Noname manuscript No.
(will be inserted by the editor)

A note on using the F-measure for evaluating data linkage algorithms

David Hand · Peter Christen

Received: date / Accepted: date

Abstract Record linkage is the process of identifying and linking records about the same entities from one or more databases. Record linkage can be viewed as a classification problem where the aim is to decide if a pair of records is a match (i.e. two records refer to the same real-world entity) or a non-match (two records refer to two different entities). Various classification techniques — including supervised, unsupervised, semi-supervised and active learning based — have been employed for record linkage. If ground truth data in the form of known true matches and non-matches are available, the quality of classified links can be evaluated. Due to the generally high class imbalance in record linkage problems, standard accuracy or misclassification rate are not meaningful for assessing the quality of a set of linked records. Instead, precision and recall, as commonly used in information retrieval and machine learning, are used. These are often combined into the popular F-measure, which is the harmonic mean of precision and recall. We show that the F-measure can also be expressed as a weighted sum of precision and recall, with weights which depend on the linkage method being used. This reformulation reveals that the F-measure has a major conceptual weakness: the relative importance assigned to precision and recall should be an aspect of the problem and the researcher or user, but not of the particular linkage method being used. We suggest alternative measures which do not suffer from this fundamental flaw.

Keywords Record linkage · Entity resolution · Classification · Precision · Recall · Class imbalance

Peter Christen is partially supported by a grant from the Simons Foundation. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Data Linkage and Anonymisation where this work was conducted (EPSRC grant EP/K032208/1).

David Hand
Imperial College London and Winton Group Limited London, UK
E-mail: d.j.hand@imperial.ac.uk

Peter Christen
The Australian National University, Canberra, Australia
E-mail: peter.christen@anu.edu.au

1 Introduction

Tools and systems for linking records from multiple sources are becoming more and more important (see, for example Herzog et al. (2007) [14], Christen (2012) [2], and Harron et al. (2015) [13]). Such linkages allow inferences beyond those possible from the individual databases, as well as facilitating error correction and the imputation of missing values. Applications of record linkage range from the health domain, national security and crime and fraud detection, to national censuses and social science research.

Linking records is a non-trivial exercise for several reasons, including errors in the variables used for calculating the similarities between records, variations in the way information is recorded (e.g. variations in name or address format), different information being recorded at different dates, and different databases covering non-identical populations. Because of its importance and the technical difficulty, record linkage is a topic of intensive research in various domains, including statistics, computer science, and health informatics.

Central to such work is the ability to measure the effectiveness of a proposed linkage method or algorithm (Christen and Goiser 2007) [5]. This is important so that we can know if an approach is good enough for some purpose, determine if it is better than alternative methods, and decide if proposed alterations to an algorithm improve or degrade it.

There are two fundamental aspects to assessing a record linkage method, namely the quality and scalability of the method. In some contexts, the extent of privacy-preservation may also be important (Domingo-Ferrer and Torra 2003; Vatsalan et al. 2013) [7,20].

Quality refers to how effective the method is in correctly classifying pairs of records as matches if they correspond to the same real-world entity, while avoiding classifying pairs of records as matches if they do not refer to the same real-world entity. Popular measures used, to be defined below, include precision, recall, and the F-measure. As we discuss in more detail in Section 2, record linkage classification has much in common with machine learning, data mining, and information retrieval systems (Christen and Goiser 2007) [5].

Scalability refers to how well the method scales to real-world problems where the potential number of comparisons to be conducted between the records in two databases scales as the product of the sizes of the two databases. In practical applications, to reduce the quadratic space of record pair comparisons, blocking or indexing techniques (Christen 2012b) [3] are commonly applied. These techniques in some way split the databases into smaller (potentially overlapping) blocks such that only records within the same block are to be compared. For example, using a 'postcode' attribute for blocking would mean only records with the same postcode value will be compared with each other. Records that are not compared are implicitly classified as non-matches (i.e. referring to two different entities).

This paper is concerned solely with quality and we do not consider scalability. For recent work on the influence of blocking on record linkage quality the reader is referred to Murray (2016) [17]. For simplicity of exposition, we restrict ourselves to the linking of just two databases, taken to have m and n records, respectively.

