Probability and Equality: A Probabilistic Model of Identity Uncertainty

Rita Sharma and David Poole

Department of Computer Science, University of British Columbia
Vancouver, BC V6T 1Z4, Canada
rsharma,poole@cs.ubc.ca

Abstract. Identity uncertainty is the task of deciding whether two descriptions correspond to the same object. In this paper we discuss the identity uncertainty problem in the context of the person identity uncertainty problem – the problem of deciding whether two descriptions refer to the same person. We model the inter-dependence of the attributes using a similarity network representation. We present results that show that our method outperforms the traditional approach for person identity uncertainty which considers the attributes as independent of each other.

1 Introduction

Identity uncertainty has been studied independently under various names by different user communities. Within the statistics community, this problem has been studied as record linkage [2]. The Fellegi-Sunter method [2] is the standard probabilistic method for solving this problem. In computer science literature the same problem has been studied under various names, duplicate detection [5], merge/purge problem [4], identity uncertainty [6], or unsupervised classification [7]. With the exception of [7], in all of the above approaches, an independence assumption is made: i.e., matching of one attribute doesn't depend on other attributes. However, this assumption is often faulty. For example, people living in the same household have the same address, phone number and often the same last name. In this situation, the independence assumption can cause a "false positive match". As an another example, when a person moves to a different city, his address, phone number, and postal code all change together. In this situation, the independence assumption can cause a "false negative match". In this paper we discuss the identity uncertainty problem in the context of person identity uncertainty. We model the dependence/independence between attributes using a similarity network representation [3]. To deal with data entry errors, we use different error models. To test the proposed approach, as real databases are confidential, we model a reasonably realistic distribution of attribute values by modelling the people in a set of households.

2 Probabilistic Modelling of Person Identity Uncertainty

X and Y are two records, which refer to the people to be compared and $Desc_X$ and $Desc_Y$ denote their corresponding descriptions. There are two hypotheses for records

X and Y given their descriptions: X and Y refer to the same person ($X = Y$), or X and Y refer to different persons ($X \neq Y$). The odds, $Odds$, for hypotheses

$$Odds = \frac{P(X = Y)}{P(X \neq Y)} \times \frac{P(Desc_X \wedge Desc_Y | X = Y)}{P(Desc_X \wedge Desc_Y | X \neq Y)}$$

The ratio $\frac{P(Desc_X \wedge Desc_Y | X=Y)}{P(Desc_X \wedge Desc_Y | X \neq Y)}$ is a likelihood ratio (LR). The decision can be made using decision theory [1], given LR and the cost of false positive and negative matches.

To identify a person we consider the following seven attributes: *Social insurance number (SIN)*, *first name (Fname)*, *last name (Lname)*, *date of birth (DOB)*, *gender (Gen)*, *phone number (PH)*, and *postal code (PC)*. We model the inter-dependence between the attributes using a similarity network representation [3].

2.1 The Model of Attribute Dependence for Hypothesis $\mathbf{X} \neq \boldsymbol{Y}$

The statistical dependence among the attributes that we assume is shown in Fig. 1 (a). Propositions *twins*, *relative*, *samehousehold*, and *samelastname* represent that X and Y are twins, relatives, living in the same household, or have the same last name. Attribute SIN doesn't depend on the other attributes. However, we cannot assume that the SIN of two different people is independent. Knowing a different person's SIN changes our belief in X's SIN, because, we expect that they shouldn't be the same; see [9] for details.

2.2 The Model of Attribute Dependence for Hypothesis $\mathbf{X} = \boldsymbol{Y}$

If records X and Y refer to the same person, we expect that the attributes values should be the same for both X and Y. However, there may be differences because of errors, for example: typing errors, nick names, and so on. We model the dependence among attributes using their actual values, the sloppiness of the data entry person (*SloppyX*, *SloppyY*), and the possibility of movement (*move*). The dependence between attributes is shown in Fig. 1 (b). The proposition *Afname* represents the actual first name. The proposition *EFx* represents the error in first name for record X. To make this paper more readable, we consider only the following errors[1] (values of *EFx*): *copy error* (ce), an error where a person copies a correct name, but from the wrong row of a table, *single digit/letter error* (sde), and the lack of any errors ($noerr$). The random variables $Fnamex$, $Fnamey$, and $Afname$ have, as domains, all possible first names. We assume that we have a procedural way for generating the prior probabilities of the variables that have very large domains (even unbounded); see [9] for details. For the probability $P(Afname|Sex)$, we use name lists available from the U.S. Census Bureau[2]. The conditional probability $P(Fnamex|Afname \wedge Sex \wedge EFx)$ cannot be represented in a tabular form because the domains of $Afname$ and $Fnamex$ are very large. To reason in an efficient manner we need a compact representation for the large CPTs.

