Semi-Supervised Classification of Hyperspectral Images for Brain Tumours detection

Patricia Beltran Alonso, Gustavo Marrero Callico, Samuel Ortega Sarmiento and Beatriz Martinez Vega
Institute for Applied Microelectronics (IUMA), University of Las Palmas Gran Canaria (ULPGC), Spain
{pbeltran, gustavo, sortega, bmartinez}@iuma.ulpgc.es

***Abstract*— In the medical field, hyperspectral (HS) images have represented a technological breakthrough due to their non-invasive nature and because they provide useful information for the diagnosis of diseases. However, in many practical classification applications the number of available unlabelled samples is large, since the collection of labelled samples is complicated. For this reason, it is interesting to develop algorithms able to exploit both labelled and unlabelled samples in the classification process to obtain high-performance classifiers. Semi-Supervised Learning (SSL) is a powerful tool to generate learning models when the number of labelled samples is low. This paper describes different methodologies of the design of semi-supervised algorithm for brain tumour detection. For the evaluation of these designs, the Support Vector Machines (SVM) and the Random Forest (RF) classifiers were employed.**

Keywords-Hyperspectral imaging; Semi-supervised Learning; Support Vercor Machines; Random Forest; k-means; Machine Learning.

I. INTRODUCTION

Hyperspectral (HS) Imaging (HSI) is also known as imaging spectroscopy. The word "imaging" stand for the representation of the appearance or morphology of the object, while the term "spectroscopy" indicates the study of the interaction of electromagnetic radiation with different materials. HS technology is able to acquire hundreds of contiguous spectral bands, obtaining the spectral signature of any material. The spectral signature can be used to identify different types of materials [1] by measuring the radiation reflected by each material at each wavelength.

The medical field faces problems when applying the technology of the HS images for the diagnosis diseases. This technology represents a great advance due to its non-invasive nature and because it provides a lot of useful information about the pathology to be evaluated [2]. The principal problems are to collect samples labelled is a costly process that requires effort and human experience. For cases where the number of unlabelled samples is large, algorithms capable of including both labelled and unlabelled samples in the classification process have been developed [3]. These algorithms use semi-supervised learning (SSL) to generate models from a low number of samples. In this paper, the use of SSL techniques for the classification of medical HS data is proposed.

II. METHODOLOGY

To perform this work, a database obtained at the University Hospital Doctor Negrín with European HELICOiD project was employed [4]. This HS database is composed by 26 HS cubes belonging to a total of 16 different patients diagnosed with Glioblastoma primary brain tumour. The HS images were captured during surgical procedures. The images are labeled with 4 different classes: normal tissue, tumour tissue, hypervascularized tissue, and background. In the next sections, the different semi-supervised classification proposed methods are explained.

A. Proposed methodology

The motivation of this work is to simulate a realistic case in the operating room, where there is a previously labelled database and the new acquired data of the patient who is going to receive the intervention. The objective is to include this current patient data in the database with which to train the supervised classifiers. The methodology proposed to develop the semi-supervised classification of HS images of brain tumours is as follows: It starts from a database that consist of pre-processed and previously labelled HS images. With this database, the labelling of a new patient is performed by using the distance of each pixel of the new patient with respect to the mean spectral signatures of the complete database (composed by labelled data from previous patients). First, an evaluation of the most suitable distance metric was performed. Second, an evaluation was made to select which value of the k parameter best fits our database in the k-means clustering [8]. The database of the previous patients without the new patients used in the k-means to get the different clusters. Once it is known which cluster belongs to which class, the labels of the new patient are generated by calculating the minimum distance to the centroids (Figure 1).

Figure 1 Block diagram corresponding to the proposed procedure.

The new patient data labelled with this methodology and the dataset of the previous patients are fed into the classifiers, in order to train them, generate a model, and finally evaluate its performance. The supervised classifiers selected for this work are the Support Vector Machines (SVM) [5], [6] and Random Forests (RF) [7]. For this purpose, the following evaluation metrics have been employed: accuracy, the confusion matrix, specificity, sensitivity and, finally, the kappa coefficient. To obtain these metrics, a data partition based on the leave-one-out cross-validation technique was used.

III. RESULTS

Once the most suitable distance metric was the cosine distance, after the distance was selected, the value of the input parameter *k* is calculated, with which the k-means method divides the samples into *k* groups. The range evaluated was from 4 to 24. To make this decision, the k-means algorithm is executed with that range of values. The decision was based on observing what the percentage of each class was grouped in each cluster. This evaluation has been performed using the dataset belonging to the previous patients. No value of *k* can group in such a way that in at least one cluster its majority class is the tumour tissue. The data are evaluated and k=10 (Figure 2) is enough to be able to identify the other three classes in at least one cluster, a higher *k* would increase the computation time unnecessarily.

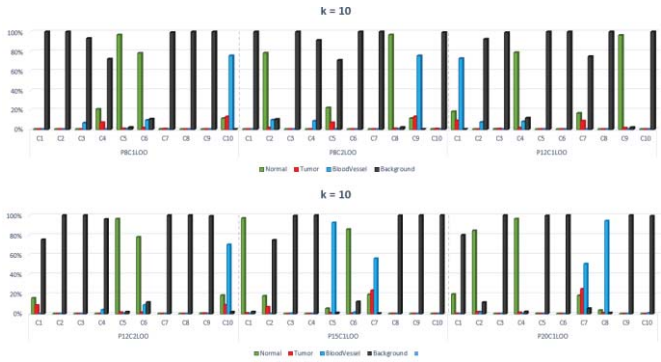

Figure 2 Percentage graph of the of the classes contained in each each cluster (k = 10).

