A Trainable Model to Assess the Accuracy of Probabilistic Record Linkage

Robespierre Pita[1], Everton Mendonça[1], Sandra Reis[2],
Marcos Barreto[3], Spiros Denaxas[3]

[1] Computer Science Department – Federal University of Bahia (UFBA), Brazil
Email: {pierre.pita,evertonmj}@gmail.com
[2] Centre for Data and Knowledge Integration for Health (CIDACS)
Oswaldo Cruz Foundation (FIOCRUZ), Brazil
Email: ssreis02@gmail.com
[3] Farr Institute of Health Informatics Research
University College London, UK
Email: {m.barreto,s.denaxas}@ucl.ac.uk

Abstract. Record linkage (RL) is the process of identifying and linking data that relates to the same physical entity across multiple heterogeneous data sources. Deterministic linkage methods rely on the presence of a set of common uniquely identifying attributes across all sources while probabilistic approaches use non-unique attributes and calculates similarity indexes for pairs of records. A key component of record linkage is accuracy assessment, the process of manually verifying and validating matched pairs to further refine linkage parameters and increase its overall accuracy. This process however is time-consuming and impractical when applied to large administrative data sources where millions of records are being linked. Additionally, it is potentially biased as the gold standard used is often the intuition of the reviewer. In this paper, we discuss the evaluation of different self-training approaches (decision trees, naïve Bayes, logistic regression, random forest, linear support vector machines and gradient boosted trees) for assessing and refining the accuracy of probabilistic linkage. We used data sets extracted from large (more than 100 million individuals) Brazilian socioeconomic and public health care data sources. These models were evaluated using receiver operating characteristic plots, sensitivity, specificity and positive predictive values collected from a ten-fold cross-validation method. Results show that logistic regression outperforms other classifiers and enables the creation of a generalized model achieving very accurate results.

1 Introduction

Record linkage (RL) is a methodology to aggregate data from disparate data sources believed to pertain to the same entity [21]. It can be implemented using deterministic and probabilistic approaches, depending on the existence (first case) or the absence (second case) of a common set of identifier attributes in

all data sources. In both cases, these attributes are compared through some similarity function that decides if they match or not.

The number of comparisons performed by record linkage methods is represented by a quadratic function, as defined in [13]. Therefore, a RL execution using databases A with a' records and B with b' records is expected to result in a data mart D with at most $a' \times b'$ pairs.

Literature has a wide range of sequence- and set-based similarity check functions providing very accurate results. On the other hand, there are no gold standards widely assumed to assess the accuracy of probabilistic linkage, as the resulting data marts are specific to each domain and influenced by different factors, such as data quality and choice of attributes. So, manual review is frequently used in these cases, being dependent of common sense or the reviewer experience and, as such, prone to misunderstanding and subjectivity [9].

Our proposal is to use a set of supervised machine learning techniques to build a trainable model to assess the accuracy of probabilistic linkage. We aim at to eliminate manual review as it is limited by the amount of data to be revised, as well we believe it is less reliable than a computer-based solution. In order to choose the most appropriate techniques, we made experiments with decision trees, naïve Bayes, logistic regression, random forest, linear support vector machine (SVM), and gradient boosted trees.

Training data came from an ongoing Brazil-UK project in which we built a huge population-based cohort comprised by 114 million individuals receiving cash transfer support from the government. This database is probabilistically linked with several databases from the Public Health System to generate data marts for various epidemiological studies. Accuracy is assessed through established statistical metrics (sensitivity, specificity and positive predictive value) calculated during the manual review phase. So, these data marts together with their accuracy results are used to train our models. Our results show that linear SVM outperforms the other methods.

The main contribution of our proposal is a workflow to preprocess data marts obtained through probabilistic linkage and use them as training data sets for different machine learning classifiers. Scenarios where fuzzy, approximate and probabilistic decisions on matching, such as record linkage and deduplication, can benefit from this workflow to reduce or even eliminate manual review specially in big data applications.

This paper is structured as follows: Section 2 presents some related works focusing on accuracy assessment and different approach to improve it. Section 3 presents some basic concepts related to accuracy assessment and details on our data linkage scenario. Section 4 briefly describes the machine learning techniques used in this work. Section 5 presents the proposed trainable model targeted to eliminate the manual review during the probabilistic linkage of huge data sets. Our experimental results are discussed in Section 6 and some concluding remarks and future work are given in Section 7.

