Record Linkage using Probabilistic Methods and Data Mining Techniques

Ogerta Elezaj

Faculty of Economy, University of Tirana

Gloria Tuxhari

Faculty of Economy, University of Tirana

Doi:10.5901/mjss.2017.v8n3p203

Abstract

Nowadays corporations and organizations acquire large amounts of information daily which is stored in many large databases (DB). These databases mostly are heterogeneous and the data are represented differently. Data in these DB may simply be inaccurate and there is a need to clean these DB. The record linkage process is considered to be part of the data cleaning phase when working with big scale surveys considered as a data mining step. Record linkage is an important process in data integration, which consists in finding duplication records and finding matched records too. This process can be divided in two main steps Exact Record Linkage, which founds all the exact matches between two records and Probabilistic Record Linkage, which matches records that are not exactly equal but have a high probability of being equal. In recent years, the record linkage becomes an important process in data mining task. As the databases are becoming more and more complex, finding matching records is a crucial task. Comparing each possible pair of records in large DB is impossible via manual/automatic procedures. Therefore, special algorithms (blocking methods) have to be used to reduce computational complexity of comparison space among records. The paper will discuss the deterministic and probabilistic methods used for record linkage. Also, different supervised and unsupervised techniques will be discussed. Results of a real world datasets linkage (Albanian Population and Housing Census 2011 and farmers list registered by Food Safety and Veterinary Institute) will be presented.

Keywords: *Record linkage, data cleaning, data matching, blocking algorithms, data mining, data integration, clustering*

1. Introduction

Record linkage is a process which is mostly used in creating a frame where the duplications are removed for the file. This process is both a science and an art. This process is a difficult process and a time consuming one. In the past years the work of record linkage has been done manually with a lot of human intervention or using some predefined user rules. In the last decades different algorithms are discovered and used to accomplish this process based on mathematical models and mostly in the probability theory. Record linkage compares the objects using some common variables that are considered to be common identifiers. The goal of this process is to classify two objects as the same objects or different objects. If the datasets that we are going to link have a common identifier (variable key) the linkage is done based on this identifier. On the other hand, if the common identifier is missing, a group of variables will be used during the matching process. The variables used should exist in both datasets. This process faces two main challenges, determination of similarity functions and computational complexity. The similarity function should first be applied to the fields of the pair records for defining the field similarity and next the results of all the fields should be combined for calculating the overall similarity of both records. Regarding the time complexity, the blocking technique that will be described deals with this issue. Instead of calculating the Cartesian product of all possible combination among records of datasets, this technique reduce the comparison space to compare pairs that agree on some basic criteria specified by the nature of datasets that are going to be linked. This technique improves the performance of matching but on the other hand it has an impact in the matching results. For that reason different blocking scenarios should be evaluated. The process of matching results in calculating the high-dimensional similarity/distance Euclidian spaces.

Often the datasets that will be linked have sensitive information. During the data pre-processing phase the identification variables are removed and the datasets are anonymised. Usually, as many public and private institutions share data among them, the data should be encrypted using some encryption algorithms or the identifiers can be replaced with other random identifiers. Therefore, privacy preserving record linkage requires special algorithms.

2. Linkage Methods

In the record linkage field, the most used methods are Deterministic and probabilistic record linkage methods. Different researchers apply even a combination of both methods in the same process. The deterministic method uses deterministic rules which instruct the system to match records based on those rules. This method has a lower level of accuracy compared to the probabilistic ones. Such methods fit better in not large and high dimensional databases. In cases when the organisation decides to use this approach it has to be specified a list of matching rules in any programming language. But when the databases are dynamic and new information is stored, the deterministic rules should be updated and making this method to suffer from low scalability.

The probabilistic matching is based on the probability theory during the comparison phase. As this method lies in statistical theory and data analyses, it can be used in heterogeneous datasets when there is presence of noise and errors. This methods end up with defining a probability value for indicating the probability of a match. The most used probabilistic algorithms are based on the linkage formulas defined by Fillegi and Sunter [5].

Both methods are compared based on the sensitivity which measures the proportional of valid matches by linkage protocols. Usually the sensitivity of deterministic method is lower than probabilistic ones. The software used in the case study, (The Link King software) integrates both methods. The datasets that are processed using this software are linked by deterministic and probabilistic algorithms based on user defined links.

