Classification of Multiple Sclerosis patients using a histogram-based K-Nearest Neighbors algorithm

Sana Rebbah, Daniel Delahaye, Stéphane Puechmorel, Pierre Maréchal, Florence Nicol, Isabelle Berry

▶ **To cite this version:**

Sana Rebbah, Daniel Delahaye, Stéphane Puechmorel, Pierre Maréchal, Florence Nicol, et al.. Classification of Multiple Sclerosis patients using a histogram-based K-Nearest Neighbors algorithm. OHBM 2019, 25th annual meeting of Organization for Human Brain Mapping, Jun 2019, Rome, Italy. hal-02156448

HAL Id: hal-02156448
https://enac.hal.science/hal-02156448v1

Submitted on 14 Jun 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Classification of Multiple Sclerosis patients using a histogram-based KNN algorithm

Sana REBBAH[1,2], Daniel Delahaye[2], Stéphane Puechmorel[2], Pierre Maréchal[3], Florence Nicol[2], Isabelle Berry[1,4]

[1]INSERM Toulouse NeuroImaging Center, Toulouse, France, [2]Ecole Nationale de l'Aviation Civile, Toulouse, France, [3]Institut de Mathématiques de Toulouse, Toulouse, France, [4]Centre Hospitalier Universitaire de Toulouse, Toulouse, France

Introduction:

Multiple Sclerosis (MS) is a demyelinating neurodegenerative disease. Due to diffuse aspect of the disease several studies focused on histogram-analysis to quantify the diffuse pathological changes of the disease (Cercignani, 2001; Dehmeshki, 2001). A common drawback of these studies is that the entire information included in the histogram is not used, only arbitrary measures are chosen to describe histogram; these include mean, median, percentiles, peak height, peak location, skewness and kurtosis. In our study, we propose an alternative way to use histograms by including the entire histogram information and not just a few local histogram descriptors, in a k-nearest neighbors classifier, with the aim to improve classification performance of MS population.

Methods:

Subjects and MRI acquisition: Our study included 111 subjects, 71 Healthy Control subjects from Alzheimer's Disease Neuroimaging Initiative (ADNI) database and 40 patients with Progressive Multiple Sclerosis from an MRI substudy of MS-SPI clinical trial (Tourbah, 2016). The groups are age- and gender-matched. The MR scans are 3D T1-weighted MR images and were acquired on a 1.5T scanner for HC subjects and 3T scanner for MS patients.
Anatomical MRI measure: Gray matter atrophy is a crucial marker of neurodegeneration and has therefore been used in several MS studies (Steenwijk, 2016). Cortical Thickness (CTh) was measured using the Matlab Toolbox CorThiZon and computed on the entire cortical ribbon using a Laplace's-equation-based algorithm (Querbes, 2009). Thereby, a 3D cortical thickness map was obtained, from which a CTh histogram have been extracted (cf. Figure 1). The histograms are normalized by dividing each histogram value by the sum of all the histogram values.
Histogram-based KNN algorithm: K-nearest neighbors (KNN) is a lazy learning algorithm that classify objects simply by assigning to the label of its K nearest neighbors (i.e. K number of neighbors). The main difference between histogram-based KNN and the classical KNN algorithm is that distances/dissimilarities are measured between CTh histograms and not between single descriptors such as CTh mean. Therefore, different distance/dissimilarity measures between histograms as presented by Cha (2007) were used, including Euclidean, Manhattan, Jaccard, Canberra, Pearson, Chi-squared, Kullback-Leibler and Jensen-Shannon and thus we have obtained a distance matrix for each metric. From there, it's easy to find the K-nearest neighbors and conclude on the object predicted label.
Two groups (HC and MS) were classify using stratified 5-fold cross-validated histogram-based KNN. Classification performance was described using accuracy (ACC), sensitivity (SE), specificity (SP) and Area Under the Curve (AUC).

