $p(X|c_i)$ or $p_i(X)$
- Unconditional density function or mixture density function

$$p(X) = \sum_{i=1}^{L} P_i \, p_i(X)$$

- A posteriori density function $P(c_i|X)$ or $q_i(X)$
- *Bayes rule*

Bayes Rule for Minimum Error

- Let X a comparison vector.
- Determine whether X belongs to M or U.
- If the a posteriori probability of M given X is larger than the probability of U, X is classified to M, and vice versa.

Fellegi-Sunter Model

- Order X's based on their likelihood ratio

$$l(X) = \frac{p_M(X)}{p_U(X)}$$

- For a pair of error levels (μ, λ), choose index values n and n' such that:

$$\sum_{i=1}^{n-1} p_U(X_i) < \mu \leq \sum_{i=1}^{n} p_U(X_i)$$

$$\sum_{i=n'}^{N} p_M(X_i) \geq \lambda > \sum_{i=n'+1}^{N} p_M(X_i)$$

Minimum Cost Model

- Minimizing the probability of error is not the best criterion to design a decision rule because the misclassifications of M and U samples may have different consequences.
- The misclassification of a cancer patient to normal may have a more damaging effect than the misclassification of a normal patient to cancer.
- Therefore, it is appropriate to assign a cost to each situation.

Decision Costs

Cost	**Decision**	**Class**
C_{1M}	A_1	M
C_{1U}	A_1	U
C_{2M}	A_2	M
C_{2U}	A_2	U
C_{3M}	A_3	M
C_{3U}	A_3	U

Mean Cost (I)

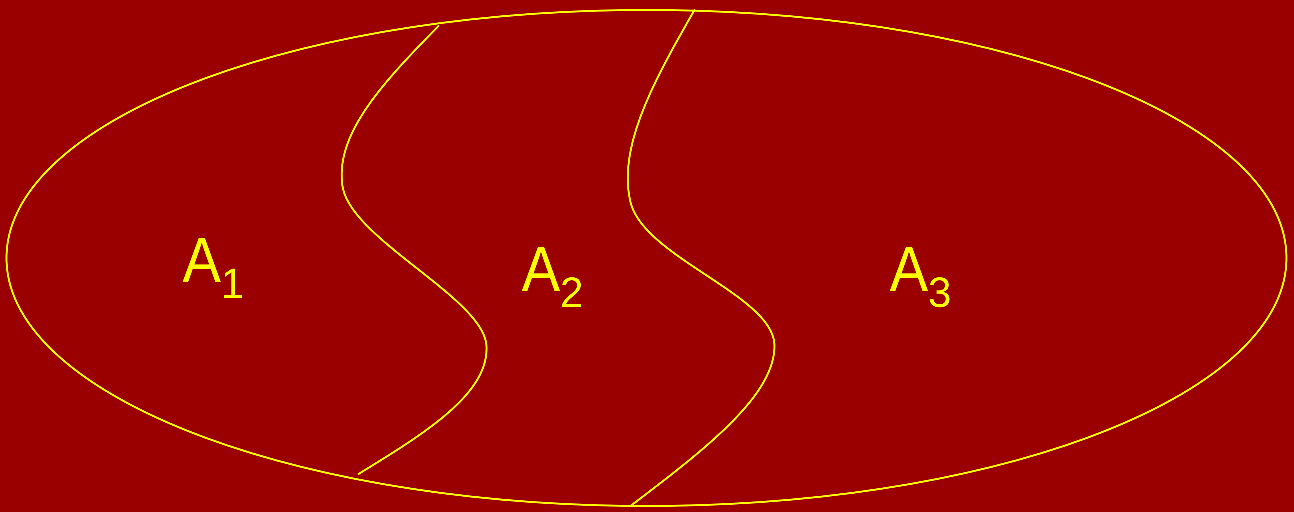
$$\bar{c} = c_{1M} \cdot P(d=A_1, c=M) + c_{1U} \cdot P(d=A_1, c=U) +$$
$$c_{2M} \cdot P(d=A_2, c=M) + c_{2U} \cdot P(d=A_2, c=U) +$$
$$c_{3M} \cdot P(d=A_3, c=M) + c_{3U} \cdot P(d=A_3, c=U)$$