2 Related Work

Record linkage is a research field with significant contributions present in the literature, covering from data acquisition and quality analysis to accuracy assessment and disclosure methods (including a vast discussion on privacy). In this section, we emphasize some works presenting different approaches to validate the accuracy of probabilistic data linkage as well as the use of machine learning techniques on linkage applications.

The authors in [28] established a validation procedure based on three steps and two teams. A stratified sampling approach was adopted by team A on the first step, due to the number of 23,352 linked pairs. The team B were responsible for collecting external data to build a full admission history on the second step. This external information allowed the retrieval of a gold standard to identify the data quality and make the manual verification easier. As result, 64% of specificity and 73% of sensitivity was achieved on established gray area and 100% of accuracy on certain linked and non-linked area.

Some approaches apply machine learning to improve the pairwise classification [12, 32, 33], their validation using synthetic and real-world data achieves accuracy, precision and recall measures above 90%. The work described on [26] explores the potential use of machine learning techniques in record linkage applied to epidemiological cancer registries. The authors have used neural networks, support vector machines, decision trees and ensemble of trees to classify records. Ensemble techniques outperformed the others approaches by achieving 95% of classification rate.

Learned models are also used to scale up record linkage by using blocking schemes [6, 20]. In [30], neural networks were applied to record linkage and the results compared to a naïve Bayes classifier, measuring the accuracy and concluding they outperform Bayesian classifiers in this task.

The need of using data mining techniques for ease or eliminate manual review was pointed by [10]. A Unsupervised learning approach has been adopted to analyze a record linkage result [17]. The author established a gold standard by running a deterministic merge of the involved databases before the record linkage procedure. The transformed first name, last name, gender date of birth and a common primary key between the bases were submitted to several iterations of Expectation Maximization algorithm in order to improve the agreement of true positive pairs. The estimated review has shown results very similar to manual observed verification.

We have been involved with probabilistic data linkage and subsequent accuracy assessment for more than three years. We have discussed the implementation of our first probabilistic linkage tool in [23], followed by a deeper discussion on different ways to implement probabilistic linkage routines and their accuracy assessment in controlled (databases with known relationships) and uncontrolled scenarios [22]. These works used socioeconomic and public health data from Brazilian governmental databases. The dataset used to train our models is derived from the results of the data integration reported on these works. Our proposal differences from these works falls into two categories. The first one is

characterized by the establishment of a workflow which can be used to assess accuracy of either record linkage or deduplication procedures in a way to reduce or eliminate the manual effort of this validation process as well as the subjectivity often associated to this verification phase. In the other hand, the lack of a gold standard and external data for RL validation makes the establishment of a cutoff point more challenging, requiring supervised approaches to build models.

3 Assessing the accuracy of record linkage

Since Fellegi and Sunter [13] provided a formal foundation for record linkage, several ways to estimate match probabilities raised [31]. One way to do matching estimation is using similarity indexes capable of dealing with different kinds of data (e.g. nominal, categorical, numerical). These indexes provide a measure, which can be probability-based [11] or cost-based [16], between attributes from two or more data sets.

The attributes are assumed to be a "true match" if their measure pertains to a given interval or a "true unmatch" if their measure pertains to another interval. These intervals are delimited by upper and lower cut-off points: a similarity index above the upper cut-off point means a true positive (matched) pair of records, while an index below the lower cut-off point means a true negative (unmatched) pair of records. All pairs of records classified in between these cut-off points (the so-called "gray area") are subject to a manual review for reclassification.

Sensitivity, specificity and positive predictive values (PPV) are summary measures used to evaluate record linkage results [27]. These measures take into consideration the number of pairs classified as true positive (TP), true negatives (TN), false positives (FP), and false negatives (FN). Thus, the accuracy function is depicted by $(\textit{true pairs})/(\textit{all pairs})$

The PPV measure, calculated by the equation $TP/(TP + FP)$, brings the proportion of true positive matches against all positive predictions, representing the ability of a given method to raise positive predictions [3].

Sensitivity represents the proportion of pairs correctly identified as true positives, as depicted by equation $TP/(TP + FN)$. In other hand, specificity represents the proportion of pairs correctly identified as true negatives [1], defined by $TP/(TP + FN)$.

