Bio-Medical Materials and Engineering 26 (2015) S1821–S1828
DOI 10.3233/BME-151484
IOS Press

HClass: Automatic classification tool for health pathologies using artificial intelligence techniques

Yolanda Garcia-Chimeno* and Begonya Garcia-Zapirain
DeustoTech-LIFE, University of Deusto, Avda Universidades, 24, 48007, Bilbao, Spain

Abstract. The classification of subjects' pathologies enables a rigorousness to be applied to the treatment of certain pathologies, as doctors on occasions play with so many variables that they can end up confusing some illnesses with others. Thanks to Machine Learning techniques applied to a health-record database, it is possible to make using our algorithm. hClass contains a non-linear classification of either a supervised, non-supervised or semi-supervised type. The machine is configured using other techniques such as validation of the set to be classified (cross-validation), reduction in features (PCA) and committees for assessing the various classifiers. The tool is easy to use, and the sample matrix and features that one wishes to classify, the number of iterations and the subjects who are going to be used to train the machine all need to be introduced as inputs. As a result, the success rate is shown either via a classifier or via a committee if one has been formed. A 90% success rate is obtained in the ADABoost classifier and 89.7% in the case of a committee (comprising three classifiers) when PCA is applied. This tool can be expanded to allow the user to totally characterise the classifiers by adjusting them to each classification use.

Keywords: Classification, machine learning, PCA, cross-validation, committee

1. Introduction

Pathology is the scientific study of illnesses, describing the cause, evolution and term of the illness. To determine certain pathologies, doctors base their work on accurate observations made of patients subject to study. There was a time when pathologies were classified by logical deductions by the doctors themselves, although thanks to new technologies and the creation of algorithms, this classification can now be automatic with a major degree of success. Therefore, by combining processing of the features of pathologies and the automation of classification processes, doctors are able to swiftly ascertain whether patients have a specific pathology or otherwise.

Nowadays there are numerous types of classification which are encompassed under the term Machine Learning [1, 2]. Thanks to this technique, patterns and hidden relations in data can be identified which enable a better diagnosis of the pathology itself to be made [3]. To ensure optimum classification, the features of the pathology deemed to be determining factors need to perfectly

* Address for correspondence: Yolanda Garcia-Chimeno, DeustoTech-LIFE, University of Deusto, Avda Universidades, 24, 48007, Bilbao, Spain. Tel.: 94 413 90 00-2980; Fax: 94 413 90 01; E-mail: yolanda.garcia@deusto.es.

selected. Thus be able to reach a conclusion as to whether the patient subject to study has the pathology in question or otherwise. To this end, doctors are of great importance as they are fully aware of the features that may determine the pathology, although they also have algorithms at their disposal that enable certain features [4] to be reduced by disregarding those considered redundant or because they play no major role.

There are different types of classifier, among which are supervised [5], non-supervised [6] and semi-supervised [7, 8] learning. The first type of classification supervised learning performs a mathematical function with which, by means of training samples that have already been labelled, the classifier deduces which class or type the set of samples to be classified belong.

As regards non-supervised learning, there is no set of already-labelled training samples, but rather, clustering techniques [9] have to be used to be able to classify the samples in groups that share similar features, provided that the resemblance between the groups to be classified is low.

The last type of classification, semi-supervised learning, is aimed at those cases in which a large set of samples needs to be classified and where manually labelling single supervised cases would otherwise constitute a very laborious task.

However, features can be reduced prior to classification, either owing to their being considered redundant or of limited significance within the set of samples. PCA (Principal Component Analysis) [10] can be applied to achieve an optimum classification. The cross-validation technique [11] can also be used to ensure optimum classification of subjects, as this enables the results obtained in the classification to be kept totally separate from the division between the training and testing group. The tool created can be validated in this way. Lastly, it is also possible to apply committees which enable the results from several classifiers to be obtained and an end result or conclusion for each sample drawn.

Other related tools that integrate machine learning algorithms are predominantly libraries to use in Weka [12] and Torch [13]. Also there is an integrated support for implementation and analysis in machine learning, Gestalt, focusing on source code and data to allow implementing a classification pipeline [14]. Moreover, other authors developed machine-learning-based target prediction tools (miTarget, MirTarge2 and NBmiRTar), that have not been through rigorous independent assessment [15].

