A Machine Learning Trainable Model to Assess the Accuracy of Probabilistic Record Linkage

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

[1] Computer Science Department,
Federal University of Bahia (UFBA), Salvador, Brazil
pierre.pita@gmail.com, evertonmj@gmail.com
[2] Centre for Data and Knowledge Integration for Health (CIDACS),
Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil
ssreis02@gmail.com
[3] Farr Institute of Health Informatics Research,
University College London, London, UK
{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 common uniquely identifying attributes across all sources while probabilistic approaches use non-unique attributes and calculates similarity indexes for pair wise comparisons. 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 effectiveness. This process however is time-consuming and impractical when applied to large administrative data sources where millions of records must be linked. Additionally, it is potentially biased as the gold standard used is often the reviewer's intuition. In this paper, we present an approach for assessing and refining the accuracy of probabilistic linkage based on different supervised machine learning methods (decision trees, naïve Bayes, logistic regression, random forest, linear support vector machines and gradient boosted trees). We used data sets extracted from huge 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 10-fold cross-validation method. Results show that logistic regression outperforms other classifiers and enables the creation of a generalized, very accurate model to validate linkage 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

L. Bellatreche and S. Chakravarthy (Eds.): DaWaK 2017, LNCS 10440, pp. 214–227, 2017.
DOI: 10.1007/978-3-319-64283-3_16

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

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" (domain-specific data) 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 were used to train our models. Our results show that SVM presents better sensitivity but logistic regression outperforms the remaining methods presenting better overall results.

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

This paper is structured as follows: Sect. 2 presents some related work focusing on accuracy assessment and different approaches to improve it. Section 3 presents some basic concepts related to accuracy assessment and details on our data linkage scenario. Section 4 describes the machine learning techniques used in this work. Section 5 presents the proposed trainable model targeted to eliminate manual review during the probabilistic linkage of huge data sets. Our experimental results are discussed in Sect. 6 and some concluding remarks and future work are given in Sect. 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] have proposed a generalizable method to validate probabilistic record linkage consisting of three phases (sample selection, data collection and data analysis) performed by different teams on a double-bind manner. They used more than 30.000 records from a newborn registry database linked against 408 records produced by pediatricians based on external data sources. The results obtained showed a high accuracy rate with less than 1% of errors.

Some approaches have applied machine learning to improve pairwise classification [12,32,33], presenting accuracy, precision and recall measures above 90% using synthetic and real-world data. The work described in [26] explores the use of machine learning techniques in linking 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 other approaches by achieving 95% of classification rate.

Learned models were 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]. An unsupervised learning approach has been adopted to analyze record linkage results [17]. The author established a gold standard by running a deterministic merge over the involved databases before the record linkage procedure. Transformed attributes (first name, last name, gender, date of birth and a common primary key between the bases) were submitted to several iterations of the Expectation Maximization algorithm in order to improve the agreement of true positive pairs. The estimated review showed results very similar to manual observed verification.

We have been involved with probabilistic data linkage and subsequent accuracy assessment for more than four 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 in this work is derived from the results reported on these previous works. Our proposal comprises 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 the subjectivity often associated to this verification phase.

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 indices capable of dealing with different kinds of data (e.g. nominal, categorical, numerical). These indices provide a measure, which can be probability-based [11] or cost-based [16], between attributes from two or more data sets.

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, while an index below the lower cut-off point means a true negative (unmatched) pair. All pairs of records classified in between these cut-off points (the "gray area") are subject to manual review for reclassification.

Sensitivity, specificity and positive predictive values (PPV) are summary measures commonly 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 usually defined as $(\textit{true pairs})/(\textit{all pairs})$.

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 the equation $TP/(TP + FN)$. In contrast, specificity represents the proportion of pairs correctly identified as true negatives [1], defined by $TN/(TN + FP)$.

Validation of accuracy in deterministic scenarios is relatively easy due to existence of common key attributes and well-known relationships between the data sources being linked. This favors the definition of gold standards or other forms of validation even if some uncertainty is present. Probabilistic data linkage faces two major challenges regarding accuracy ascertainment: the first is to establish a gold standard, which may use external data to validate linked pairs, whereas the second refers to defining cut-off points in order to enhance the ability of finding true positive and true negative pairs.

Given a cut-off 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 indices, as well a set of unmatched pairs undoubtedly classified as such. The gray area (dubious records) appears in situations where two or more cut-off points are used, leading to the need of manual review or other form of reclassification over these dubious records.

