Journal of Neurology (2018) 265:2745–2748
https://doi.org/10.1007/s00415-018-8990-9

NEUROLOGICAL UPDATE

Machine learning in neurology: what neurologists can learn from machines and vice versa

Rose Bruffaerts[1,2]

Received: 20 July 2018 / Revised: 26 July 2018 / Accepted: 27 July 2018 / Published online: 2 August 2018

Abstract
Artificial intelligence is increasingly becoming a part of everyday life. This raises the question whether clinical neurology can benefit from these novel methods to increase diagnostic accuracy. Several recent studies have used machine learning classifiers to predict whether subjects suffer from a neurological disorder. This article discusses whether these methods are ready to make their entrance into clinical practice. The underlying principles of classification will be explored, as well as the potential pitfalls. Strengths of machine learning methods are that they are unbiased and very sensitive to patterns emerging from small changes spread across a large number of variables. Potential pitfalls are that building reliable classifiers requires large amounts of well-selected data and extensive validation. Currently, machine learning classifiers offer neurologists a new diagnostic tool which can aid in the diagnosis of cases with a high degree of uncertainty.

Keywords Artificial intelligence · Machine learning · Support vector machines · Diagnostic accuracy · Classification

Introduction

Machine learning is an increasingly popular technique in clinical research, but is there a place for artificial intelligence in clinical practice? In Alzheimer's disease (AD), a seminal multicentric study from 2008 demonstrated that classifiers based on MRI images performed as well as clinicians [1]. Numerous studies have followed since (for reviews in AD, which will be used as an illustrative example: see [2, 3]), but why hasn't machine learning found its way to the bedside 10 years later? This article discusses the strengths and weaknesses of artificial intelligence classifiers. I will focus on classifiers used for automated diagnosis, i.e. the classifier will propose a diagnosis. This goes beyond the use of biomarkers for the confirmation of a clinical suspicion, where the clinician interprets the result in a specific clinical context.

✉ Rose Bruffaerts
rose.bruffaerts@kuleuven.be

[1] Laboratory for Cognitive Neurology, Department of Neurosciences, KU Leuven, Herestraat 49, 3000 Leuven, Belgium

[2] Neurology Department, University Hospitals Leuven, Leuven, Belgium

Relevant machine learning principles

Machine learning classifiers perform pattern recognition. They use a large number of variables as input, which means they are essentially multivariate analysis techniques. The classifier "learns" to predict the outcome (or "class", e.g. healthy or AD) based on a large set of variables (e.g. voxels of an MRI, age, APOE carrier status, health records…). Using a training dataset with a definite diagnosis, the classifier is trained ("learns") to separate between the two classes based on a combination of the input variables (e.g. pattern of atrophy, pace of decline…). An important question is which gold standard to use for classification of the training set (e.g. neuropathology, amyloid PET). Since a bigger training set (including disease mimics, healthy controls, …) increases reliability, sometimes a tradeoff will be necessary between sample size and diagnostic gold standard.

Support vector machines (SVM), currently a popular method, can illustrate how machine learning works. The subsequent discussion of strengths and limitations can be extrapolated to other methods, e.g. neural networks (NN), …. SVMs are typically used for binary classification (e.g. healthy versus AD) whereas NN map onto a large possible number of outcomes. However, multiclass SVMs do exist. To illustrate the underlying principles, I will focus on the binary SVM case. Theoretically, the input variables of the

SVM classifier can be projected to a high-dimensional space, in which each dimension represents one variable. For a classifier trained to separate subjects with AD from healthy controls, input variables such as age, education level, neuropsychological test scores, voxels of the subject's MRI, lumbar puncture results, etc. can be mapped to a high-dimensional space, with each of these variables representing one dimension. In this high-dimensional space, the classifier is the hyperplane which separates the data points belonging to the 2 classes. In two dimensions, this hyperplane can be visualized as a line (Fig. 1). The type of hyperplane is determined by the user. Most frequently, SVMs are set to generate linear hyperplanes (a straight line in 2D—Fig. 1a, or a flat surface in 3D). Allowing non-linear solutions can result in the generation of oddly shaped hyperplanes (a curvy line in 2D—Fig. 1b, or a rippled surface in 3D).

