Impact of the Discretization of VOCs for Cancer Prediction Using a Multi-Objective Algorithm

Sara Tari[1(✉)], Lucien Mousin[1,2], Julie Jacques[1,2], Marie-Eleonore Kessaci[1], and Laetitia Jourdan[1]

[1] University of Lille, CNRS, UMR 9189 CRIStAL, 59650 Villeneuve d'Ascq, France
{sara.tari,marie-eleonore.kessaci,laetitia.jourdan}@univ-lille.fr
[2] Lille Catholic University, Faculté de Gestion, Economie et Sciences, Lille, France
{lucien.mousin,julie.jacques}@univ-catholille.fr

Abstract. Volatile organic compounds (VOCs) are continuous medical data regularly studied to perform non-invasive diagnosis of diseases using machine learning tasks for example. The project PATHACOV aims to use VOCs in order to predict invasive diseases such as lung cancer. In this context, we propose to use a multi-objective modeling for the partial supervised classification problem and the MOCA-I algorithm specifically designed to solve these problems for discrete data, to perform the prediction. In this paper, we apply various discretization techniques on VOCs data, and we analyze their impact on the performance results of MOCA-I. The experiments show that the discretization of the VOCs strongly impacts the classification task and has to be carefully chosen according to the evaluation criterion.

Keywords: Supervised classification · Medical data · Multi-objective optimization

1 Introduction

Human bodies emit a wide range of volatile organic compounds (VOCs), some of which are odorous. The composition of VOCs produced by a given individual corresponds to a unique signature odor. Age, sex, diet are among many factors that can influence this unique fingerprint, as well as diseases. These modifications often result in smell changes and explain what allowed Hippocrates to report changes related to the presence of certain diseases in the smell of urine and sputum. Nowadays, the composition of VOCs produced by individuals is regularly studied as a non-invasive way to detect pathologies [5,7,8]. The project PATHACOV[1] aim at designing a classifier based on VOCs data in order to predict invasive diseases, with a major focus on lung cancer. Thus, we propose to

[1] This project is funded by the Interreg France-Wallonie-Vlaanderen program, with the support of the European Regional Development Fund see www.pathacov-project.com for more information.

I. S. Kotsireas and P. M. Pardalos (Eds.): LION 14 2020, LNCS 12096, pp. 151–157, 2020.
https://doi.org/10.1007/978-3-030-53552-0_16

use an approach based on the Pittsburgh representation and where the classification task is modeled as a multi-objective optimization problem. The medical datasets have specific characteristics; in particular, the number of attributes is significantly higher than the number of individuals, and the classes are regularly imbalanced. Most frequent disease like diabetes only occurs on less than 6% of the population. These characteristics strongly impact on the performance of classification techniques. Therefore, the algorithm MOCA-I (Multi-Objective Classification Algorithm for Imbalanced data) [3], designed for a multi-objective modeling and these types of characteristics, has been chosen to identify the relevant VOCs. However, MOCA-I requires discrete attributes, while VOCs are continuous data.

This paper presents our resolution approach for the detection of diseases using VOCs and an experimental study where various discretization techniques and their impact on the performance of MOCA-I to produce good models are analyzed. The experiments are conducted on three different medical datasets with VOCs.

The outline of the paper is as follows. Section 2 presents the proposed approach and various data discretization techniques. Section 3 describes the datasets and the experimental protocol before giving and analyzing the results. Finally, Sect. 4 provides a discussion about this study and points out future work.

2 Proposed Resolution Approach

Bronchopulmonary cancer is often discovered late. The objective of the PATHACOV project is to detect it earlier by non-invasive means with a low-cost breath test, by measuring exhaled VOCs. For each individual, we can measure the VOCs produced and their quantities. They may vary significantly from an individual to another. Moreover, none of the individuals emit all the VOCs present in the dataset. This task can be seen as a supervised partial classification problem, where we want to identify which VOCs can predict Bronchopulmonary cancer.

2.1 Description

This problem can be modelized as a multi-objective optimization problem. Since the VOCs profile may vary from an individual to another, we opted for a Pittsburgh modelization, where each solution is a ruleset. Hence, several profiles can fit into several rules. Moreover, Pittsburgh is a white box modelization, which means it is compatible with November 2018 CCNE[2] (French National Consultative Ethics Committee)' recommendations about AI and health, suggesting to use AI approaches that the care team can criticize or challenge.

For this problem, three objectives are considered. The *sensitivity* – to maximize – will measure the ability of the model to detect a high proportion of

[2] https://www.ccne-ethique.fr/en/.

patients with the disease. The *confidence* – to maximize – will measure if the predicted patients are correctly identified. Moreover, *sensitivity* and *confidence* are two classical machine learning complementary metrics that are adapted to deal with imbalanced and medical data [6]. We also want to minimize the number of VOCs used in each model: this will generate models easier to understand.

