Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210823

Distribution-Based Similarity Measures Applied to Laboratory Results Matching

Martin COURTOIS[a,1], Alexandre FILIOT[a] and Gregoire FICHEUR[a,b]
[a] *CHU Lille, INCLUDE: Integration Center of the Lille University hospital for Data Exploration, F-59000, Lille, France*
[b] *Univ. Lille, CHU Lille, ULR 2694 - METRICS, Public health dept, F-59000, Lille, France*

Abstract. The use of international laboratory terminologies inside hospital information systems is required to conduct data reuse analyses through inter-hospital databases. While most terminology matching techniques performing semantic interoperability are language-based, another strategy is to use distribution matching that performs terms matching based on the statistical similarity. In this work, our objective is to design and assess a structured framework to perform distribution matching on concepts described by continuous variables. We propose a framework that combines distribution matching and machine learning techniques. Using a training sample consisting of correct and incorrect correspondences between different terminologies, a match probability score is built. For each term, best candidates are returned and sorted in decreasing order using the probability given by the model. Searching 101 terms from Lille University Hospital among the same list of concepts in MIMIC-III, the model returned the correct match in the top 5 candidates for 96 of them (95%). Using this open-source framework with a top-k suggestions system could make the expert validation of terminologies alignment easier.

Keywords. ontology matching, health informatics, probability distribution, probability metrics

1. Introduction

The use of international laboratory terminologies (e.g. LOINC) inside hospital information systems is required to conduct data reuse analyses through inter-hospital databases. Well-known strategies from the ontology matching field of computer science have already been proposed, like string-based or language-based models [1], to standardize local terminologies toward an international reference.

Distribution matching is an *instance-based matching technique* [2] that performs terms matching based on the statistical similarity of their respective sets of instances. This technique is an extension of two-sample hypothesis testing to compare distributions.

Recent developments on *distribution matching* include a comprehensive evaluation of schema matching techniques [3], where a Wasserstein distance-based *distribution matching* algorithm [4] competes with state-of-the-art *schema-based matching*

[1] Corresponding Author, Martin COURTOIS, CHU Lille, INCLUDE, Institut Coeur-Poumon Boulevard du Professeur Jules Leclercq, 59000 Lille, France; E-mail: martin.courtois@chru-lille.fr, martin.courtois@protonmail.com.

techniques. To the best of our knowledge, *distribution matching* has only been applied once to healthcare terminologies alignment through the use of expert data preprocessing [5].

In this work, we design and assess a structured, reproducible framework to perform *distribution matching* in a real-world setting. This framework aims at aligning laboratory terminologies described by continuous variables without any required preprocessing while using *f*-divergences as similarity measures (e.g. the Hellinger distance).

2. Methods

2.1. The Distribution Matching Framework

In this section we propose a generic framework for distribution matching, applied on two experimental scenarios. This framework is based on a machine learning classification model which is trained on features describing the similarity of distributions. This model outputs the probability of a match for a given pair of terms from two distinct terminologies. Our framework consists in the following steps:

1. The set of all possible pairs is defined by the cartesian product of two terminologies. The equivalence (hence disjointness) of those pairs is defined manually.
2. For each pair, we compute distribution-based features from the measurements: (a) the Kolmogorov-Smirnov statistic, (b) the Hellinger distance, (c) the absolute difference of the means and (d) the absolute difference of the standard deviations.
3. We train and fine-tune a random forest classifier on the previous set of features.
4. For each pair, we use the model's probability of correct match to predict equivalence or disjointness. The model's predictive ability is then assessed using the following metrics: *Precision*, *Recall*, F_1 or *Precision-Recall AUC*. In practice, we are interested in producing mappings from one source to another. In this case, we can also use the *Mapping Score*, which measures the ability of the matching technique to provide a correct match in the top 5 ranked candidates.