The general record linkage process, as illustrated in Figure 1, consists of several major steps (Christen 2012) [2]. First, the databases to be linked need to be cleaned and standardised in the same way to ensure the actual values in the

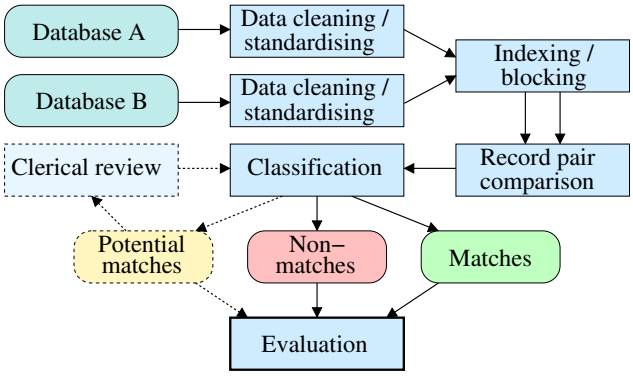

Fig. 1 Outline of the general record linkage process as described below.

attributes (fields) to be compared between records are consistent and in the same formats. This pre-processing includes, for example, converting all letters into lower-case characters, expanding abbreviations, and replacing known misspellings with their correct spellings (Christen 2012) [2]. The second step is blocking or indexing (Christen 2012b) [3], as discussed above, which is aimed at improving scalability by grouping together records that are potential matches into smaller blocks.

In the third step, pairs of records in the same block are compared in detail using appropriate comparison functions over a set of attributes. Most commonly, the attributes to be compared between records are textual strings (such as names and addresses) or dates, and comparison functions specific to the content of such attributes are employed. Approximate string comparators, such as edit distance, Jaro-Winkler, or the Jaccard coefficient, are commonly used (Christen 2012; Herzog et al. 2007) [2,14]. Each of these calculates a normalised similarity, s, between two attribute values (strings), where two completely different values (like 'david' and 'peter') result in a similarity of $s = 0$, two values that are the same in a similarity of $s = 1$, and somewhat similar values (like 'peter' and 'pedro') in similarities $0 < s < 1$. Note that different approximate string comparators can result in different similarities for the same pair of string values (Winkler 2004) [21].

In the next step, for each compared record pair, the similarities for the different compared attributes are combined into a *similarity vector*. These vectors are then used to classify each compared record pair as a *match* (the two records refer to the same entity) or a *non-match* (the two records refer to two different entities). Depending upon the classification technique used, a record pair might also be classified as a *potential match* and given to a domain expert for manual clerical review (i.e. to decide manually whether the pair is a match or a non-match). In this paper we assume a binary classifier is employed that classifies record pairs into matches and non-matches only.

Both supervised and unsupervised classifiers have been employed in record linkage (Christen 2012) [2]. The simplest classifier is based on a similarity threshold t, and classifies a record pair as a match if its overall similarity (for example, calculated as a weighted sum of its attribute similarities) is equal to or above t, and as a non-match otherwise. The final step of the record linkage process is evaluation, which is the topic of the remainder of this paper.

		True link status	
		Match	Non-match
Predicted link status	Match	d (true match)	b (false match)
	Non-match	c (false non-match)	a (true non-match)

Fig. 2 Data table (confusion or error matrix) for evaluating record linkage quality.

2 Performance data and current measures

We assume the linking of two databases, D_A and D_B, containing $m = |D_A|$ and $n = |D_B|$ records, respectively. Each pair of records, with one record from each of the two databases, can correspond to a correct link (called a *match*) or an incorrect link (a *non-match*) (Christen 2012) [2]. Thus, assuming no blocking, we can produce a two-by-two table, as in Figure 2, with each of the potential $m \times n$ compared pairs of records being allocated to one of the four cells. The columns show whether a pair corresponds to the same entity, and so should be a match or do not correspond to the same entity and so should not be a match, while the rows show if a pair is linked by the linkage procedure or not linked by the procedure.

The number of pairs which should be linked and which the procedure does link is given by the count d in the top left corner. In the context of classification, these are the *true positives* (correctly linked matches). The bottom right corner, count a, shows the number of pairs which are not linked and which the procedure (correctly) does not link. These are the *true negatives* (correctly not linked non-matches). In general, the number in this cell will far outstrip the number of pairs which should be linked. For example, in two identical databases of size n, a perfect procedure would give $d = n$ and $a = n(n-1)$.

