Evaluation of a Binary Semi-supervised Classification Technique for Probabilistic Record Linkage

D. Nasseh; J. Stausberg
IBE, Ludwig-Maximilians-Universität München, Munich, Germany

Methods Inf Med 2016; 55: 136–143
http://dx.doi.org/10.3414/ME14-01-0087
received: August 25, 2014
accepted: March 25, 2015
epub ahead of print: April 20, 2015

1. Introduction

Collections of medical data are growing at an increasing rate which directly results in new opportunities for research and science [1]. While many research projects focus on one specific data set others might need to combine patient data belonging to different data sources in order to contribute to their research goals [2]. The standard technique of matching patient data of different data sets to each other is commonly referred to as medical record linkage [3, 4]. This process can be divided into four steps consisting of preprocessing, indexing, matching and classification [5]. Usually, the matching process itself is performed by measuring the agreement between a selection of identifying attributes like first name, date of birth and gender of a patient within one data set to identifying patient data within another data set. Although this might be perfunctory regarded as an easy task, handling medical data has to be performed while conforming to specific laws of privacy protection which require a one way encryption of all identifying attributes [3–9]. This consequently increases the complexity of the method since the linking cannot be performed on clear text attributes any more. Thus, the just mentioned adaption is also referred to as privacy preserving record linkage and exists in different variations, most prominent, probabilistic [10] or approximate record linkage [6]. The privacy preserving variant of probabilistic record linkage utilizes frequencies of hash values of the original attribute values [10, 11] while the more recent approaches of approximative techniques approximate the string distances between the clear text values by utilizing bloom filters [6, 12]. Both techniques share, that they rate the agreement of two
compared patient records as a weight of either negative or positive value or as a value between zero and one. The higher the value the more likely that the two compared records refer to the same entity representing a true link [13].

The differentiation between true and false links is conducted by using classification techniques based on the height of a link's weight. Optimally, the final result of the matching process should be a list of exclusively true matches. Each link within such a list can be represented as a pair of data source specific IDs [13]. Practically though, a perfect classification is often unfeasible for instance due to patient records having similar attributes but not belonging to the same entities which can be mistaken as true links.

In order to minimize the amount of false matches mistaken as true links (false positives) and true matches mistaken as false links (false negatives) the classification needs to be as precise as possible. In many scenarios without data protection obligations this problem can be solved by introducing a third class, which is the class of uncertain links [13]. Links in this area can be sorted to the classes of true or false links by manually comparing all the matched attributes. This comparison easily allows identifying small errors like variants of names describing the same person (e.g. "Rob" and "Robert"). Unfortunately, this only reliably works for unprotected clear text attributes. In the context of privacy preserving record linkage, often, the only solution might be to abandon the class of uncertain links and to attempt to split the set of links into the two classes of true and false links at a specific weight value as accurate as possible [14]. While binary classification methods might be important for privacy preserving record linkage, they are in general required in all fully automated record linkage systems which base their matching results on weights representing the attribute agreement [15–17].

While there are rule based classification systems [18, 19], which usually utilize highly scenario-specific rules, supervised or unsupervised classification systems might be considered as a more unspecific approach of classifying matching data. Typical unsupervised classification techniques like clustering base their classification decision completely on the given unlabeled matching data, meaning the different weight values within the set of weights [5, 20, 21]. If matching data would be displayed as a histogram the graphic typically would display two large elevations (each representing a class) and unsupervised classification systems would usually tend to place a border splitting the classes in-between. Still, there might be more than two obvious elevations or the class border might just be too blurred to permit an easy classification.

In contrast, supervised classification systems in context of record linkage use training data consisting of two training subsets and the knowledge about the true matches between these subsets which have similar characteristics to the given original data [22, 23]. Based on the knowledge about the true matches within the training data, an optimized cut-off value on the given training set can be calculated. This cut-off-value can be used for separating the links within the original data into the classes of true and false links. While these methods might often not be as accurate as the unsupervised systems they should be unaffected by misleading gaps and data noise in the original data set, thus being more stable and reliable. While they reportedly yield better classification results than unsupervised methods, they require training data which displays a burden for automating the process [17].