We will use the MOCA-I (multi-objective classification algorithm for imbalanced data) algorithm, which implements the preceding modelization. It uses a multi-objective local search (MOLS) to tackle the resulting problem. MOCA-I was initially developed for handling discrete medical data. Thus, each VOC amount will be discretized, and the objective of this paper is to determine which is the impact of discretization on the cancer prediction. Since a classification task generates only one model and MOCA-I produces a Pareto set of equivalent solutions, the solution of best *G-mean* is selected among this set.

2.2 Data Discretization Techniques

In this work, we consider nine discretization techniques, that are briefly described in Table 1, following the taxonomy of [2].

Table 1. Description of discretization techniques.

Method	Static	Supervised	Separation	Global	Direct	Measure
10-bin	Yes	No	Yes	Yes	Yes	Bin.
1R	Yes	Yes	Yes	Yes	Yes	Bin.
CAIM	Yes	Yes	Yes	Yes	No	Stat.
Chi2	Yes	Yes	No	Yes	No	Stat.
ChiMerge	Yes	Yes	No	Yes	No	Stat.
Fayyad	Yes	Yes	Yes	No	No	Info.
FUSINTER	Yes	Yes	No	Yes	No	Info.
ID3	No	Yes	Yes	No	No	Info.
Zeta	Yes	Yes	Yes	Yes	Yes	Stat.

Following this taxonomy, a discretization technique can be *static* or *dynamic*, depending on when it is applied respectively before or during the learning algorithm. A *supervised* method takes into account the class to construct the intervals. For the *separation* approach, a single initial interval is produced and is then progressively split into several intervals. The opposite approach is *fusion*, where many intervals are produced and then merged. A *global* method may use the entirety of the available data for the discretization process, whereas a *local* one only uses a subset of the data. *Direct* approaches define a single interval at each iteration, while incremental approaches create many intervals at each

step. The evaluation measure is used to select the best solution produced by the discretization technique.

In the following, we will test these techniques to discretize VOCs data in our resolution approach.

3 Experiments

This section presents the datasets and the experimental protocol of our approach. Then the results of these experiments are given and an analysis is drawn.

3.1 Datasets

In this study, we use three medical datasets with VOCs (see Table 2). The datasets T3 and T4 have been provided by our partners of the PATHACOV project and come from dialysis patients while P1 has been taken from the literature [4]. Note that T3 and T4 contain the VOCs of respectively 36 and 37 patients before and after dialysis, meaning that a given individual provides two samples (a positive one and a negative one) and that the extraction of biomarkers is probably easier to perform on these datasets.

Table 2. Description of real datasets resulting from patients samples.

Name	Diagnosis	#individuals	#positive	#attributes
T3	Dialysis	72	36	346
T4	Dialysis	74	37	341
P1	Prostate cancer	103	59	137

3.2 Experimental Protocol

The purpose of this work is to predict a class. Since we have only three datasets, we use a 5-fold cross-validation protocol to limit *overfitting* as follows. Each dataset is separated in five same-size folds, then four folds are combined into a training set, while the remaining one corresponds to the test set. This process is repeated for each fold's combinations and creates five training sets associated with 5 test sets. For each discretization method, we conduct 6 independent runs of MOCA-I on each training set, leading to 30 runs per dataset.

We used the software KEEL [1] to discretize the datasets. Note that in order to reduce the bias when assessing the efficiency of the discretization methods, we limit the risk to overfit the data by discretizing each training set independently.

MOCA-I parameters correspond to the default parameters proposed by [3]: initial population of 100 solutions, 10 rules maximum per ruleset, a maximal archive size of 500. At each iteration, the multi-objective local search under

consideration selects one solution in the archive and explores the whole neighborhood of this solution. Note that, the non-dominated neighbors are considered, which explains the use of a bounded archive.

We compare the effect of the discretization methods according to four machine learning metrics: *sensitivity*, *specificity*, *geometric mean* (G-mean), and Matthew's correlation coefficient (MCC). MCC is comprised between -1 and 1, where 1 corresponds to the best performance and 0 to the theoretical performance of a random classifier. The other metrics' values are comprised between 0 and 1, where 1 corresponds to the highest performance and 0.5 to a performance that is not better than a random classifier.

3.3 Results

Table 3 presents the ranks of the nine discretization techniques according to the four considered measures (Sensitivity, Specificity, G-Mean and MCC) for each dataset. Bold types means that the discretization techniques are statistically equivalent according to the statistical test of Friedman.

Table 3. Ranking of the discretization methods in function of the average sensitivity (top-left), specificity (top-right), G-mean (bottom-left) and MCC (bottom-right).