The accuracy assessment falls into two challenges when external data is not available to support the validation of RL results. The first is to establish a gold standard, which may use external data to validate the pairs resulted from RL. And the second refers to define a cutoff point in order to enhance the ability to find true positive pairs. Given a cutoff point, all linked pairs are separated as matched or unmatched. The expected behavior of probabilistic linkage results is to contain a significant number of matched pairs with higher similarity indexes, as well a set of unmatched pairs undoubtedly classified as such. The gray area (or dubious records) appears in situations where we must use two or more cutoff points, leading to the need of manual review or any other form of reclassification.

Probabilistic record linkage, specially in big data scenarios, lacks of gold standards as they are hard to set up considering the idiosyncrasies of each application and its data. Common scenarios do not provide additional data to reviewers do their verifications, which makes this process based on common sense, intuition and personal expertise [9]. Manual (or clerical) review is also limited by the amount of data to be revised.

In our experimental scenario, we assess the accuracy of our probabilistic linkage tool through the use of data marts generated by linking individuals (from a huge socioeconomic database) to their health outcomes (using databases from the Public Health System). These data marts are used in several epidemiological studies assessing the impact of public policies, so accuracy of integrated data is really a huge concern.

4 Machine learning algorithms

Usually, machine learning algorithms can be divided in three categories: *supervised learning*, where a training data set is used to train the classification algorithm; *unsupervised learning (or clustering)*, where the algorithm does not have a prior knowledge (labeled data) about the data and relies on similar characteristics to perform classification; and *semi-supervised learning*, where some parts of data are labeled and some are not, being a mixture of the two previous methods. Our trainable model was developed using some supervised classification methods, which are described in this section.

4.1 Decision trees

Decision trees are used to classify instances by splitting their attributes from the root to some leaf node. They use some *if-then* learned rules to provide disjunctions of conjunctions on the attribute values [19].

Let C be a number of classes and f_i be a frequency of some class in a given node. The Gini impurity, given by

$$Gini = \sum_{i=1}^{C} f_i(1 - f_i), \tag{1}$$

refers to the probability of some sample be correctly classified. The entropy, given by

$$Entropy = \sum_{i=1}^{C} -f_i \log_2(f_i), \tag{2}$$

measures the impurity within a set of examples. The most popular implementations of decision trees use either Gini or entropy impurity measures to calculate the data information gain, mostly getting similar results [25].

The information gain determines the effectiveness of some attribute to classify the training data [19]. Splitting data using this measure may reduce impurity

of samples. The information gain calculation considers some attribute A in a sample S, where Imp can be either the Gini or entropy impurity measure of S, $Values(A)$ represents all possible values of A, and S_v is the subset of S in which the attribute A has the value v [19]. So, the information gain can be obtained by

$$IG(S, A) = Imp(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Imp(S_v). \tag{3}$$

4.2 Gradient boosted trees

Gradient boosted trees (GBT) refers to iteratively train different random subsets of training data in order to build an ensemble of decision trees and minimize some loss function [14]. Lets N be the number of instances in some subsample, y_i the label of an instance i, x_i keeps the features of an instance and $F(x_i)$ brings a predicted label, for instance, i by the model. So, the equation

$$logloss = 2 - \sum_{i=1}^{N} log(1 + exp(-2y_i F(x_i))) \tag{4}$$

illustrates the log loss function used by GBT on classification problems

4.3 Random forests

Random forests combine a number of tree-structured classifiers to vote for the most popular class of an instance [7]. The training of each classifier takes an independent, identically distributed random subset of the training data to decide about the vote. This randomness often reduce over-fitting and produce competitive results on classification in comparison to other methods [7].

4.4 Naïve Bayes

The naïve Bayes assumes that a target value is the product of the probabilities of the individual attributes because their values are conditionally independent [19]. It is calculated as shown in Equation 5.

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j). \tag{5}$$

4.5 Linear support vector machine

Given a training data set with n points $(\overrightarrow{x_1}, y_1), ..., (\overrightarrow{x_n}, y_n)$, where y_1 may assume 1 or -1 values to indicate which class the point $\overrightarrow{x_1}$ belongs to, and $\overrightarrow{x_1}$ is a p-dimensional vector $\in \mathbb{R}$, the linear support vector machine (LSVM) aims to find a hyperplane that divides these points with different values of y [8].