The other two cells of the table show the number of missed true matches, i.e. false non-matches (bottom left, a count of c) and the number of incorrect or false matches (top right, a count of b).

To be able to compare and evaluate linkage procedures we must summarise the four numbers in the table to a single number, yielding a score on a quality continuum. Since the four numbers necessarily sum to $m \times n$, we have only three degrees of freedom to reduce to a single number. In other contexts, various methods have been proposed and indeed are used for this summarisation — see Christen and Goiser (2007) [5] and Christen (2012) [2] for a review in record linkage and Hand (1997, 2012) [9,12] for more general overviews — with different measures finding favour in different domains. For example, *error rate* (or its complement *accuracy*, $(a+d)/(a+b+c+d)$, the proportion correctly classified) is overwhelmingly the most popular measure in machine learning. In the realm of record linkage, as in information retrieval, the most popular measures are *precision* (also known as *positive predictive value*), *recall* (also known as *sensitivity* or *true predictive rate*), and their combination in the *F-measure* (Manning et al. 2008; van Rijsbergen 1979) [15, 19]. Precision, P, and recall, R, are defined as follows:

- $P = d/(b+d)$, the proportion of compared record pairs classified as matches that are true matches.
- $R = d/(c+d)$, the proportion of true matching record pairs that are classified as matches.

Precision and recall can be regarded as conditional probabilities:

$$Precision = P(true\ match \mid predicted\ match)$$
$$Recall = P(predicted\ match \mid true\ match)$$

The main reason for the popularity of precision and recall is that record linkage is commonly a very unbalanced classification problem. If $m \times n$ record pairs are being compared across two databases, and assuming there are no duplicates in each database (i.e. no two records in a single database refer to the same real-world entity), then the maximum number of true matches will be $d = min(m, n)$. This number (d) grows linearly with the size of the databases being linked, while the comparison space, $m \times n$, grows quadratically. Using the error rate or accuracy, which contains a (the number of true non-matches), as evaluation measure would likely lead to a meaningless result. For large databases, an accuracy of 99.99% can easily be obtained by classifying all record pairs as non-matches (Christen and Goiser 2007) [5].

Neither one of precision or recall alone, however, completely captures the performance of a record linkage procedure. For example, we can obtain a perfect recall of $R = 1$, meaning that all true matches have been correctly predicted as matches, if we simply predict all record pairs to be matches. But this would be useless. Its uselessness would be picked up by the precision, which would be low because of all the true non-matches that are classified as matches, b, in the denominator. It is therefore necessary to take both precision and recall into account when evaluating linkage procedures.

Taken as a pair, precision and recall reduce the three degrees of freedom remaining in the table to just two. Combining them will yield the univariate score we need. The most common way of combining these two measures in the record linkage literature is through the F-measure (Christen and Goiser 2007; Christen 2012; Getoor and Machanavajjhala 2012; Naumann and Herschel 2010) [5, 2, 8, 18]. This is the harmonic mean of precision and recall:

$$F = 2\left[P^{-1} + R^{-1}\right]^{-1} = \frac{2PR}{P+R} = \frac{2d}{c+b+2d}. \tag{1}$$

The value of F is high if both P and R are high. For low values of the similarity classification threshold t (as discussed in Section 1), R is higher than P, while for high threshold values P is higher than R.

From the above, we see that the F-measure can be rewritten as

$$F = \frac{c+d}{c+b+2d} \times \frac{d}{c+d} + \frac{b+d}{c+b+2d} \times \frac{d}{b+d} = pR + (1-p)P, \tag{2}$$

where

$$p = \frac{c+d}{c+b+2d}. \tag{3}$$

That is, as well as being the harmonic mean, the F-measure is also a weighted arithmetic mean with weight p given to recall and weight $(1-p)$ given to precision.

Using a weighted arithmetic mean has a sensible justification — the weights would be the relative importance assigned to each of precision and recall. However, the weights p and $(1-p)$ in the arithmetic mean interpretation of the F-measure are not chosen on the grounds of relative importance of precision and recall. They

will vary from linkage method to linkage method[1], since the cell counts in Figure 2 will vary (this is precisely why they need comparing). The measure being used to evaluate performance therefore depends on the thing being evaluated. In short, when looked at from the weighted arithmetic mean perspective, we see that the F-measure is equivalent to using different performance criteria for different linkage methods.

