BLOCKING CONSIDERATIONS FOR RECORD LINKAGE UNDER CONDITIONS OF UNCERTAINTY

Robert Patrick Kelley, U.S. Bureau of the Census

I. INTRODUCTION

Record linkage, and its associated statistical problems, are a special case of a larger area of concern. This area makes use of various mathematical and statistical techniques to study the problems involved in the classification of observed phenomema. Discriminant analysis, discrete discriminant analysis, pattern recognition, cluster analysis and mathematical taxonomy are some of the specific fields which study various aspects of the classification problem. While record linkage contains its own specific set of problems it also has a great deal in common with these other fields.

The basic unit of study in the linking of two files F1 and F2 is F1xF2, the set of ordered pairs from F1 and F2. Given F1xF2, our job is to classify each pair as either matched or unmatched. This decision will be based on measurements taken on the record pairs. For example, if we are linking person records a possible measurement would be to compare surnames of the two records, and assign the value 1 for those pairs where there is agreement and 0 for those pairs where there is disagreement. These measurements will yield a vector, Γ, of observations on each record pair.

The key fact which will allow us to link the two files is that Γ behaves differently for matched and unmatched pairs. Statistically we model this by assuming that Γ is a random vector generated by $P(\cdot \mid M)$ on matched pairs and $P(\cdot \mid U)$ on unmatched pairs. Thus, the Γ value for a single randomly selected record pair is generated by $pP(\cdot \mid M) + (1-p)\, P(\cdot \mid U)$ where p is the proportion of matched records.

This model for the problem is basically the same as the one used in discriminant analysis. In particular, as Γ is almost always discrete, the literature on discrete discriminant analysis is extremely useful (see for example Goldstein and Dillon (1978). There are however, several areas of concern that seem to be a great deal more important for record linkage than for the other classification techniques.

Our topic of discussion in this paper, blocking, arises from consideration of one of these problem areas. That area concerns the extreme size of the data sets involved for even a relatively small record linkage project. The size problem precludes our being able to study all possible record pairs. So, we must determine some rule which automatically assigns some pairs the match status of unmatched without further investigation. Such a rule is referred to as a blocking scheme since the resulting subset of record pairs often forms rectangular blocks in F1xF2.

Before we go on to dicuss the details of blocking we need to look at some background information on record linkage.

II. BACKGROUND

Again, our job in linking the two files F1 and F2 is to classify each record pair as either matched or unmatched. In practice, however, we usually include a clerical review decision for tricky cases. So, our set of possible decisions is

A_1: the pair is a match
A_2: no determination made - review
A_3: the pair is not a match.

Now, consider the class of decision functions $D(\cdot)$ which transform our space of comparison vector values, elements of which we will denote by γ, to the set of decisions $\{A_1, A_2, A_3\}$. Given two or more decision function in this class, what criterion will we use to choose between them?

In Fellegi and Sunter (1969) the argument is put forward that, as decision A_2 will require costly clerical review, we should pick a decision procedure which will minimize the expected number of A_2 decisions while keeping a bound on the expected number of pairs which are classified in error. Since the comparison vector values computed on the record pairs are identically distributed, this reduces to picking that decision procedure which will minimize $P(A_2)$ subject to $P(A_1 \mid U) \leq \mu$ and $P(A_3 \mid M) \leq \lambda$.

Given that you know $P(\cdot \mid M)$ and $P(\cdot \mid U)$ Fellegi and Sunter prove that the decision procedure which solves this problem is of the form

$$(1) \quad D(\gamma) = \begin{cases} A_3 & \text{if } \ell(\gamma) \leq t_1 \\ A_2 & \text{if } t_1 < \ell(\gamma) < t_2 \\ A_1 & \text{if } \ell(\gamma) \geq t_2 \end{cases}$$

where $\ell(\gamma) = P(\gamma \mid M)/P(\gamma \mid U)$ and t_1, t_2 are the least extreme values in the range of $\ell(\gamma)$ which satisfy the constraints.

It is this decision procedure that forms the basis for our study of the blocking problem.

III. BLOCKING

In the past sections we have outlined the more general aspects of record linkage and defined the blocking problem. In this section we will discuss blocking in the context of the decision procedure given in section II.