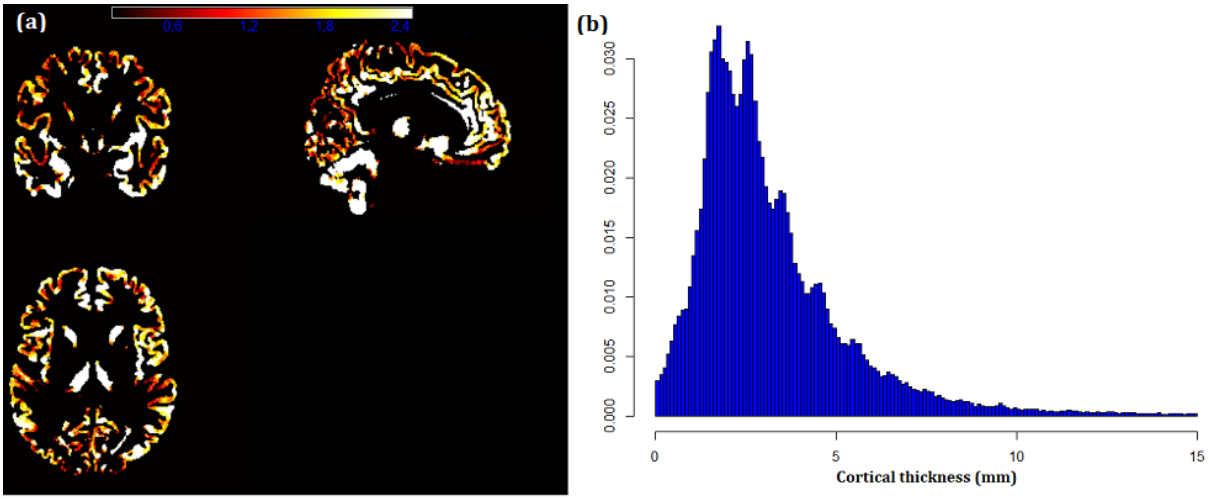

Figure 1 - Cortical Thickness
(a) represent the 3D Cortical Thickness map and (b) the normalized histogram extracted from the Cortical Thickness map

Results:

Since the value of K can influence the accuracy of the overall classification, we tested different K value, between 1 and 21 (only odd values because it's a binary classification). Overall the best value for K is 3. Table 1 presents the performance of the proposed method for different distance/dissimilarity measures. The performances are mostly very satisfying, indeed our approach distinguish HC and MS patients with 83% accuracy using Kullback-Leibler divergence followed by Jensen-shannon divergence with 81% ACC, which is not surprising since they are both from the Shannon's entropy family.

Distances/dissimilarities	ACC	SE	SP	AUC
Euclidean	79.3	87.3	65	0.76
Manhattan	80.2	87.3	67.5	0.77
Jaccard	78.4	85.9	65	0.76
Canberra	69.4	83.1	45	0.64
Pearson	80.2	87.3	67.5	0.77
Chi-squared	80.2	87.3	67.5	0.77
Kullback-Leibler	83	90.1	70	0.80
Jensen-Shannon	81.1	87.3	70	0.79

Table 1 – Results of histogram-based KNN classification of HC and MS patients with K=3 for different distance/dissimilarity measures. ACC, SE and SP are in % (Abbreviations: ACC, accuracy; SE, sensitivity; SP, specificity and AUC, area under the curve)

Conclusions:

The histogram-based KNN approach achieved comparable and even higher classification performances than previous studies based on histogram features or using region of interest-based approaches (Kwok, 2012; Wottschel, 2017). Indeed, classification based on local and arbitrary histogram features are unlikely to be optimum as much potential information is ignored. Moreover, using the entire histogram is a powerful aid to the study of diffuse diseases because of its ability to detect subtle changes early in the course of the disease, which increase significantly the performances of the disease classification.

References:

Cha, S.-H. (2007) 'Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions', International Journal Of Mathematical Models And Methods In Applied Sciences, 1(4), pp. 300–307

Cercignani, M. (2001), 'Mean diffusivity and fractional anisotropy histograms of patients with multiple sclerosis', AJNR. American journal of neuroradiology, 22(5), pp. 952–958.

Dehmeshki, J. (2001), 'Magnetisation transfer ratio histogram analysis of primary progressive and other multiple sclerosis subgroups', Journal of the Neurological Sciences, 185(1), pp. 11–17

Kwok, P. P. (2012), 'Predicting Clinically Definite Multiple Sclerosis from Onset Using SVM'. In Lecture Notes in Computer Science (pp. 116–123)

Querbes, O. (2009), 'Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve', Brain: A Journal of Neurology, 132(Pt 8), pp. 2036–2047

Steenwijk, M. D. (2016), 'Cortical atrophy patterns in multiple sclerosis are non-random and clinically relevant', Brain: A Journal of Neurology, 139(Pt 1), pp. 115–126

Tourbah, A. (2016), 'MD1003 (high-dose biotin) for the treatment of progressive multiple sclerosis: A randomised, double-blind, placebo-controlled study', Multiple Sclerosis (Houndmills, Basingstoke, England), 22(13), pp. 1719–1731

Wottschel, V. (2017), 'Supervised machine learning in multiple sclerosis: applications to clinically isolated syndromes'. Doctoral. UCL (University College London)