While there are different approaches for classifying matching data, there is a lack of extensive comparisons of these techniques giving insights about both the compared efficiency of a technique as well as when to use a specific technique in dependence of the given scenario [5]. In this work we therefore intended comparing advanced unsupervised classification to an own implementation of semi-supervised classification optimized for automated probabilistic record linkage on a broad variety of test sets with different characteristics like size, amount of matches (overlap) and data quality.

As an unsupervised technique we planned on measuring the effectiveness of the two-step seeded nearest neighbor [15] approach by Peter Christen which has been shown to be capable of outperforming similar established techniques like the TAILOR classifier [24]. It has to be noted that the just mentioned approach should be categorized as active learning and combines aspects of both unsupervised and supervised learning [25]. Still, since it is completely dependent on the given unlabeled matching data it should be considered as an advanced unsupervised technique [15, 16].

In the following we compared it to our own model of an automatable semi-supervised system optimized on probabilistic record linkage. Semi-supervised techniques only utilize a small amount of labeled data within a large amount of unlabeled data [26]. Therefore the contribution of this work is not only the comparison of unsupervised and semi-supervised techniques in terms of probabilistic record linkage but also the introduction of a model for an easy to use fully automatable classification system.

2. Methods

2.1 Creation of Test Sets

In context of simulated tests, the use of test sets offer the opportunity to benchmark the performance of different methods. A test set within a simulation can be understood as an equivalent to a real data set within a real use case. In the field of record linkage there is a lack of test sets based on real world data. Therefore tests in this field are often performed on artificial test sets [15]. In this work tests were performed on a total of 400 artificial test sets, based on SHA-2 encrypted patient data supplied by the university hospital of the LMU in Munich consisting of a sample of about 10% (n = 46,629) of all registered patients throughout the years 2008–2012 [27].

Each of the 400 newly constructed test sets consisted of two subsets (referred to as subset A/large and subset B/small) with varying combinations of size and overlap. According to the predefined size combination each of the two subsets of a test set were filled with random copies of either 100, 1000, 10,000 or 20,000 patients of the supplied base data. An overlap was created by replacing either 5%, 25%, 50% or 75% of
placed affected attribute values by empty strings while value transformations resulted in the replacement with another value already present in the set (e.g. 'George' → 'Christian'). According to the frequency of their occurrence within different attribute groups the errors are meant to simulate different levels of data quality (Q1–Q10) within the test sets and allow to analyze the influence of data quality on classification in the later steps of this work. The label Q1 represents test sets with a very high data quality, while Q10 represents the worst level of data quality. ▶ Figure 1 summarizes the general process of the test set creation while ▶ Table 1 gives further information about the error occurrences in a test set according to a specific value of data quality (Q1–Q10). Within this table, the group of naming attributes for patient records consisted of firstname1, firstname2, firstname3, surname1, surname2, surname3 (so called fillers like ‘van’ or ‘von’) while date-related attributes included the birthdate split into day, month and year. The address consisted of only postcode and city with gender being a single letter code. Finally, it should be mentioned that the number of 400 test sets was not chosen arbitrarily but based on the combinatorial amount of test sets with a unique combination of size, overlap and data quality.

2.2 Specification of Record Linkage System

A requirement for the desired analysis was the existence of a record linkage system. Therefore an implementation of a privacy preserving probabilistic record linkage system based on the algorithm of Fellegi and Sunther with modifications as suggested by Martin Meyer has been written and implemented in Java [10, 11]. The system has extensively been tested and used in the context of a study on family based cancer in Germany [2]. The matching variables used for this work were restricted to 12 standardized and one way encrypted elements consisting of components of 'first name' (3 components), 'last name' (3 components), 'date of birth' (3 components), 'gender' (1 component) and 'address' (2 components) using standardization.
rules according to the German UNICON rule set [28].

