Data Clustering for Improving the Selection of Donors for Data Imputation

Clustering di Dati per Migliorare la Selezione dei Donatori per l'Imputazione dei Dati

Gianpiero Bianchi
Istat, Direzione Centrale Censimenti della Popolazione, Territorio e Ambiente (DCCE)
Via A. Ravà 150, 00142 Roma, Italy, gianbia@istat.it

Renato Bruni
Dipartimento di Informatica e Sistemistica dell'Università degli Studi di Roma
"La Sapienza", Via M. Buonarroti 12, 00185 Roma, Italy, bruni@dis.uniroma1.it

Rino Nucara
Dipartimento di Informatica e Sistemistica dell'Università degli Studi di Roma
"La Sapienza", Via M. Buonarroti 12, 00185 Roma, Italy, rino@nucara.it

Alessandra Reale
Istat, Direzione Centrale Censimenti della Popolazione, Territorio e Ambiente (DCCE)
Via A. Ravà 150, 00142 Roma, Italy, reale@istat.it

Riassunto: Il presente lavoro si inserisce nell'ambito dell'imputazione automatica dei dati effettuata per mezzo di dati esatti, detti donatori. Per ogni record errato, occorre selezionare un numero di donatori aventi particolari caratteristiche. Quando tale selezione deve essere effettata all'interno di serbatoi di potenziali donatori molto ampi, come nel caso di un censimento della popolazione, i tempi di calcolo possono rivelarsi troppo elevati. Al fine di ridurre il numero di potenziali donatori da esaminare, è qui proposto l'innovativo utilizzo di una procedura di clustering. L'insieme dei potenziali donatori viene diviso in numerosi sottoinsiemi, in modo che elementi dello stesso sottoinsieme abbiano caratteristiche simili. È stato in particolare sviluppato un algoritmo per il clustering di dati demografici. I risultati sono molto soddisfacenti, dal punto di vista sia della qualità dei dati, sia computazionale.

Keywords: Clustering, Data Imputation, Nearest Neighbourhood, Household Data

1. Introduction

This paper is concerned with the problem of automatic *detection* and *correction* of inconsistent or out of range data in a general process of statistical data collecting. Our attention will be particularly focused on the problem of automatic imputation of the hierarchical demographic data of a Population Census. Census data are collected at the household level with information gathered for each person within the household. Such data records may contain errors and missing values, and there exist several methodologies to impute them (see e.g. Winkler, 1999). The problem of error detection is generally approached by formulating a set of rules that each household must respect in order to be declared *correct*. Households that do not respect such rules are declared *erroneous*. Hence, the editing process classifies records as correct or erroneous. Afterwards, in the correction process, the incorrect values of erroneous records should be replaced by new correct ones with the purpose of restoring their unknown

original values. The *adjusted* records are therefore obtained.
By combining and revising two main imputation approaches, the probabilistic one (Fellegi and Holt, 1976) and the data driven one (e.g. Bankier *et al.*, 2000), a new imputation methodology, implemented in the software system DIESIS (Bruni *et al.*, 2002), has been recently developed. DIESIS has been successfully used for the correction of demographical data of the 2001 Italian Population Census.

