IDENTIFYING DISCRIMINATORY MODELS IN RECORD-LINKAGE

Yves Thibaudeau, U.S. Bureau of the Census
Room 3000, Federal Building 4, Washington DC 20233

Abstract

The direction of the research presented in the paper is toward the development of simple techniques to identify the statistical process underlying a given record-linkage process. A good identification of the process leads to a decision rule with accrued discriminatory power in record linkage (Thibaudeau, 1991). The paper presents a particular situation: the unduplication of the mailing list for the 1992 agricultural census. In this context the paper suggests an approach to formulate models representing record-linkage processes, and diagnostic tools to assess their discriminatory power.

Key Words: Underlying Probabilistic Model; Record-Linkage Rule; Discrimination Power.

1. Introduction

The paper gives loose directives for the construction of a most discriminatory record-linkage rule. An example of a record-linkage process is introduced in section 2. In section 3 the theory of record-linkage is reviewed. This theory points to a method for the construction of most discriminatory record-linkage rules to interpret a record-linkage process, when the underlying model is known. In section 4, the example of section 2 is analyzed, in light of the theory of section 3. Two possible underlying models are suggested. These models are carefully compared in section 5, using ad-hoc diagnostic tools, and one model is selected for having the best discriminatory power. In the conclusion, the techniques used in this paper are discussed in term of future applications.

2. Unduplicating a List of Businesses

In 1992, the Census Bureau conducts an agricultural census. To do so it needs to construct a complete mailing list. That is a list containing the names, addresses, and other characteristics of all the agricultural businesses in the United States. Each business should not be listed more than one time; i.e. there should be no duplicate. To produce this complete and unduplicated list, several partial lists of agricultural businesses are merged together to obtain a primitive mailing list. To identify the duplicates, the records of the list are brought together in pairs. Two records in a same pair are compared over the following comparison variables: last name, first name, middle initial, box number, rural route, street, phone, social security number.

The comparison of two records in a same pair yields some information about the status of that pair. The status of a pair of records is either a "match" or a "non-match". The status of a pair is a match, whenever the two records represent the same agricultural business. Otherwise it is a non-match. In most situation, the status of a pair is not known and it must be inferred. If the inferred status is a match, then one of the records in the pair is designated a duplicate, and it is deleted from the list. If the inferred status is a non-match, then the list is left intact.

There are physical limitations in the search for duplicates. For instance, the ideal situation is to have enough memory to keep the entire list in core memory, but of course, this is typically not possible. Accordingly, the records are grouped by zip groups. A zip

group is a set of one or more contiguous entire zip codes. The computer is fed one zip group at the time, and as a result, only the duplicates in the same zip group as the original are retrieved. A specific example is presented later. In this example, 390412 records are processed in 2061 separate zip groups.

3. A Most Discriminatory Decision Rule

A record-linkage rule is an inference rule associating a status to all the pairs in the universe considered. Fellegi and Sunter (1969) apply the Neyman-Pearson lemma in the record-linkage context to define a "most discriminatory" record-linkage rule. Before defining the concept of "most discriminatory" some notation in needed.

Let γ be the agreement-disagreement pattern generated by the comparison of two records in a same pair. γ is a binary vector of dimension 8 (the number of comparison variables), whose entries are 0 or 1. The entry is 0 if the two records don't agree over the corresponding variable and 1 if they agree. In the paper, only rules that are function of γ are considered. Let $m(\gamma)$ be the probability of observing γ, given that the pair generating γ has a match status, and let $u(\gamma)$ be the probability of observing γ, given that the pair generating γ has a non-match status. The overall probability of observing γ is the following mixture:

$$Pr(\gamma) = p\,m(\gamma) + (1-p)\,u(\gamma) \qquad (1)$$

p is the probability of selecting a match (a pair whose true status is match), when choosing a random pair of records.

In this set-up, Fellegi and Sunter (1969) show that the pairs most likely to be matches correspond to the value of the vector γ maximizing the ratio $m(\gamma)/u(\gamma)$. Furthermore, if the pairs are ordered in a series, by decreasing order of the ratio $m(\gamma)/u(\gamma)$ (the order is arbitrary if some pairs have a same value of $m(\gamma)/u(\gamma)$), the most discriminatory decision rule is to infer a match status for the first N pairs of the series. The value of N depends on the tolerance on the rate of false matches. The false matches are the pairs whose inferred status is a match status but whose real status is a non-match status. The Fellegi-Sunter rule is most discriminatory in the sense that for any other rule with the same tolerance on the proportion of false matches, the number of matches retrieved is not bigger.