Therefore, a tool has been created by combining all these techniques that enables a range of pathologies to be classified and in turn enabling committees to be applied or not and making it possible for certain minimum features of one of the classifiers to be selected, namely neural networks. As a result, the tool will generate a report in which the success rate of each classification made will be established.

2. Material and methods

2.1. Classifiers

Five classifiers have been selected from all possible ones in order to create this tool.

- SVM: this classifier also corresponds to the supervised type aimed at classification and regression problems. In this way, training is carried out with the set of training samples, creating a model to classify each new sample. These new samples are classified in one class or another depending on their proximity. In this case, SVM classifier has been trained only with the training samples and in which group fits each subject.

- Neural Networks: this classifier, of the supervised type, is based on the functioning of the neural system in the human body. In their learning phase, the samples are used as patterns by obtaining the values of weights of the connections and adjusting them to a specific criterion. Furthermore, training is also based on learning about the network according to its own error. For the test of this classifier, it has used neural network with one hidden layer and two layers. Moreover, it has been chosen a neuron number range in each layer to take the best result (for one layer between 20 and 40, for two layers: the first layer between 30 and 50, and second layer between 10 and 20).
- K-means: on this occasion, this classifier is of the non-supervised type in which samples are grouped together according to the similarity existing between them. Samples are therefore classified from the number of previously-known classes, i.e. the number of classes needs to be introduced in which one wishes to classify the set of samples. The features of this classifier are: 500 maximum number of iterations, Correlation distance, which each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to zero mean and unit standard deviation, other parameter is the action to take if cluster loses all member observation that it is singleton to create a new cluster consisting of the one point furthest from its centroid, and 100 replicates, that indicates the number of times to repeat clustering using new initial cluster centroid positions.
- K Nearest Neighbour: the purpose of this (supervised) classifier is, together with the training set, to create the classes of nearest neighbours by using a metric distance. This means that this algorithm classifies according to the majority of its neighbours, whereby each sample is assigned the most common class among its nearest "K" neighbours. This classifier is highly efficient in vectors with low dimensionality features. The features chosen for this classifier are: a Euclidean distance and 10 number of nearest neighbors used.
- AdaBoost: this final classifier is of a semi-supervised type which uses a combination of learning algorithms to improve its performance. It is an adaptive classifier, as the classifiers created in each iteration are adjusted so as to be able to improve samples that have been poorly classified by previous ones. This classifier consists of two parts: weak classifier tries to find the best threshold in one of the data dimensions to separate the data into two classes, and the boosting part calls the classifier iteratively, that after classification step it changes the weights of miss-classified samples. This creates a cascade of "weak classifiers" which behaves like a "strong classifier".

2.2. Reduction in features

PCA is a method which, among others, is used to disregard features considered redundant or of little significance within the feature matrix itself. It is a mathematical procedure that uses orthogonal transformation to convert a set of possibly correlated features into a set of values for linearly correlated samples (main components). This number of main components is either the same as or less than the set of original features. To find these components, the variability of the set of features is ascertained and ordered according to importance while disregarding those features that fail to meet minimum variance.

The dimensionality of the set of features can be reduced to ensure that success rates are as optimum as possible when introducing the matrix into the classifier.

2.3. Committees

There is no perfect classifier nowadays, with techniques still existing to try and deal with

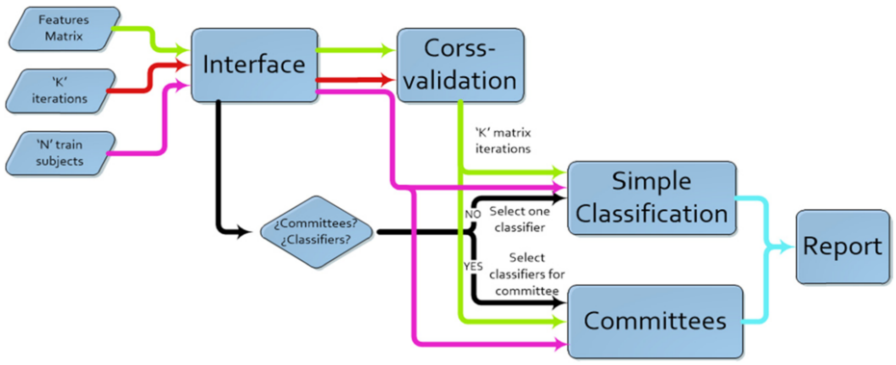

Fig. 1. High Level Design of the tool, with all blocks, inputs and outputs.

classification problems. The real problem is to select a classifier that entails finding the optimum division for separating classes. One option is to combine several classifiers [16], i.e. to seek diverse opinions, process them and combine them so as to then obtain a final conclusion that provides the best contrast and is the most reliable. Algorithms can be used to obtain the best classification combining different classifiers.