[1] Although, we consider many more errors in the experiment.

[2] http://www.census.gov/genealogy/names/

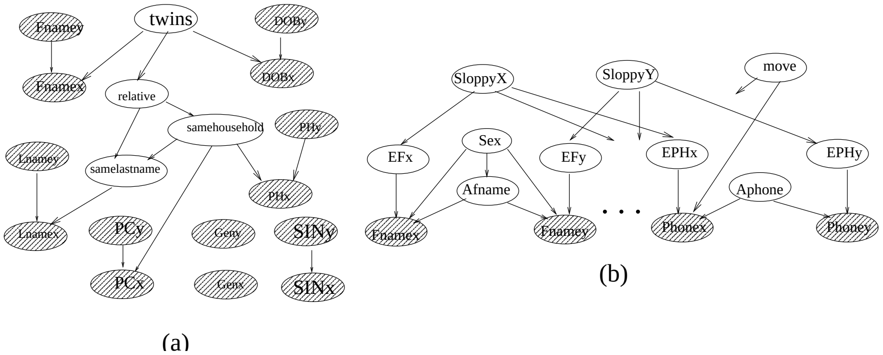

Fig. 1. Similarity network representation of attribute dependency

3 Representation of Large CPTs

We can represent the large CPTs in a compact form using both *intensional* and *extensional* representation. For example, the CPT $P\left(Fnamex|Afname \wedge Sex \wedge EFx\right)$ can be represented in a decision tree form by conditioning on the values of EFx as shown in Fig. 2. The predicate $equal$ tests whether variables $Fnamex$ and $Afname$ have the same value or not, predicate $singlet$ tests whether the values for variables $Fnamex$ and $Afname$ are a single letter apart or not, and predicate $intable$ tests whether the value of $Fnamex$ exist is in the male (or female) name file or not. The function $prsing$ is used to compute the probability when the data entry person makes the "single digit error" (sde). For example, if $EFx = sde$, $Fnamex = dave$ then $prsing(dave) = \frac{1}{100}$. The function $lookup(Fnamex, male)$ computes the probability of $Fnamex$ by looking in the male name file. We assume here that we have the procedures that can compute these predicates and functions in an efficient manner.

4 Inference

To compute the likelihood ratio we need to condition on the observations and marginalize over the unobserved variables in the Bayesian networks shown in Fig. 1. We can marginalize over the unobserved variables for Bayesian network shown in Fig. 1(a) using the Variable Elimination (VE) algorithm. We get the likelihood of the observed data given the hypothesis $X \neq Y$. The marginalization for the network shown in Fig. 1 (b) is complicated. The standard inference algorithms do not allow the intensional representation. To overcome this, we use the *Large Domain VE* algorithm [8] that allows us to make inference with intensional representation. The main challenges of applying the *Large Domain VE* algorithm to the "person identity uncertainty" problem are in the computation of *intensional functions* and predicates that arise in this problem. Due to space constraints, we omitted these details from this short paper; for details see the full version of the paper [9].

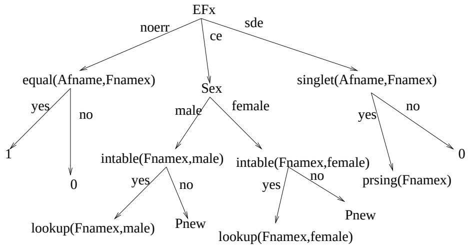

Fig. 2. A Decision Tree Representation of the CPT $P(Fnamex|Afname \wedge Sex \wedge EFx)$

5 Experimental Evaluation

To test our approach for the person identity uncertainty (as real databases are confidential), we model a a small town of 1500 households. Persons living in the same household have the same address and phone number. The probability that a *single person* lives in a house is 0.4. The probability that a person is living with a *partner* is 0.6. For a *single person* there is a 30% chance of having one child[3]. The chances for a subsequent child is 10%. The probability that partners have the same last name is 0.5. For partners there is a 70% chance of having one child. The chances for a subsequent child is 30%. When both partners have different last names then the probability that the child will have any of the parent's last name is the same. Each record of the population contains seven fields as mentioned in Section 2. Personal first names and last names are chosen according to the distribution from U.S. census file[4].

