Validating Distance-Based Record Linkage with Probabilistic Record Linkage

Josep Domingo-Ferrer[1] and Vicenç Torra[2]

[1] Universitat Rovira i Virgili, Dept. of Computer Science and Mathematics,
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain,
jdomingo@etse.urv.es
[2] Institut d'Investigació en Intel·ligència Artificial,
Campus de Bellaterra, 08193 Bellaterra, Catalonia, Spain
vtorra@iiia.csic.es

Abstract. This work compares two alternative methods for record linkage: distance based and probabilistic record linkage. It compares the performance of both approaches when data is categorical. To that end, a distance over ordinal and nominal scales is defined. The paper shows that, for categorical data, distance-based and probabilistic-based record linkage lead to similar results in relation to the number of re-identified records. As a consequence, the distance proposed for ordinal and nominal scales is implicitly validated.

1 Introduction

Re-identification procedures are tools developed to detect the presence of the same individual in different data files. Record linkage is a particular strategy for re-identification which links records in separate files that correspond to the same individual. While these procedures can be developed under various assumptions (see [4] and [9] for a detailed description of the alternative approaches: e.g. files share or not share variables; two or more files to be re-identified; types of the variables) we restrict here the work to the case of linking only two data files that share a set of variables.

Note that re-identification in the case of common variables is far from trivial because it is usually the case that a matching procedure among pairs of records is not enough to establish links between them. This is so due to the presence of errors in the files (the usual case).

In the case of re-identification assuming common variables, the two most successful re-identification methods are probabilistic record linkage and distance-based record linkage. [6] describes both approaches and includes a comparison for numerical data files. An alternative promising method based on clustering techniques is describe in [5]. The rationale of the latter proposal is similar to the one of [4].

In this work we focus on probabilistic and distance-based record linkage. The characteristics of these methods are as follows:

M.T. Escrig Monferrer and F. Toledo Lobo (Eds.): CCIA 2002, LNAI 2504, pp. 207–215, 2002.

Probabilistic record linkage: Probabilistic record linkage applied to files A and B is based on the computation of an index for each pair of records (r_A, r_B) where r_A and r_B are records of file A and B, respectively. Then, some index thresholds are used to label the pair as a linked pair (LP), a clerical pair (CP) or a non-linked pair (NP). A clerical pair is one that cannot be automatically classified as linked or non-linked; human inspection is needed to classify it.

To use probabilistic record linkage in an effective way, we need to set the thresholds (*e.g.* the values *linkThreshold* and *nonLinkThreshold*) and the conditional probabilities used in the computation of the indices. In plain words, thresholds are computed from: (i) the probability of linking a pair that is an unmatched pair (a *false positive* or *false linkage*) and (ii) the probability of not linking a pair that is a match pair (a *false negative* or *false unlinkage*). Conditional probabilities used for the indices are usually estimated using the EM algorithm [10].

For a detailed description of this method see [7], [8] and [2].

Distance-based record linkage: In general, for each record in file A, the distance to every record in file B is computed. Then the *nearest* record in file B is considered. A record in file B is labeled as *linked* when the nearest record in file A turns out to be its corresponding original record (the one that generated the distorted record). In all other cases, records are not linked. Details on this method are given in [11].

These two approaches are radically different. The following aspects can be underlined:

- Distance-based record linkage methods are simple to implement and to operate. The main difficulty consists of establishing appropriate distances for the variables under consideration. In particular, distances for categorical variables (in ordinal and nominal scales) are required. On the other hand, distance-based record linkage allows the inclusion of subjective information (about individuals or variables) in the re-identification process.
- Probabilistic record linkage methods are less simple. However, they do not assume rescaling or weighting of variables and require the user to provide only two probabilities as input: the probabilities of false positive and false negative.

For numerical data, it has been proven (see [6]) that both approaches lead to similar re-identification results. For categorical data, no comparison is available in the literature, probably because distances over categorical data are less straightforward than distances over numerical data.

In this work we consider two distances for categorical data (one for ordinal sccales and the other for nominal scales). We compare then probabilistic and distance-based record linkage. The comparison is based on extensive experimentation. Results show that the behavior of both approaches is similar as in the numerical case, and thus it validates the proposed distances.

The structure of the rest of this paper is as follows. In Section 2, we describe the methodology to compare both record linkage approaches. In particular, this section proposes distances for categorical data. Section 3 describe the results obtained and Section 4 contains some conclusions and mentions future work.

Table 1. Variables used in the analysis.

