Multiple Instance Learning for Viral Pneumonia Chest X-ray Classification

(Discussion Paper)

Matteo Avolio[1,3], Antonio Fuduli[1,3], Eugenio Vocaturo[2,3] and Ester Zumpano[2,3]

[1]Department of Mathematics and Computer Science - University of Calabria, Rende, Italy
[2]Department of Computer Engineering, Modeling, Electronics, and Systems Sciences - University of Calabria, Rende, Italy
[3]CNR-NANOTEC National Research Council, Rende, Italy

Abstract

At the end of 2019 anew coronavirus, SARS-CoV-2, was identified as responsible for the lung infection, now called COVID-19 (coronavirus disease 2019). Since then there has been an exponential growth of infections and at the beginning of March 2020 the WHO declared the epidemic a global emergency. An early diagnosis of those carrying the virus becomes crucial to contain the spread, morbidity and mortality of the pandemic. The definitive diagnosis is made through specific tests, among which imaging tests play an important role in the care path of the patient with suspected or confirmed COVID-19. Patients with serious COVID-19 typically experience viral pneumonia. This paper uses the Multiple Instance Learning paradigm to classify pneumonia X-ray images, considering three different classes: radiographies of healthy people, radiographies of people with bacterial pneumonia and of people with viral pneumonia. The proposed algorithms, which are very fast in practice, appear promising especially if we take into account that no preprocessing technique has been used.

Keywords

Pneumonia imaging Classification, Multiple Instance Learning, Machine Learning

1. Introduction

The world is coping with the COVID-19 pandemic. COVID-19 is caused by a Severe Acute Respiratory Sindrome Coronavirus (SARS-CoV-2) and its common symptoms are: fever, dry cough, fatique, short breathing, vanishing of taste, loss of smell. The most effective way to limit the spread of COVID-19 and the number of deaths is to identify infected persons at an early stage of the disease and many different proposals have been investigated for the development of automatic screening of COVID-19 from medical images analysis. COVID-19 has interstitial pneumonia as the predominant clinical manifestation. Radiological imaging is able to highlight any pneumonia: in the case of Coronavirus (Sars-Cov2) infection, it is possible to see an opacity on the radiograph, called *thickening* and a greater extension of the pulmonary thickening [1]. Therefore, while huge challenges need to be faced, medical imaging analysis arises as a key factor in the screening of viral pneumonia from bacteria pneumonia. In reasoning on the

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy
matteo.avolio@unical.it (M. Avolio); antonio.fuduli@unical.it (A. Fuduli); e.vocaturo@dimes.unical.it (E. Vocaturo); e.zumpano@dimes.unical.it (E. Zumpano)

assessment of COVID-19 chest radiography (CXR) and computed tomography (CT) are used. CT imaging shows high sensitivity, but X-ray imaging is cheaper, easier to perform and in addition (portable) X-ray machines are much more available also in poor and developing countries [2, 3]. The idea underlying this work arises within the described scenario, characterized by an intense activity of scientific research, aimed at supporting fast solutions for diagnostics on COVID-19, which is a special case of viral pneumonia. Considering that there are recurring features that characterize the radiographs of patients affected by viral pneumonia, we propose a chest X-ray classification technique based on the Multiple Instance Learning (MIL) approach.

We have considered a subset of images taken from the public Kaggle chest X-ray dataset [4] from which we have randomly extracted 50 images related to radiography of healthy people, 50 of people with bacterial pneumonia and 50 of people with viral pneumonia. This data set is widely used in the literature in connection with specific COVID-19 data sets, as reported in [5].

