3.1 Bayesian Decision Models

Both the FS theory and [Newcombe and Kennedy, 1962] are based on the Bayes theorem to calculate suitable probabilities used to decide whether or not two records refer to the same entity according to user-set thresholds. The hypothesis under which the following decision models can be applied is that the conditional *Probability Density Function* (PDF) and the *a priori* matching probabilities must be known.

3.1.1 Error Based

This model is *error based* since it calculates the decision thresholds T_μ and T_λ by minimizing the error of incorrectly classify a record in either M or U. The record pairs in $A \times B$ are sorted according to their composite weights and indexed according to such order. Instead of referring to the record pairs, one can refer to the elements of comparison vector $\gamma \in \Gamma$ without losing generality.

The method ensures that the level of user-defined errors (μ, λ) are admissible. Given the above defined ordering, two indexes are chosen n and n' such that the action taken among A_1, A_3 ensures the minimum error; if no better decision is achievable then A_2 is chosen. This intuition is formalized in the following condition for choosing n and n':

$$\sum_{i=n'}^{|\Gamma|} m_i \geq \lambda > \sum_{i=n+1}^{|\Gamma|} m_i$$

$$\sum_{i=1}^{n-1} u_i < \mu \leq \sum_{i=1}^{n} u_i$$

Under the assumption of mutual statistical independency of the components of γ w.r.t. each of the conditional distributions, the above conditions defines the decision function for each γ_i:

$$d(\gamma_i) = \begin{cases} (1,0,0) & 1 \leq i \leq n \\ (0,1,0) & n < i < n' \\ (0,0,1) & n' \leq i \leq |\Gamma| \end{cases}$$

In practice, when the data sets do not represent real random samples of the whole population, $m(\gamma)$ and $u(\gamma)$ (and thus, weights) can be calculated by two different methods proposed in [Fellegi and Sunter, 1969]. Furthermore, the authors also recall that such probability values can be used on subsequent linkage processes working on sub-populations $A' \subseteq A$, $B' \subseteq B$ if the underlying process which generates the records is the same (i.e., if the data sets are drawn from the same source of information).

Using prior information This method assumes that *a priori* information is available. In other words, it assumes that the probability distributions of both (i) the errors contained in the original records and (ii) the comparison characteristics are known for A and B.

Indeed, the method estimates the respective error-free frequencies of each record field in A and B, denoted as $f_{(\cdot)}$.

For instance, if the field under consideration is “address”, it is required to count all the records in which such field is reported correctly. Using the respective counts $N_A, N_B, N_{A\cap B}$, the frequencies of each distinct address are estimated $f_{A_1}, f_{A_2}, \ldots, f_{A_m}$, $f_{B_1}, f_{B_2}, \ldots, f_m$, $f_{(A\cap B)_1}, f_{(A\cap B)_2}, \ldots, f_{(A\cap B)_m}$. Each A_k correspond to a record in A where the field “address” is identified by “1”, while similarly A_k correspond to a record where the field “address” is identified by “1”. The same holds for each $(A \cap B)_k$ but the counts span to the intersection of the two record sets.

Given the above frequencies and total counts, the authors provide examples on how to estimate the probabilities $m(\cdot)$ and $u(\cdot)$; however, the following *a priori* probabilities of error are required:

- $\varepsilon_A, \varepsilon_B$ probability of misreporting an address into either the two files, A or B;
- $\varepsilon_{A^-}, \varepsilon_{B^-}$ probability of not reporting an address into either the two files, A or B;
- ε_{AB} probability of reporting the address in the wrong set, regardless of the correctness of the value itself.

This point is critical in our opinion. [Fellegi and Sunter, 1969] assume that all the addresses (i.e., different values for a field) have the same probability of being reported erroneously. However, it is not uncommon that complicated addresses are more likely to be mistyped/misreported; in addition, there are many factor, more or less related to the data itself, influencing the probability of error which is all but uniform among different values. Considering the following comparison vector:

$$\gamma = [\text{“addresses disagree”}, \text{“addresses missing on either file”}]$$

the actual probabilities are composed by means of the above defined error rates:

$$\begin{aligned}
m(\gamma^1) &= [1-(1-\varepsilon_A)(1-\varepsilon_B)(1-\varepsilon_{AB})](\varepsilon_{A^-})(1-\varepsilon_{B^-}) = \\
&= \varepsilon_A + \varepsilon_B + \varepsilon_{AB} \\
m(\gamma^2) &= 1-(1-\varepsilon_{A^-})(1-\varepsilon_{B^-}) = \\
&= \varepsilon_{A^-} + \varepsilon_{B^-} \\
u(\gamma^1) &= \left[1-(1-\varepsilon_A)(1-\varepsilon_B)(1-\varepsilon_{AB})\sum_j \frac{f_{A_j}}{N_A}\frac{f_{B_j}}{N_B}\right](1-\varepsilon_{A^-})(1-\varepsilon_{B^-}) = \\
u(\gamma^2) &= 1-(\varepsilon_{A^-})(1-\varepsilon_{B^-}) = \\
&= \varepsilon_{A^-} + \varepsilon_{B^-}
\end{aligned}$$