Bayes Theorem

$$P(d=A_i, c=j) = P(d=A_i | c=j) \cdot P(c=j)$$

$$\text{where } i = 1, 2, 3 \text{ and } c = M, U$$

Conditional Probability



$$P(d=A_i|c=j)=\sum_{X\in A_i} p_j(X), \text{ where } i=1,2,3 \text{ and } c=M,U$$

$$P(c=M)=\pi_0 \text{ and } P(c=U)=1-\pi_0$$

Mean Cost (II)

Using the Bayes theorem:

$$\bar{c} = c_{1M} \cdot P(d{=}A_1 | c{=}M) \cdot P(c{=}M) + c_{1U} \cdot P(d{=}A_1 | c{=}U) \cdot P(c{=}U) + c_{2M} \cdot P(d{=}A_2 | c{=}M) \cdot P(c{=}M) + c_{2U} \cdot P(d{=}A_2 | c{=}U) \cdot P(c{=}U) + c_{3M} \cdot P(d{=}A_3 | c{=}M) \cdot P(c{=}M) + c_{3U} \cdot P(d{=}A_3 | c{=}U) \cdot P(c{=}U)$$

Using the definition of the conditional probability:

$$\bar{c} = c_{1M} \cdot \sum_{X \in A_1} p_M(X) \cdot \pi_0 + c_{1U} \cdot \sum_{X \in A_1} p_U(X) \cdot (1 - \pi_0) + c_{2M} \cdot \sum_{X \in A_2} p_M(X) \cdot \pi_0 + c_{2U} \cdot \sum_{X \in A_2} p_U(X) \cdot (1 - \pi_0) + c_{3M} \cdot \sum_{X \in A_3} p_M(X) \cdot \pi_0 + c_{3U} \cdot \sum_{X \in A_3} p_U(X) \cdot (1 - \pi_0) +$$

Mean Cost (III)

$$\bar{c} = \sum_{X \in A_1} [p_M(X) \cdot c_{1M} \cdot \pi_0 + p_U(X) \cdot c_{1U} \cdot (1 - \pi_0)] +$$

$$\sum_{X \in A_2} [p_M(X) \cdot c_{2M} \cdot \pi_0 + p_U(X) \cdot c_{2U} \cdot (1 - \pi_0)] +$$

$$\sum_{X \in A_3} [p_M(X) \cdot c_{3M} \cdot \pi_0 + p_U(X) \cdot c_{3U} \cdot (1 - \pi_0)]$$

Decision Areas

- Every sample X in the decision space A, should be assigned to only one decision class: A_1, A_2 or A_3.
- We should thus assign each sample to a class in such a way that its contribution to the mean cost is minimum.
- This will lead to the optimal selection for the three sets which we denote by A_1^0, A_2^0, A_3^0.

Decision Making

- A sample is assigned to the optimal areas as follows:

To A_1^0 if:

$$p_M(X) \cdot c_{1M} \cdot \pi_0 + p_U(X) \cdot c_{1U} \cdot (1-\pi_0) \le p_M(X) \cdot c_{2M} \cdot \pi_0 + p_U(X) \cdot c_{2U} \cdot (1-\pi_0)$$

$$p_M(X) \cdot c_{1M} \cdot \pi_0 + p_U(X) \cdot c_{1U} \cdot (1-\pi_0) \le p_M(X) \cdot c_{3M} \cdot \pi_0 + p_U(X) \cdot c_{3U} \cdot (1-\pi_0)$$

To A_2^0 if:

$$p_M(X) \cdot c_{2M} \cdot \pi_0 + p_U(X) \cdot c_{2U} \cdot (1-\pi_0) \le p_M(X) \cdot c_{1M} \cdot \pi_0 + p_U(X) \cdot c_{1U} \cdot (1-\pi_0)$$

$$p_M(X) \cdot c_{2M} \cdot \pi_0 + p_U(X) \cdot c_{2U} \cdot (1-\pi_0) \le p_M(X) \cdot c_{3M} \cdot \pi_0 + p_U(X) \cdot c_{3U} \cdot (1-\pi_0)$$

To A_3^0 if:

$$p_M(X) \cdot c_{3M} \cdot \pi_0 + p_U(X) \cdot c_{3U} \cdot (1-\pi_0) \le p_M(X) \cdot c_{1M} \cdot \pi_0 + p_U(X) \cdot c_{1U} \cdot (1-\pi_0)$$

$$p_M(X) \cdot c_{3M} \cdot \pi_0 + p_U(X) \cdot c_{3U} \cdot (1-\pi_0) \le p_M(X) \cdot c_{2M} \cdot \pi_0 + p_U(X) \cdot c_{2U} \cdot (1-\pi_0)$$