3. System design

3.1. High design

Block architecture is shown in high-level design that follows on from the tool created. As can be seen in Figure 1, the first block is the interface, in which the user introduces all the data required for classification purposes. A decision is also taken in the interface as to whether to carry out a single classification using a sole classifier or, conversely, to resort to a combination of several classifiers or committees.

The second block corresponds to cross-validation, which is necessary for classification as a classification cannot be taken into consideration that has not involved several iterations of that classification by changing the training and testing samples.

The following two blocks, which are parallel, correspond to the classification itself, which may be single using a sole classifier or using committees. The user will be the one who selects the type of classification.

3.2. Low design

Each block that makes up the tool is shown in greater detail in the case of low-level design.

3.2.1. Interface

The tool interface is clear and simple in terms of the attributes required that need to be introduced for classification purposes. The inputs required are the feature matrix of the subjects to be classified, the number of iterations for validation (cross-validation), the number of samples that one wishes to use as training, with which the remaining ones are set aside for testing and lastly how one wishes to classify needs to be selected. This decision implies that if what is required is a single classification, the

user will have to select the classifier to be used, whereas if they select the option to use committees, they will need to select the classifiers to create that committee.

3.2.2. PCA and cross-validation

The decision to apply PCA is one for the user of the tool to make themselves. If they select the tool, they will calculate the covariance matrix and its eigenvectors and eigenvalues. A vector is created from these eigenvalues, which arranged in order from greatest to least value. A characteristic value will subsequently be formed by rearranging the eigenvectors from greatest to least according to the magnitude of the eigenvalues. It is at this point when it is a good idea to reduce dimensions, whereby the threshold will be to select the number of features deemed sufficient to explain the 93% [17] of total variance of all the components.

The "K" type iteration has been selected for cross-validation purposes, in which the process for the validation itself is repeated during "K" iterations. Lastly, the arithmetic mean of the results of each iteration is calculated and a single result obtained.

3.2.3. Classifiers

This block of classifiers contains all those available within the tool. The classifier will receive the subject matrix to be classified, which contains all the features. If one wishes to carry out a supervised or semi-supervised classification, the number of samples will need to be stipulated that one wishes to user for training purposes, whereas the remaining ones will be used for testing the classifier itself. In the case of K-means, the value of subjects to be classified will not be needed because, as has been stated previously, the technique used is clustering.

The result of classification will be compared with the real values corresponding to which group each sample belongs to, and the success rate will be issued.

3.2.4. Committees

This block is parallel to that of the classifiers in which the user needs to select the number of classifiers they wish to use for the committee and which ones. The technique that will be used is that the end result of the committee will be obtained from most of the classification results that coincide. For instance, if a committee is formed with 5 classifiers, the threshold is that three of them will have to coincide in the same result, whereby the conclusion drawn from the classification will be that value.

4. Results

To check whether the tool is efficient for classification purposes, a data set free of MRI structural images was chosen (The Machine Learning Challenge (MLC 2014) (https://www.nmr.mgh.harvard.edu/lab/laboratory-computational-imaging-biomarkers/miccai-2014-machine-learning-challenge)), made up of 150 subjects and 180 features that constitute reference points for different parts of the brain [18]. Within these 150 subjects can be found two different groups: control subjects and clinical patients.

In addition, some variables were selected to verify the results and so as to ensure that the classification was the same in all cases.

100 iterations were made for validation (cross-validation) purposes. For practically all classifiers (except K-means) that need a training group, 100 out of the 150 subjects making up the Dataset were chosen to train the classifier, with the 50 remaining ones being used to assess the tool.