If two files are large enough and they are drawn from the same population one may assume that $\frac{f_{A_j}}{N_A} = \frac{f_{B_j}}{N_B} = \frac{f_{(A\cap B)_j}}{N_{AB}}$. From a quantitative point of view, positive weighs contribute to a “match” decision while negative ones contribute to the opposite. Also, it must be noticed that the weight of each field represents somehow the “rarity” of a value: the rarer the value, the larger the weight. Finally, missing values tend to zero out the weights.

To make this algorithm feasible in practice the authors point out that it is not required to list all possible values for each field but, for the reason outlined above, the portion of the most common one will lead to an optimal approximation.

Probability estimation This method estimates $m(\gamma)$ and $u(\gamma)$ from the available records and it is presented in [Fellegi and Sunter, 1969]. Not only it requires the independence assumption but the two data sets must be large enough to make the estimates valid and statistically significant. Beside the estimates of the probabilities, the algorithm also outputs the number N of linked records.

The algorithm proposed by the author is direct and can be applied simply by instantiating the given formulae with certain frequencies parameters which can be automatically calculated from data. The assumption is that m and u must be such that:

$$\begin{aligned} m(\gamma) &= m_1(\gamma^1) \cdot m_2(\gamma^2) \cdots m_k(\gamma^K);\ K \geq 3 \\ u(\gamma) &= u_1(\gamma^1) \cdot u_2(\gamma^2) \cdots u_k(\gamma^K);\ K \geq 3 \end{aligned}$$

which means that γ must have at least three components and they have to be independent to each other. The frequencies of each different configuration of γ is calculated by direct comparison of A against B; the only frequencies of interest are those of “agreements” configurations, denoted with $\Gamma_h^+ \in \Gamma$ for the hth component. More precisely:

- $\hat{F}_{M_{h-}}$ = frequency of agreements in all components except the hth and any configuration in the hth component. The associated probability is denoted as $m_h = \sum_{\gamma \in \Gamma_h^+} m(\gamma)$.
- $\hat{F}_{U_h}$ = frequency of agreements in the hth component and any configuration in all but the hth. The associated probability is denoted as $u_h = \sum_{\gamma \in \Gamma_h^+} u(\gamma)$.
- $\hat{F}_M$ = frequency of agreements in all components;

The authors have proven that given the frequencies expressed in terms of m and u

$$\begin{aligned} \hat{F}_{M_{h-}} &= \frac{N}{N_A N_B} \prod_{j=1, j\neq h}^{3} m_j + \frac{N_A N_B - N}{N_A N_B} \prod_{j=1, j\neq h}^{3} u_j \quad h = 1, 2, 3 \\ \hat{F}_{U_h} &= \frac{N}{N_A N_B} m_h + \frac{N_A N_B - N}{N_A N_B} u_h \quad h = 1, 2, 3 \\ \hat{F}_M &= \frac{N}{N_A N_B} \prod_{j=1}^{3} m_j + \frac{N_A N_B - N}{N_A N_B} \prod_{j=1}^{3} u_j \end{aligned}$$

	Error based	
	Prior Information (PI)	**Probability Estimation (PE)**
Hypotheses	Known PDF	$\lvert\gamma\rvert \geq 3$ (independent)
Input	error probabilities	frequency of values
Output	$m(\cdot), u(\cdot), N$	$m(\cdot), u(\cdot)$

Table 2: Brief comparison of differences, similarities and peculiarities of the two methods for weights calculation in error based record linkage techniques. Note: N here indicates the number of matching records.

and solving such equations in m_h, u_h and N, the estimates of $m(\gamma^k)$ and $u(\gamma^k)$ can be computed after the direct observation of $\hat{F}_{M_{h-}}$, $\hat{F}_{U_1}$, $\hat{F}_{U_2}$ and $\hat{F}_{U_3}$ for the specific configurations γ_i^k, γ_i^1, γ_i^2 and γ_i^3, respectively. As the authors stress, this method requires the sample to be large and representative of the whole population.

Comparison In this paragraph we compare the two techniques that have been reported in Section 3.1.1 and Section 3.1.1, summarized in Table 2.