Recently, researchers have used machine learning algorithms to match datasets. They used both supervised and unsupervised learning. The supervised learning consists in classifying the records. This techniques treat each pair <α,β> independently as in case of probabilistic methods. The CART algorithm and a vector quantization approach are used by Cochinwala in 2001[2] resulting in a small percentage of errors. Also, Bayesian classifier is proposed to be used for record matching. The decision rules defined by this classifier reduce the error rates and increased the classification of an object to the right class. Because the misclassification of different samples may have different consequences, their decision model minimises the cost of making a decision rather than the probability of error in a decision [8]. Several authors used decision trees for matching between records and neural networks which improved the process of matching compared to the probabilistic methods [3], [4].

On the other hand, clustering techniques, from the unsupervised learning can also be used. The used of bootstrapping approaches proposed by Verykios based on clustering can be used for matching records. The comparison space can be clustered resulting in clusters with result vectors with similar characteristics.

Also, other supervised methods such as decision trees and support vector machines can be applied to match records. The multilayer perceptron neural network can be used for finding hidden patterns among records and to find out matched records. These methods make a huge improvement in accuracy. [6] Based on literature review, no recent applications in official statistics of supervised machine learning techniques to matching in record linkage have been found.

3. Data Linkage Process

The process of linkage follows different steps as shown in figure 1.

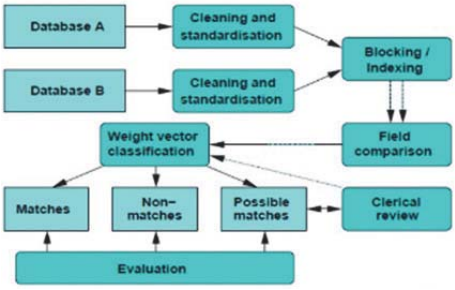


Figure 1: Record Linkage Process

The decision of which variables to use as linkage variables, is the first and the most important step in the process. The identification variables differ in different databases and depend on the information stored in them. These variables can be

classified in three main categories:

- Primary variables- Identification Numbers
- Personal identifiers- name, date of birth, sex, address etc.
- Content variables- number of children, education level, disability status etc.

If both datasets that will be matched contains common primary variables, the best approach is to use them during the matching process. In case that these variables do not exist or cannot be used for different reasons, personal identifiers and content variables can be used as an alternative. The variables used for linkage should be time invariant and should have the same definition in all the databases that we will match. The linkage variables can be of different types as follows:

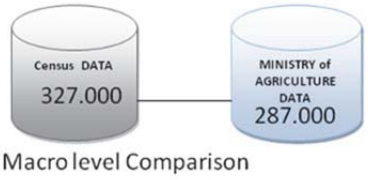
- String (such as name, address, etc.);
- Numeric (age, date of birth etc.);
- Categorical (gender, ethnic group, education level, marital status, etc.).

In some situations, the patterns of errors or inconsistencies in the records are specific to the type of variables. The string variables are prone to spelling errors, the numeric ones deals with precision and rounding issues.

Since the process of record linkage needs to compare each record from database A and database B, scalability is an issue. Let us assume that each dataset that we want to link has 10,000 records. The linkage process requires 100 million comparisons between records. This process is not practical, time consuming and is related with hardware issue for the system where we will run this process. To avoid this problem we use the concept of blocking which consists in using a subset of comparisons. In the blocking phase are produced candidate matches from a fast approximate comparison between datasets. The reason why this process is called blocking is because of division of the full Cartesian product of the datasets into mutually exclusive blocks [7]. For example if you want to do a block by one attribute we sort or cluster the datasets by that attribute and then we apply the comparison method to only a single member of that block. After the process of blocking the candidate matches are evaluated for finding out the true matches. There is a trade off in using the blocking techniques, while it facilitates in terms of computational complexity, it can increase the rate of false not matches because there can be pairs of records that do not match on the blocking criteria specified by users at the beginning of the process.

4. Experimental Evaluations

The process of record linkage is widely used for decades for matching statistical surveys with administrative data. In the national level, government institutions have a strategy to use administrative registers and to combine them with surveys data because it is increased the administrative data capacity but even their quality. Conducting surveys or censuses is a complex process that requires financial costs. As administrative data are collected and can be available for statistical purposes, institutions should profit from their use. The record linkage software King Link is used to link the records of the farmers list generated by the Population and Housing Census with the farmers list registered by ISUV (Food Safety and Veterinary Institute, Albania). So we have to link a statistical database with an administrative register.



Macro level Comparison

		a	b	c=(a-b)	c/a %
District_name	District Code	PHC Albani	F2011 MICRODATA X not adjusted	Difference	% diff
01 Berat	1	15,115	16,218	(1,103)	-7%
02 Bulqizë	2	4,250	4,060	190	4%
03 Delvinë	3	2,126	2,727	(601)	-28%
04 Devoll	4	5,643	5,853	(210)	-4%
05 Dibër	5	8,731	7,798	933	11%
06 Durrës	6	16,158	11,587	4,571	28%
07 Elbasan	7	24,274	20,930	3,344	14%
08 Fier	8	26,067	22,760	3,307	13%

Figure 2: Macro Level Comparison

The census database has 327.000 records and the administrative database has 287.000 records. Before doing a record by record matching, an aggregated comparison among records is done in the district level.