The pivotal information in the application of the Fellegi-sunter rule is the ratio $m(\gamma)/u(\gamma)$. In practice this ratio is unknown and must be estimated. In that respect, a model, defining a probabilistic structure for $m(\gamma)$ and $u(\gamma)$, is conjectured. Then $m(\gamma)$ and $u(\gamma)$ are estimated. The next section shows how the choice of the model can have serious repercussions on the Fellegi-Sunter rule.

4. A Most Discriminatory Model for the Agricultural Data

The goal of this section is to identify the model which best reflects the underlying probabilistic trends in the record-linkage process described in section 2. Two simplifying assumptions are often made regarding the underlying probabilistic model. The first is to assume that, if a pair is a match, the agreement or disagreement over each comparison variable is independent of the agreement or disagreement over any other comparison variable. Similarly, under the second assumption, if a pair is a non-match, the agreement or disagreement over a comparison variable is independent of the agreement or disagreement over any other comparison variable. These two assumptions together are the "conditional independence

assumptions". The corresponding statistical model is the "conditional independence model". This model is quite popular in the analysis of categorical cross-classification data in the presence of latent variables (Goodman, 1974; Haberman 1979). This model is worth exploring. However, there is no guaranty that the conditional independence assumptions are applicable in this case and other models should also be experimented.

In order to assess the validity of the conditional independence assumptions, a more thorough inspection is conducted. The correlation matrix between the agreements/disagreements of two comparison variables simultaneously, for a sample of zip groups, is computed and presented in Table 1. Moderate correlations can be observed between the social security number, the box number, the phone, and the last name. A good model should account for these correlations. Therefore, a log-linear model incorporating interaction factors of degree 2 and 3 is proposed. Attempts to estimate higher degree interaction factors were unsuccessful. The only third degree factor that could be estimated was the interaction factor between box number, phone, and social security number. The second order interaction factors involving the six pairs of variables that can be formed with the four variables listed above, are integrated in the model. The interaction factors enter into play only when the two records compared form a match. The model assumes independence between all the comparison variables whenever the two records involved in the comparison form a non-match.

The next section evaluate and compare the two models for a particular instance of the example presented in section 2, in terms of their associated Fellegi-Sunter decision rules.

5. Comparing Two Models

The attention is centered on the performances induced by the two probabilistic models introduced in the preceding section. Before each model can be used in applications of the Fellegi-Sunter record-linkage rule, the parameters of these models must be estimated. The conditional independence model has an advantage here. Indeed, it is generally straightforward to estimate the parameters of this model. A simple expectation-maximization algorithm is used to maximize the likelihood. The maximum likelihood estimate can be substituted in the model and the Fellegi-Sunter record-linkage rule is derived.

For the model with dependencies, the estimation of the parameters is more difficult. In the case of the agriculture data, the parameters were estimated with Newton's method, adapted to latent-class models (Haberman, 1979, pp. 547-552). This method requires a good starting point and then converges quickly to a local maximum of the log-likelihood.

Table 2 gives an account of some differences between the two set of estimated parameters, under each model. It shows the marginal probabilities of agreement of two records in the same pair, over each variable, given the status of the pair, for both models. Overall, the conditional probabilities are more or less the same for the two models, with the exception of the probability of agreement on the social security number, conditional on a non-match status. This probability under the model with interaction is thirty times bigger than the same conditional probability under the conditional independence model. This means that more discrimination power is placed on the social security number, under the conditional independence model.

Further insight is gained by examining carefully the ordering of the agreement/disagreement patterns induced by each model. Because of the shortage of space, the eight variables problem must be reduced to a four variables problem before presenting the orderings. The natural choice for the four variables is the set of the four variables involved in the dependency structure, namely last name, box number, phone, and social security number. The ordering of the agreement-disagreement patterns, as prescribed by the Fellegi-Sunter record-linkage rule, assuming each models in turn, is given in table 3. A shorthand notation is used. For instance "L,B,P,." is the pattern with agreement on last name, box number, and phone and disagreement on social security number.

From table 3, the two orderings are not equivalent. Because of its structure, the conditional independence model excessively promotes the discrimination power of the social security number. This explains why, under the conditional independence model, the pattern with agreement over phone and social security number only (.,.,P,S) is given preference over the pattern with agreement over last name, box number, and phone (L,B,P,.) contrarily to the ordering under the model with interactions. The same situation is repeated when the conditional independence model allocates a higher rank to the pattern showing agreement on the social security number alone (.,.,.,S) than to the pattern with agreement on last name and phone. (L,.,P,.).