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 verification work, 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 data linkage tool through the use of data marts generated by linking individuals from a huge socioeconomic database to their health outcomes. These data marts are used in

several epidemiological studies assessing the impact of public policies, so their accuracy 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 Eq. 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 Eq. 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 Comparative Analysis

All these methods have different advantages and disadvantages when applied to different scenarios. By using decision trees, the user do not need to worry with data normalization as it does not highly affect tree construction. Also, decision trees are easier to visualise, explain and manipulate, and do 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 provides 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, being unable to learn interactions between features.

Linear support vector machine (SVM) has a regularization parameter that helps the developer to reduce the impact of overfitting and get good results. SVMs use kernels, so it is possible to build expert knowledge by adjusting these kernels. 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, ot requires a large data set than other methods to achieve stability and works better with a single decision boundary. Also, logistic regression is less prone to overfitting.

5 Proposed Trainable Model

The input data of our 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 (i) how different are the nominals, and (ii) 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.

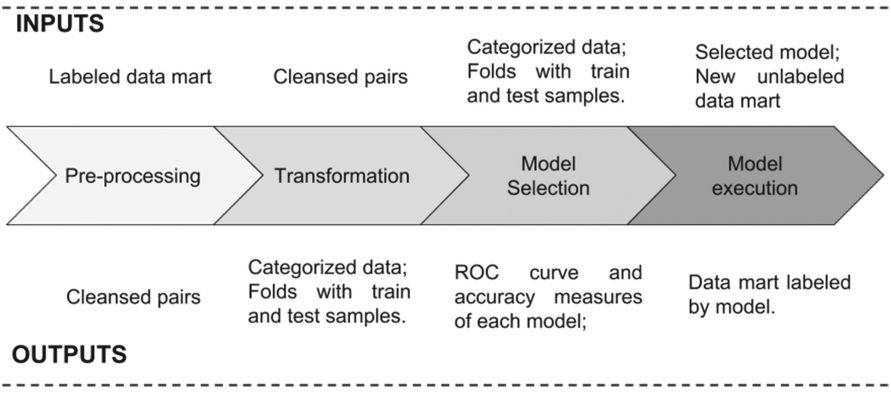

Fig. 1. Workflow for the proposed trainable model.

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's name, date of birth, gender* and *municipality of residence*. 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 date of birth, gender and municipality code. Converging all different formats of date of birth, 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 in given name and surname is to observe the influence of each part of the name on pair verification.

As an evidence of the common sense applied on accuracy assessment, a reviewer usually tolerates less errors on common names than on less popular names. To map this reviewer's empirical behavior, we use two new attributes to associate with each first name a given probability of occurence. These probabilities come from a greater data repository containing socioeconomic and census data.

A categorization using medians of distances (from the attribute name) and probabilities is made 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 comprising: the similarity index, the distance among full, given name and surname (the same approach for mother's name), the probability of first names, equality of date of birth (day, month and year) and gender.

5.3 Model Selection

The model selection phase refers to find the best classifier to our data set. One of the best methods to evaluate and select classifiers is 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 test model. 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 a cross-validation process. Furthermore, the balance between specificity and sensitivity and their interpretation by ROC curve plots [2] 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 supervised learning models, we used a data sample containing 13.300 pairs resulting from the linkage of a Brazilian longitudinal

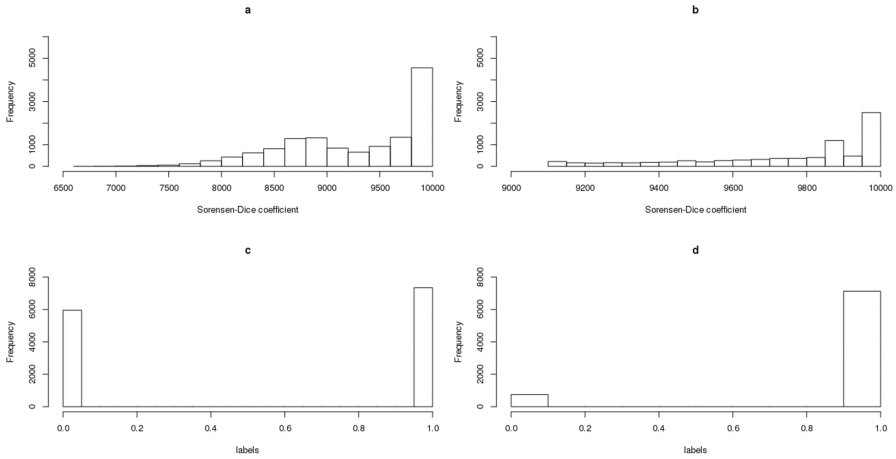

Fig. 2. Figures *a* and *b* refer to the similarity index distribution before and after the establishment of the cut-off point, respectively. As well, Figures *c* and *d* illustrate the difference with labels distribution after cut-off.

socioeconomic database with more than 100 million records (CadastroÚnico or CADU) with records from hospitalization, disease notification and mortality databases. For each pair, there is a similarity index calculated by the linkage algorithm and a label to determine if the pair is a true or false match. This label also indicates this pair already passed by a statistician evaluation [5] and can be used to train the models.