To detect duplications of Census records the Link King software has been used. In this software two features are included:

1. Deterministic Record Linkage- exact matches between two entities

A match is considered deterministic when some identifiers agree between two records where all identifiers are required to agree. But it has some limitations. The limitation of deterministic linkage is that each identifier is given the same weight in terms of quality. In practice, identifiers fields can contains missing values or even incorrect values which can cause a link to fail or to be a fake link.

The fields used for detecting duplications are the selected the entire fields that can identify a person in the census database such as first name, father name, last name, date of birth, gender and the districts where the person lives.

2. Probabilistic Record Linkage, which - high probability of being the same entity

The FREQ Procedure

level of blocking criteria where match was initially found

blocking_crit	Frequency	Percent
NYSIIS last name and dob	131713	78.15
NYSIIS firt name, dob	14863	8.82
NYSIIS fn & ln, birth year	15063	8.94
NYSIIS nickname & ln, YOB	2	0.00
fn & ln 3 char, DOB similar	1295	0.77
fn & ln 2 char, minit, DOB sim	645	0.38
NYSIIS fn & ln, birth month	4481	2.66
name only (missing ssn and dob)	29	0.02
enhanced MN processing	453	0.27

Figure 3: Blocking results

certainty	Frequency	Percent
Level 1: Highest Possible	150398	89.23
Level 2: Very High	173	0.10
Level 3: High	17973	10.66

Figure 4: Blocking results

89% of the records were matched with highest probability, 0.1% with very high and only 10.66 % of the records were matched with high probability.

After the matching process, we discovered 20.000 units that were in the administrative dataset but not in the census dataset. So, at the end we extended the list including these new records and making it more consistent and to cover the entire units in the country.

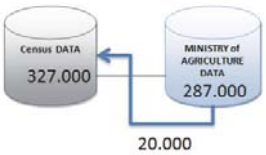


Figure 5: Census Databases

As the dataset to be matched are going to be larger with higher dimensions, there is a need to use data mining techniques to predict based on learning models.

5. Conclusions

In order to improve the performance of record-linkage programs and algorithms, large training and reference data sets should be produced. This should be real-life datasets, containing linkage variables. The record linkage process can also, be seen as a supervised machine learning problem, more precisely, classification of pairs of records as matching or non-

matching However, many traditional matching techniques are not from machine learning [1], and the quality of record linkage is believed to be more sensitive to the quality of pre-processing and standardization than that of matching [9]. Consequently, innovations in record linkage are believed to come more probably from improvements in pre-processing and standardization, as well as scaling up to ever increasing file sizes, rather than from innovations in matching. Due to computational complexity, there is a need for special hardware (massive parallel processors), a team trained in record linkage and the data protection facilities necessary to act as a data trustee for large scale projects Multi-disciplinary research group of computer scientists, lawyers, linguists, historians and social scientists is needed to solve the problems of privacy-preserving record linkage using standard identifiers like names and surnames. As no recent applications in official statistics of supervised machine learning techniques to matching in record linkage have been found, further tests should be done using neural networks and decision trees.

References

[1] Christen P., Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, New York: Springer, 2012.
[2] Cochinwala, M., V. Kurien, G. Lalk and D. Shasha, 2001. Efficient data reconciliation. Inform. Sci., 137: 1-15.
[3] Cohen W., "The WHIRL approach to data integration," IEEE Intelligent Systems, vol. 13, no. 3, pp. 20-24, 1998.
[4] Elfeky M., Verykios V. and Elmagarmid A., "TAILOR: A record linkage toolbox," IEEE ICDE, pp. 17-28, 2002.
[5] Fillegi, I. P. and Sunter, A. B. "A Theory for Record Linkage," Journal of the American Statistical Association, vol. 64, 1183-1210, 1969.
[6] Michie D., Spiegelhalter D., and Taylor C.. Machine Learning, Neural and Statistical Classification, Ellis
[7] Newcombe, H. B. 1967. Record linkage: The design of efficient systems for linking records into individual and family histories. American Journal of Human Genetics 19(3)
[8] Verykios V., Moustakides G.V., and Elfeky M.G.. A Bayesian decision model for cost optimal record matching. The VLDB Journal, 2002.
[9] Winkler W., "Matching and record linkage," WIREs Comput Stat, vol. 6, pp. 313-325, 2014.