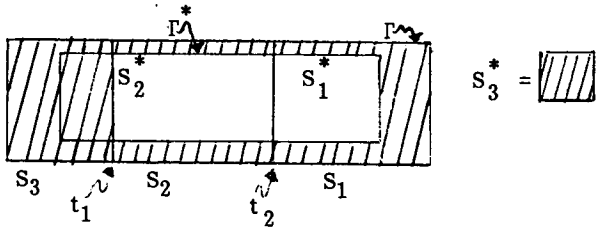
Our general blocking strategy is based on the fact that, for the type of files we work with, the number of matched pairs is considerably smaller then the number of unmatched pairs. So want try to restrict our investigation to pairs which have a good chance of being a match. The rest of the pairs will automatically be classified as unmatched. This will result in a reduction of the number of false matches and referrals at the expense of an increase in the number of false non-matches.

In Fellegi-Sunter (1969) this is accomplished by restricting the set of Γ vectors we are willing to study to a subspace Γ^*. Now, there are several possible ways to pick the best blocking subspace Γ^*. But we will restrict our attention to two methods.

The first method is suggested by the following amended decision procedure:

$$(2) \quad D'(\gamma) = \begin{cases} A_3 \text{ if } \ell(\gamma) \leq t_1 \text{ or } \gamma \in \Gamma^{*c} \\ A_2 \text{ if } t_1 < \ell(\gamma) < t_2 \text{ and } \gamma \in \Gamma^* \\ A_1 \text{ if } \ell(\gamma) \geq t_2 \text{ and } \gamma \in \Gamma^* \end{cases}$$

A Venn diagram of this situation is given by



where S_i and S_i^* are the regions of Γ values for which we make decision A_i under decision functions given by (1) and (2), respectively.

The error levels for this amended decision rule are given by

$$P(S_3^* \mid M) = P(S_3 \mid M) + P(S_3^* - S_3 \mid M)$$

$$= \lambda + P(S_3^* - S_3 \mid M).$$

and

$$P(S_1^* \mid U) = P(S_1 \mid U) - P(S_1 \cap S_3^* \mid U)$$

$$= \mu - P(S_1 \cap S_3^* \mid U).$$

Further, $P(A_2) = P(S_2) - P(S_2 \cap S_3^*)$.

These equations give us a means to compute a loss incurred by blocking on the subspace Γ^*. Namely, $P(S_3^* - S_3 \mid M)$, the increase in probability of a false non-match. The benefit gained from blocking on Γ^*, as opposed to using all of the Γ vector space to make our decision, takes the form of a decrease in the expected number of pairs which will have to be processed. Based on these calculations, we define the best blocking scheme to be the one which minimizes $P(\Gamma^*)$ subject to $P(S_3^* - S_3 \mid M) \leq \omega$.

The second method of comparing blocking schemes makes use of the conditional decision function D^* which is defined as follows:

$$D^*(\gamma) = \begin{cases} A_3 & \text{if } \ell^*(\gamma) \leq t_1^* \\ A_2 & \text{if } t_1^* < \ell^*(\gamma) < t_2^* \\ A_1 & \text{if } \ell^*(\gamma) \geq t_2^* \end{cases}$$

where $\gamma \in \Gamma^*$, $\ell^*(\gamma) = P(\gamma \mid \Gamma^*, M)/P(\gamma \mid \Gamma^*, U)$ and false non-match and match rates are λ^* and μ^* respectively. Now suppose we use D^* to form a new decision function, say D^{**}, on the whole space of Γ values. Let

$$D^{**}(\Gamma) = \begin{cases} D^*(\Gamma) & \text{if } \Gamma \in \Gamma^* \\ A_3 & \text{if } \Gamma \in \Gamma^{*c} \end{cases}$$

then the overall error rates for D^{**} are $P(\Gamma^* \mid M)\ \lambda^* + P(\Gamma^{*c} \mid M)$ for false non-match and $P(\Gamma^* \mid U)\ \mu^*$ for false match. Also, D^{**} has a total probability of an A_2 decision of $P(\Gamma^*)\ P(A_2 \mid \Gamma^*)$. To pick the best Γ^* we select the subspace which gives rise to the D^{**} decision which minimizes the probability of an A_2 decision subject to $P(A_1 \mid M) \leq \omega_1$ and $P(A_3 \mid U) \leq \omega_2$.

It is obvious that these two methods are related but it is unclear as to whether or not they are equivalent. At this point it would be benefical to consider an example. But before we do lets look at some of the practical aspects of blocking.

The previous decisions provide a general framework for studying, blocking; however, it does not give us any insight into the actual determination of a blocking subspace Γ^*. At first glance it is obvious that not all Γ^* will be feasible, since for many of them a Γ vector must actually be computed on each record pair before we can classify that pair as within or outside Γ^*. This would totally defeat the purpose of blocking. One solution for this problem is to block by using certain fields on the record (such as city, or state) or fields which we might add prior to matching (such as soundex code on surname or address range) as sort keys. The blocks would be determined by those record pairs with equal keys. Restricting our study to blocking schemes which are determined by sort keys implies that the comparison vector we want to use will consist of dichotomous components measuring agreement on the record identifier fields.