3 How to use the F-measure fairly

To make a fair comparison between linkage methods, the same pair of weights, p and $(1-p)$, must be used for all methods. This, of course, means that the ratio of weights must be the same for all methods. In the F-measure, when viewed as a weighted arithmetic mean, the ratio of weights is

$$\frac{p}{1-p} = \frac{c+d}{b+d} = \frac{number\ of\ true\ matches}{number\ of\ predicted\ matches}. \tag{4}$$

Since the number of true matches is the same for all methods (for a given linkage problem, i.e. the same pair of data sets to be linked) this means that the F-measure is a fair measure only if the similarity thresholds t (as discussed in Section 1) of different linkage methods are chosen so that each method produces the same number of predicted matches, $b+d$.

In summary, there are two separate issues here:

1. For the F-measure to make legitimate comparisons, the similarity thresholds t must be chosen so that each linkage method makes the same number of predicted matches.
2. The ratio of weights accorded to recall and precision is equal to $(c+d)/(b+d)$, so the similarity thresholds should be chosen such that this ratio reflects the relative importance accorded to recall and precision by the particular researcher or user for the particular problem.

The second point is often difficult to satisfy in practice: Researchers or users of a record linkage system typically find it difficult to give a precise number to the relative values they accord to precision and recall.

The problem here is precisely the same as that arising when choosing a classification threshold in supervised learning — in applications such as medical diagnosis, fraud detection, credit scoring, spam filtering, and so on. For medical diagnosis, for example, the classification threshold should be chosen so as to strike an appropriate balance between the two kinds of misclassification: misclassifying someone with the disease as disease-free, versus misclassifying someone without the disease as having it. Experience shows that medical doctors find it very difficult to choose an appropriate balance: it clearly depends on many factors, including the severities ('costs') of the two kinds of misclassification. Even in financial domains, with an apparent common numéraire of some currency, experience shows it is very difficult to do. Extensive discussions of this problem, along with possible solutions, are given in Hand (1997, 2012) [9,12].

[1] And of course also from data set to data set, however it is generally not meaningful to evaluate and compare linkage results across different data sets.

Table 1 Summary of the data sets used in the empirical evaluation. As detailed in Section 4, we employed blocking during the linkage of these data sets, resulting in the shown number of record pairs, and corresponding numbers of true matches and true non-matches.

Data set	Number of records	Record pairs	True matches	True non-matches	Imbalance M:N-M
Cora	1,295	286,141	16,532	269,609	1:16.3
NCVR	224,073 / 224,061	3,495,580	124,597	3,370,983	1:27.1

As a partial way of tackling this, to show the impact of possible choices of a threshold, one often sees plots of F-measure values against the similarity threshold, t, corresponding to a range of values of the ratio of the importance of precision to recall. However, these plots can be misleading. Similarity measures are generally heuristic constructs, so that a given similarity value using one measure is not related to the same value using another measure: it is meaningless to note that the F-measure value for linkage method A is greater than that for linkage method B at a particular similarity threshold, since the similarity values may be unrelated.

This has important implications. For example, one way in which F-measure plots are used is to see where the curves for different linkage methods cross: in terms of the F-measure, one method is apparently superior below such a crossing point, and the other method above. However, the fact that different methods use intrinsically non-comparable similarity measures means that, in plots of F-measure against similarity, such crossing points are meaningless. The relative weight given to precision and recall corresponding to a particular similarity threshold t for one method may be quite different from the relative weight which corresponds to t for another method.

What is needed is some transformation of the similarity scales so that the F-measure plots for different methods can be compared. And this, is achieved if we transform the similarity values to represent the relative weights accorded to precision and recall. That is, instead of plotting raw similarity values on the horizontal axis, we plot the ratio $p/(1-p) = (c+d)/(b+d)$. Alternatively, since $p/(1-p)$ is monotonically related to p, we could just use p as the horizontal axis, or the logarithm of the ratio, $log(p/(1-p))$. Yet another alternative, since $(c+d)$ (the number of true matches) is fixed for any given problem, is to use $(b+d)$, the number of predicted matches, as the horizontal axis.

The various alternatives of p, $p/(p-1)$, $log(p/(1-p))$ or $(b+d)$ are all equivalent and meaningful, in the sense that, unlike raw similarity, we can legitimately compare F-measure values at each value of the horizontal axis. We can choose between them on grounds of visual separation between the F-measure plots of different linkage methods.