Correct validation of the classifier

A reliable classifier needs to generalize to unseen data. The capacity to generalize implies that the learned differences between the classes in the training dataset are likely due to the underlying pathology or disease process. The classifier should be tested on data that was not part of the training dataset: testing on training data would lead to an overoptimistic classification accuracy [4]. Cross validation is a frequently used technique to demonstrate that a classifier can generalize to new data. During cross validation, all subjects from a given dataset are split up into a number of groups (i.e. "folds", e.g. 5). Each fold is then used once as a test dataset, while the training dataset consists of the other folds. Classification accuracy is based on the average accuracy across the folds.

Along with the classification accuracy, a test of significance should be reported, e.g. a p value. This is necessary because the absolute value of classification accuracy does not provide evidence that the classifier performs better than expected by chance. Random permutation labelling can be used to calculate a p value: the class labels of the dataset are randomly assigned and the classification procedure is repeated over and over (e.g. 10,000 times) to generate a distribution of all theoretically possible accuracies. The true classification accuracy is then compared to the distribution obtained with random labels to derive a p value. Obviously, classifiers used in clinical practice should often perform a lot better than merely being better than expected by chance.

Another important measure of the diagnostic ability of a classifier is the Area Under the Curve (AUC) of the receiver operating characteristic (ROC) curve of the classifier. This is particularly relevant for rare diseases. Recall that the ROC curve plots the false positive rate (= 1 − specificity) versus the true positive rate (=sensitivity). A bad classifier might fail to detect the few true positive cases of a rare disease and label them as negative. This classifier would then have a high accuracy (low absolute number of misclassified cases), but a low sensitivity and small AUC.

A high classification accuracy combined with a significant p value and a large AUC is only one of the basic requirements which determine whether a classifier can generalize to the extent that it can be used in clinical practice. Often, cross validation implies that the test dataset is closely related to the training dataset (same scanner, same population, …). Ideally, the test dataset should be acquired

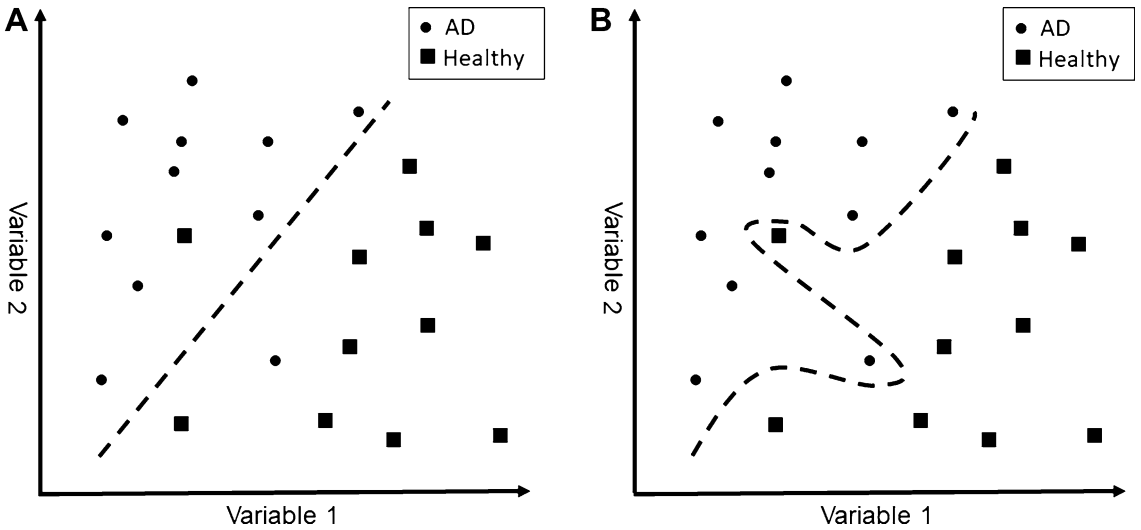

Fig. 1 Linear versus non-linear binary SVMs. The hyperplane (dotted line) visualizes how the classifier separates healthy subjects from AD patients in an example with only 2 input variables (1 variable could be age, 1 voxel of an MRI, a neuropsychological test score, …). The separation can be either (**a**) linear, which is visualized as a straight line in 2 dimensions or (**b**) non-linear

in another center, to demonstrate reliable classification performance. The SVM's settings can also influence generalization: if a non-linear hyperplane is allowed (Fig. 1b), the risk of overfitting is increased. Overfitting means that during training, a solution is found which can optimize classification in the training dataset but cannot generalize to unseen data. Intuitively, one can imagine that the use of non-linear hyperplanes can lead to solutions that are idiosyncratic to the training dataset (Fig. 1b: the hyperplane is locally curved to perfectly separate all data points).