4.6 Logistic regression

The logistic regression classifier aims to model the probability of the occurrence of an event E depending on the values of independent variables x [24]. The Equation 6

$$p(x) = Pr\{E|x\} = 1/[1 + \exp\{-\alpha - \beta' x\}] \tag{6}$$

can be used to classify a new data point x with a vector of independent variables w, being (α, β) estimated from the training data. Let z be the odds ratio of positive or negative outcome class given x and w. If $z > 0.5$, the outcome class is positive; otherwise is negative.

$$f(z) = \frac{1}{1 + e^{-z}} \tag{7}$$

4.7 Methods comparison

All these methods have different advantages and disadvantages when applied to different scenarios. Using decision trees, the user do not need to worry with data normalization as it does not highly affect the tree construction. Also, decision trees are easy to visualise, explain and manipulate, and does not require a large data set for training.

Gradient boosted trees usually have a good performance, but require a bigger time to learn because the trees are built sequentially. Usually, they are more prone to overfitting, so it is important to be careful in the pre-processing stage.

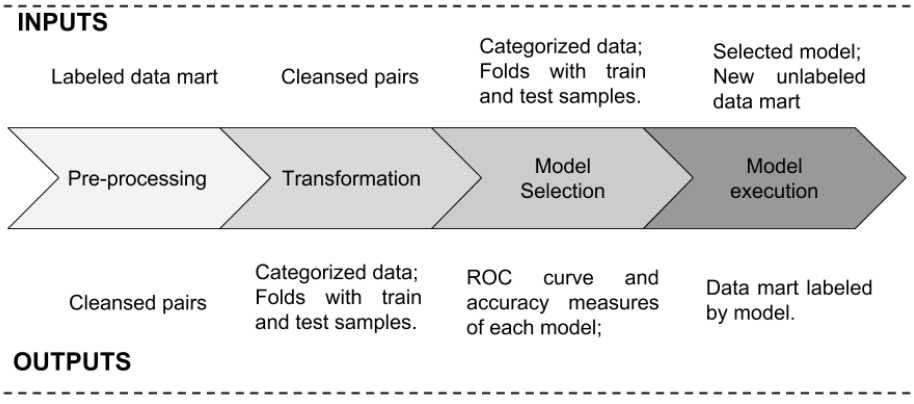
Random forests have a good performance and are more robust than single decision trees, giving more accurate results. Also, they suffer less from overfitting and can handle thousands of input variables without variable deletion. For categorical data with more than one level, random forests could be biased to the attributes with a bigger number of levels.

Naïve Bayes classifiers are fast and highly scalable. The classifier provide good results and is simple to implement, well fit with real and discrete data and is not sensitive to irrelevant features. As main disadvantage, this classifier assumes independence of features on training data. but it is not able to learn interactions between features.

Linear support vector machine (SVM) has a regularization parameter that helps the developer to avoid overfitting. This technique will not avoid the overfitting problem, that in general SVM suffer from, but help the user to optimize this value and get good results. SVMs uses kernels, so it is possible to build expert knowledge by adjusting the kernel. SVM is defined by a convex optimization problem and there are different efficient methods to deal with this, for example, the Sequential Minimal Optimization (SMO)

Logistic regression is a simple method and is very fast. Usually requires a large data set than other methods to achieve stability and work better with a single decision boundary. Also, logistic regression is less prone to overfitting.

Fig. 1. Proposed workflow to build a trainable model.



5 Proposed trainable model

The input data of trainable model must contain features that can simulate what a statistician often use to evaluate linkage results. Our methodology consists of construct a data set to show how different are the nominals and the equality of either categorical and numerical attributes used by the linkage algorithm. A categorization based on medians is made in order to assure some data balance.

Figure 1 shows the proposed pipeline to build a trainable model to accuracy assessment of probabilistic record linkage. This pipeline submits a data mart produced by the linkage tool to data cleansing, generation of a training data set to build models, evaluation and use. There is a possibility to rearrange some pre-processing, transformation and model selection settings by re-executing these steps.

5.1 Pre-processing

The pre-processing step consists of i) providing a descriptive analysis of data to select eligible common attributes within pairs and discard their missing values; ii) select attributes to be used to build the model, usually the same attributes used by the linkage algorithm; and iii) data cleansing and harmonization to guarantee that those selected attributes will have the same format.