The next section illustrates these various points on two real data sets.

4 Empirical evaluation and discussion

We now present an empirical evaluation using two data sets that have commonly been used in record linkage research. Table 1 summarises these data sets, showing their numbers of records, the number of compared record pairs after blocking was employed (as described below), the corresponding numbers of true matches and

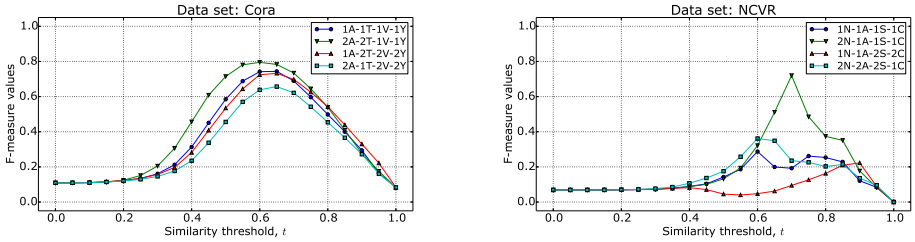

Fig. 3 Traditional F-measure plots for the 'Cora' (left) and 'NCVR' (right) data sets as described in Section 4. For 'Cora', the attributes (fields) compared are: author (A), title (T), venue (V) and year (Y); while for 'NCVR' they are: first and last name (N), age (A), street address (S) and city (C). Different weights, shown in the numbers in the legend, were assigned to different attribute similarities to simulate different linkage classifiers.

true non-matches, and the class imbalance (M:N-M) between true matches (M) and true non-matches (N-M).

The 'Cora'[2] data set (McCallum et al. 2000; Naumann and Herschel 2010) [16, 18] contains 1,295 records of 112 unique machine learning publications, where each record contains 12 fields or attributes (authors, titles, venues, years, pages, etc.). The linkage task is to identify duplicate records about the same publication (deduplication).

The North Carolina Voter Registration ('NCVR')[3] data set (Christen 2014) [4] contains real voter records including names and addresses. We used two sub-sets of this data set collected at different points in time, and extracted records about individuals that had some of their values changed between the two sub-sets (such as changed surnames or addresses, or corrected name variations and misspellings).

For both data sets we used the *Febrl* open source record linkage system (Christen 2009) [1] to conduct the linkage. For the 'Cora' data set we generated blocks using five criteria (using phonetic encoding (Christen 2012) [2] for the title, author, venue, and publisher attributes, as well as exact year values), and compared attribute values using a character q-gram based approximate string comparison function (Christen 2012) [2] on author, title, venue and year values. For the 'NCVR' data set, we blocked on pairs of phonetically encoded first name and last name, street address, and city values, and compared the two name attribute values using the Jaro-Winkler approximate string comparison function, street address and city using a q-gram based comparison function, and on age values we used edit distance (Christen 2012) [2]. The applied blocking approaches led to 286,141 compared record pairs for 'Cora' and 3,495,580 record pairs for 'NCVR', each comparison resulting in a similarity vector which we then used in the evaluation.

To simulate different linkage classifiers (i.e. linkage methods), we used four different schemes of how individual attribute similarities were weighted in the calculation of the final record similarities, as described in the caption of Figure 3. For both data sets the true link status of all record pairs is known. We normalised the record pair similarity values into $[0, 1]$. We implemented all evaluation programs in the Python programming language (version 2.7). The program codes and data sets are available from the authors.

[2] `http://secondstring.sourceforge.net`

[3] `http://dl.ncsbe.gov/`

In Figure 3 we present the traditional F-measure (Eqn. 1) plots for different values of the similarity threshold t. As can be seen, with different attribute weighting schemes (i.e. different classifiers), different F-measure results are obtained. Also, for a given data set, for different classifiers the best F-measure results are obtained for different values of the similarity threshold.

Figure 3 is fundamentally misleading. For example, the left plot for 'Cora' shows that two of the F-measure performance curves ('1A-1T-1V-1Y' and '1A-2T-2V-2Y') cross at a similarity value of around 0.7. However, as explained above, a similarity value of 0.7 for one curve may correspond to a different pair of weights for precision and recall to that of another curve. To be able to compare curves, we must transform the similarity thresholds so they are commensurate across linkage methods.