2. Multiple Instance Learning

Multiple Instance Learning [6] is a classification technique consisting in the separation of point sets: such sets are called *bags* and the points inside the sets are called *instances*. The main difference of a MIL approach with respect to the classical supervised classification is that in the learning phase only the class labels of the bags are known, while the class labels of the instances remain unknown. A particular role played by MIL is in medical image and video analysis, as shown in [7]. Diagnostics by means of image analysis is an important field in order to support physicians to have early diagnoses [8, 9, 10]. We focus on binary MIL classification with two classes of instances, on the basis of the the so-called standard MIL assumption, which considers positive a bag containing at least a positive instance and negative a bag containing only negative instances. Such assumption fits very well with diagnostics by images: in fact a patient is non-healthy (i.e. is positive) if his/her medical scan (bag) contains at least an abnormal subregion and is healthy if all the subregions forming his/her medical scan are normal. In [11] a MIL approach has been used for melanoma detection on color dermoscopic images, with the aim to discriminate between melanomas (positive images) and common nevi (negative images). The obtained results encourage to investigate possible use of MIL techniques also in viral pneumonia detection by means of chest X-rays images. In particular, using binary MIL classification techniques, our aim is to discriminate between X-rays images of healthy patients versus patients with bacteria pneumonia, healthy patients versus patients with viral pneumonia and patients with bacteria pneumonia versus patients with viral pneumonia.

3. The MIL-RL algorithm

MIL-RL algorithm [12] is an instance-level technique based on solving, by Lagrangian relaxation [13], the Support Vector Machine (SVM) type model proposed by Andrews et al. in [14]. Such model, providing an SVM separating hyperplane of the type

$$H(w,b) \stackrel{\triangle}{=} \{x \in \mathbb{R}^n \mid w^T x + b = 0\}, \tag{1}$$

is the following:

$$\begin{cases} \min\limits_{y,w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{m}\sum\limits_{j\in J_i^+}\xi_j + C\sum\limits_{i=1}^{k}\sum\limits_{j\in J_i^-}\xi_j \\ & \xi_j \geq 1 - y_j(w^T x_j + b) \quad j \in J_i^+,\ i = 1,\dots,m \\ & \xi_j \geq 1 + (w^T x_j + b) \quad j \in J_i^-,\ i = 1,\dots,k \\ & \sum\limits_{j\in J_i^+}\frac{y_j+1}{2} \geq 1 \quad i = 1,\dots,m \\ & y_j \in \{-1,+1\} \quad j \in J_i^+, \quad i = 1,\dots,m \\ & \xi_j \geq 0 \quad j \in J_i^+,\ i = 1,\dots,m \\ & \xi_j \geq 0 \quad j \in J_i^-,\ i = 1,\dots,k, \end{cases} \tag{2}$$

where: m is the number of positive bags; k is the number of negative bags; x_j is the j-th instance belonging to a bag; J_i^+ is the index set corresponding to the instances of the i-th positive bag; J_i^- is the index set corresponding to the instances of the i-th negative bag.

Variables b and w correspond respectively to the bias and normal to the hyperplane, variable ξ_j gives a measure of the misclassification error of the instance x_j, while y_j is the class label to be assigned to the instances of the positive bags. The positive parameter C tunes the weight between the maximization of the margin, obtained by minimizing the Euclidean norm of w, and the minimization of the misclassification errors of the instances. Finally, the constraints

$$\sum_{j\in J_i^+}\frac{y_j+1}{2} \geq 1 \quad i = 1,\dots,m \tag{3}$$

impose that, for each positive bag, at least one instance should be positive (i.e. with label equal to +1). Note that, when $m = k = 1$ and $y_j = +1$ for any j, problem (2) reduces to the classical SVM quadratic program. The core of MIL-RL is to solve, at each iteration, the Lagrangian relaxation of problem (2), obtained by relaxing constraints (3):

$$\begin{cases} \min\limits_{y,w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{m}\sum\limits_{j\in J_i^+}\xi_j + C\sum\limits_{i=1}^{k}\sum\limits_{j\in J_i^-}\xi_j + \sum\limits_{i=1}^{m}\lambda_i\left(1 - \sum\limits_{j\in J_i^+}\frac{y_j+1}{2}\right) \\ & \xi_j \geq 1 - y_j(w^T x_j + b) \quad j \in J_i^+,\ i = 1,\dots,m \\ & \xi_j \geq 1 + (w^T x_j + b) \quad j \in J_i^-,\ i = 1,\dots,k \\ & y_j \in \{-1,+1\} \quad j \in J_i^+, \quad i = 1,\dots,m \\ & \xi_j \geq 0 \quad j \in J_i^+,\ i = 1,\dots,m \\ & \xi_j \geq 0 \quad j \in J_i^-,\ i = 1,\dots,k, \end{cases} \tag{4}$$