The eligible common attributes to be used are: *name, mother name, birth date, gender* and *municipality of residence*. The attributes are chosen by their capacity of identifying an individual and their potential use by statisticians to manually verification about pairwise matching. Nominal attributes usually have a more discriminative power to determine how different two records are, followed by birth date, gender and municipality code. Converge all different formats of birth date, gender and municipality code into an unique one is an important task due to the diversity and heterogeneity of Brazilian information systems.

The approach applied to nominal attributes is to deal with special characters, double spaces, capitalization, accentuation and typos (imputation errors).

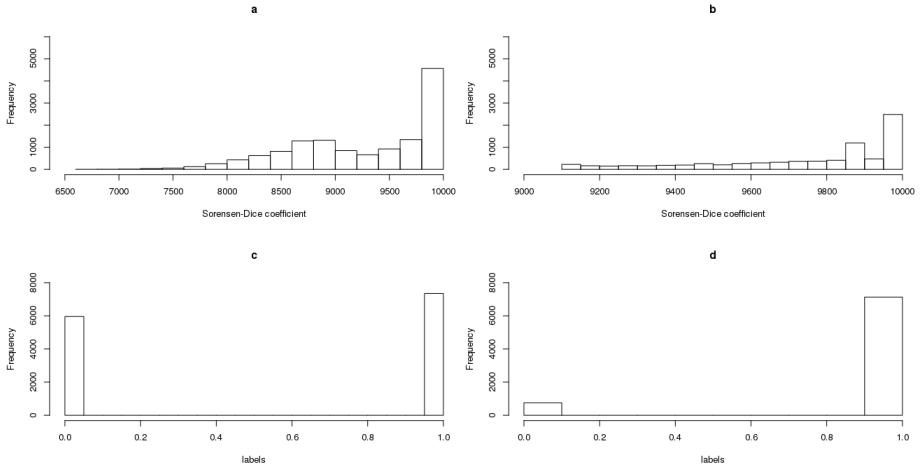
5.2 Transformation

Statisticians verify the differences between attribute values in each pair during the accuracy assessment step despite the use of the similarity values provided by the linkage algorithms. In order to simulate this verification, a data set must reflect either equality, dissimilarity or cost between linked records. Both categorical and numerical attribute types output a binary value that represents their equality. A different approach is taken with nominal values which the degree of the dissimilarity may be useful.

A Levenshtein distance metric [16] is used to calculate how much deletions and insertions need to be done to equalize two strings. In the transformation step, this metric calculates the distance between the first, the last and the whole names in linked pairs. The approach of splitting the name attribute is to allow the study of each part of the name on pair verification.

As a show of the common sense applied on accuracy assessment, a reviewer usually tolerate less errors on common names than when unknown or less popular names. To map this empirical behavior of reviewer, two new attributes brings the probability of the first names occurrence at the greater available data repository, such as socioeconomic or census databases.

Fig. 2. The graphs a and b refers to similarity index distribution on data before and after the establishment of cutoff point, respectively. As well, the c and d illustrates the difference within labels distribution after cutoff establishment.



A categorization using medians of distances (from the attribute name) and probabilities is made in order to promote data balance and prevent bias. Therefore, the transformation step is responsible for making a shallow descriptive

analysis of data before categorization. The transformation step results on 12 features representing: the similarity index; the distance between both full, first and surnames (the same approach for the name of mother), the probability of first names, equality of birthday, month, year and gender.

5.3 Model selection

The model selection phase refers to find the best classifier to our data set. One of best methods to evaluate and select classifiers is the ten-fold cross-validation [15]. The general idea of this method is to split the data set into n-folds and make n iterations setting a different fold as the model test. The remaining folds are set as training data to be used by different models and their several parameters. Accuracy measures are collected to evaluate the model at each iteration.

In addition to general accuracy, the capacity of correctly classify true positive pairs is the most important part to this work. Thus, accuracy, PPV and sensitivity become the main measures to be collected from each iteration of ten-fold cross-validation. Furthermore, the balance between specificity and sensitivity and their study by ROC curve plots [2] interpretation may be useful to model selection.

5.4 Model execution

The model execution phase allows the reuse of some evaluated method with a new input data mart. This step outputs the classification as true or false based on the selected learned model. Also, the results from this step could increase the training data after some verification effort.

A high performance processing approach can be required due to the size of the databases involved. To meet this requirement, we use the distributed implementation of classification algorithms available in the Spark MLlib [18] tool.