2.2. Datasets and Scenarios

Our framework is assessed using data from the freely accessible database *MIMIC-III* and from the Lille University Hospital's laboratory terminology, the reuse of which for medical research purposes has been authorized by the CNIL in 2019 (reference number 2202081). All models were fine-tuned using 5-fold cross-validation on a training dataset and evaluated on a testing dataset. The training and testing datasets both contain the complete terminologies (i.e. the whole terms). Splitting is performed at random on the series of measurements according to a 70% (train) / 30% (test) ratio. We propose two experimental scenarios to assess the model's behaviour :

Case 1: We worked on a subset of 54 terms from the *MIMIC-III* database and 101 terms from the Lille University Hospital's laboratory terminology. These two sets share the same exact concept domain. A reference alignment was manually produced by a biologist with expertise in laboratory terminology. This alignment is composed of 5340 disjoint and 114 equivalent pairs for a total of 5454 pairs. In this use case, we computed the *Mapping Score* from the Lille University Hospital terminology to *MIMIC-III*, in addition

to the *Precision-Recall AUC*. We used separate univariate logistic regressions to perform classification based on single features (e.g. KS statistic). A random forest classifier (our proposed model) was used to classify pairs through the combination of multiple features. *Case 2:* We propose a complementary analysis which is a positive control. In this use case, we tried to match *MIMIC-III* with itself on a subset of 195 terms. The reference alignment was produced automatically using terms' ids to identify the correct pairs.

3. Results

Our *distribution matching* framework is implemented in *Python* version 3.8 and is available under Apache-2.0 License at https://github.com/mcrts/dmatch. The developed package provides through a CLI the required tools to extract and prepare the data in order to train and evaluate a decision model between two given terminologies.

3.1. Experimental Results

Table 1 displays the evaluation metrics computed on the testing dataset from Lille University Hospital to *MIMIC-III* terminologies as part of use case 1. We provide *PrecisionRecall AUC* and the *Mapping Score* along with 95% confidence intervals for each model. Figure 1 shows the features' importance derived from the random forest classifier. Those are computed as the average sums of impurity decrease within each tree.

Table 1. Evaluation metrics of the decision models on the testing dataset (use case 1)

Model	*Precision-Recall AUC*	*Mapping Score*
Kolmogorov-Smirnov (KS) statistic	0.63 [0.53, 0.71]	0.97 [0.92, 0.99]
Hellinger distance	0.51 [0.41, 0.61]	0.81 [0.72, 0.88]
L^1 norm of the means	0.18 [0.13, 0.24]	0.82 [0.73, 0.89]
L^1 norm of the standard deviations	0.08 [0.05, 0.12]	0.46 [0.36, 0.56]
Our model (random forest)	0.67 [0.58, 0.77]	0.95 [0.89, 0.98]

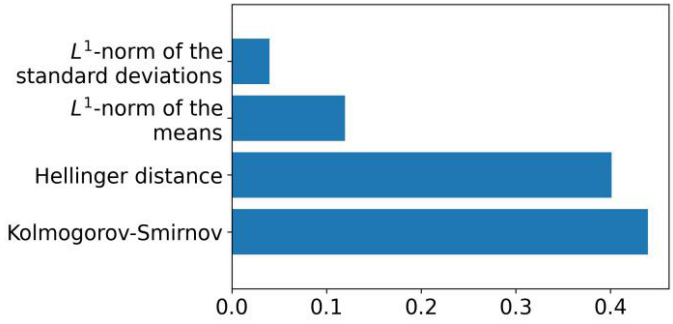

Figure 1. Random forest features' importance (use case 1)

3.2. Positive Control Use Case

When trained for the specific task of matching the *MIMIC-III* terminology with itself, the selected random forest model reaches a *Precision-Recall AUC* of 0.96[0.94,0.98] on testing pairs (use case 2). Table 2 shows the five model's suggestions with highest probability for the case of Monocytes cell count in cerebrospinal fluid (LOINC 26486-1).

Table 2. Top 5 candidates for Monocytes cells count in cerebrospinal fluid 26486-1

Laboratory Term in MIMIC-III	Nature	Probability
51120 — Monocytes — Ascites — 26488-7	False	0.988
51355 — Monocytes — CSF — 26486-1	True	0.912
50801 — Alveolar-arterial Gradient — Blood — 19991-9	False	0.000
51130 — Absolute CD3 Count — Blood — 8124-0	False	0.000
51332 — Absolute CD8 Count — Blood — 8138-0	False	0.000

4. Discussion

In this work, we applied distribution analysis to match laboratory terminologies between hospitals. The objective was to explore the use of distribution-based similarity measures for terminology matching, implement and benchmark this technique against uncurated laboratory data from the *MIMIC-III* database and the Lille University Hospital. The selected model was able to give the correct correspondence among the 5 best candidates for 95% of the 101 terms considered. As illustrated by the overall *PrecisionRecall AUC* and features' importance, distribution-based similarity measures such as the KS statistic and the Hellinger distance strongly improve the performance of the decision model compared to the absolute difference of the means or standard deviations.