After discarding pairs with missing values and defining a cut-off point (Sorensen-Dice similarity index) [11] of 9100 (0.91%), the data sample was reduced to 7.880 pairs. This value was chosen based on several previous works and analyzes we done during these four years working on probabilistic linkage and taking into account the characteristics of our Brazilian databases.

Figures 2*b* and *d* show the data balancing of similarity index and labels. Experiments with different cut-off points obtained lower accuracy results than those showed in Figs. 3 and 4.

Several executions of machine learning 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]. Figure 3 shows the accuracy, PPV, sensibility and specificity results of tested models. These measures are described in Sect. 3 and their interpretation may serve to assess the performance of these models.

Boxplots are used to allow the study of results variation for each fold in cross-validation. These 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.

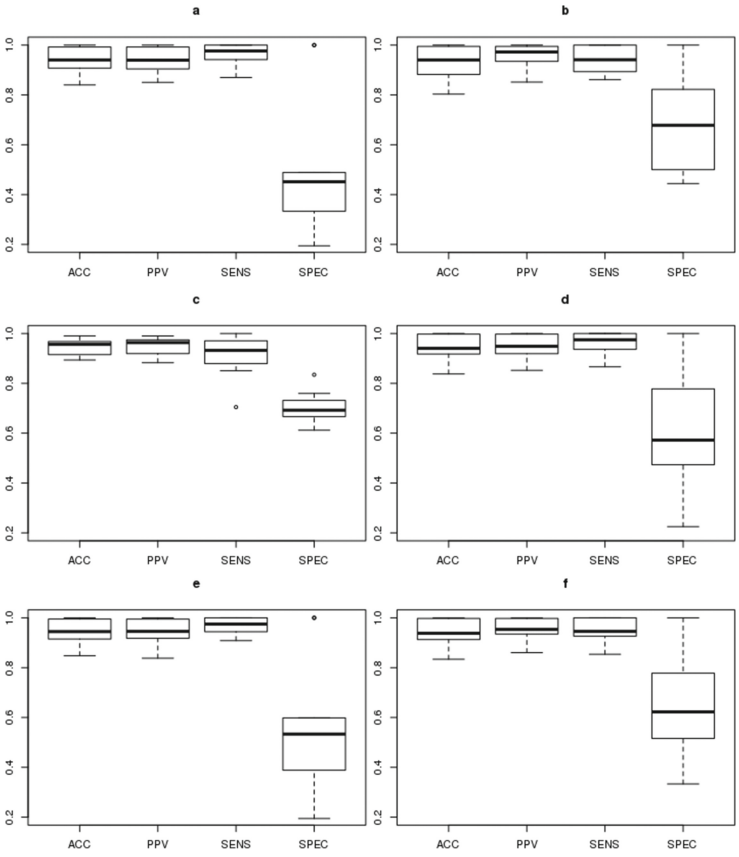

Fig. 3. Boxplots of 10-fold accuracy, PPV, sensibility and specificity measures in different machine learning algorithms: a = decision trees, b = naïve Bayes, c = logistic regression, d = random forest, e = linear support vector machine, f = gradient boosted trees.

Figure 3 shows the best results of each model. The use of entropy to split data and set the maximum depth of trees as 3 achieves the best results, showed in Fig. 3*a*. Results of naïve Bayes classifier are showed in Fig. 3*b*. Figure 3*c* presents logistic regression results with 1.000 iterations. Random forest achieved best results by setting 1.000 trees for voting, Gini impurity to split data and the maximum depth of tree as 5, as shown in Fig. 3*d*. LSVM results with 50 iterations to well fit the hyperplane are illustrated in Fig. 3*e*. Figure 3*f* brings the results for gradient boosted trees with a depth of at most 3 and 100 iterations 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.