After evaluating what distance and what value of k best fits our database, four case studies are presented. In the case of SVM (cosine distance and k value of 10 for the k-means method), the automatic generation of labels for the current patient is performed first, without any conditions (all clusters are used) and then with a condition, only the clusters that have a high presence of a certain class (more than 60%) are used.

For RF it is decided to use a k = 15, since it is the only one that meets the same criteria as k = 10 but with the difference that the condition no longer needs to be applied. The procedure will only have to be performed once, since the criterion used was that all clusters should consist of at least 60% of a class. Finally, when the RF algorithm (cosine distance and k value of 15 for the k-means method and 100 trees for RF) was used for the evaluation. For this classifier, two case studies were presented: using all clusters for the generation of labels or selecting only the three clusters that best represent the healthy tissue, blood vessel and background class.

SVM and RF algorithms were trained without semi-supervised vision in order to compare if using semi-supervised classification improves the results. Table 1 shows the average and standard deviation of the evaluation metrics. The highest success rate was obtained for the semi-supervised approach with RF (RF without condition) with 46.56%. However, the sensitivity obtained from normal and tumour tissues are lower than for the rest of the classifiers, except for the background class, which has a sensitivity of 98.77%.

Table 1 Results obtained in all semi-supervised process.

		Sensitivity				Specificity				
	OA	Normal	Tumour	Blood Vessel	Background	Normal	Tumour	Blood Vessel	Background	**Kappa**
Supervised process										
SVM	**78.77%**	93.00%	28.03%	87.44%	95.39%	81.96%	98.09%	93.32%	93.94%	-
RF	**76.99%**	97.04%	9.91%	89.79%	91.73%	76.92%	99.88%	95.67%	91.09%	**0.67**
Semi-supervised process										
SVM without Condition	**45.57%**	45.18%	1.90%	46.34%	64.69%	64.53%	99.62%	86.34%	57.10%	**0.25**
SVM with condition	**44.70%**	48.05%	1.43%	33.26%	87.28%	74.95%	99.30%	97.64%	46,.6%	**0.28**
RF without Condition	**46.56%**	44.80%	0.07%	33.96%	98.77%	96.55%	99.92%	97.67%	27.48%	**0.27**
RF (evaluating with three clusters)	**45.89%**	64.20%	0.03%	63.87%	70.99%	62.45%	100.00%	61.55%	77.32%	**0.19**

The highest sensitivity value for tumour class was obtained with the SVM without condition classifier with an 1.90%, which is still a too low value. If focusing on the rest of the data, perhaps the RF evaluation approach with three clusters gives the best results for all class types except tumour. The proposed processing method may not be adequate to improve the results. The semi-supervised algorithm proposal worsens the classification results compared to the non-semi-supervised.

IV. Conclusions

The main problem in this field is working with a limited database. According to the exposed problem, the objective of this paper was to design a semi-supervised classification system. This type of classification was intended to increase the existing database for the supervised classification, and to improve the baseline results using the data for the patient which is being classified.

The proposed methodology consists in that, when the patient is in the operating room, the images taken by the surgeon will be automatically labelled by a SSL algorithm and then, together with the existing database, the model can be generated. To perform the labelling of the current patient samples, we proposed a method which rely in the k-means algorithm. When evaluating all the results, it is seen that most of the clusters belong to the background class. This likely caused by the great variability within this class data. To avoid this, the last proposed procedure is performed again but using only 3 clusters, one cluster per class, background, blood vessel and normal tissue class. It is assumed with them that there is no cluster identified to the tumour class. When analysing these last results, it is seen how it is possible to improve the sensitivity of these three classes.

It is considered that the image used in the semi-supervised to automatically label it and thus increase the database with which the model is generated, must be an image that does not include any tumour pixels. In this way we can ensure that when the automatic labelled is generated, there will be no mislabelled tumour pixels. If we improve the balance of specificity and sensitivity of the rest of the classes, we will also be able to improve it for the tumour class. Finally, although a method to improve the classification has been proposed, this goal has not been achieved. The proposed methods worsen the original classification (without semi-supervised). Further work will be carried out in order to find an appropriate SSL approach to improve the original classification results.

References

[1] G. ElMasry and D.-W. Sun, “Principles of Hyperspectral Imaging Technology,” in *Hyperspectral Imaging for Food Quality Analysis and Control*, Elsevier, 2010, pp. 3–43.

[2] D. R. C. G. Samuel Ortega Sarmiento, Dr. Gustavo Marrero Callicó, “Técnicas de Reconocimiento Automático de Patrones Aplicadas a Imágenes Hiperespectrales Médicas.” 2016.

[3] Z.-H. Zhou and M. Li, “Semi-supervised learning by disagreement,” *Knowl Inf Syst*, vol. 24, pp. 415–439, 2010.

[4] H. Fabelo *et al.*, “HELICoiD project: a new use of hyperspectral imaging for brain cancer detection in real-time during neurosurgical operations,” 2016, p. 986002.

[5] M. P. S. Brown *et al.*, “Knowledge-based analysis of microarray gene expression data by using support vector machines,” *Proc. Natl. Acad. Sci.*, vol. 97, no. 1, pp. 262–267, Jan. 2000.

[6] E. J. Carmona Suárez, “Máquinas de Vectores Soporte (SVM),” *Dpto. Intel. Artif. ETS Ing. Informática, Univ. Nac. Educ. a Distancia*, pp. 1–25, 2014.

[7] D. Bertsimas and J. Dunn, “Optimal classification trees,” *Mach. Learn.*, vol. 106, no. 7, pp. 1039–1082, Jul. 2017.

[8] A. A. Aldino, D. Darwis, A. T. Prastowo, and C. Sujana, “Implementation of K-Means Algorithm for Clustering Corn Planting Feasibility Area in South Lampung Regency,” *J. Phys. Conf. Ser.*, vol. 1751, no. 1, 2021.