Decision Support Systems 57 (2014) 160–171

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Identity matching and information acquisition: Estimation of optimal threshold parameters

Pantea Alirezazadeh [a], Fidan Boylu [a], Robert Garfinkel [a], Ram Gopal [a], Paulo Goes [b]

[a] Department of Operations and Information Management, University of Connecticut, United States
[b] Department of Management Information Systems, University of Arizona, United States

ARTICLE INFO

Article history:
Received 3 July 2012
Received in revised form 20 June 2013
Accepted 27 August 2013
Available online 9 September 2013

Keywords:
Data quality
Statistical estimation
Sampling distributions
Record matching
Information acquisition
Type I and Type II errors

ABSTRACT

With the growing volume of collected and stored data from customer interactions that have recently shifted towards online channels, an important challenge faced by today's businesses is appropriately dealing with data quality problems. A key step in the data cleaning process is the matching and merging of customer records to assess the identity of individuals. The practical importance of this research is exemplified by a large client firm that deals with private label credit cards. They needed to know whether there existed histories of new customers within the company, in order to decide on the appropriate parameters of possible card offerings. The company incurs substantial costs if they incorrectly "match" an incoming application with an existing customer (Type I error), and also if they falsely assume that there is no match (Type II error). While there is a good deal of generic identity matching software available, that will provide a "strength" score for each potential match, the question of how to use the scores for new applications is of great interest and is addressed in this work. The academic significance lies in the analysis of the score thresholds that are typically used in decision making. That is, upper and lower thresholds are set, where matches are accepted above the former, rejected below the latter, and more information is gathered between the two. We show, for the first time, that the optimal thresholds can be considered to be parameters of a matching distribution, and a number of estimators of these parameters are developed and analyzed. Then extensive computations show the effects of various factors on the convergence rates of the estimates.

1. Introduction

With the growing volume of collected and stored data from customer interactions that have recently shifted towards online channels, an important challenge faced by today's businesses is appropriately dealing with data quality problems. A key step in the data cleaning process is the matching and merging of customer records to assess the identity of individuals. In a variety of situations, organizations use high-speed searching and matching capabilities integrated into their processes to identify existing relationships and historical customer information, to aid in critical decision making applications. For example, our client firm in dealing with private label credit cards needed to know whether there existed histories of new customers within the company, in order to decide on the appropriate parameters of possible card offerings.

If the records to be matched pertain to people, the social security number (SSN) would be a highly desirable identifier. However, faced with the recent growing problem of identity theft, a number of states are actively enacting legislation that prevents businesses from seeking SSN information from applicants. More importantly, federal law restricts most businesses to collect SSNs only when a transaction requires the Internal Revenue Service to be notified, or when it is a financial transaction that is subject to the federal Customer Identification Program rules (www.privacyrights.org). Consequently, despite their importance in aiding 'identity matching', many businesses are forced to operate without such unique identifiers. This makes identity matching a challenge in situations such as cross-checking customer information with other data sources.

A viable approach to lower the risks associated with wrong decisions is to implement a matching algorithm. There is a good deal of commercially available software (e.g. www.datalever.com/matching.html, www.dataqualityapps.com) that provides record comparison capabilities to assess whether two records pertain to the same individual. In this work we do not focus on the design of such software, but focus on the best way to implement it, for their particular client database, as requested by our client company.

In the typical matching framework each new record (usually called an enquiry) is fed to the algorithm. For each incoming record, it finds the most closely matching record resident in the database and assigns it a score (Y). Without loss of generality, let the possible scores be denoted by $\mathbf{S} = (S_1,\ldots,S_K)$ where $S_1 < S_2 < \cdots < S_K$, and higher scores indicate more likely matches. That is, scores can be considered to be proxies for the probabilities of true matches, although the actual relationship

E-mail addresses: pantea.alirezazadeh@business.uconn.edu (P. Alirezazadeh), fidan.boylu@business.uconn.edu (F. Boylu), rgarfinkel@business.uconn.edu (R. Garfinkel), ram.gopal@business.uconn.edu (R. Gopal), pgoes@eller.arizona.edu (P. Goes).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.08.014

between scores and probabilities can only be determined based on the actual data in the company's database. That is, they will differ among different databases for a generic software. In many cases, for instance the experience of our client, Y is virtually universally unique in that it applies to exactly one resident record. (Later in this work we indicate how to deal with the case of multiple "best" matches.) These scores are then compared to context-specific lower and upper thresholds $(S_{\hat{L}}, S_{\hat{H}})$, set by the organization, to determine how to proceed.

The resulting rules are of the following form: accept the match if $Y \geq S_{\hat{H}}$; reject the match if $Y < S_{\hat{L}}$; and gather additional information if $S_{\hat{L}} \leq Y < S_{\hat{H}}$. Notice that the rule is structured so that in the special case of a single threshold, e.g. where the cost of acquiring additional information is prohibitive, the solution is well-defined when $Y = S_{\hat{L}} = S_{\hat{H}}$, namely to accept the match. Also observe that the 'hats' over L and H represent the observation that organizations calibrate the appropriate thresholds for particular applications, through sampling from historical data. That is, they attempt to estimate the optimal thresholds (S_L, S_H), which can be considered to be parameters of the population probability distribution of correct matches given scores.