What classifiers can learn

Once a classifier is built which can classify unseen data from another study center or population, can one reliably use it in clinical practice? I would argue against this, and rather seize the opportunity to dissect the classifier. A machine learning classifier becomes more trustworthy if the internal workings, i.e. which variables lead to successful classification, are clear. For instance, age is a risk factor for AD, so it is reasonable to include age as a contributing variable. However, the weight assigned to age cannot be excessively high: even in the case that classification would reach significance based solely on age, this is clearly undesirable. For SVMs, the weight each variable contributes to the classification of the training set can be determined. Similar strategies exist for other machine learning methods, although some are less straightforward, e.g. Deep Taylor decomposition for NN [5]. Scientifically, the inspection of the weights is an opportunity to discover new contributing variables and to discard others. Misclassified cases should be examined: systematic errors (e.g. classification of all subjects under 60 years old as not having AD) are a cause for concern. Misclassification can also reveal that critical input variables are missing. For instance, an amnestic syndrome with hyperacute onset has a low probability to arise due to AD, but misclassification might occur if a "speed of onset" variable is not included.

What neurologists can learn from classifiers

Human supervision is obviously still required to generate a reliable diagnostic classifier. Next, several scenarios are conceivable where missing data undermine the use of the classifier: e.g. a subject who cannot have an MRI scan, an illiterate subject unable to perform certain cognitive tests etc. Does this mean that it is still too premature to implement machine learning classifiers into clinical practice? No, since classifiers can provide useful additional information to the clinical neurologist. A strength of machine learning is its sensitivity to pick up patterns arising from a combination of small effects spread across a high number of variables. Such patterns can be hard to detect for a human observer. Classifiers might facilitate decision making when a neurologist is faced with a high degree of uncertainty, e.g. a suspicion of in AD in a patient with comorbidities in the absence of CSF/PET biomarkers. In such a situation, a classifier might be useful to translate the clinician's suspicion into objective measures by quantifying the presence or absence of variables contributing to a diagnosis. For clinical research purposes, studies often report contributing variables and their respective weights from the training dataset. For implementation in clinical practice, it would be very useful to have the classifier explicitly list which contributing variables are driving classification for the index patient. Second, humans are inherently prone to a certain amount of bias, which does not affect artificial intelligence. Relevant to diagnostics is the concept of "confirmation bias", i.e. the tendency to interpret new findings (for instance the result of a diagnostic test) in a way that confirms one's hypothesis [6].

Conclusion

Machine learning classifiers are ready to be developed as a tool which can assist clinical neurologists to make a diagnosis, but they cannot replace clinical judgement. If clinicians were to introduce classifiers to aid them in decision making, they are advised to examine the validation method and internal workings of the classifier to judge its reliability. An important step in quality control entails elucidating which input variables drive classification. Currently, the lack of large multicentric samples and the required amount of validation are slowing down the introduction of classifiers in everyday clinical practice. However, in combination with the advancement in data mining of electronic health records, the diagnostic accuracies of classifiers will continue to rise in the next decade. Taking everything into account, machine learning techniques have the potential to become very powerful diagnostic tools and reliable aids to clinical neurologists.

Acknowledgements The author thanks Bruno Bergmans, M.D., Ph.D., for useful comments on prior versions of this paper.

Funding RB is a postdoctoral fellow of the Research Foundation Flanders (F.W.O.)

Compliance with ethical standards

Conflicts of interest Dr. Bruffaerts reports no disclosures.

References

1. Kloppel S, Stonnington CM, Chu C et al (2008) Automatic classification of MR scans in Alzheimer's disease. Brain 131:681–689. https://doi.org/10.1093/brain/awm319
2. Rathore S, Habes M, Iftikhar MA et al (2017) A review on neuroimaging-based classification studies and associated feature

extraction methods for Alzheimer's disease and its prodromal stages. NeuroImage 155:530–548. https://doi.org/10.1016/j.neuroimage.2017.03.057
3. Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017) Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. NeuroImage 145:137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079
4. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540. https://doi.org/10.1038/nn.2303
5. Montavon G, Lapuschkin S, Binder A et al (2017) Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 65:211–222. https://doi.org/10.1016/j.patcog.2016.11.008
6. Duncan J (2010) How intelligence happens. Yale University Press, New Haven