where $\lambda_i \geq 0$ is the i-th Lagrangian multiplier associated to the i-th constraint of the type (3). [12] shows that, considering the Lagrangian dual of the primal problem (2), in correspondence to the optimal solution there is no dual gap between the primal and dual objective functions.

4. The mi-SPSVM algorithm

Algorithm mi-SPSVM has been introduced in [15] and it exploits the good properties exhibited for supervised classification by the SVM technique in terms of accuracy and by the PSVM (Proximal Support Vector Machine) approach [16] in terms of efficiency. It computes a separating hyperplane of the type (1) by solving, at each iteration, the following quadratic problem:

$$\begin{cases} \min\limits_{w,b,\xi} \quad \frac{1}{2} \left\| \begin{array}{c} w \\ b \end{array} \right\|^2 + \frac{C}{2} \sum\limits_{j \in J^+} \xi_j^2 + C \sum\limits_{j \in J^-} \xi_j \\ \\ \qquad \xi_j = 1 - (w^T x_j + b) \quad j \in J^+ \\ \\ \qquad \xi_j \geq 1 + (w^T x_j + b) \quad j \in J^- \\ \\ \qquad \xi_j \geq 0 \quad j \in J^-, \end{cases} \tag{5}$$

by varying of the sets J^+ and J^-, which contain the indexes of the instances currently considered positive and negative, respectively. At the initialization step, J^+ contains the indexes of all the instances of the positive bags, while J^- contains the indexes of all the instances of the negative bags. Once an optimal solution, say (w^*, b^*, ξ^*), to problem (5) has been computed, the two sets J^+ and J^- are updated in the following way:

$$J^+ := J^+ \setminus \bar{J} \qquad \text{and} \qquad J^- := J^- \cup \bar{J}$$

where $\bar{J} = \{j \in J^+ \setminus J^* \mid w^{*T} x_j + b^* \leq -1\}$,
with $J^* = \{j_i^*,\ i = 1, \ldots, m \mid w^{*T} x_{j_i^*} + b^* \leq -1\}$ and $j_i^* \stackrel{\triangle}{=} \arg\max_{j \in (J_i^+ \cap J^+)} \{w^{*T} x_j + b^*\}$.

Some comments on the updating of the sets J^+ and J^- are in order. A particular role in the definition of the set $\bar{J}$ is played by the set J^*, introduced for taking into account constraints (3). We recall that such constraints impose the satisfaction of the standard MIL assumption, stating that, for each positive bag, at least one instance must be positive. At the current iteration, the set J^* is the index set (subset of J^+) corresponding to the instances closest, for each positive bag, to the current hyperplane $H(w^*, b^*)$ and strictly lying in the negative side with respect to it. If an index, say $j_i^* \in J^*$, corresponding to one of such instances entered the set J^-, all the instances of the i-th positive bag would be considered negative by problem (5), favouring the violation of the standard MIL assumption. This is the reason why the indexes of J^* are prevented from entering the set J^-: in this way, for each positive bag, at least an index corresponding to one of its instances is guaranteed to be inside J^+.