2.3 Design of the Semi-supervised Classification Technique

The main objective of this work was to design an easy to use binary semi-supervised classification technique specified towards automated probabilistic record linkage in order to compare its performance to other advanced automated unsupervised techniques. The method is based on the creation of \( n \) training sets created with a formula set of rules \( X \) based on the original data within a test set. This means that each training set (consisting of two sub-training sets) is a direct, slightly modified copy of a given test set (consisting of its two subsets). For each training set an optimal classifier for itself can be calculated. The arithmetic mean of those single classifiers can then be used as a cut-off value for classifying the actual test set. This approach basically resembles the technique of bagging where several single, training set-based classifiers are merged into a combined classifier in order to achieve more stable classification results [22, 29]. In the following, numbers within parentheses reference to the current step illustrated within Figure 2. The value \( n \) can be based on the performance of the underlying hardware system and time limitations. The rule set \( X \) was defined as following: 1) As a first rule, subset \( A'_1 \) as well as subset \( B'_1 \) of a training set with index \( i \) are direct copies of subset \( A \) and subset \( B \) of an original test set which has to be classified in the following: 1.1) Since the overlap between the original subsets is exceedingly not known, each patient record has to be marked with a new unique identifier. 1.2) Subsequently, a fixed number \( O_i \) of random unique patients of subset \( A'_1 \) is copied into subset \( B'_1 \), while replacing former random patient records in subset \( B'_1 \) which can be understood as a new introduced overlap. 1.3) This labeled overlap can later be used to calculate the best possible classification border for the training set. In preliminary analysis concerning this technique we were experimentally testing whether and in what way the size of the overlap influences the classification quality. Therefore, we tested different classification variants with artificial overlaps of a size of 3%, 30% and 90% on all 400 test sets. The result was that the smaller the artificial overlap the better the classification quality. The reason for this is that the artificial overlap is created by replacing records in the smaller subset \( B'_1 \) with records from the larger subset \( A'_1 \). Consequently, a large overlap replaces larger proportions of the original records within subset \( B'_1 \) altering its original distribution and decreasing the training set-based classification quality. Therefore in this work we assigned a fixed value of \( O = 3\% \) of the size of \( B'_1 \). It is important to note that this is a suggested value and could probably be adjusted depending on the given test set.

When creating training sets we were also considering if it was necessary to conserve error rates of the original test set data. Analysis with variants including error rates equal to those in the original data as well as variants without any errors at all did not significantly vary in performance of classification therefore conservation of errors was not added to the final set of construction rules \( X \). It has to be noted, that ignoring the errors seems to work for probabilistic record linkage, but won’t work for approximate record linkage because the weight of perfect matching links in probabilistic record linkage varies in dependence of abundance of matching fields, while links using the approximate approach will always display a weight of 1.0 prohibiting any attempt of classification.

After the construction of a specific training set, a record linkage matching \( A'_1 \) and \( B'_1 \) has to be performed (2). The new ID within the now known new artificial overlap allows to determine which records of \( A'_1 \) represent the same records in \( B'_1 \), therefore it is possible to calculate the optimal cut-off value (max. F-measure value) for a class-split within the given training set (3). The F-measure is the harmonic mean of precision and recall and has been shown to be a good quality measure in context of record linkage scenarios [15]. Since only the links within the overlap of the training sets contribute to the amount of true or false positives, while original true positives remain unknown this method should be considered as a semi-supervised methodology [26]. Finally the average of the single classifiers (4) can be used as a classifier for the actual testing data.

2.4 Seeded Nearest Neighbor Approach (SNN)

Even though the seeded nearest neighbor approach should primarily be regarded as an active learning approach [30], its methodology is completely reliant on the given matching data and therefore should be considered as an unsupervised technique [15, 16]. Since it showed a reported relatively high classification quality it has been implemented for this work. The technique consists of two major steps.

<table>
<thead>
<tr>
<th>Type of attribute</th>
<th>Error per Quality-level (%)</th>
<th>Occurrence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1</td>
<td>Q2</td>
</tr>
<tr>
<td>Name</td>
<td>2.5</td>
<td>5.0</td>
</tr>
<tr>
<td>Date</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Address</td>
<td>5.0</td>
<td>10.0</td>
</tr>
<tr>
<td>Gender</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Structural errors were excluded from consideration at this stage of the analysis.

Downloaded from www.methods-online.com on 2017-08-10 | IP: 54.191.40.80
For personal or educational use only. No other uses without permission. All rights reserved.
maining yet still unclassified links. For this work, the seeds were created by using specific thresholds (a negative as well as a positive border). We also tried out a second variant using a formula as presented in the original publication for this technique which allows calculating a ratio for positive and negative seeds (applied on 5% of all links) but the classification results of this approach did not match the quality of the threshold variant.

After defining the set of seeds the second major step is the classification of yet unclassified links by using the k-nearest neighbor approach which iteratively sorts each link into the set of positive or negative seeds [15, 31]. While implementing the SNN we stuck closely to the description within its original publication [15]. Since the value k (amount of k-nearest neighbors) as well as the distance measure were not compulsory defined we set the value k to 3 and used standard Euclidian distance for all calculations.