4.1. Single classification

To be able to assess each classifier separately, the 100 iterations and 100 subjects were selected to train the classifiers. The resulting percentages that can be seen in Table 1, correspond to each classifier that contains the tool. Moreover, another variant was used in one case without applying PCA and another one applying it so as to reduce the features and disregard any redundant or less important ones.

4.2. Classification with committees

Additionally and so as not just to obtain the result from each classifier separately, a decision was made to apply a committee, which comprised 3 classifiers in this case. Thus, a final conclusion would be able to be drawn from classification using the 3 classifiers that make up the committee that would constitute an improvement on individual classification.

In this case, it was decided that the committee should comprise 3 classifiers, with the ones selected being: SVM, K Nearest Neighbour and Neural Network with a hidden layer. Although these three classifiers individually account for quite a high success rate, neither did they represent the best rates obtained. Table 2 shows the success rates for the committee, once again dividing the result into which PCA was applied and to which it was not applied.

5. Discussion and conclusion

A tool has been created which is of great use in classifying patients who have been undergone a brain scan (MRI), and three types of classifier have also been included (supervised, non-supervised

Table 1

Classifiers success rate applying PCA5 and not

Classifier		Success rate (%)
SVM	No PCA	89.5
	PCA	89.3
K-means	No PCA	51.6
	PCA	53.7
KNN[a]	No PCA	86.6
	PCA	86.9
ADABoost	No PCA	90
	PCA	90
NN(1)[b]	No PCA	82.8
	PCA	85.9
NN(2)[c]	No PCA	89.4
	PCA	89.4

Note: [a]K Nearest Neighbor; [b]Neural Network with one hidden layer; [c]Neural Network with two hidden layers.

Table 2

Committee with three classifiers: SVM, K Nearest Neighbor and Neural Network with one hidden layer

Classifier	Success rate (%)
No PCA	88.4
PCA	89.7

and semi-supervised). As cross-validation is included, a stricter classification is undertaken as different groups are trained and other different ones assessed, i.e. not only the same training or validation group is introduced, whereby a final percentage is calculated when introducing different samples for each iteration, this being obtained from the mean of the iterations. Additionally, a selection can also be made as to whether to apply the reduction in features or not (PCA) and lastly, there is also the possibility of creating committees by selecting the number of classifiers that comprise them, and which ones. The success rate is obtained once the classification has been made. Therefore, it is not necessary to be an expert to be able to classify a set of samples, as the initial parameters which are requested are the basic ones and the tool can be used without any problem simply by having a certain amount of knowledge about the subject.

Moving on to the resulting classifier percentages, good results were obtained except for in the case of the K-mean classifier, which is the only non-supervised type. The 51.5% without PCA and 53.7% percentage indicates that classification was undertaken at random, as one half was being properly classified and the other half poorly classified. However, in the other cases, the worst percentage obtained was 82.8% for the Neural Network (supervised) classifier with a hidden internal layer without applying PCA, while the best was the ADABoost (semi-supervised) classifier, which obtained a 90% percentage [19] both with and without PCA. It makes sense that this classifier obtained the best percentage, as it learned the error that had been committed in each iteration from the classifier itself and adjusted this so as to improve the classification. When applying PCA, a mean of 1.97% in improvement was obtained in the 6 cases of classification, in the best case improving by 5.8% for the SVM classifier [20].

Lastly, it has been ascertained that applying committees substantially improves the percentage, as it improved the individual percentage obtained by the classifiers that make up the committee. 88.4% was obtained in the case where PCA was not applied and 89.7% when it was applied.

Therefore, the conclusion can be drawn that this tool is a good option for classifying MRI images, permitting a variety of different types of classification and characterising certain classification features, such as validation iterations and the number of subjects to be trained [21, 22]. Furthermore, high classification percentages were obtained by improving them when applying the PCA feature reduction technique, although future lines of research will be able to allow the user of the tool to characterise each classifier, thus enabling the tool to be further optimised.

Moreover, it can conclude that the results obtained are a good basis for enhancing the tool. To improve hClass, more pathologies databases are necessary to obtain a consistent success percentage. Also, for future developments, the tool will provide more opportunities to the user to characterize the classifiers algorithms inside. With that, the tool will be more robust and a perfect tool to classify any pathology.