A second use case (positive control) consisted in matching the *MIMIC-III* dataset against itself. The model built as part of this use case gave near perfect results which illustrates the general feasibility of our framework.

4.1. Methodological Issues

As opposed to conventional language-based matching techniques, *distribution matching* does not rely on the quality and richness of terminologies. Indeed, it showed to be resilient to data anomaly when tested on uncurated datasets. However, in its current state, our framework remains sensitive to mismatching unit systems between data sources.

In practice, the Hellinger distance relies on kernel density estimates which are sensitive to ill-behaving data sample and requires cpu intensive numerical integration. In spite of this limitations, our model still remains accurate thanks to the combination of other distribution-based features using ensemble learning. At last, our framework yet supports only univariate distribution of continuous variables.

4.2. Perspectives

In this work, we focused on the Kolmogorov-Smirnov statistic and the Hellinger distance. Other distribution-based similarity measures can also be used [6], especially the Integral Probability Metrics for which efficient computation techniques exist [7].

To further evaluate our *distribution matching* framework, we intend to benchmark it against regular language-based technique using only publicly available data such as AmsterdamUMCdb [8]. A composite model combining language and distribution analysis could then be trained to reach better performances. In particular, we believe that using top-*k* suggestions could make the expert validation of terminologies alignment easier. Through the alignment of *MIMIC-III* terminology with itself, our framework can be used for assessing the quality and consistency of a single terminology. Thus, we believe that such a tool could be used to detect terminology's anomalies, e.g. the modification of the identifier of a concept over time (especially for local terminologies).

As part of the operational setting of hospital data processing, a single concept usually has different identifiers in a single terminology for each laboratory or production site. This issue could also be addressed by the proposed framework.

5. Conclusion

In this study, we proposed a framework that combines *distribution matching* and machine learning techniques for terminology matching in a clinical setting. We trained and evaluated an algorithm on two scenarios and identified operational use cases. Finally, we provided a frame of reference that will pave the way for future improvements.

References

[1] Khan AN, Griffith SP, Moore C, Russell D, Rosario AC, Bertolli J. Standardizing laboratory data by mapping to LOINC. Journal of the American Medical Informatics Association. 2006;13(3):353-5.
[2] Euzenat J, Shvaiko P. Ontology matching. 2nd ed. Heidelberg (DE): Springer-Verlag; 2013.
[3] Koutras C, Siachamis G, Ionescu A, Psarakis K, Brons J, Fragkoulis M, et al. Valentine: Evaluating Matching Techniques for Dataset Discovery. In: 2021 IEEE 37th International Conference on Data Engineering (ICDE); 2021. p. 468-79.
[4] Zhang M, Hadjieleftheriou M, Ooi BC, Procopiuc CM, Srivastava D. Automatic Discovery of Attributes in Relational Databases. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. SIGMOD '11. New York, NY, USA: Association for Computing Machinery; 2011. p.109–120. Available from: https://doi.org/10.1145/1989323.1989336.
[5] Ficheur G, Chazard E, Schaffar A, Genty M, Beuscart R. Interoperability of medical databases: construction of mapping between hospitals laboratory results assisted by automated comparison of their distributions. In: AMIA Annual Symposium Proceedings. vol. 2011. American Medical Informatics Association; 2011. p. 392.
[6] Cha SH. Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions. Int J Math Model Meth Appl Sci. 2007 01;1.
[7] Sriperumbudur BK, Gretton A, Fukumizu K, Scholkopf B, Lanckriet G. Hilbert space embeddings and¨ metrics on probability measures. The Journal of Machine Learning Research. 2010;11:1517-61.
[8] Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration. Critical Care Medicine. 2021 Feb;Publish Ahead of Print. Available from: https://doi.org/10.1097/ccm.0000000000004916.