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 paper provides an overview of two methods for improving quality. The first is record linkage 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 record linkage methods are suitable for files with hundreds of millions of records (Winkler, 2004a, 2008). The fastest edit/imputation methods are suitable for files with millions of records (Winkler, 2004b, 2007a).

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. 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 organization 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 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 OF THE CHAPTER

This section provides an overview of record linkage and of statistical data editing and imputation. The cleaning-up and homogenizing of the files are preprocessing steps prior to data mining.

Record Linkage

Record linkage is also referred to as *entity resolution* or *object identification*. Record linkage was 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** ´ **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:

$$R = P(\gamma \in \Gamma \mid M) / P(\gamma \in \Gamma \mid U) \quad (1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ over all the pairs in **A** ´ **B**. For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. The distinct patterns in Γ partition the entire set of pairs in **A** ´ **B**. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as “Smith”, “Zabrinsky”, “AAA”, and “Capitol” oc-

cur. 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:

If $R > T_{\mu}$, then designate pair as a match.

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

If $R < T_{\lambda}$, then designate 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 it is intuitive that $\gamma \in \Gamma$ would 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 *unique identifiers* such as a verified social security number that identifies an individual record or entity. *Quasi identifiers* such as name, address and other non-uniquely identifying information are used. The combination of quasi identifiers can determine whether a pair of records represents the same entity. If there are errors or differences in the representations of names and addresses, then many duplicates can erroneously be added to a warehouse. For instance, a business may have its name 'John K Smith and Company' in one file and 'J K Smith, Inc' in another file. Without additional corroborating information such as address, 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 pre-processing 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 there is minor typographical error in individual fields, then string comparators that account for typographical error can allow effective comparisons (Winkler, 2004b; Cohen, Ravikumar, & Fienberg, 2003). Individual fields might be first name, last name, and street name that are delineated by extraction and standardization software. Rule-based methods of standardization are available in commercial software for addresses and in other software for names (Winkler, 2008). The probabilities in equations (1) and (2) are referred to as *matching parameters*. If training data consisting of matched and unmatched pairs is available, then a *supervised method* that requires training data can be used for estimation matching parameters. Optimal matching parameters can sometimes be estimated via unsupervised learning methods such as the EM algorithm under a conditional independence assumption (also known as *naïve Bayes* in machine learning). The parameters vary significantly across files (Winkler, 2008). They can even vary significantly within a single file for subsets representing an urban area and an adjacent suburban area. If two files each contain 10,000 or more records, than it is impractical to bring together all pairs from two files. This is because of the small number of potential matches within the total set of pairs. *Blocking* is the method of only considering pairs that agree exactly (character-by-character) on subsets of fields. For instance, a set of blocking criteria may be to only consider pairs that agree on U.S. Postal ZIP code and first character of the last name. Additional blocking passes may be needed to obtain matching pairs that are missed by earlier blocking passes (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 look-up table may show that a given value in not in an acceptable set of values. For multiple fields, an edit might require that an individual of less than 15 years of age must have 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. Alternatively, an edit might require than the ratio of the total payroll divided by the number of employees at a company within a particular industrial