Variable	Meaning	u	l	s	m	o	N. Categ.
BUILT	year structure was built	X	X			X	25
DEGREE	long-term average degree days			X		X	8
GRADE1	highest school grade	X	X			X	21
METRO	metropolitan areas			X			9
SCH	schools adequate			X			6
SHP	shopping facilities adequate			X			6
TRAN1	means of transportation to work	X			X		12
WHYMOVE	primary reason for moving	X	X				18
WHYTOH	main reason for choice of house	X			X		13
WHYTON	id. for choosing this neighborhood	X			X		13

2 Methodological Considerations

To compare the probabilistic and distance-based re-identification methods, and due to the lack of benchmarks for this purpose, we have performed a set of experiments based on the ones used by National Statistical Offices to evaluate masking procedures and to determine the re-identification risk for a particular data file prior to its publication.

Thus, we have applied re-identification procedures between an original data file and some masked data files obtained through application of several masking methods on the original file. This is consistent with the methodology proposed in [6] for the case of continuous variables.

Several re-identification experiments were performed in order to mitigate the dependency of results on a single dataset. Thus, different sets of variables, different masking methods and different method parameterizations were considered. In this section, we detail the experiments obtained so far. We first describe the original data file used (a publicly available data file). Then, we describe how masking methods were applied to obtain different masked data files. We then propose some distances for categorical. Results of the experiments are reported in the next section.

2.1 Test Data Collection

Data from the *American Housing Survey 1993* were used (these data can be obtained from the U.S. Census Bureau using the Data Extraction System at

http://www.census.gov/DES/www/welcome.html). A set of 10 categorical variables (see Table 1). Five groups of variables were defined over the set of selected variables, and the same analysis was performed for each of them. First, three groups were defined by grouping variables with a similar number of categories. Let 's', 'm' and 'l' denote the groups of variables with small, medium and large number of categories, respectively. A fourth group denoted by 'u' was defined that corresponds to the union of the groups 'm' and 'l'; thus, 'u' corresponds to the group of variables with medium or large number of categories. Finally, a fifth group 'o' was defined as the subset of ordered variables (variables that range in an ordinal scale). This latter group was defined after analyzing the meaning of each category in the range of variables. Table 1 gives for each variable, in which groups is present and also the number of categories.

To allow a substantial amount of experiments in reasonable time, we took only the first 1000 records from the corresponding data file.

2.2 Generation of File Pairs for Re-identification

The generation of pairs of files to perform re-identification was achieved by masking the original data file. Each pair was formed by the original data set and one masked version of it. To generate masked versions of the original data, several masking methods were applied to the original 1000-record file containing the 10 variables in Table 1. Four masking methods were considered and for each one nine different parameterizations were applied. Masking methods were selected among those commonly used by National Statistical Offices to protect data files. Different parameterizations were taken so that different levels of data protection were experimented with. The consideration of both aspects led to $4 * 9 = 36$ different masked data files.

The following masking methods were considered (see [13] for a detailed description of masking methods): Top and Bottom coding, Global recording and PRAM (Post-Randomization Method). 9 different parameterizations (with parameter $p = 1, \cdots, 9$ – the larger the p, the larger the distortion) where considered for each method. The application of these four masking methods above with nine different parameterizations per method led to 36 different masked data files.

2.3 Re-identification Experiments

For each record linkage method (probabilistic record linkage and distance-based record linkage) and for each pair of *(original-file, masked-file)*, five re-identification experiments were performed. More specifically, each of the five experiments corresponded to one of the five groups of variables 'u', 'l', 's', 'm' and 'o' defined in Table 1. Since there were 36 different file pairs, 36*5=180 re-identification experiments were performed for each record linkage method.

The implementation of probabilistic record linkage used in the experimentation was the U.S. Census Bureau software provided by W. Winkler [15], [3] with some additions. The EM algorithm was used for the estimation of the probabilities. The implementation of distance-based record linkage was especially written

Table 2. Re-identification results for the 's' group of variables using probabilistic (left) and distance-based (right) record linkage

parameter	Bottom	Global	PRAM	Top	Bottom	Global	PRAM	Top
1	1000	1000	966	1000	502	986	978	853
2	699	1000	921	891	263	861	938	617
3	577	917	897	749	176	651	916	447
4	447	730	881	493	101	326	905	200
5	279	835	843	429	66	103	882	95
6	161	695	803	458	43	83	828	78
7	79	355	789	688	36	9	792	42
8	51	51	759	45	3	3	780	42
9	51	51	734	188	3	3	753	42

in C for the experimental work reported in this paper. An essential point was to define a distance for categorical variables, which was done as follows:

Definition 1. *1. For a nominal variable V, the only permitted operation is comparison for equality. This leads to the following distance definition:*

$$d_V(c, c') = \begin{cases} 0 \text{ if } c = c' \\ 1 \text{ if } c \neq c' \end{cases}$$

where c and c' correspond to categories for variable V.