The behavior of the ordering induced by the conditional independence mode suggests that the information given by the variables other than the social security number is not exploited. At this point the interaction model is more general and gives a better account of the trends at work in the data. From all evidence, this model yields more discrimination power in the application of the Fellegi-Sunter rule.

6. Conclusion

In the unduplication example it is shown how to extract some information from the process, through the correlation matrix. Combined with some experimentation, this information leads to the delineation of more credible decision rules, such as the one based on a model incorporating dependencies between some of the variables. At this point there is no formal elicitation procedure for general record-linkage situations. However, as more research is done and the behavior of the models involved in record-linkage is better understood, it is realistic to expect the emergence of programmable techniques finding the most discriminatory models systematically.

References

Fellegi, I. P. and Sunter, A. B. (1969), "A Theory for Record Linkage" *Journal of the American Statistical Association,* **40**, 1183-1210.

Goodman, L. A. (1974), "Analyzing Qualitative/Categorical Data" *Abt Books.*

Haberman, S. J. (1979), "Analysis of Qualitative Data, Volume 2" *Academic Press.*

Thibaudeau, Y. (1991), "The Discrimination Power of Dependency Structures in Record Linkage" *Proceedings of the 23rd Symposium on the Interface,* 415-418.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

Table 1: Correlation between the Comparison Variables in the Agriculture Unduplication Project

	First	Mid.	Box	Strt	RR	Phone	SSN
Last	.215	.049	.454	.242	.136	.318	.308
First	1.00	.113	.229	.115	.039	.142	-.005
Midin	.113	1.00	.043	.015	.009	.015	-.011
Box	.229	.043	1.00	.423	.082	.491	.556
Strt	.115	.015	.423	1.00	-.029	.274	.328
RR	.039	.009	.082	-.029	1.00	.074	.075
Phone	.142	.015	.491	.274	.074	1.00	.492
SSN	-.005	-.011	0.556	.328	.075	.492	1.00

Table 2: Conditional Probabilities of Agreement for the Model with Interactions and the Conditional Independence Model.

Variable	**Probability of Agreement Given a Match Status**	**Probability of Agreement Given a Match Status**	**Probability of Agreement Given a Non-Match Status**	**Probability of Agreement Given a Non-Match Status**
	Model with Interactions	**Conditional Independence Model**	**Model with Interactions**	**Conditional Independence Model**
Last Name	0.9473	0.9465	.06862	.08208
First Name	0.2206	0.2854	0.01045	0.01112
Middle Init	0.1870	0.2201	0.03050	0.03160
Box	0.4470	0.6154	0.0008839	0.001115
Rural Route	0.5557	0.5441	0.1360	0.1428
Street	0.1624	0.2130	0.004060	0.0045
Pnone	0.2533	0.3531	0.0002677	0.0002344
Soc. Sec. No.	0.2211	0.3084	0.0000350	0.0000011

Table 3: Pattern Orderings Induced by the Model with Interactions and the Conditional Independence Model. The Orderings are by Decreasing value of the Fellegi-Sunter Ratio

$$\left(\frac{m(\gamma)}{u(\gamma)} \right).$$

Ordered Patterns	Ordered Patterns	Number of Pairs	Number of Pairs
Model with Interactions	Conditional Independence Model	Model with Interactions	Conditional Independence Model
L,B,P,S	L,B,P,S	30887	30887
.,B,P,S	.,B,P,S	1363	1363
L,.,P,S	L,.,P,S	8996	8996
L,B,.,S	L,B,.,S	25208	25208
L,B,P,.	.,.,P,S	27595	416
.,B,.,S	L,B,P,.	832	27595
.,.,P,S	.,B,.,S	416	832
.,B,P,.	L,.,.,S	2476	9103
L,.,.,S	.,B,P,.	9103	2476
L,B,.,.	L,.,P,.	61770	16069
L,.,P,.	.,.,.,S	16069	619
.,B,.,.	L,B,.,.	11508	61770
.,.,.,S	.,.,P,.	619	2348
.,.,P,.	.,B,.,.	2348	11508
L,.,.,.	L,.,.,.	584690	584690
.,.,.,.	.,.,.,.	5883794	5883794