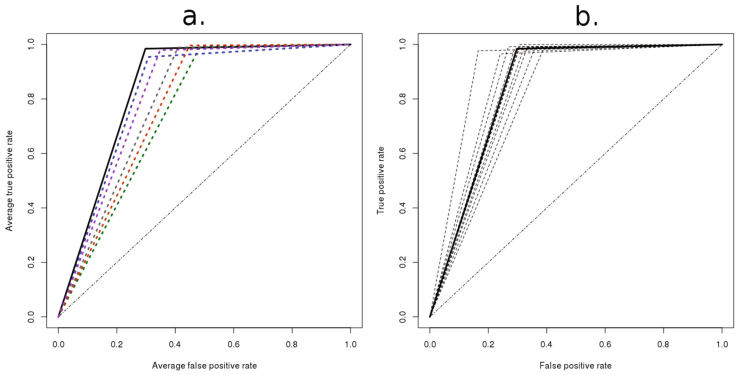

Fig. 4. ROC curves depicting the average true and false positive rates of 10-fold cross-validation. Different curve color represents an algorithm: dark green for decision trees, blue for naive Bayes, black for logistic regression, gray for random forests, orange for linear support vector machine and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

ROC curves allow the accuracy study by drawing the relation between true and false positive rates. Figure 4*a* shows the average true and false positive rates of each fold in cross-validation. The unbroken black line in Fig. 4*a* brings the average ROC curve to the 10-fold cross-validation. This curve shows the logistic regression superiority in comparison to other curves in terms of sensibility and sensitivity. The performance variation of all logistic regression curves on folds are showed in Fig. 3*b*.

7 Conclusions and Future Work

Accuracy assessment of record linkage 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 validation process can be assumed as a binary classification problem [9]. The proposed pipeline has initial steps capable of establishing a dataset with features used to build and evaluate models. The final steps allow building models by using different machine learning classifiers and their settings in order to evaluate and use them to validate new data marts.

The logistic regression outperformed others classifiers using the available dataset under a 10-fold cross-validation approach. Other models may achieve better results due to new preprocessing, transformation and categorization approaches. Different results may also occur depending on the increase or decrease of data size.

The proposed workflow is suitable to be used either in record linkage or deduplication scenarios where fuzzy, approximate and probabilistic decisions about pairs of record matching should be made. However, a trainable model could not

always eliminate the manual review, mainly in situations with tiny train data sets or with lower accuracy results from cross-validation. It is possible to adopt a feedback behavior of the proposed workflow, where newly submitted data marts can increase the training data set since this new result becomes labeled.

The use of deep learning classification algorithms such as artificial neural networks with several hidden layers may achieve better model accuracy results. Increasing iterations of gradient boosted trees, random forest and SVM can also provide better results. New classical and novel classifiers may be used to verify their performance within the proposed pipeline. 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 Br. Med. J. **308**(6943), 1552 (1994)
2. Altman, D.G., Bland, J.M.: Diagnostic tests 3: receiver operating characteristic plots. BMJ Br. Med. J. **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: Sixth International Conference on Data Mining, ICDM 2006, 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, p. 12. Oxford (2016)
6. Bilenko, M., Kamath, B., Mooney, R.J.: Adaptive blocking: Learning to scale up record linkage. In: Sixth International Conference on Data Mining, ICDM 2006, pp. 87–96. IEEE (2006)
7. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
8. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**(2), 121–167 (1998)
9. Christen, P., Goiser, K.: Quality and complexity measures for data linkage and deduplication. In: Guillet, F.J., Hamilton, H.J. (eds.) Quality Measures in Data Mining, pp. 127–151. Springer, Heidelberg (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, pp. 1-27 (2002). Project web page: http://datamining.anu.edu.au/linkage.html
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: 18th International Conference on Data Engineering, 2002, Proceedings, pp. 17–28. IEEE (2002)
13. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. J. Am. Stat. Assoc. **64**(328), 1183–1210 (1969)
14. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **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. Sov. Phys. Dokl. **10**, 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. J. Mach. Learn. Res. **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, Heidelberg (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. J. Am. Stat. Assoc. **73**(364), 699–705 (1978)
25. Raileanu, L.E., Stoffel, K.: Theoretical comparison between the gini index and information gain criteria. Ann. Math. Artif. Intell. **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 Trans. Multimed. **18**(10), 1929–1941 (2016)
27. da Silveira, D.P., Artmann, E.: Accuracy of probabilistic record linkage applied to health databases: systematic review. Rev. 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 Inf. Med. **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. Ann. Intern. Med. **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. 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: Proceedings of Section on Survey Research Methods, American Statistical Association, pp. 20–29 (2000)