Optimal Decision Areas

- We thus conclude from the previous slide:

$$A_1^0 = \left\{ X : \frac{p_U}{p_M} \le \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{3M} - c_{1M}}{c_{1U} - c_{3U}} \text{ and, } \frac{p_U}{p_M} \le \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{2M} - c_{1M}}{c_{1U} - c_{2U}} \right\}$$

$$A_2^0 = \left\{ X : \frac{p_U}{p_M} \ge \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{2M} - c_{1M}}{c_{1U} - c_{2U}} \text{ and, } \frac{p_U}{p_M} \le \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{3M} - c_{2M}}{c_{2U} - c_{3U}} \right\}$$

$$A_3^0 = \left\{ X : \frac{p_U}{p_M} \ge \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{3M} - c_{1M}}{c_{1U} - c_{3U}} \text{ and, } \frac{p_U}{p_M} \ge \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{3M} - c_{2M}}{c_{2U} - c_{3U}} \right\}$$

Threshold Values

$$c_{1M} \le c_{2M} \le c_{3M}, \quad c_{1U} \ge c_{2U} \ge c_{3U}$$

$$\kappa = \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{3M} - c_{1M}}{c_{1U} - c_{3U}}$$

$$\lambda = \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{2M} - c_{1M}}{c_{1U} - c_{2U}}$$

$$\mu = \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{3M} - c_{2M}}{c_{2U} - c_{3U}}$$

Threshold Values

- In order for A_2^0 to exist:

$$\lambda = \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{3M} - c_{1M}}{c_{1U} - c_{3U}} \leq \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{3M} - c_{2M}}{c_{2U} - c_{3U}} = \mu$$

- We can easily prove now, that threshold κ lies between λ and μ.

Threshold Values

$$\lambda = \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{2M} - c_{1M}}{c_{1U} - c_{2U}} \Rightarrow \lambda \cdot (c_{1U} - c_{2U}) = \frac{\pi_0}{1-\pi_0} \cdot (c_{2M} - c_{1M})$$

$$\lambda \leq \frac{\pi_0}{1-\pi_0} \cdot \frac{c_{3M} - c_{2M}}{c_{2U} - c_{3U}} \Rightarrow \lambda \cdot (c_{2U} - c_{3U}) \leq \frac{\pi_0}{1-\pi_0} \cdot (c_{3M} - c_{2M})$$

- Adding by parts the relationships above, we can easily show that $\lambda \leq \kappa$
- Similarly we can prove that $\kappa \leq \mu$.

Optimality of the Model

$$\bar{c} = \sum_{X \in A_1} z_1(X) + \sum_{X \in A_2} z_2(X) + \sum_{X \in A_3} z_3(X) =$$

$$\sum_{X \in A} [z_1(X) \cdot I_{A_1}(X) + z_2(X) \cdot I_{A_2}(X) + z_3(X) \cdot I_{A_3}(X)] \geq$$

$$\sum_{\mathrm{X} \in \mathrm{A}} \min\{z_1(X), z_2(X), z_3(X)\} \overset{\mathrm{def}}{=}$$

$$\sum_{X \in A_1^0} z_1(X) + \sum_{X \in A_2^0} z_2(X) + \sum_{X \in A_3^0} z_3(X)$$

Probabilities of Errors

- Type I

$$P(d=A_3, r=M) = P(d=A_3 | r=M) \cdot P(r=M)$$
$$= \pi_0 \cdot \sum_{X \in A_3} p_M(X)$$

- Type II

$$P(d=A_1, r=U) = P(d=A_1 | r=U) \cdot P(r=U)$$
$$= (1-\pi_0) \cdot \sum_{X \in A_1} p_U(X)$$