References

[1] K.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, Cambridge MA, US, 2012.
[2] K. Yoshida and A. Sakurai, Machine learning, in: Encyclopedia of Information Systems, H. Bidgoli, ed., Elsevier, New York, 2003, pp. 103-114.
[3] E. Zhang, F. Wang, Y. Li and X. Bai, Automatic detection of micro-calcifications using mathematical morphology and a support vector machine, Bio-Medical Materials and Engineering **24** (2014), 53-59.
[4] R.W. Swiniarski, Rough sets methods in feature reduction and classification, International Journal of Applied Mathematics and Computer Science **11** (2001), 565-582.
[5] S.B. Kotsiantis, I. Zaharakis and P. Pintelas, Supervised machine learning: A review of classification techniques, Informatica **31** (2007), 249–268.

[6] J.A. Richards, Clustering and unsupervised classification, in: Remote Sensing Digital Image Analysis, Springer, 2013, pp. 319-341.
[7] B. Krishnapuram, D. Williams, Y. Xue, L. Carin, M. Figueiredo and A.J. Hartemink, On semi-supervised classification, Advances in Neural Information Processing Systems **17** (2004), 721-728.
[8] S. Basu, A. Banerjee and R.J. Mooney, Semi-supervised clustering by seeding, Proceedings of the International Conference on Machine Learning **2** (2002), 27-34.
[9] T. Næs, P.B. Brockhoff and O. Tomic, Cluster analysis: Unsupervised classification, statistics for sensory and consumer science, in: Front Matter, John Wiley & Sons, Ltd, Corporate Headquarters, United States, 2010, pp. 249-261.
[10] H. Abdi and L.J. Williams, Principal component analysis, Wiley Interdisciplinary Reviews, Computational Statistics **2** (2010), 433-459.
[11] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of International Joint Conference on Artificial Intelligence **14** (1995), 1137-1145.
[12] G. Holmes, A. Donkin and I.H. Witten, Weka: A machine learning workbench, Intelligent Information Systems, Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, 1994, pp. 357-361.
[13] R. Collobert, S. Bengio and J. Mariéthoz, Torch: A modular machine learning software library, Technical Report IDIAP-RR 02-46, IDIAP, Martigny, Switzerland, 2002, EPFL-REPORT-82802.
[14] K. Patel, N. Bancroft, S.M. Drucker, J. Fogarty, A.J. Ko and J. Landay, Gestalt: Integrated support for implementation and analysis in machine learning, Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, New York, USA, 2010, pp. 37-46.
[15] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao and T. Li, miRecords: An integrated resource for microRNA-target interactions, Nucleic Acids Research **37** (2009), 105-110.
[16] I. Bonet, A. Rodriguez, M.M. Garcia and R. Grau, Combinacion de clasificadores para bioinformatica, Computacion Sistemas **16** (2012), 191-201.
[17] M.S. Alvarez, L.B. Moraña and M.M. Salusso, Evaluation of Water Quality Analysis by Main Components, Localidad de Vaqueros, Salta, Laboratorio de Calidad de Agua, Facultad de Ciencias Naturales, Universidad Nacional de Salta.
[18] S. Teicher and A. Martinez, Diagnosing and Segmentation Brain Tumors and Phenotypes using MRI Scans, CS229 Final Project, 2014.
[19] P. Coupé, S.F. Eskildsen, J.V. Manjón, V.S. Fonov, D.L. Collins and Alzheimer's disease Neuroimaging Initiative, Simultaneous segmentation and grading of anatomical structures for patient's classification: Application to Alzheimer's disease, NeuroImage **59** (2012), 3736-3747.
[20] J. Zhang, W. Cheng, Z. Wang, Z. Zhang, W. Lu, G. Lu and J. Feng, Pattern classification of large-scale functional brain networks: identification of informative neuroimaging makers for epilepsy, PloS One **7** (2012), e36733.
[21] J. Gao, S. Yue, J. Chen and H. Wang, Classification of normal and cancerous lung tissues by electrical impendence tomography, Bio-Medical Materials and Engineering **24** (2014), 2229-2241.
[22] M.A. Fernandez-Granero, D. Sanchez-Morillo, A. Leon-Jimenez and L.F. Crespo, Automatic prediction of chronic obstructive pulmonary disease exacerbations through home telemonitoring of symptoms, Bio-Medical Materials and Engineering **24** (2014), 3825-3832.