Using Structured Neural Networks for Record Linkage

Burdette Pixton and Christophe Giraud-Carrier
Department of Computer Science, Brigham Young University
Provo, UT 84602

1 Introduction

This paper reports on our continuing work on pedigree-based record linkage. Our earlier work, MAL4:6 - Using Data Mining for Record Linkage, presented at FHTW 2005, introduced a heterogeneous similarity metric for genealogical records, and showed preliminary results of its performance on individual-only and pedigree-based record linkage. Although these results were encouraging, two main challenges remained: the similarity metric was unweighted and the data was naturally skewed. In this paper, we present a solution to both of these challenges in the form of Filtering Structured Neural Networks.

2 Structured Neural Networks

To facilitate our exposition, we being with the following definitions.

- Let $A = \{A_1, A_2, \ldots, A_n\}$ be the set of attributes that characterize individuals. In practice, each A_i represents some piece of information about individuals, e.g., first name, last name, date of birth, place of birth, etc.

- For each attribute A_i, let sim_{A_i} denote the type-dependent similarity metric associated with A_i (e.g., Jaro-Winkler, Euclidean, etc.).

- Let $x =< A_1 : a_1^x, A_2 : a_2^x, \ldots, A_n : a_n^x >$ denote an individual, where each a_j^x is the value of attribute A_j for x. The individual John Smith, for example, is represented by the tuple $< firstname : John, lastname : Smith, ... >$.

- Let $R = \{R_0, R_1, \ldots, R_m\}$ be a set of functions, $\forall i, R_i : Individual \leftarrow Individual$, that map an individual to one of its relatives (e.g., Father(x)). In practice, each $R_i(x)$ represents a member of x's pedigree. By convention, we letR_0 denote the identity function (i.e., self).

In the context of pedigree-based record linkage, where we wish to take into account similarity between both individuals and their relatives, it is rather easy to see that a composite similarity measure would benefit from being weighted in at least two complementary ways, as follows.

1. Across two individuals, the attributes are likely to carry different weights when considering overall similarity between individuals. For example, it

seems reasonable to expect that matching surnames may be more relevant in determining an overall match than would be matching birth places.

2. Across two pedigrees, the similarities between corresponding individuals (e.g., grandmothers) are likely to carry different weights when considering overall similarity between pedigrees. For example, it seems reasonable to expect that similarity between mothers may be more relevant in determining an overall match than would be similarity between great-grandfathers.

In other words, we wish our similarity metric to have the following form.

$$sim(x, y) = \sum_{i=0}^{p} \omega_i \sum_{j=1}^{n} \alpha_j sim_{A_j}(A_j^{R_i(x)}, A_j^{R_i(y)}) \qquad (1)$$

The α_j's capture the first set of weights as discussed above, and the ω_i's the second. The remaining question is that of how to find an optimal set of weights, i.e., one that results in high precision and recall in record linkage. Our approach is to learn the weights from data, specifically using a structured neural networks.

A structured neural network is a neural network whose architecture or topology is constrained in some way to bias its learning, as well as possibly to facilitate knowledge extraction. Here, there is a straightforward 1-to-1 mapping between equation (1) and a structured neural network, that can in turn be trained using standard backpropagation to learn a near-optimal set of weights. We will describe this structured neural network and report on its performance on data provided by the Family History Department of the Church of Jesus-Christ of Latter-day Saints.

3 Filtering

One of the inherent characteristics of record linkage data sets is their (sometimes extreme) skewness, arising from the fact there are far more mismatches than there are matches, i.e., the probability that 2 records taken at (almost) random represent the same person is very small. Skewness is a major challenge for any learning algorithm.

We will present a method, called filtering, that allows a kind of successive reproportioning of the data through delegation among structured neural networks. At each stage, the structured neural network classifies the data items it is most confident about, and passes the rest to another structured neural network. The delegation is such that skewness decreases at each stage, thus simplifying the learning task along the way. We will describe the approach in details and present results that demonstrate its effectiveness and the performance improvement it produces.

Although beyond the scope of this paper, we mention here that filtering structured neural networks also allow us to compare the network at different levels of distribution skewness, as well as the ability to examine the impact of various pieces of information on record linkage through well separated weights. This is the subject of further work.