For each of the two data sets, the top row of Figure 4 shows such transformations of the four similarity schemes to scales needed to obtain corresponding number of matches, $(b + d)$, so that corresponding values of the F-measures can be properly compared. As expected, the transformations for the different linkage methods are different, demonstrating clearly that the raw similarity measures do not yield F-measure plots which can be directly compared.

The second to fifth rows of this figure show the F-measure curves corresponding to $p/(1-p)$, $log(p/(1-p))$, p, and $(b+d)$, respectively. As explained in the preceding section, the F-measure curves for these plots can be compared: a particular value on the horizontal axis means recall and precision have the same pair of weights, p and $(1-p)$, in the weighted arithmetic mean interpretation of the F-measure, whichever similarity scheme is used. The curves in the plots of these four rows are equivalent, in the sense that the horizontal axes are (nonlinear monotonic) transformations of each other — with all curves in any one plot using the same transformation. This equivalence means we can choose whichever is more convenient for understanding or exposition.

For the 'Cora' data set (left column), the clearest set of plots seem to occur when the p or $log(p/(1-p))$ horizontal axes are used (third and fourth rows). From this it is easy to see that similarity scheme (linkage classifier) '2A-2T-1V-1Y' is best (in terms of the F-measure) when the weight accorded to recall is less than about $p = 0.55$ (and the weight accorded to precision is $1 - p = 1 - 0.55 = 0.45$). Above that weight, for a short interval of p values, '1A-1T-1V-1Y' is best, before being beaten by '1A-2T-2V-2Y' for slightly larger weights on recall. The method using similarity measure '2A-1T-2V-2Y' is never best, for any weight put on recall.

The 'NCVR' data set (right column) shows rather less clear behaviour. Once again, Figure 3 is meaningless, and cannot be interpreted in terms of the relative weight given to precision and recall in a weighted sum, since each of the F-measure curves implicitly uses a different horizontal axis. The second row of Figure 4, right hand side, shows that the ratio $p/(1 - p) = (c + d)/(b + d)$ has a very large range, so that it is useless for interpretation. On a logarithmic scale, shown in the third row, the differences between the four similarity schemes become much more clear. Plotting the curves using the weight p, in the fourth row, also shows the relative behaviour of the F-measure results of the four similarity schemes much more clearly. Plotting using $(b + d)$ directly, in the bottom row, also yields clear results.

The odd shaped curves for the 'NCVR' data set both in Figures 3 and 4 require some further explanation. The reason for these curious shapes is the non-monotonic