### 2.5. Comparison of Classification Techniques

After the creation of 400 test sets, each consisting of a subset A and a subset B, the test sets were matched to each other by using the record linkage system as presented under 2.2. The resulting links, each with an individual weight, were classified by using the newly implemented semi-supervised methodology as well as the SNN. The performance of each method was measured by calculating the F-measure of the classification for each test set based on the given IDs (Figure 1). Due to the different combinations of characteristics within the test sets, it was possible to conduct a comparative analysis of performance of classification in dependence of the given test set specific attributes in particular data quality.

### 3. Results

For all matched links of a test set a classification with each of the given classification methods had been performed. For a comparative analysis, these results were either grouped by size of the subsets, size of overlap or data quality. While results on test

---

**Figure 2** Schematic summary of the proposed semi-supervised classification system
sets grouped by size of the subsets and size of the overlap showed that the semi-supervised method has an overall better classification performance than the SNN, the results of performance in dependence of data quality of the given test set allowed an even more differentiable view allowing further conclusions. Since there was a total amount of 400 test sets and 10 levels of data quality, each group of data quality consisted of exactly 40 test sets with the same data quality but differing in size and overlap.

Figure 3 displays the mean F-measure on test sets within each of the groups in dependence of classification method. Based on the given IDs within the test sets, which can be used to identify the true positive links, a third graph with the best possible mean classification values for each group has been added as a reference.

Both compared methods performed comparably well on test sets with a high data quality. As already mentioned in the methods of this work, in order to decrease a negative bias, we tried out different variants of the SNN, differing in the initial set of seeds. Comparing those variants the threshold based variant outperformed the formula based SNN.

It appears that in general, classification on test sets with a low data quality leads to worse classification results than on test sets with high data quality. Especially the semi-supervised technique seems not as affected by this effect as the unsupervised technique, particularly for test sets with a data quality level of Q8–Q10. The F-measure for all test sets with a quality value of Q10 for the semi-supervised technique averages at 0.964 (standard deviation (SD): 0.03) while for the SNN it averages at 0.803 (SD: 0.32).

Overall the semi-supervised system performs better, with a classification performance often close to the maximal best possible classification even at higher levels of data quality (Q1: 0.996 (SD: 0.01) for semi-supervised classification/0.993 (SD: 0.02) for the SNN).

4. Discussion and Conclusion

In the field of privacy preserving record linkage, the process of classification seems to be overshadowed by the scientific progression regarding the matching process [6, 12]. In this scenario, the gold standard approach, which is manual classification of uncertain links [13, 32], is often not possible. While human decision making based on histogram data might be a classification option [11] there are also automated binary approaches as for instance the SNN, an algorithm of the family of active learning or more general unsupervised classification, which has been shown to be capable of supplying comparably good classification results [15, 22]. The mentioned algorithm has been adapted to the field of record linkage. In comparison to standard unsupervised methods like basic clustering approaches it more robustly handles noisy data and outliers [33].

While there exist automated options for unsupervised classification, supervised classification is hard to adapt to automated systems due to the need of training data. Sariyar et al. devoted work in this area [22, 23] but techniques require a deep understanding, user knowledge (in particular in terms of configuration) and are often focusing on approximate record linkage. The objective of this paper therefore was to design an easy to use semi-supervised binary classification method, inspired by supervised techniques, with an unambiguous configuration step particularly designed for probabilistic record linkage and its comparison in performance to the SNN algorithm.

The method itself was optimized by testing and comparing different variants in a preliminary analysis with the main conclusion that every aberration of the train-

![Figure 3](https://example.com/figure3.png)

Figure 3  a) Classification quality of the semi-supervised approach, and the SNN measured by average F-measure of all test sets with the given data quality (n = 40 per quality level) in comparison to a best possible classification (optimal). b) Closer zoom to better display differences in test sets with higher data quality
ing set distribution from the distribution of the test sets might adversely impact the classification. Therefore, size as well as major components of the distributions in the created training sets are identical to the given test sets. In order to decrease the differences in distribution, a relatively small overlap (3%) has been suggested. The aim of conserving the distribution between test set and training set might raise the question of the danger of overfitting [5, 33]. Since training sets are exclusively used for one specific test set and their cut-off values are not applied on other test sets overfitting should not be an issue.