3.1.2 Cost Based

The cost based model can be seen as a generalization of the classical purely Bayesian decision model we mentioned in the previous section. Instead of relating a link between records only with a probability, this models attaches a *cost function* to each decision (i.e., match vs. non-match). Thus, instead of minimizing the error, this method give hints in designing decision rules based on the minimization of a cost.

Generally speaking, the "cost" models the fact that a misclassification has different impacts on the organization data, depending on many factors influencing the whole data flow.

Linear loss method [Tepping, 1968] proposed a method based on a linear loss function, $g(A_i, (a, b))$, defined for each action A_i on the pair (a, b). Given the conditional probability $P(M|\gamma) = P((a, b) \in M|\gamma[\alpha(a), \beta(b)])$, defined as above, the authors define the expected loss G as a function of the action and the conditional probability: $G(A_i, P(M|\gamma)) = \mathrm{E}[g(A_i, (a, b))]$. Hence, the total expected loss is $\sum P(\gamma) \cdot G(A_i, P(M|\gamma))$, which is minimized in order to obtain the optimal linkage rule.

The authors have shown that under the assumption of linearity of G, the interval $(0, 1)$ for the probability of a match is partitioned into a fixed number of possible actions (e.g., $4, A_1, A_2, A_3, A_4$ but it could be any number). The so called *action interval* is the interval in which the loss function G is minimal w.r.t. the same function evaluated in all other action:

$$G(P) = \min_{A_i} G(A_i, P(M|\gamma))$$

	Cost based	
	Linear Loss Function (LLF)	**Cost Matrix (CM)**
Hypotheses	Loss as a linear function	Known PDF
Input	Conditional probabilities	Per-action cost
Output	Decision intervals	Decision thresholds

Table 3: Brief comparison of differences, similarities and peculiarities of the two methods for cost based record linkage techniques. Note: N here indicates the number of matching records.

Skipping the details reported in [Tepping, 1968], it can be proved that three thresholds exist in the decision space. In particular, referring to λ, μ defined by the FS theory the thresholds are:

$$\lambda = \frac{\pi_U}{\pi_M} \cdot \frac{c_{A_2M} - c_{A_1M}}{c_{A_1U} - c_{A_2U}} \quad \kappa = \frac{\pi_U}{\pi_M} \cdot \frac{c_{A_3M} - c_{A_1M}}{c_{A_1U} - c_{A_3U}} \quad \mu = \frac{\pi_U}{\pi_M} \cdot \frac{c_{A_3M} - c_{A_2M}}{c_{A_2U} - c_{A_3U}}.$$

According to the actual values of such thresholds, the decision space depicted in Figure 1 is divided into two or three areas. In particular, the authors show that the sufficient and necessary condition for A_2 to exist is that $\lambda \leq \mu$. Furthermore, if it holds, they have shown that $\lambda \leq \kappa \leq \mu$. Otherwise (i.e., $\lambda > \mu$) $\lambda > \kappa > \mu$ but κ is such that A_2 disappears and there are only two decision areas. This happens because A_2 have a higher cost w.r.t. A_1 and A_3 because it results in manual classification.

However, the method is proven by the authors to be optimal w.r.t. the cost, in the sense that it minimizes a cost function; but on the other hand, no proof is given of its optimality in general. In other words, no clues are given to prefer this method over the others available.

Comparison In this paragraph we compare the two techniques that have been reported in Section 3.1.2 and Section 3.1.2, summarized in Table 3.

The method by [Tepping, 1968] requires slightly stronger assumption w.r.t. to the one recently proposed in [Verykios et al., 2003]. However, the idea of using cost as a criterion was originally due to [Tepping, 1968], in which the theoretical framework has been defined and detailed. The rigorous theory proposed in [Fellegi and Sunter, 1969] was significant in the construction of the more complete and directly applicable approach by [Verykios et al., 2003].

3.1.3 Comparing Error Based and Cost Based Decision Models

It is interesting to compare the two decision models at a generic level. In particular, it could be proven that the former is a special case of the latter. First of all, it must be remarked that they are both likelihood ratio tests with thresholds computed on the basis of the available information, that is the *a priori* probabilities.

in the presence of unobservable variables or missing data. Given these premises, it is straightforward to notice that the EM algorithms perfectly fits the probabilistic model defined by the FS theory.