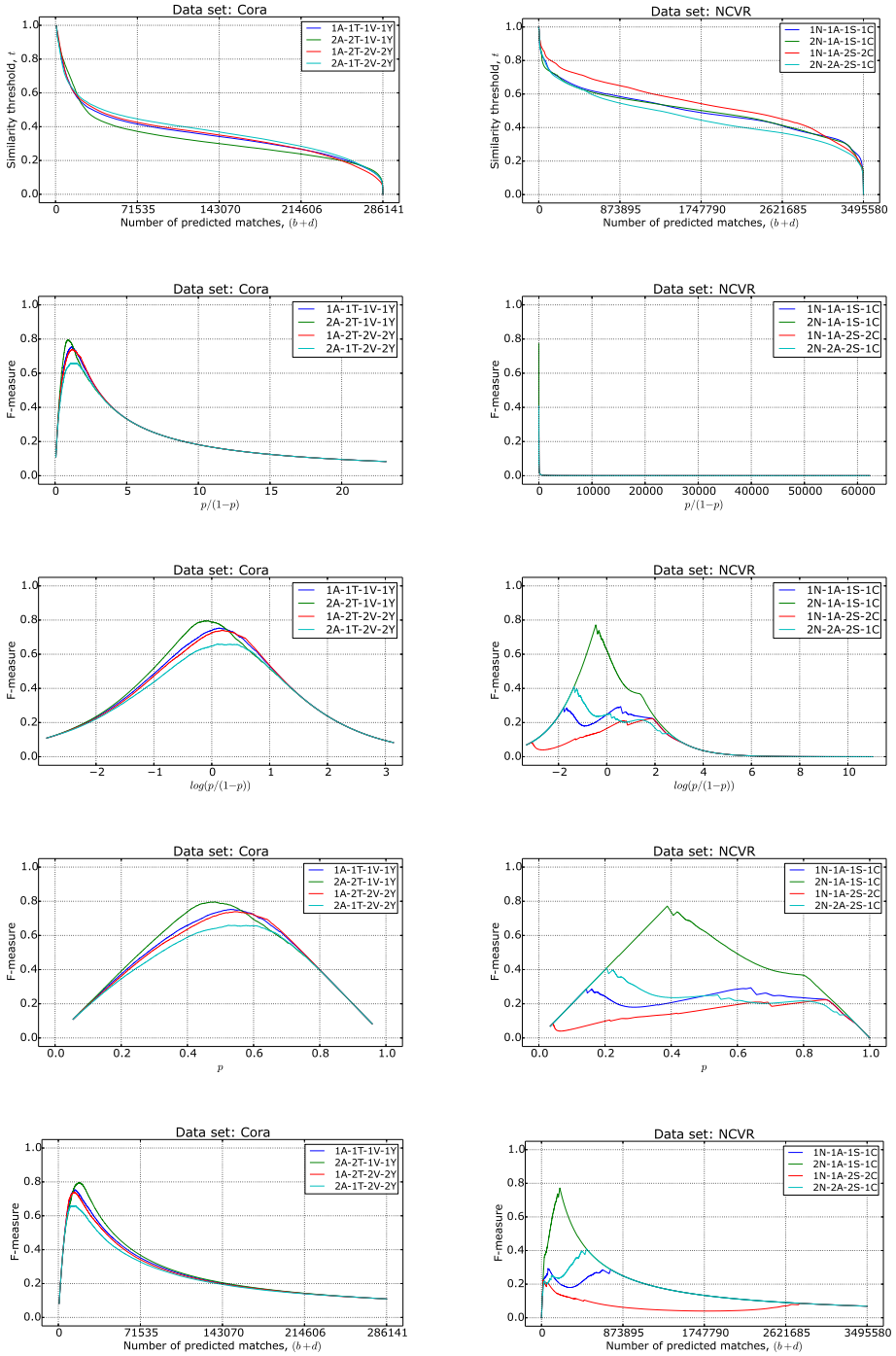

Fig. 4 Top row shows the number of predicted matches for different values of the similarity threshold; the following four rows show the transformations proposed in Section 3.

distribution of matches and non-matches over different values of the similarity threshold t (Christen et al. 2015) [6]. The age attribute specifically leads to a small number of discrete similarity values that result in a sudden significant increase in the number of false non-matches (and a corresponding significant decrease in the number of true matches) at similarities of around $t = 0.6$.

5 Discussion

To summarise, the F-measure can be used to compare linkage methods, but for a fair comparison we must use the same weights for precision and recall for all methods. To achieve this, each method must classify the same number of record pairs as matches. We can accomplish this by *different choices of the similarity threshold t for different methods* so they classify the same number of record pairs as matches.

The plots in Figure 4 show how the different linkage methods perform for the entire range of different possible weights given to precision (and, by implication, to recall). However, they naturally give no guidance on what particular choice of weights should be made. Once again, the problem is analogous to choosing the misclassification cost ratio or severity ratio in supervised classification problems.

And, also once again, the answer must depend on the particular problem and the particular researcher's or user's aims. We would expect different problems to result in different choices of p, and different researchers and users to make different choices. The important thing is that researchers and users must report the choice they made when reporting comparative performance of linkage methods.

Having said all that, it might sometimes be useful to report results corresponding to a conventional value of p. This would be analogous to reporting supervised classification comparisons using accuracy or misclassification rate, which implicitly weights the two possible kinds of misclassification equally.

In our context, one such possible conventional value might be $p = 1/2$, weighting recall and precision equally. That would mean choosing the similarity thresholds, t, of the different methods so that

$$\frac{c+d}{c+b+2d} = \frac{b+d}{c+b+2d} \tag{5}$$

for all methods. That is, for each method, choose the threshold so that the number of predicted matches equals the number of true matches.

Alternative, and more sophisticated, methods could average F-measures over a distribution of possible values for p. This would be analogous to the H-measure in supervised classification — see Hand (2009, 2010) [10,11].

6 Conclusions

In this paper we have discussed issues related to the evaluation of classification of record pairs in the context of record linkage systems. We specifically highlighted problems with the commonly used F-measure, which is calculated as the harmonic mean of precision and recall. We have shown that the F-measure is also a weighted arithmetic mean of precision and recall, with the relative importance given to

precision and recall depending upon the number of predicted matches (i.e. the linkage method (classifier) used and similarity threshold chosen).

