Data Quality in Data Warehouses

William E. Winkler
U.S. Bureau of the Census, USA

INTRODUCTION

Fayyad and Uthursamy (2002) have stated that the majority of the work (representing months or years) in creating a data warehouse is in cleaning up duplicates and resolving other anomalies. This article provides an overview of two methods for improving quality. The first is data cleaning for finding duplicates within files or across files. The second is edit/imputation for maintaining business rules and for filling in missing data. The fastest data-cleaning methods are suitable for files with hundreds of millions of records (Winkler, 1999b, 2003b). The fastest edit/imputation methods are suitable for files with millions of records (Winkler, 1999a, 2004b).

BACKGROUND

When data from several sources are successfully combined in a data warehouse, many new analyses can be done that might not be done on individual files. If duplicates are present within a file or across a set of files, then the duplicates might be identified. Data cleaning or record linkage uses name, address, and other information, such as income ranges, type of industry, and medical treatment category, to determine whether two or more records should be associated with the same entity. Related types of files might be combined. In the health area, a file of medical treatments and related information might be combined with a national death index. Sets of files from medical centers and health organizations might be combined over a period of years to evaluate the health of individuals and discover new effects of different types of treatments. Linking files is an alternative to exceptionally expensive follow-up studies.

The uses of the data are affected by *lack of quality* due to the duplication of records and missing or erroneous values of variables. Duplication can waste money and yield error. If a hospital has a patient incorrectly represented in two different accounts, then the hospital might repeatedly bill the patient. Duplicate records may inflate the numbers and amounts in overdue-billing categories. If the quantitative amounts associated with some accounts are missing, then the totals may be biased low. If values associated with variables such as billing amounts are erroneous because they do not satisfy edit or business rules, then totals may be biased low or high. Imputation rules can supply replacement values for erroneous or missing values that are consistent with the edit rules and preserve joint probability distributions. Files without error can be effectively data mined.

MAIN THRUST

This section provides an overview of data cleaning and of statistical data editing and imputation. The cleanup and homogenization of the files are preprocessing steps prior to data mining.

Data Cleaning

Data cleaning is also referred to as record linkage or object identification. Record linkage was introduced by Newcombe, Kennedy, Axford, and James (1959) and given a formal mathematical framework by Fellegi and Sunter (1969). Notation is needed. Two files, A and B, are matched. The idea is to classify pairs in a product space, **A x B**, from two files A and B into M, the set of true matches, and U, the set of true nonmatches. Fellegi and Sunter considered ratios of conditional probabilities of the form

$$R = P(\gamma \in \mathbf{\Gamma} \mid M) / P(\gamma \in \mathbf{\Gamma} \mid U) \qquad (1)$$

where γ is an arbitrary agreement pattern in a comparison space $\mathbf{\Gamma}$. For instance, $\mathbf{\Gamma}$ might consist of eight patterns representing simple agreement or disagreement on the largest name component, street name, and street number. Alternatively, each $\gamma \in \mathbf{\Gamma}$ might additionally account for the relative frequency with which specific values of name components such as “Smith,” “Zabrinsky,” “AAA,” and “Capitol” occur. Ratio *R,* or any monotonely increasing function of it, such as the natural log, is referred to as a *matching weight (or score).*

The decision rule is given by the following statements:

- If $R > T_\mu$, then designate the pair as a match.
- If $T_\lambda \leq R \leq T_\mu$, then designate the pair as a possible match and hold it for clerical review. (2)

- If $R < T_\lambda$, then designate the pair as a nonmatch.

The cutoff thresholds T_μ and T_λ are determined by a priori error bounds on false matches and false nonmatches. Rule 2 agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then $\gamma \in \Gamma$ would intuitively be more likely to occur among matches than nonmatches, and Ratio 1 would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then Ratio 1 would be small. Rule 2 partitions the set $\gamma \in \Gamma$ into three disjoint subregions. The region $T_\lambda \leq R \leq T_\mu$ is referred to as the no-decision region or clerical-review region. In some situations, resources are available to review pairs clerically.

Linkages can be error prone in the absence of strong or unique identifiers such as a verified social security number that identifies an individual record or entity. Weak identifiers such as name, address, and other nonuniquely identifying information are used. The combination of weak identifiers can determine whether a pair of records represents the same entity. If errors or differences exist in the representations of names and addresses, then many duplicates can erroneously be added to a warehouse. For instance, a the name of a business may be "John K Smith and Company" in one file and "J. K. Smith, Inc." in another file. Without the additional corroborating of information such as addresses, it is difficult to determine whether the two names correspond to the same entity. With three addresses such as "123 E. Main Street," "123 East Main St.," and "P.O. Box 456" and the two names, the linkage can still be quite difficult. With suitable preprocessing methods, it may be possible to represent the names in forms in which the different components can be compared. To use addresses of the forms "123 E. Main Street" and "P.O. Box 456," it may be necessary to use an auxiliary file or expensive follow up that indicates that the addresses have at some time been associated with the same entity.

If individual fields have a minor typographical error, then string comparators that account for such errors can allow effective comparisons (Winkler, 1995, 2004b; Cohen, Ravikumar, & Fienberg, 2003). Individual fields might be first name, last name, and street name, which are delineated by standardization software. Rule-based methods of standardization are available in commercial software for addresses and in other software for names (Winkler, 1995, 1999b). The probabilities in Equations 1 and 2 are referred to as matching parameters. If training data consisting of matched and unmatched pairs are available, then a supervised method requiring training data can be used for estimation of the matching parameters. Optimal-matching parameters can sometimes be estimated via unsupervised learning methods, such as the EM algorithm. The parameters are known to vary significantly across files (Winkler, 1999b). They can even vary significantly across similar files representing an urban area and an adjacent suburban area. If two files each contain 1,000 or more records, than bringing together all pairs from two files is impractical,due to the small number of potential matches within the total set of pairs. Blocking is the method of considering only pairs that agree exactly (character by character) on subsets of fields. For instance, a set of blocking criteria may be to consider only pairs that agree on the U.S. Postal zip code and the first character of the last name. Additional blocking passes may be needed to obtain matching pairs that are missed by earlier blocking passes (Newcombe et al., 1959; Hernandez & Stolfo, 1995; Winkler, 2004a).

Statistical Data Editing and Imputation

Correcting inconsistent information and filling in missing information needs to be efficient and cost effective. For single fields, edits are straightforward. A lookup table may yield correct diagnostic or zip codes. For multiple fields, an edit might require that an individual younger than 15 years of age must have a marital status of unmarried. If a record fails this edit, then a subsequent procedure would need to change either the age or the marital status.

Editing has been done extensively in statistical agencies since the 1950s. Early work was clerical. Later computer programs applied if-then-else rules with logic similar to the clerical review. The main disadvantage was that edits that did not fail, for a record would initially fail as the values in fields associated with edit failures were changed. Fellegi and Holt (1976) provided a theoretical model. In providing their model, they had three goals:

1. The data in each record should be made to satisfy all edits by changing the fewest possible variables (fields).
2. Imputation rules should derive automatically from edit rules.
3. When imputation is necessary, it should maintain the joint distribution of variables.

Fellegi and Holt (1976; Theorem 1) proved that implicit edits are needed for solving the problem of Goal 1. Implicit edits are those that can be logically derived from explicitly defined edits. Implicit edits provide information about edits that do not fail initially for a record but may fail as the values in fields that are associated with failing edits are changed. The following example illustrates some of the computational issues. An edit can be considered as a set of points. Let edit E =