Conditionally Independent Binary Components

$$X = [x_1 \; x_2 \; \cdots \; x_n]$$

$$p_j(X) = p_j(x_1) \cdot p_j(x_2) \cdots p_j(x_n),$$

where $j = M, U$

$$p_M(x_i = 1) = p_i$$

$$p_M(x_i = 0) = 1 - p_i$$

$$p_U(x_i = 1) = q_i$$

$$p_U(x_i = 0) = 1 - q_i$$

Conditionally Independent Binary Components

$$\log \frac{p_U}{p_M}(x_1, x_2, \ldots x_n) = \log \frac{p_U(x_1) \cdot p_U(x_2) \cdots p_U(x_n)}{p_M(x_1) \cdot p_M(x_2) \cdots p_M(x_n)}$$

$$\log \frac{p_U}{p_M}(x_1, x_2, \ldots x_n) = \log \frac{p_U(x_1)}{p_M(x_1)} + \log \frac{p_U(x_2)}{p_M(x_2)} + \cdots + \log \frac{p_U(x_n)}{p_M(x_n)}$$

$$= \sum_{i=1}^{n} \log \frac{p_U(x_i)}{p_M(x_i)}$$

Conditionally Independent Binary Components

- Note, that since x_i can only assume the values of 0 or 1:

$$\log\frac{p_U}{p_M}(x_i) = x_i \cdot \log\frac{p_i}{q_i} + (1-x_i)\cdot\log\frac{1-q_i}{1-p_i}$$

$$= x_i \cdot \log\frac{q_i \cdot (1-p_i)}{p_i(1-q_i)} + \log\frac{1-q_i}{1-p_i}$$

$$\log\frac{p_U}{p_M}(X) = \sum_{i=1}^{n} x_i \cdot \log\frac{q_i \cdot (1-p_i)}{p_i(1-q_i)} + \sum_{i=1}^{n}\log\frac{1-q_i}{1-p_i}$$

Example

- Records are being compared.
- Three attributes: last name, first name and sex.
- Two possible outcomes: agree and disagree.
- Comparison vector contains eight 3-component vectors.

Probabilities of Agreement and Disagreement

Attribute	Under M		Under U	
	p_i	$1-p_i$	q_i	$1-q_i$
Last Name	0.90	0.10	0.05	0.95
First Name	0.85	0.15	0.10	0.90
Sex	0.95	0.05	0.45	0.55

Comparisons and Costs

$$X = (x_1, x_2, x_3)$$

if attrbitute values agree

then $x_i = 1$ else $x_i = 0$

$$c_{1M} = 0,\ c_{2M} = 0.2,\ c_{3M} = 1$$

$$c_{1U} = 1,\ c_{2U} = 0.2,\ c_{3U} = 0$$

$$\pi_0 = 1 - \pi_0 = 0.5$$

Decisions Made

i	X	$\text{Log}(p_U/p_M)$	Decision
1	(0,0,0)	2.795	A_3
2	(0,0,1)	1.429	A_3
3	(0,1,0)	1.088	A_3
4	(0,1,1)	-0.272	A_2
5	(1,0,0)	0.562	A_2
6	(1,0,1)	-0.804	A_1
7	(1,1,0)	-1.145	A_1
8	(1,1,1)	-2.511	A_1

Experiments

Attribute	Under M		Under U	
	p_i	$1\text{-}p_i$	q_i	$1\text{-}q_i$
SSN	1.00	0.00	0.35	0.65
FNAME	0.96	0.04	0.29	0.71
MINIT	0.95	0.05	0.05	0.95
LNAME	0.97	0.03	0.30	0.70
STREET#	1.00	0.00	0.00	1.00
SADDRESS	0.77	0.23	0.01	0.99
APRT#	1.00	0.00	0.00	1.00
CITY	0.89	0.11	0.06	0.94
STATE	1.00	0.00	0.00	1.00
ZIPCODE	0.97	0.03	0.43	0.75

Percent of Error VS. No of Records in A_2

GID	c_{2M}	c_{2U}	λ	μ	%Error	% of Recs in A_2
A	0.50	0.50	-0.2126	-0.2126	1.0013	0.0000
	0.40	0.60	-0.2126	-0.2126	1.0013	0.0000
B	0.50	0.25	-0.3887	0.0884	1.0013	0.0000
	0.50	0.05	-0.4914	0.7874	1.0013	0.0062
	0.50	0.005	-0.5115	1.7884	0.3650	1.1692
	0.50	0.0005	-0.5134	2.7874	0.3602	1.5797
C	0.25	0.25	-0.6897	0.2645	0.9890	0.0186
	0.1	0.1	-1.1668	0.7416	0.9890	0.0186
	0.05	0.05	-1.4914	1.0661	0.9836	0.0995
	0.005	0.005	-2.5115	2.0862	0.3471	1.4553
	0.0005	0.0005	-3.5134	3.0882	0.2028	1.8720

Concluding Remarks

- Efficiency
- Time optimal models
- Prototype implementation