Unless equal weights are used for the different systems being compared, this is clearly nonsensical — it means that different linkage methods are being evaluated using different measures, measures with different importance accorded to precision (and recall) when viewed from the weighted arithmetic mean perspective.

It is a fundamental tenet of performance evaluation that one must use the same instrument to evaluate competing systems: in the same way that we measure people's height using rulers with the same graduations. (We do not say: "my height, in centimetres, is larger than your height, in inches, so I am taller than you.")

The problem can be overcome by choosing the similarity thresholds for the different linkage methods so that all systems being compared classify the same number of pairs as matches. This is equivalent to choosing the same weight pairs, p and $(1-p)$, for all systems being compared.

It remains to decide what that common weight pair should be. Ideally, this should be chosen by the researcher or user for a particular problem — and this choice should be reported in any comparative study. However, it may also be useful to report a conventional standard. We suggest that one possible convention would be to weight recall and precision equally, in their simple unweighted mean.

Acknowledgements We like to thank David Hawking and Paul Thomas for their advice on the use of the F-measure in information retrieval; and Mark Elliot, Ross Gayler, Yosi Rinott, Rainer Schnell, and Dinusha Vatsalan for their comments during the development of this paper.

References

1. Christen, P.: Development and user experiences of an open source data cleaning, deduplication and record linkage system. SIGKDD Explorations **11**(1), 39–48 (2009)
2. Christen, P.: Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Springer (2012)
3. Christen, P.: A survey of indexing techniques for scalable record linkage and deduplication. IEEE Transactions on Knowledge and Data Engineering **24**(9), 1537–1555 (2012)
4. Christen, P.: Preparation of a real temporal voter data set for record linkage and duplicate detection research. Tech. rep., The Australian National University (2014)
5. Christen, P., Goiser, K.: Quality and complexity measures for data linkage and deduplication. In: F. Guillet, H. Hamilton (eds.) Quality Measures in Data Mining, *Studies in Computational Intelligence*, vol. 43, pp. 127–151. Springer (2007)
6. Christen, P., Vatsalan, D., Wang, Q.: Efficient entity resolution with adaptive and interactive training data selection. In: IEEE ICDM, pp. 727–732. Atlantic City (2015)
7. Domingo-Ferrer, J., Torra, V.: Disclosure risk assessment in statistical microdata protection via advanced record linkage. Statistics and Computing **13**(4), 343–354 (2003)
8. Getoor, L., Machanavajjhala, A.: Entity resolution: theory, practice and open challenges. VLDB Endowment **5**(12), 2018–2019 (2012)
9. Hand, D.J.: Construction and assessment of classification rules. Wiley (1997)
10. Hand, D.J.: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine learning **77**(1), 103–123 (2009)
11. Hand, D.J.: Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. Statistics in Medicine **29**(14), 1502–1510 (2010)
12. Hand, D.J.: Assessing the performance of classification methods. International Statistical Review **80**(3), 400–414 (2012)
13. Harron, K., Goldstein, H., Dibben, C.: Methodological Developments in Data Linkage. John Wiley & Sons (2015)
14. Herzog, T., Scheuren, F., Winkler, W.E.: Data quality and record linkage techniques. Springer Verlag (2007)

15. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008)
16. McCallum, A., Nigam, K., Ungar, L.H.: Efficient clustering of high-dimensional data sets with application to reference matching. In: ACM SIGKDD, pp. 169–178. Boston (2000)
17. Murray, J.S.: Probabilistic record linkage and deduplication after indexing, blocking, and filtering. Journal of Privacy and Confidentiality **7**(1), 2 (2016)
18. Naumann, F., Herschel, M.: An introduction to duplicate detection, *Synthesis Lectures on Data Management*, vol. 3. Morgan and Claypool Publishers (2010)
19. van Rijsbergen, C.: Information Retrieval. Butterworth (1979)
20. Vatsalan, D., Christen, P., Verykios, V.S.: A taxonomy of privacy-preserving record linkage techniques. Information Systems **38**(6), 946–969 (2013)
21. Winkler, W.E.: Methods for evaluating and creating data quality. Information Systems **29**(7), 531–550 (2004)