Now lets consider our example.

IV. AN EXAMPLE

Suppose our comparison vector consists of the agreement-disagreement pattern of three fields. Further, lets assume that the fields act independently under both

$P(\cdot \mid M)$ and $P(\cdot \mid U)$.

$$\text{So, } P((\gamma_1\ \gamma_2\ \gamma_3) \mid M) = \prod_{i=1}^{3} (m_i)^{\gamma_i}(1-m_i)^{1-\gamma_i}$$

$$\text{and } P((\gamma_1\ \gamma_2\ \gamma_3) \mid U) = \prod_{i=1}^{3} (u_i)^{\gamma_i}(1-u_i)^{1-\gamma_i}$$

where m_i equals the probability that the ith component agrees for a matched pair and u_i equals the probability that the ith component agrees for an unmatched pair.

This implies that

$$\ell(\gamma) = \prod_{i=1}^{3} \left(\frac{m_i}{u_i}\right)^{\gamma_i} \left(\frac{1-m_i}{1-u_i}\right)^{1-\gamma_i}$$

or

$$L(\gamma) = \ln(\ell(\gamma))$$

$$= \sum_{i=1}^{3} \gamma_i \ln\left(\frac{m_i}{u_i}\right) + (1-\gamma_i) \ln\left(\frac{1-m_i}{1-u_i}\right).$$

Now suppose that

$$m_1 = .90 \qquad u_1 = .05$$

$$m_2 = .85 \qquad u_2 = .10$$

$$m_3 = .95 \qquad u_3 = .45 .$$

So,

$$\frac{m_1}{u_1} = \frac{.90}{.05} = 18, \quad \frac{1-m_1}{1-u_1} = \frac{.10}{.95} = .1053$$

$$\frac{m_2}{u_2} = \frac{.85}{.10} = 8.5, \quad \frac{1-m_2}{1-u_2} = \frac{.15}{.90} = .1667$$

$$\frac{m_3}{u_3} = \frac{.95}{.45} = 2.11, \quad \frac{1-m_3}{1-u_3} = \frac{.05}{.55} = .091.$$

The possible patterns along with their corresponding ℓ and L values and their rank is as follows:

Pattern	ℓ	L	rank
(0,0,0)	.0015974	-6.44	1
(0,0,1)	.0370379	-3.30	2
(0,1,0)	.0814496	-2.51	3
(0,1,1)	1.8885555	.64	5
(1,0,0)	.2730546	-1.30	4
(1,0,1)	6.331266	1.85	6
(1,1,0)	13.923	2.63	7
(1,1,1)	322.83	5.78	8

Thus, using the decision

unmatched for patterns 1-4
review for pattern 5
matched for pattern 6-8

would give us the Fellegi-Sunter decision procedure with P(false non-match) = .026 and P(false match) = .02525.

Now lets consider some different blocking schemes. In particular we will look at three blocking schemes, $\bar{B}_i$ i=1,2,3, where $\bar{B}_i$ denotes blocking on the ith component. We will compare these schemes using the two methods outlined in the last section.

To make use of Method 1 we must first calculate $P(S_3^* - S_3 \mid M)$ for each blocking scheme. To facilitate this calculation consider the following Venn diagram:

S_3	S_2	S_1
(0,0,0) (0,0,1) (0,1,0) (1,0,0)	(0,1,1)	(1,1,1) (1,1,0) (1,0,1)

From this we see that for B_1

$$P(S_3^* - S_3 \mid M) = P((0,1,1) \mid M) = .08075.$$

Similarly,

$$P(S_3^* - S_3 \mid M) = P((1,0,1) \mid M) = .12825$$

for B_2 and

$$P(S_3^* - S_3 \mid M) = P((1,1,0) \mid M) = .03825$$

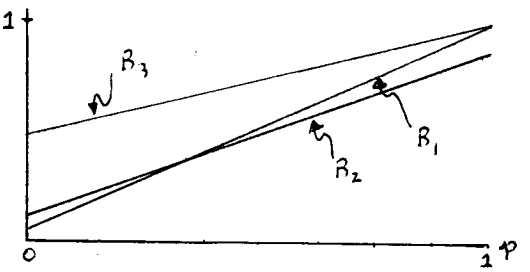
for B_3.

Further $P(\Gamma^*)$ is given by

.90p + .05
.75p + .10
.50p + .45

for blocking schemes 1 through 3, respectively.