6 Experimental Results

To train and evaluate models, a datavset containing 13,300 RL resulting pairs of Brazilian longitudinal socioeconomic database with more than 100 million records, called *Cadastro Único* (CadUnico), with hospitalization, disease notification and mortality databases. For each pair, there is a similarity index calculated by RL algorithm and a label to determine if the pair is a true or false match. This label indicates that this pair already passed on statistician evaluation [5] and can be used to train the models.

After discard pairs with missing values and establishment of a cutoff point as 9100 of Sorensen-Dice similarity [11] the data was reduced to 7,880 pairs. The Figures 2.*b* and 2.*d* show the data balancing of both similarity index and labels. Experiments with different cutoff points obtained lower accuracy results than those showed on Figures 3 and 4.

Several runs of ML algorithms with different settings are necessary to select the best model. Accuracy estimation and ROC curves may be used to choose the best model with available training data [4, 15]. In the context of this work, the capacity of well classifies TP pairs. The Figure 3 shows the accuracy, PPV, sensibility and specificity results of the built models. This measures are described on Section 3 and their interpretation may fit to assess the performance of models. Boxplots are used in order to allow the study of results variation for each fold on cross-validation. This plots can summarize and make comparisons between groups of data by using medians, quartiles and extremes data points [29]. A good model must get uppermost boxplots with closest quartiles, which means either a low variation of results on each fold or satisfactory model generalization.

The Figure 3 shows the best results of each model. The use of entropy to split data and set the maximum depth of tree as 3 achieves the best results, showed on Figure 3.*a*. The results of naive Bayes classifier are in Figure 3.*b*. Figure 3.*c* presents logistic regression results with 1000 iterations. Random forest achieved best results by setting 1000 trees for voting, Gini impurity to split data and the maximum depth of tree as 5, as shown in Figure 3.*d*. LSVM results with 50 iterations to well fit the hyperplane are illustrated by Figure 3.*e*. Figure 3.*f* brings the gradient boosted trees results with at most depth 3 and 100 iterations in order to minimize the log loss function.

Figure 3.*c* shows that logistic regression outperforms the other models by comparing accuracy, PPV and specificity medians. Despite the better sensibility performance of LSVM, the best specificity result is achieved by logistic regression.

ROC curves allow the accuracy study by drawing the relation between true and false positive rates. The Figure 4.*a* shows the average true and false positive rates of each fold on cross-validation. The unbroken black line on Figure 4.*a* brings the average ROC curve ten-fold cross-validation. This curve shows the logistic regression superiority in comparison to others curves in terms of sensibility, either sensitivity. The variation of all logistic regression curves performance on folds are showed on Figure 3.*b*.

7 Conclusions

Accuracy assessment of RL refers to a time-consuming process that becomes impractical when huge databases are involved. This manual review may be reduced or even eliminated by using trainable models since this RL validation process can be assumed as a binary classification problem [9]. A proposed pipeline has initial steps capable of establishing a dataset with features used to build and evaluate models. Final steps allow building models by using different ML classifiers and their settings in order to evaluate and use them.

The logistic regression outperformed others classifiers using the available dataset under a ten-fold cross-validation approach. Others models may achieve better results due to new preprocessing, transformation and categorization approaches. Different results also may occur if increase or decrease of data size.

The proposed workflow is suitable to be used either in RL and deduplication scenarios, since the fuzzy, approximate and probabilistic decisions about pairs of record matching are made. However, a trainable model not always exempt a manual review of results, mainly on situations with tiny train dataset or with lower accuracy results on cross-validation. Is possible to adopt a feedback behavior of the proposed workflow, where newly submitted datamarts can increase the training dataset since this new result becomes labeled.

8 Future Work

The use of deep learning classification algorithms such as artificial neural networks with several hidden layers may achieve better model accuracy results. Also increase iterations of gradient boosted trees, random forest and LSVM to explore this algorithm. New classical and novel classifiers may be used as well to verify their performance with the proposed dataset. New attributes and dissimilarity metrics may be proposed in order to get more accurate results.