2. Clustering of the Set of Donors

The correction methodology already adopted in DIESIS is based on the use of correct records as *donors*. A household record r, denoted in particular by e, d, c in the cases, respectively, of an erroneous, a donor, and an adjusted one, consists in a set of values, one for each demographic variable: $r = \{v_1, \ldots, v_p\}$. For each erroneous record e, a number k of donors records $\{d_1(e), \ldots, d_k(e)\}$ having the minimum distance from e are selected, by searching them within the set of all possible donors D. The distance function $f(e, d) \in [0,1]$ is based on the joint distributions of the demographic variables, that can be both qualitative and quantitative, and consists in a weighted sum of the distances for each of such variables. Such latter distances are given by dissimilarity tables, determined using the whole set of current erroneous and correct data, by computing the distance between each couple of values. Subsequently, DIESIS selects the imputation action by minimising the weighted sum of the changes and respecting the original frequency distributions. In particular, for each erroneous record e , the aim is to choose, among the selected donors $\{d_1(e), \ldots, d_k(e)\}$, the one $d_o(e)$ that allows the adjusted record c to preserve the largest weighted set of values from e (minimum weighted change, Bruni *et al.*, 2001).
However, in the described approach, when D is very large, as in the case of a Census, the iterative comparison between every single erroneous record e and all $d \in D$ could require unacceptable computational time. A solution often adopted (Bankier *et al.*, 2000) consists in arresting such search before examining the entire set D, according to some stopping criterion. This obviously may lower the imputation quality, since in this case the selection of the set of donors $\{d_1(e), \ldots, d_k(e)\}$ having minimum distance is not guaranteed at all.
Therefore, we propose here a new approach for reducing the number of donors that must be examined. This is obtained by preventively dividing the large set of donors D into a collection of smaller subsets $\{D_1, \ldots, D_n\}$ in such a way that $D_1 \cup \ldots \cup D_n = D$, and that all elements of the same subset D_j have similar characteristics. Such subdivision is here obtained by solving a *clustering* problem (see e.g. Hastie *et al.*, 2001, Jain *et al.*, 1999 for a review on clustering). Since no a priori information about such subdivision is known, we are in the case of unsupervised clustering. The search for the donors is now conducted, for each erroneous record e, by examining only the cluster(s) containing the donors which are more similar to e.

3. The Proposed Clustering Algorithm

The clusterization of the set D is obtained by progressively selecting some donors, and by considering around each of them a sphere of radius r using the above defined distance function f. The proposed algorithm has been called *algorithm of spherical neighbourhoods*. With more detail, the algorithm is composed by an *initial phase*, which is composed by only one step, and by a *subsequent phase*, which may be composed by a number of steps, as follows:

1) Initial phase: iteratively select a donor $d_s \in D$ and form a cluster D_s for d_s by taking all other donors

$d \in D$ having distance $f(d_s, d) \leq r$ (the spherical neighbourhood) until the cardinality of D_s reaches a maximum value m or the set D has been completely examined. Record d_s will be the *centroid* of the cluster D_s. Each donor which is not a centroid may in this phase belong to more than one cluster, since the spherical neighbourhoods may overlap.
2) Subsequent phase, step i-th: given a radius $r_i < r$ and a maximum cardinality $m_i < m$, iteratively subdivide each cluster D_h having cardinality $> m_i$. The subdivision is obtained by iteratively selecting a donor $d_s \in D_h$ and forming a cluster $D_s \subset D_h$ by taking all other donors $d \in D_h$ which have a distance $f(d_s, d) \leq r_i$ until D_h has been completely examined. Record d_s will be the *centroid* of the cluster D_s. Donors lying in more than one sphere within D_h are in this phase assigned to only one cluster, by selecting the minimum distance centroid within D_h . Note that such donors may still belong to other clusters not originated by the subdivision of D_h.

Therefore, a clusterization $\{D_1, \ldots, D_n\}$ of the set of donors is obtained. Each donor may belong to more than one cluster. During the various steps of the subsequent phase, it is convenient to progressively reduce the maximum cardinality allowed m_i , otherwise the following steps would produce no effect, and to increase the radius r_i , remaining however $< r$, since during the various steps progressively less dense clusters are being subdivided. The number of steps required for the subsequent phase should be set on the basis of the desired cardinalities of the final clusterization. The above algorithm is computationally inexpensive, and may be used for very large data sets. The availability of centroids representing each cluster is useful for the attribution of erroneous records to clusters.

4. Experimental Results

The described procedure has been implemented in C++. Tests have been conducted on large data sets of household records having the same number of individuals. Individuals within the household have been ordered in decreasing age. Two types of test have been conducted:

1) Comparison between the imputations obtained by: (*i*) exhaustive search within all D of the set of the minimum distance donors $\{d_1(e), \ldots, d_k(e)\}$, and (*ii*) the above search guided by the described clustering approach.
2) Comparison between the selections of the set of the minimum distance donors $\{d_1(e), \ldots, d_k(e)\}$ obtained by: (*i*) searching by allowing a number of computations of $f(e, d)$ corresponding to 2% of the cardinality of D, and (*ii*) the above search guided by the described clustering approach with the same limitation on the number of computations of $f(e, d)$.