To compare these blocking schemes lets first consider the graphs of $P(\Gamma^*)$ for each of the schemes.



The first thing we note from this graph is that B_1 and B_2 are both uniformly better than B_3. So, even if B_3 is admissable according to loss, it isn't admissable according to benefit. However, lets go ahead with a formal application of Method 1 and let $\omega = .1$. This would elimiate B_2 from consideration. Of the two remaining schemes we see that B_1 is uniformly best.

Next we will study schemes B_1, B_2 and B_3 using Method 2. To make the necessary computations we note that if S_i^* is the set on which decision A_i is made then the error rates are given by

$$P(S_3^* \mid M) = P(S_3^* \cap \Gamma^* \mid M) + P(\Gamma^{*c} \mid M)$$

and

$$P(S_1^* \mid U) = P(S_1^* \cap \Gamma^* \mid U).$$

So we select $S_3^* \cap \Gamma^*$ so that

$$P(S_3^* \cap \Gamma^* \mid M) \leq \omega_1 - P(\Gamma^{*c} \mid M)$$

and $S_1^* \cap \Gamma^*$ so that

$$P(S_1^* \cap \Gamma^* \mid U) \leq \omega_2.$$

Suppose we let $\omega_1 = .2$ and $\omega_2 = .005$.

For B_1 we have

$$P(S_3^* \cap \Gamma^* \mid M) \leq .2 - .1 = .1$$

and

$P(S_1^* \cap \Gamma^* \mid U) \leq .005$

So, $S_3^* \cap \Gamma^* = \{(1,0,0)\}$.

and, $S_1^* \cap \Gamma^* = \{(1,1,1), (1,1,0)\}$.

Thus, $S_2^* = \{(1,0,1)\}$, and so, $P(S_2^*) = .108\ p + .02025$.

For B_2 we have $P(S_3^* \cap \Gamma^* \mid M) \leq .2 - .15 = .05$

and $P(S_1^* \cap \Gamma^* \mid U) \leq .005$.

So, $S_3^* \cap \Gamma^* = \{(0,1,0)\}$.

and, $S_1^* \cap \Gamma^* = \{(1,1,1), (1,1,0)\}$.

Thus, $S_2^* = \{(1,1,0)\}$, and so, $P(S_2^*) = .0355\ p + .00275$.

For B_3 we have

$P(S_3^* \cap \Gamma^* \mid M) \leq .2 - .05 = .15$

and $P(S_1^* \cap \Gamma^* \mid U) \leq .05$

So $S_3^* \cap \Gamma^* = \{(0,0,1), (0,1,1)\}$.

and, $S_1^* \cap \Gamma^* = \{(1,1,1)\}$

Thus $S_2^* = \{(1,0,1)\}$, and so, $P(S_2^*) = .108\ p + .02025$.

On reviewing these calculations we see that according to Method 2 B_1 and B_3 are equivalent while B_2 is uniformly better than both.

V. CONCLUSIONS

Of the many possible methods we could construct to compare different blocking schemes we selected two for study.

Method 1 is based on the change which is induced in the original Fellegi-Sunter procedure by blocking. It's strength lies in its specific expression of the loss and benefit due to blocking.

Method 2 is based on the conditional Fellegi-Sunter procedure given the blocking subspace Γ^* . It's strength lies in the ability to specify an overall bound for both error rates through its use. Also, it is appealing because it maintains the same gobal objective as the unconditional Fellegi-Sunter procedure.

In comparing these two methods, as they pertain to the example of section IV, we see that they clearly are not equivalent (they, in fact, give quite different results). Further, the results given by Method 1 are more intuitively appealing than those given by Method 2. This is since Method 1 chooses the blocking scheme based on the component with the most discriminatory power [1] while Method 2 chooses the blocking scheme based on the component with the least discriminatory power. While this does not invalidate Method 2, it certainly causes us to question when and how we might use it.

There is a great deal more that needs to be done on this topic. We are currently working on some simulation studies which will allow us to relax some of the assumptions (for example, the assumption of component independence). It is hoped that these studies will lead to greater insight into the blocking problem and into the use of various models needed for its solution.

FOOTNOTE

[1] Discriminatory power of a component is not a well defined term; however, its connotation is fairly clear. One possible way to give numeric substance to this concept would be to use the Divergence function in Kullback (1959).

REFERENCES

1. Goldstein M. and Dillon, W. R. (1978). Discrete Discriminant Analysis, Wiley.

2. Fellegi, I. and Sunter, A. (1969), "A Theory for Record Linkage." Journal of the American Statistical Association, 64: 1183-1210.

3. Kullback, S. (1959). Information Theory and Statistics, Wiley.