References

1. Altman, D.G., Bland, J.M.: Diagnostic tests. 1: Sensitivity and specificity. BMJ: British Medical Journal 308(6943), 1552 (1994)
2. Altman, D.G., Bland, J.M.: Diagnostic tests 3: receiver operating characteristic plots. BMJ: British Medical Journal 309(6948), 188 (1994)
3. Altman, D.G., Bland, J.M.: Statistics notes: Diagnostic tests 2: predictive values. Bmj 309(6947), 102 (1994)
4. Antonie, M.L., Zaiane, O.R., Holte, R.C.: Learning to use a learned model: A two-stage approach to classification. In: Data Mining, 2006. ICDM'06. Sixth International Conference on. pp. 33–42. IEEE (2006)
5. Barreto, M.E., Alves, A., Sena, S., Fiaccone, R.L., Amorim, L., Ichihara, M., Barreto, M.: Assessing the accuracy of probabilistic record linkage of huge brazilian healthcare databases. vol. 1, pp. 12–12. Oxford (2016)
6. Bilenko, M., Kamath, B., Mooney, R.J.: Adaptive blocking: Learning to scale up record linkage. In: Data Mining, 2006. ICDM'06. Sixth International Conference on. pp. 87–96. IEEE (2006)
7. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
8. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2(2), 121–167 (1998)
9. Christen, P., Goiser, K.: Quality and complexity measures for data linkage and deduplication. In: Quality Measures in Data Mining, pp. 127–151. Springer (2007)
10. Christen, P., et al.: Parallel techniques for high-performance record linkage (data matching). Data Mining Group, Australian National University, Epidemiology and Surveillance Branch, Project web page: http://datamining. anu. edu. au/linkage. html pp. 1–27 (2002)
11. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26(3), 297–302 (1945)
12. Elfeky, M.G., Verykios, V.S., Elmagarmid, A.K.: Tailor: A record linkage toolbox. In: Data Engineering, 2002. Proceedings. 18th International Conference on. pp. 17–28. IEEE (2002)

13. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of the American Statistical Association 64(328), 1183–1210 (1969)
14. Friedman, J.H.: Stochastic gradient boosting. Computational Statistics & Data Analysis 38(4), 367–378 (2002)
15. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. vol. 14, pp. 1137–1145. Stanford, CA (1995)
16. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710 (1966)
17. McDonald, C.J.: Analysis of a probabilistic record linkage technique without human review (2003)
18. Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., et al.: Mllib: Machine learning in apache spark. Journal of Machine Learning Research 17(34), 1–7 (2016)
19. Michalski, R.S., Carbonell, J.G., Mitchell, T.M.: Machine learning: An artificial intelligence approach. Springer Science & Business Media (2013)
20. Michelson, M., Knoblock, C.A.: Learning blocking schemes for record linkage. In: AAAI. pp. 440–445 (2006)
21. Newcombe, H.B., Kennedy, J.M., Axford, S., James, A.P.: Automatic linkage of vital records. Science 130(3381), 954–959 (1959)
22. Pinto, C., Pita, R., Melo, P., Sena, S., Barreto, M.: Correlação probabilística de bancos de dados governamentais pp. 77–88 (2015)
23. Pita, R., Pinto, C., Melo, P., Silva, M., Barreto, M., Rasella, D.: A spark-based workflow for probabilistic record linkage of healthcare data. In: EDBT/ICDT Workshops. pp. 17–26 (2015)
24. Press, S.J., Wilson, S.: Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association 73(364), 699–705 (1978)
25. Raileanu, L.E., Stoffel, K.: Theoretical comparison between the gini index and information gain criteria. Annals of Mathematics and Artificial Intelligence 41(1), 77–93 (2004)
26. Siegert, Y., Jiang, X., Krieg, V., Bartholomus, S.: Classification-based record linkage with pseudonymized data for epidemiological cancer registries. IEEE Transactions on Multimedia 18(10), 1929–1941 (Oct 2016)
27. Silveira, D.P.d., Artmann, E.: Accuracy of probabilistic record linkage applied to health databases: systematic review. Revista de saúde pública 43(5), 875–882 (2009)
28. Tromp, M., Ravelli, A., Meray, N., Reitsma, J., Bonsel, G., et al.: An efficient validation method of probabilistic record linkage including readmissions and twins. Methods of information in medicine 47(4), 356–363 (2008)
29. Williamson, D.F., Parker, R.A., Kendrick, J.S.: The box plot: a simple visual method to interpret data. Annals of internal medicine 110(11), 916–921 (1989)
30. Wilson, D.R.: Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage. In: The 2011 International Joint Conference on Neural Networks. pp. 9–14 (July 2011)
31. Winkler, W.E.: The state of record linkage and current research problems. In: Statistical Research Division, US Census Bureau. Citeseer (1999)
32. Winkler, W.E.: Methods for record linkage and bayesian networks. Tech. rep., Technical report, Statistical Research Division, US Census Bureau, Washington, DC (2002)
33. Winkler, W.E., et al.: Machine learning, information retrieval and record linkage. In: Proc Section on Survey Research Methods, American Statistical Association. pp. 20–29 (2000)