In the case of the search guided by clustering, for each erroneous record e, the search is performed by examining only the cluster D_e containing the donors which are more similar to e, or, if the cardinality of D_e is not adequate, only the clusters D_e , D_e‘ , D_e“ , … in increasing distance order from e until the cardinality of their union is adequate. The first test is intended to study whether the use of clustering would decrease the imputation quality with respect to the “ideal” search. Such evaluation has been conducted by considering, for the whole data set, the following statistical indicators (Manzari, Reale, 2001): percentage of not modified values erroneously imputed; percentage of modified values not imputed; percentage of imputed values for which imputation is a failure; average absolute deviation between imputed and original values; dissimilarity index between the relative distributions of imputed values and the relative distributions of the original values. The above indicators assume no sensible difference for the two donor selection methods. This demonstrates that the reduction of the search

guided by clustering does not lower data quality, although drastically reduces the number of computations of $f(e, d)$, and hence computational times. The above holds both for common and uncommon households. The second test is intended to study whether the use of clustering would increase the donor selection quality with respect to the "practical" search. Such evaluation has been conducted by considering the percentages of the (theoretical) set of the minimum distance donors $\{d_1(e), \ldots, d_k(e)\}$ which has been correctly selected by the two donor selection methods. Results show relevant differences. In particular, for common households, a percentage of ~100% of the above set of minimum distance donors can be obtained by using clustering, percentage which decreases to ~70% of such set when no clustering is used. On the other hand, for uncommon household, a percentage of ~95% of the above set of minimum distance donors can be obtained by using clustering, percentage which decreases below 5% of such set when no clustering is used. Note also that the different types of uncommon households represent, for households with 4 individuals, about 40% of the data set, and such percentage increases when increasing the number of individuals in the household.

5. Conclusions

In the case of a very large set of donors, the search for the donor records having minimum distance from each erroneous record may require unacceptable computational times. The preventive subdivision of the set of donors into many smaller subsets, in such a way that elements of the same subset have similar characteristics, here proposed as a novel point, allows to limit such search only to the subset(s) having minimum distance form the current erroneous record. A noteworthy reduction of number of donors that must be examined is made possible. Such subdivision is here obtained by solving a clustering problem by means of the spherical neighbourhood algorithm. The proposed algorithm has, in the considered case, several advantages on other clustering approaches. Tests prove that the search for the donors guided by the described clustering approach is able to sensibly reduce computational times without lowering imputation quality. This especially holds in the case of uncommon household records.

References

Bankier M., Lachance M., Poirier P. (2000) 2001 Canadian Census Minimum Change Donor Imputation Methodology, *Proceedings of the Workshop on Data Editing*, Cardiff, UK.

Bruni R., Reale A., Torelli R. (2001) Optimization Techniques for Edit Validation and Data Imputation, *Proceedings of Statistics Canada Symposium 2001*, Ottawa, Canada.

Bruni R., Reale A., Torelli R. (2002) DIESIS: a New Software System for Editing and Imputation, *Proceedings of 41st Riunione Scientifica SIS 2002*, Milan, Italy.

Fellegi I.P., Holt D. (1976) A Systematic Approach to Edit end Imputation, *Journal of the American Statistical Association*, 71, 17-35.

Hastie T., Tibshirani R., Friedman J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York, US.

Jain A.K., Murty M.N., Flynn P.J. (1999) Data Clustering: A Review, *ACM Computing Surveys*, 31:3.

Manzari A., Reale A. (2001) Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology, *Proceedings of the 53rd Session of the International Statistical Institute*, Seoul, Korea.

Winkler W.E. (1999) State of Statistical Data Editing and current Research Problems, *Proceedings UN/ECE Work Session on Stat. Data Edit*., Working Paper 29, Rome, Italy.