Another finding in the preliminary analysis was that a conversion of error rates from test sets into training sets does not strongly influence the classification quality. It is important to note that this only applies for probabilistic record linkage, since even perfectly matched links show variance in the height of their weight due to different frequencies in the matched attributes. Therefore, the conservation of error rates was not incorporated due to the increase in usability. It should be possible to estimate and include the error rates and therefore adapt the method to be useable for approximate record linkage by using the expectation maximization algorithm [34]. This would need additional verification through further analysis.

The SNN as well as the semi-supervised method were tested and compared on a set of 400 test sets based on real clinical data. While using artificial data has been shown to be a valid way of analysis [15], clinical data as the base of artificial test sets also includes family relations of patients and local characteristics which are abundant when using completely artificial data which can be generated by for instance using the FEBRL toolbox [35]. These relations pose a special burden for the record linkage process and are therefore interesting to maintain.

Furthermore, in comparison to similar analyses [15] the vast amount of test sets allowed the detailed analysis of behavior of classification quality in dependence of size, overlap and data quality. In general, even though the resulting F-measure numbers for the comparison analyses, especially for test sets with a good data quality might appear high, they stay within reported range-values [36]. The semi-supervised method was capable to surpass the classification quality of the SNN. While the differences of the methods on test sets with high data quality where relatively small, they increasingly diverged on test sets with lower data quality. For these test sets, the semi-supervised classification maintained a good classification quality close to the optimal classification value while the SNN was not capable to compete anymore. Another finding was that only one of the two tested variants of the SNN showed comparably good results stressing the fact, that an unambiguous configuration of uncertain factors like the set of seeds can drastically alter the classification quality. Since the semi-supervised classification system only has one slightly unambiguous factor, the choice of overlap (here 3%), the danger of a misconfiguration should be comparably low.

4.1 Limitations

The methodology is specifically designed for automated privacy preserving probabilistic record linkage. While probabilistic record linkage might still be considered as a standard approach, more modern techniques as the approximate record linkage will most probably replace probabilistic techniques within the upcoming years. Therefore, it could be an open research task to adapt the basic methodology to also work with matching data of approximate record linkage.

An open issue is the specification of the amount of records within the overlap. While the preliminary analysis pointed out, that an overlap should be relatively small, it is hard to define a value which fits to every scenario. Therefore, it has to be noted, that the value of O = 3% should not be considered as the unchallenged optimal value since in very small test sets this might lead to overlaps consisting of only a couple of links and therefore inducing a strong bias with the danger of potential increased misclassification. This is one of the reasons why the system does not only rely on the suggested cut-off value of one training set but on the mean cut-off value of several training sets [29].

While we invested high efforts into examining the best configuration for the presented semi-supervised methodology, we did not, to the same degree, examine different configurations for the SNN. Parameter tuning, like increasing the k-value, changing the size of the set of seeds or changing the used distance measure (e.g. Hamming distance instead of Euclidian distance) might increase the SNN’s performance and therefore decrease the performance differences between both algorithms.

Aside from that, the classification system is based on the optimal F-Measure value within the training sets. Although it has been expressed that the F-Measure value is generally a fitting quality measure for the scenario of record linkage [15], there might be scenarios which require other quality measure, e.g. measures which result in higher values of sensitivity. While the analysis has been performed on a large set of test sets, the 400 test sets were based on a shared pool of patient records. More comprehensive analysis could include data sets of different domains. Also, since the comparison was restricted to only two methods, an even more comprehensive analysis might include rule based and fully supervised approaches.

Acknowledgments

We would like to thank Stefan Heinisch for giving technical access to the clinical base data as well as Gerhard Meyer, the hospital’s data protection commissioner for allowing to conduct the presented analyses based on the given data.

References

4. Sauleau EA, Paumier JP, Buemi A. Medical record linkage in health information systems by approximate string matching and clustering. BMC
Evaluation of a Binary Semi-supervised Classification Technique for Probabilistic Record Linkage

D. Nasseh, J. Stausberg

Medical Informatics and Decision Making 2005; 5: 32.
Christen P. Febrl – a freely available record linkage system with a graphical user interface. HDKM’08, CRPIT, vol. 80, 2008.