Figure 1: Examples of X-ray chest images of healthy people

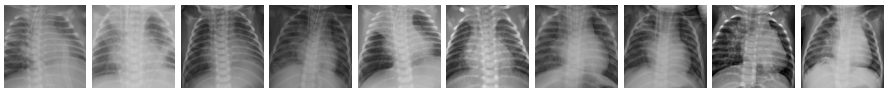

Figure 2: Examples of X-ray chest images of people with bacterial pneumonia

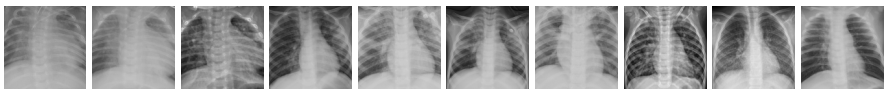

Figure 3: Examples of X-ray chest images of people with viral pneumonia

5. Numerical results

No pre-processing step is performed in this paper. This assumption allows us to attribute the results only to the performance of the applied algorithms and not to the goodness of the pre-processing phase. A balanced dataset extracted from from the public dataset ([4]) available at https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia consists of 50 images of healthy people (Figure 1), 50 of people with bacterial pneumonia (Figure 2) and 50 of people with viral pneumonia (Figure 3) This proposal uses the Matlab implementation of MIL-RL in [11] and the Matlab implementation of mi-SPSVM in [15].

As for the segmentation process, we have adopted a procedure similar to that one used in [17]. In particular, we have reduced the resolution of each image to 128×128 pixels dimension and we have grouped the pixels in appropriate square subregions (blobs). In this way, each image is represented as a bag, while a blob corresponds to an instance of the bag. For each instance (blob), we have considered the following 10 features: the average and the variance of the grey-scale intensity of the blob: 2 features; the differences between the average of the grey-scale intensity of the blob and that ones of the adjacent blobs (upper, lower, left, right): 4 features; the differences between the variance of the grey-scale intensity of the blob and that ones of the adjacent blobs (upper, lower, left, right): 4 features.

The following X-ray chest images classification have been performed: (i) bacterial pneumonia (positive) images versus normal (negative) images (Table 1); (ii) viral pneumonia (positive) images versus normal (negative) images (Table 2); (iii) viral pneumonia (positive) images versus bacterial pneumonia (negative) images (Table 3).

In particular, in order to consider different sizes of the testing and training sets, we have used three different validation protocols: the 5-fold cross-validation (5-CV), the 10-fold cross-validation (10-CV) and the Leave-One-Out validation. As for the optimal computation of the tuning parameter C characterizing the models (2) and (5), in both the cases we have adopted a be-level approach of the type used in [18] and in [11].

	5-CV		10-CV		Leave-One-Out	
	MIL-RL	mi-SPSVM	MIL-RL	mi-SPSVM	MIL-RL	mi-SPSVM
Correctness (%)	88.00	91.00	89.00	91.00	90.00	91.00
Sensitivity (%)	94.70	94.70	93.83	93.83	94.00	94.00
Specificity (%)	82.50	87.50	83.38	86.81	86.00	88.00
F-score (%)	88.73	91.68	88.89	90.74	90.38	91.26
CPU time (secs)	0.24	0.04	0.32	0.06	0.59	0.14

Table 1
X-ray chest images: average testing values relating 50 normal vs 50 with bacterial pneumonia

	5-CV		10-CV		Leave-One-Out	
	MIL-RL	mi-SPSVM	MIL-RL	mi-SPSVM	MIL-RL	mi-SPSVM
Correctness (%)	87.00	88.00	84.00	87.00	86.00	89.00
Sensitivity (%)	86.36	93.18	90.71	90.71	88.00	94.00
Specificity (%)	88.61	83.89	77.38	83.05	84.00	84.00
F-score (%)	86.53	88.63	84.85	87.36	86.27	89.52
CPU time (secs)	0.35	0.05	0.56	0.06	0.58	0.07

Table 2
X-ray chest images: average testing values relating 50 normal vs 50 with viral pneumonia

	5-CV		10-CV		Leave-One-Out	
	MIL-RL	mi-SPSVM	MIL-RL	mi-SPSVM	MIL-RL	mi-SPSVM
Correctness (%)	72.00	67.00	75.00	74.00	71.00	74.00
Sensitivity (%)	71.74	80.23	73.98	80.83	68.00	82.00
Specificity (%)	70.83	54.44	74.86	63.10	74.00	66.00
F-score (%)	70.46	69.99	73.06	73.69	70.10	75.93
CPU time (secs)	0.36	0.06	0.89	0.06	0.93	0.15