In this setting, the decision maker (organization) is faced with three fundamental expected costs, two of which are associated with errors. These are, the expected cost of incorrectly matching a new customer with an existing customer, which can be considered to be a Type I error, where the null hypothesis is that the records are not a match, and that of not being able to find a match when the true matching record does exist in its database (Type II). The third expected cost is of gaining enough additional information to help determine the correct result. In general, that information could be stochastic (increases the reliability of the score) or deterministic (actually determines with probability one whether the two records match). In this work, we only consider the latter option and leave the former for future investigation. Thus the problem is how to balance these three costs, all of which were considered to be substantial by our client company. A number of factors create significant challenges in addressing these issues.

While a higher score generally indicates higher 'match probability', the relationship between score and match probability is context dependent (due to factors such as technology used for data collection, customer base, application setting, and business environment) and can be nonlinear. That is, an increase in strength may not imply a proportional increase in match probability. This is illustrated in Fig. 1, which shows two different possible mappings between score and match probability based on the application settings. For example, consider two possible scenarios, one where the customer base is predominantly young and mobile career professionals and the other where the customer base is older and more settled. In the latter case, deviances in address information are more indicative of a lack of a match than in the former. Assuming that 'address' is one of the fields used by the selected matching algorithm, the result of the two scenarios could reasonably be as depicted in Fig. 1.

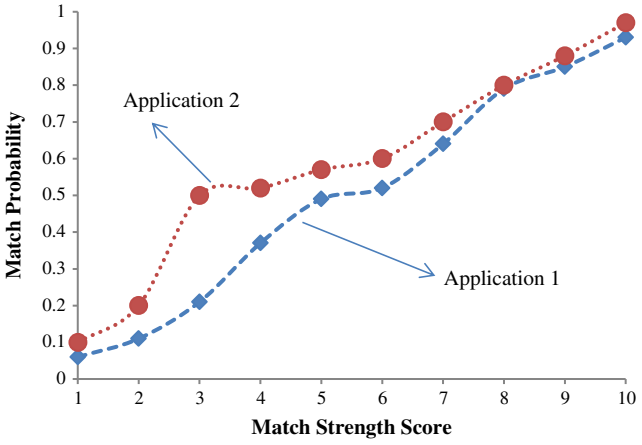

Fig. 1. Match strength score vs. match probability mapping.

Given such context specificity, organizations need to resort to calibration of the thresholds through a sampling strategy. In many business settings, such sampling can be a costly and time consuming manual endeavor, since it entails significant human effort to ascertain whether a pair of records that are matched by the software actually refers to the same individual. In our client organization, procuring sample data on about 3000 individuals was a six-month process that involved significant investigative work, ensuring compliance with regulatory guidelines, and procuring legal consent from customers.

Consequently, a major challenge that arises in this context is the development of statistically robust estimators for (S_L, S_H), and is a central issue addressed in this work. In the remainder of the paper, we compare possible estimators and find one to be dominant. We also investigate the statistical viability of the sampling strategy needed to effectively implement the estimation.

The remainder of the paper is organized as follows. Section 2 provides a brief review of the literature on record matching. Section 3 presents and analyzes some possible threshold estimators. Section 4 presents concluding remarks and directions for future research, and supplementary computational results are presented in the two appendices.

2. Related literature

Here we review the literature on record matching algorithms and on acquiring additional information to supplement the interpretation of the resulting scores. However it should be emphasized that, to the best of our knowledge, there is no extant literature comparing possible threshold estimators through sampling and analyzing their statistical characteristics.

Record matching has been extensively studied [3,8,17,21]. Many solution models have been developed ranging from probabilistic record matching models [19] to distance-based techniques [4,11]. For example, a typical matching algorithm such as a distance based technique [6,16] expresses the similarity between two records as a function of the weighted sum of the distances between their attribute values. In the database community, the same problem has been termed 'merge/purge' which is centered on the removal of duplicate data when merging multiple databases. For example, in [9], a rule-based knowledge base is used to implement a solution. In machine learning, this problem is attacked by applying supervised learning techniques to record linkage when labeled data is available [18,20].

Much attention has been devoted to the problem of how to assign scores to the matched pairs. However, less has been paid to the problem of deciding 'when' to classify records as matches, given the scores [2]. In the FS-model [8] framework, two thresholds are provided to classify a record pair as a match, a non-match or a possible match. The region between two thresholds is referred to as the clerical review region.

These thresholds are determined by a priori error bounds on false matches and false non-matches. The false match rate is defined as the number of falsely matched pairs divided by the number of declared match pairs. Intuitively, the false match rate is very sensitive to the setting of the threshold levels. Belin and Rubin [2] introduced methods for false match rate estimation. They use training data sets reviewed by human experts, and provide a predicted probability of a match as a function of the weight. Armstrong and Mayda [1] also suggest an estimation technique for acquiring these error rates and provide model-based estimators for record linkage error rates. The general problem of estimating these error rates is still an open problem, known as the regression problem [22].

Most recent research on data integration focuses on the problems caused by entity heterogeneity i.e. different representations of the same entity exist in different databases. Dey et al. [5] model the uncertainty inherent in the matching process due to errors in data collection, entry, and representation. A probabilistic model in attribute level is used to match