Sensitivity			Specificity		
T3	T4	P1	T3	T4	P1
#1 Chi2	**#1 ID3**	**#1 Fayyad**	**#1 Chi2**	**#1 ID3**	**#1 ID3**
#2 10bin	**#2 CAIM**	#2 Fusinter	**#2 ID3**	**#2 10bin**	**#2 Fusinter**
#3 Zeta	**#3 1R**	#3 Chi2	**#3 1R**	**#2 Fayyad**	**#3 1R**
#3 Fusinter	**#4 10bin**	#4 CAIM	**#3 Fusinter**	**#3 1R**	#4 CAIM
#4 ID3	**#4 Fayyad**	#5 1R	**#4 10bin**	**#4 Chi2**	#4 Zeta
#4 ChiMerge	**#5 ChiMerge**	#6 ChiMerge	**#5 CAIM**	**#5 CAIM**	#5 10bin
#5 CAIM	**#5 Zeta**	#7 ID3	#6 ChiMerge	**#6 Fusinter**	#6 Chi2
#6 Fayyad	#6 Fusinter	#8 Zeta	#7 Zeta	#7 Zeta	#7 ChiMerge
#7 1R	#7 Chi2	#9 10bin	#8Fayyad	#8 ChiMerge	#8 Fayyad
G-mean			Matthew's Correlation Coefficient (MCC)		
T3	T4	P1	T3	T4	P1
#1 Chi2	**#1 ID3**	**#1 Fusinter**	**#1 Chi2**	**#1 ID3**	**#1 Fusinter**
#2 Fusinter	**#2 1R**	**#2 ID3**	#2 ID3	**#2 1R**	**#2 1D3**
#2 ID3	**#3 Fayyad**	#3 CAIM	#3 10bin	**#3 10bin**	#3 1R
#2 10bin	**#4 CAIM**	#4 1R	#4 Fusinter	#4 Fayyad	#3 CAIM
#3 CAIM	**#5 10bin**	#5 Chi2	#5 ChiMerge	#5 CAIM	#4 Chi2
#4 1R	#6 Zeta	#6 Zeta	#6 1R	#6 1R	#5 Zeta
#5 ChiMerge	#7 ChiMerge	#7 10bin	#6 CAIM	#7 ChiMerge	#7 10bin
#8 Zeta	#8 Fusinter	#8 ChiMerge	#7 Zeta	#8 Chi2	#7 ChiMerge
#9 Fayyad	#7 Chi2	#9 Fayyad	#8 Fayyad	#9 Fusinter	#8 Fayyad

The results are heterogeneous between the datasets, the discretization techniques, and the quality measures. For example, for the sensitivity, the best-ranked techniques `Chi2` and `Fayyad` for the datasets T3 and P1 respectively are statistically different from the other techniques. In contrast, for dataset T4, seven of the nine techniques give equivalent results. For the specificity, numerous discretization techniques are equivalent for datasets T3 and T4, while only three techniques are equivalent for dataset P1. Besides, for dataset T3, `Chi2` leads to the best average score for each metric, while `ID3` leads to the most efficient rulesets for dataset T4. For dataset P1, `Fusinter` and `ID3` lead to the best specificity, G-mean, and MCC while `Fayyad` gives the best sensitivity, and it is last ranked for the three other measures. This behavior is probably due to the presence of several zeros in the samples for each attribute that leads most VOCs to have a single interval ($(-inf; +inf)$) after the application of `Fayyad`. `ID3` is among the best techniques for seven of the twelve experiments.

4 Discussion

In this work, we observed the impact of different discretization methods on the models produced by MOCA-I. In particular, we focused on real health data, where a sample corresponds to quantities of VOCs emitted by individuals. The aim was to determine which discretization method is the most suited for this type of data. The results on our datasets highlight that the ID3 discretization method seems to be suited to the case of VOCs.

In the future, we will perform these experiments on other datasets containing VOCs, in particular, datasets with more individuals provided by the PATHACOV project and imbalanced datasets. We also plan to study the impact of discretization methods with different parameters for MOCA-I, since their values may influence the quality of the resulting ruleset. In order to compare our approach to classical machine learning algorithms, we will study the impact of the discretization methods on their efficiency.

References

1. Alcalá-Fdez, J., et al.: Keel: a software tool to assess evolutionary algorithms for data mining problems. Soft. Comput. **13**(3), 307–318 (2009). https://doi.org/10.1007/s00500-008-0323-y
2. Garcia, S., Luengo, J., Sáez, J.A., Lopez, V., Herrera, F.: A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. IEEE Trans. Knowl. Data Eng. **25**(4), 734–750 (2012)
3. Jacques, J., Taillard, J., Delerue, D., Jourdan, L., Dhaenens, C.: The benefits of using multi-objectivization for mining Pittsburgh partial classification rules in imbalanced and discrete data. In: Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation, pp. 543–550. ACM (2013)
4. Khalid, T., et al.: Urinary volatile organic compounds for the detection of prostate cancer. PLoS ONE **10**(11), e0143283 (2015)

5. Leunis, N., et al.: Application of an electronic nose in the diagnosis of head and neck cancer. Laryngoscope **124**(6), 1377–1381 (2014)
6. Ohsaki, M., Abe, H., Tsumoto, S., Yokoi, H., Yamaguchi, T.: Evaluation of rule interestingness measures in medical knowledge discovery in databases. Artif. Intell. Med. **41**(3), 177–196 (2007)
7. Phillips, M., et al.: Volatile biomarkers of pulmonary tuberculosis in the breath. Tuberculosis **87**(1), 44–52 (2007)
8. Sakumura, Y., et al.: Diagnosis by volatile organic compounds in exhaled breath from lung cancer patients using support vector machine algorithm. Sensors **17**(2), 287 (2017)