Table 3
X-ray chest images: average testing values relating 50 with bacterial vs 50 with viral pneumonia

Tables 1, 2 and 3 report the average values provided by MIL-RL and mi-SPSVM in terms of correctness (accuracy), sensitivity, specificity and F-score, computed on the testing set and the average CPU time spent by the classifier to determine the optimal separation hyperplane. Observe that mi-SPSVM is clearly faster than MIL-RL and, in general, it classifies better, even if the accuracy results provided by the two codes appear comparable. In classifying bacterial pneumonia (Table 1) and viral pneumonia (Table 2) against normal X-ray chest images we obtain high values of accuracy (about 90%) and sensitivity (about 94%). We recall that the sensitivity (also called true positive rate) is a very important parameter in diagnostics since it measures the proportion of positive patients correctly identified. On the other hand, when we discriminate between the viral pneumonia and the bacterial pneumonia images (Table 3), we obtain lower results with respect to those ones reported in Tables 1 and 2, as expected since the two classes are very similar. Nevertheless, these values appear reasonable, especially in terms of sensitivity

(82% provided by mi-SPSVM) and of F-score (75.93%).

6. Conclusions and future work

This work presented some preliminary numerical results obtained from classification of viral pneumonia against bacterial pneumonia and normal X-ray chest images, by means of MIL algorithms. Results appear promising, especially considering that no preprocessing phase has been performed. Moreover our MIL techniques appear appealing also in terms of computational efficiency, since the separation hyperplane is always obtained in less than one second. Future research could consist in appropriately preprocessing the images and in considering additional features [19, 20] to be exploited in the classification process, including also COVID-19 chest X-ray images and distributing the classification algorithms [21]. Our aim goal is to create a framework that can support the diagnostics of COVID-19, possibly by expanding one of our solutions already implemented [22] in a distributed environment [23, 24, 25, 26] and also including aspects of process management from a health perspective [27]. As for further future research we plan to apply the MIL approach to other medical domains such as [28, 29, 30, 9].

References

[1] H. X. Bai, et al., Performance of radiologists in differentiating covid-19 from non-covid-19 viral pneumonia at chest ct, Radiology 296 (2020) E46–E54.

[2] E. Zumpano, A. Fuduli, E. Vocaturo, M. Avolio, Viral pneumonia images classification by multiple instance learning: preliminary results, in: IDEAS 2021, ACM, 2021, pp. 292–296.

[3] E. Vocaturo, E. Zumpano, L. Caroprese, Convolutional neural network techniques on x-ray images for covid-19 classification, in: BIBM, IEEE, 2021, pp. 3113–3115.

[4] P. Mooney, Chest X-ray images (pneumonia), 2021. URL: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia, available on line.

[5] H. S. Alghamdi, G. Amoudi, S. Elhag, K. Saeedi, J. Nasser, Deep learning approaches for detecting COVID-19 from chest X-ray images: A survey, IEEE Access 9 (2021) 20235–20254.

[6] F. Herrera, et al., Multiple instance learning: Foundations and algorithms, Springer International Publishing, 2016.

[7] G. Quellec, G. Cazuguel, B. Cochener, M. Lamard, Multiple-instance learning for medical image and video analysis, IEEE Reviews in Biomedical Engineering 10 (2017) 213–234.

[8] E. Vocaturo, E. Zumpano, Supporting the diagnosis of dysplastic nevi syndrome via multiple instance learning approaches, in: AI4H@ECAI, volume 2820, 2020, pp. 39–44.

[9] E. Vocaturo, E. Zumpano, Diabetic retinopathy images classification via multiple instance learning, in: CHASE, IEEE, 2021, pp. 143–148.