After creating the true population, we made two datasets, D_A and D_B. To create D_A we randomly took 600 records from the true population and corrupt them using the database generator of Hernandez and Stolfo [4] using typographical errors and movement into the true record. We place these corrupted records in dataset D_A. Similarly, we made D_B but we took 1500 records from the true population. We compared each record of D_A with each record of D_B. In these comparisons there were 227 duplicate cases. We compute the likelihood ratio considering both attribute dependence and independence. After computing the likelihood ratio between all pairs of records, we set the deciding threshold equal to the maximum of maximum likelihood ratio from both cases. The pair of records with likelihood ratio greater than the deciding threshold were taken as duplicates. We compute the precision and recall. We reduce the deciding threshold with a step of 1 until the deciding threshold is equal to the minimum likelihood ratio from both cases. For each value of threshold we compute the precision and recall for both cases. Figure 3 shows the precision versus recall for both cases. The recall/precision curve shows that with attribute dependence the precision of the prediction is 95% with 100% recall, while with attribute independence precision is 70% for 100% recall. Also, with attribute dependence 100% accuracy is achieved with more coverage than attribute independence.

[3] For each birth there is a 3% chance that twins will be born.

[4] http://www.census.gov/genealogy/names/

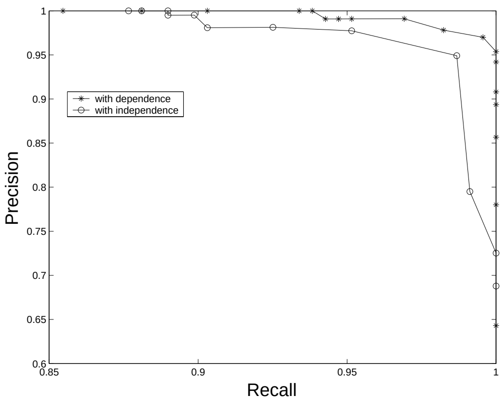

Fig. 3. Recall versus precision for both attribute dependence and attribute independence

6 Conclusion

We have presented a framework for reasoning about identity uncertainty in the context of "person identity uncertainty". The probabilistic modelling of identity uncertainty is difficult, since the domain of some of the variables is very large (even unbounded). For efficient inference in the Bayesian network we represent the big CPTs using the *intensional* and *extensional* representation. As Figure 3 shows, the proposed approach considering attribute dependence achieved a high level of accuracy over the standard approach considering the attribute independence.

References

1. R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley-Interscience Publication, John Willey and Sons Inc, second edition, 2000.
2. I.P. Fellegi and A.B. Sunter. A theory for record linkage. In *Journal of the American Statistical Association*, pages 1183–1210, 1969.
3. David Heckerman. *Probabilistic Similarity Networks*, 1990. Ph.D. thesis, Stanford University.
4. M.A. Hernandez and S.J. Stolfo. The merge/purge problem for large databases. In *Proceedings of the SIGMOD Conference, San Jose*, pages 127–138, 1995.
5. A. E. Monge and C. Elkan. An efficient domain-independent algorithm for detecting approximately duplicate database records. In *Research Issues on Data Mining and Knowledge Discovery*, 1997.
6. H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *Proceedings of the Neural Information Processing Systems (NIPS-02),* 2002.
7. P. Ravikumar and W. Cohen. A hierarchical graphical model for record linkage. In *Proceeding of Twentieth Conf. on Uncertainity in Artificial Intelligence (UAI-04),* 2004.
8. R. Sharma and D. Poole. Efficient inference in large discrete domains. In *Proceeding of Nineteenth Conf. on Uncertainity in Artificial Intelligence (UAI-03)*, 2003.
9. R. Sharma and D. Poole. *Probability and Equality: A Probabilistic Model of Identity Uncertainty,* February 2005. Technical Report TR-2005-02, Department of Computer Science, University of British Columbia.