2. For an ordinal variable V, let $\leq_V$ be the total order operator over the range of V. Then, the distance between categories c and c' is defined as the number of categories between the minimum and the maximum of c and c' divided by the cardinality of the range (denoted by $D(V)$):

$$d_V(c, c') = \frac{|c'' : min(c, c') \leq_V c'' \leq_V max(c, c')|}{|D(V)|}$$

The distance for pairs of records was computed assuming equal weight for all variables.

3 Results

Results of both re-identification procedures turn out to be similar. In Tables 2-6, the number of correctly re-identified records is displayed for the groups of variables (PRL is the probabilistic based record linkage and DBRL corresponds to distance-based one. It can be seen that for some of the experiments (some particular combination of masking method and parameterization), PRL lead to better results than DBRL and that for some of the experiments is DBRL which performs best.

The average number of re-identified records per experiment was computed as a measure of similarity between both record linkage methods. The following expression was used:

Table 3. PRL (left) and DBRL (right) results for the 'm' group

parameter	Bottom	Global	PRAM	Top	Bottom	Global	PRAM	Top
1	195	118	118	114	722	995	1000	966
2	448	104	117	88	372	984	997	878
3	428	93	114	74	340	967	997	836
4	410	83	115	54	332	917	994	805
5	386	93	109	52	323	802	990	797
6	372	148	101	38	313	493	985	758
7	367	448	96	77	302	243	981	503
8	367	576	102	356	272	202	980	275
9	593	559	98	325	156	167	981	255

$$\frac{\sum \texttt{number of correct re} - \texttt{identifications}}{\texttt{number of experiments}}$$

For distance-based record linkage, an average number of 593.06 re-identified records (over 1000) was obtained. For probabilistic record linkage, the average was 579.32 re-identified records. Thus, the performance of both methods is similar.

Note 1. It is important to point out that, even though both methods yields similar average numbers of re-identified records, the re-identified records are not the same. Furthermore, it can be seen e.g. in Tables 3 and 4 that, for a particular masking method, parameterization and set of variables, not even the number of re-identified records is similar for both approaches. Therefore, both record linkage methods should be regarded as complementary rather than as antagonistic.

Note 2. Distance-based record linkage seems to perform better for PRAM masked data while probabilistic record linkage seems to perform better for the other masking methods. As PRAM is the only non-deterministic masking method, seems that for the re-identification of data with stochastic errors, distance based record linkage is more appropriate.

Additional analyses have been carried to assess the performance of both record linkage methods ([9] describe correlation statistics between the number of correctly re-identified records and some information loss measures). These results also show that both approaches to re-identification lead to similar results.

In the experiments reported in this paper, no information is fed to record linkage procedures about the masking method applied to protect the original file. In fact, this is not the usual case in disclosure risk assessment. It can be proven that, if a distance is used which takes into account the masking method applied, the distance-based record linkage largely improves its results. A simple way for a

distance to take masking into account is as follows: assign a distance *infinity* when category c cannot be recoded as c' using the current masking method. In this case, the average number of correctly re-identified records increases to 663.49, which should be compared to 593.06 for the original distance-based record linkage and to 579.32 for probabilistic record linkage.

Table 4. PRL (left) and DBRL (right) results for the 'u' group

parameter	Bottom	Global	PRAM	Top	Bottom	Global	PRAM	Top
1	530	415	414	389	572	994	986	906
2	799	402	411	318	352	899	956	746
3	828	380	438	280	225	767	943	589
4	824	379	421	246	171	556	939	357
5	691	385	413	233	126	355	914	259
6	690	464	402	190	92	226	896	179
7	666	882	378	266	68	68	880	105
8	570	856	391	509	22	23	877	58
9	598	782	387	425	11	9	849	50

Table 5. PRL (left) and DBRL (right) results for the 'o' group

parameter	Bottom	Global	PRAM	Top	Bottom	Global	PRAM	Top
1	1000	999	988	1000	876	965	987	925
2	932	993	952	914	746	912	970	793
3	749	962	936	783	570	815	925	652
4	553	833	909	574	442	711	917	473
5	428	685	906	404	309	563	877	338
6	357	518	865	350	265	423	861	261
7	318	295	853	386	213	240	862	244
8	329	300	829	440	186	215	848	207
9	305	338	801	398	169	197	833	177