To avoid misunderstandings, we will use the $\underline{v}$ to indicate that v is a vector. The parameters of interests are $\Phi = \langle \underline{m}, \underline{u}, p \rangle$ where $\underline{m}, \underline{u}$ denotes the probability vectors $m(\cdot), u(\cdot)$, respectively; and, p denotes the proportion of the matched records w.r.t. the total: $p = \frac{|M|}{|M \cup U|}$. The data vector is defined by γ and the function g:

$$g_j = \begin{cases} (1,0) & (a,b)_j \in M \\ (0,1) & (a,b)_j \in U \end{cases}$$

The data vector is then $\underline{x} = \langle \gamma, g \rangle$; we recall that $(a,b)_j$ indicates the generic jth record pair. An independence model is assumed at this step, thus:

$$\begin{aligned} P(\gamma^j|M) &= \prod_{i=1}^{n} m_i(\gamma_i^j)(1-m_i)^{1-\gamma_i^j} \\ P(\gamma^j|U) &= \prod_{i=1}^{n} u_i(\gamma_i^j)(1-u_i)^{1-\gamma_i^j} \end{aligned}$$

Given the log-likelihood of the data vector:

$$\begin{aligned} \log f(\underline{x}|\Phi) &= \sum_{j=1}^{N} g_j \cdot (\log P(\gamma^j|M), \log P(\gamma|U))^T + \\ &+ \sum_{j=1}^{N} g_j \cdot (\log p, \log(1-p))^T \end{aligned}$$

the algorithm consists in the iteration of two steps called *Expectation* (E) and *Maximization* (M); the iteration begins with the initial (even casual) estimates $\langle \underline{\hat{m}}, \underline{\hat{u}}, \hat{p} \rangle$ continues until the required precision is not reached. The estimation of $\underline{u}$ is less difficult w.r.t. $\underline{u}$ since $|U| > |M|$, thus the u_i can be estimated by ignoring the contribution of M. Regarding m, the g function can be estimated in the (E) step as follows:

$$\begin{aligned} \hat{g}_m(\gamma^j) &= \frac{\hat{p} \cdot P(\gamma^j|M)}{\hat{p} \cdot P(\gamma^j|M) + (1-\hat{p}) \cdot P(\gamma^j|U)} \\ \hat{g}_u(\gamma^j) &= \frac{\hat{p} \cdot P(\gamma^j|U)}{\hat{p} \cdot P(\gamma^j|U) + (1-\hat{p}) \cdot P(\gamma^j|M)} \end{aligned}$$

Note that g_j is estimated by $\langle \hat{g}_m(\gamma^j), \hat{g}_m(\gamma^j) \rangle$. The (M) step, in the case of $\underline{\hat{m}}$, it is based on:

$$\hat{m}_j = \frac{\sum_{j=1}^{s^n} \hat{g}_m(\gamma^j)\gamma_i^j \hat{F}(\gamma^j)}{\sum_{j=1}^{s^n} \hat{g}_m(\gamma^j)\hat{F}(\gamma^j)}$$

	Error based		**Cost based**		**EM Based**
	PI	**PE**	**LLF**	**CM**	
Hypotheses	Known PDF	$\|\gamma\| \geq 3$ (independent)	Loss as a linear function	Known PDF	Independency
Input	error probabilities	frequency of values	Conditional probabilities	Per-action cost	frequency of values and cond. prob.
Output	$m(\cdot), u(\cdot), N$	$m(\cdot), u(\cdot)$	Decision intervals	Decision thresholds	$\hat{m}(\cdot), \hat{u}(\cdot), \hat{p}$

Table 5: Summary of the main methods for parameter estimation used in probabilistic record linkage techniques.

where $\hat{F}(\gamma^j)$ indicates the frequency count for the jth component of the comparison vector γ. The estimate of p is then $\hat{p} = \frac{\sum_{j=1}^{s^n} g_m(\gamma^j)\hat{F}(\gamma^j)}{\sum_{j=1}^{s^n} \hat{F}(\gamma^j)}$.

The author underlines that the method is extremely easy to implement, stable, and negligibly sensitive to initialization values. It is also highlighted that all the frequency counts have to be obtained after a blocking phase, but, this detail is out from the scope of this survey so we refer the reader to [Cochinwala et al., 2001] for a more general overview of all the steps of a liking algorithm.

This algorithm have been used to detect duplicates in the census data of Tampa, Florida in 1985 [Jaro, 1989] and lately on public health data [Jaro, 1995].

3.1.5 Overall comparison

Cost based methods have been already compared with error based techniques. Table 5 summarizes all the (variants of the) approaches that have been reviewed.

One may have noticed that the EM method shares slightly the same hypotheses and the same input/output w.r.t. the PE method. However, the latter also requires the comparison vector γ to have at least three components while the former does not have this limitation.

3.2 Other Models and Methods

In the previous sections we investigated and reviewed the most solid and promising methods that are also implemented into bleeding edge tools (see Section 4). For the sake of completeness, in this section we provide list of other approaches that have been proposed in the literature so far.

In particular, a different way to model the error in the data is to *explicitly* model the errors in *each* attribute, as proposed in [Copas and Hilton, 1990]. The algorithm is based on the statistical characteristics of the errors that are expected to arise; however,