Fig. 3. Boxplots of 10-fold accuracy, PPV, sensibility and specificity measures in different ML algorithms. The results of different ML algorithms are represented by letters: a = decision trees, b = naive Bayes, c = logistic regression, d = random forest, e = linear support vector machine, f = gradient boosted trees.

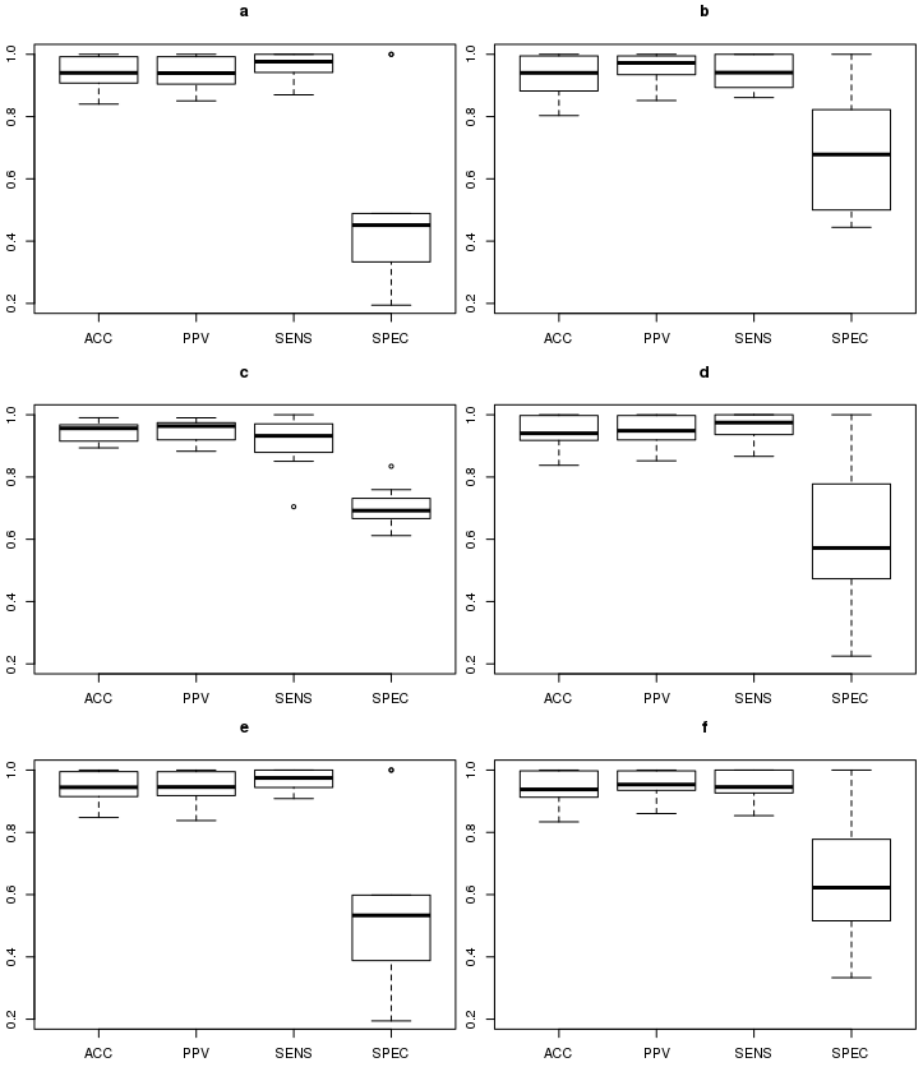


Fig. 4. ROC curves for illustrate the average true and false positive rates of 10-fold cross-validation. Different curve color represents an algorithm: dark green for decision trees, blue for naive Bayes, black for logistic regression, gray for random forests, orange for linear support vector machine and purple gradient boosted trees. Best view in color.