4 Conclusions and Future Work

Two record linkage methods for re-identification of categorical microdata have been studied in this paper: probabilistic and distance-based. Since distance-based record linkage only existed in the literature for numerical data, a distance for categorical data has been defined to extend this kind of linkage to categorical data. We have shown that the number of re-identifications is similar for both

record linkage procedures, but that the re-identified individuals/records are not the same. This is consistent with existing comparisons of both record linkage methods for numerical data. Beyond implicit validation of the proposed distance for categorical data, results in this paper show that both methods are complementary rather than antagonistic and best results are obtained if they are combined.

Table 6. PRL (left) and DBRL (right) results for the 'l' group

parameter	Bottom	Global	PRAM	Top	Bottom	Global	PRAM	Top
1	1000	999	997	997	952	996	992	970
2	993	997	993	968	937	992	985	909
3	992	993	980	950	921	969	973	874
4	983	986	975	917	907	946	973	825
5	954	979	976	862	740	912	964	772
6	936	934	967	798	723	863	975	716
7	889	894	959	730	686	824	958	661
8	826	844	961	676	608	753	960	554
9	780	785	948	571	568	687	953	476

It has also been pointed out that distance-based record linkage can substantially improve (and thus outperform probabilistic record linkage) if information about the masking method is embedded into the distance function.

Future refinements of distance-based record linkage is to give a different weight to each variable when computing the distance. This would be problem-dependent and would require a learning mechanism to adjust weights beforehand.

Acknowledgments. Partial support of the European Community under the contract "CASC" IST-2000-25069 and of the CICYT under the project "STREAMOBILE" (TIC2001-0633-C03-01/02) is acknowledged.

References

1. L. Sweeney, "Information explosion", in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. I. Lane, J. M. Theeuwes and L. M. Zayatz, Elsevier, 43–74, 2001.
   H. B. Newcombe, J. M. Kennedy, S. J. Axford and A. P. James, "Automatic linkage of vital records", *Science*, vol. 130, 954–959, 1959.
2. W. E. Winkler, "Matching and record linkage", in *Business Survey Methods*,ed. B. G. Cox, Wiley, 355–384, 1995.
   J. F. Robinson-Cox,"A record-linkage approach to imputation of missing data: analyzing tag retention in a tag-recapture experiment", *Journal of Agricultural, Biological and Environmental Statistics*, vol. 3, 48–61, 1998.

3. W. E. Winkler, "Advanced methods for record linkage", *Proc. of the American Statistical Assoc. Section on Survey Research Methods*, 467–472, 1995.
4. V. Torra, "Towards the re-identification of individuals in data files with non-common variables", in *Proceedings of ECAI'2000*, 326–330.
V. Torra, "Re-identifying individuals using OWA operators", *Proceedings of the 6th Intl. Conference on Soft Computing*, Iizuka, Fukuoka, Japan, 2000.
5. J. Bacher, S. Bender and R. Brand, "Empirical re-identification – Evaluation of a simple clustering technique", *Intl. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, forthcomming.
6. J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. I. Lane, J. M. Theeuwes, L. M. Zayatz, Elsevier, 111–133, 2001.
7. I. P. Fellegi and A. B. Sunter, "A theory of record linkage", *Journal of the American Statistical Association*, vol. 64, 1183–1210, 1969.
8. M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*, vol. 84, 414–420, 1989.
9. J. Domingo-Ferrer, V. Torra, Distance-based and probabilistic record linkage for re-identification of records with categorical variables, *Butlletí de l'ACIA*, 27, 2002.
10. A. P. Dempster, N. N. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, vol. 39, 1–38, 1977.
11. D. Pagliuca and G. Seri, *Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey*, Esprit SDC Project, Deliverable MI-3/D2, 1999.
12. F. Felsö, J. Theeuwes and G. G. Wagner, "Disclosure limitation methods in use: results of a survey", in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. I. Lane, J. M. Theeuwes and L. M. Zayatz, Elsevier, 17–42, 2001.
13. L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*, Springer-Verlag, 2001.
14. P. Kooiman, L. Willenborg and J. Gouweleeuw, *PRAM: A Method for Disclosure Limitation of Microdata*, Research Report, Voorburg NL: Statistics Netherlands, 1998.
15. U. S. Bureau of the Census, *Record Linkage Software: User Documentation*. Available from U. S. Bureau of the Census, 2000.