[10] E. Vocaturo, E. Zumpano, Multiple instance learning approaches for melanoma and dysplastic nevi images classification, in: ICMLA, IEEE, 2020, pp. 1396–1401.

[11] A. Astorino, A. Fuduli, P. Veltri, E. Vocaturo, Melanoma detection by means of multiple instance learning, Interdisciplinary Sciences: Computational Life Sciences 12 (2020) 24–31.

[12] A. Astorino, A. Fuduli, M. Gaudioso, A Lagrangian relaxation approach for binary multiple

instance classification, IEEE Transactions on Neural Networks and Learning Systems 30 (2019) 2662 – 2671.
[13] M. Guignard, Lagrangean relaxation, Top 11 (2003) 151–200.
[14] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: S. Becker, S. Thrun, K. Obermayer (Eds.), Advances in Neural Information Processing Systems, MIT Press, Cambridge, 2003, pp. 561–568.
[15] M. Avolio, A. Fuduli, A semiproximal support vector machine approach for binary multiple instance learning, IEEE Transactions on Neural Networks and Learning Systems, 32(8), pp. 3566-3577 (2021).
[16] G. Fung, O. Mangasarian, Proximal support vector machine classifiers, in: Proceedings KDD-2001: Knowledge discovery and data mining, ACM, 2001, pp. 77–86.
[17] A. Astorino, A. Fuduli, P. Veltri, E. Vocaturo, On a recent algorithm for multiple instance learning. preliminary applications in image classification, in: BIBM, 2017, pp. 1615–1619.
[18] A. Astorino, A. Fuduli, The proximal trajectory algorithm in SVM cross validation, IEEE Transactions on Neural Networks and Learning Systems 27 (2016) 966–977.
[19] P. Kukic, C. Mirabello, G. Tradigo, I. Walsh, P. Veltri, G. Pollastri, Toward an accurate prediction of inter-residue distances in proteins using 2d recursive neural networks, BMC Bioinformatics 15 (2014).
[20] G. Tradigo, S. De Rosa, P. Vizza, G. Fragomeni, P. H. Guzzi, C. Indolfi, P. Veltri, Calculation of intracoronary pressure-based indexes with jlabchart, Applied Sciences 12 (2022).
[21] N. Cassavia, S. Flesca, M. Ianni, E. Masciari, C. Pulice, Distributed computing by leveraging and rewarding idling user resources from p2p networks, Journal of Parallel and Distributed Computing 122 (2018) 81–94.
[22] E. Zumpano, P. Iaquinta, L. Caroprese, F. Dattola, G. Tradigo, P. Veltri, E. Vocaturo, Simpatico 3d mobile for diagnostic procedures, ACM, 2019.
[23] L. Caroprese, E. Zumpano, Handling preferences in P2P systems, in: FOIKS, volume 7153, Springer, 2012, pp. 91–106.
[24] L. Caroprese, E. Zumpano, Aggregates and priorities in P2P data management systems, in: IDEAS, ACM, 2011, pp. 1–7.
[25] L. Caroprese, E. Zumpano, A logic framework for P2P deductive databases, Theory Pract. Log. Program. 20 (2020) 1–43.
[26] L. Caroprese, E. Zumpano, Declarative semantics for P2P data management system, J. Data Semant. 9 (2020) 101–122.
[27] A. Guzzo, A. Rullo, E. Vocaturo, Process mining applications in the healthcare domain: A comprehensive review, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 12 (2022) e1442.
[28] E. Vocaturo, E. Zumpano, AI for the detection of the diabetic retinopathy, in: Integrating Artificial Intelligence and IoT for Advanced Health Informatics - Technology, Communications and Computing, Springer, 2022, pp. 129–140.
[29] E. Vocaturo, E. Zumpano, ECG analysis via machine learning techniques: News and perspectives, in: BIBM1, IEEE, 2021, pp. 3106–3112.
[30] E. Vocaturo, E. Zumpano, Artificial intelligence approaches on ultrasound for breast cancer diagnosis, in: BIBM, IEEE, 2021, pp. 3116–3121.