Journal of Alzheimer's Disease 27 (2011) 1–3
IOS Press

Supplementary Data

A Disease State Fingerprint for Evaluation of Alzheimer's Disease

Jussi Mattila[a,*], Juha Koikkalainen[a], Arho Virkki[a], Anja Simonsen[b], Mark van Gils[a], Gunhild Waldemar[b], Hilkka Soininen[c], Jyrki Lötjönen[a] and for The Alzheimer's Disease Neuroimaging Initiative**
[a]*VTT Technical Research Centre of Finland, Tampere, Finland*
[b]*Department of Neurology, Section 2082, The Copenhagen Memory Clinic & The Memory Disorders Research Group, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark*
[c]*Department of Neurology, Kuopio University Hospital, Kuopio, Finland*

Accepted 26 May 2011

SUPPLEMENTARY DATA

Derivation of the fitness function for scalar variables

Suppose first that x is a random variable from a distribution combining both control (e.g. healthy) and positive (e.g., disease) subjects, marked with C and P. In addition, assume that the progression of disease increases the observed values of x, making the conditional expected value $E(x|P)$ higher for positives than the corresponding value $E(x|C)$ for the controls (see Fig. 1).

Let us divide the probability density f into the components f_C and f_P, such that

$$\begin{aligned} f(x) &= f_C(x) + f_P(x) \\ &= f(x|C)p(C) + f(x|P)p(P) \end{aligned} \quad (1)$$

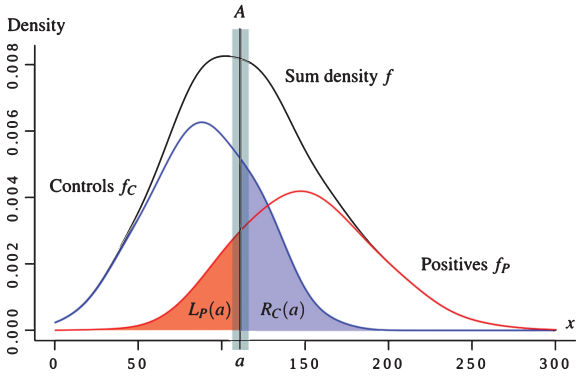

Fig. 1. Probability density function $f(x)$ and its components f_C and f_P for the control and positive groups and, respectively.

where $f(x|C)$ and $f(x|P)$ are the marginal distributions of C and P, and the probabilities $p(C)$ and $p(P)$ correspond to the overall fraction of controls and positives in the study population, respectively. These values are related by the equation $\int_R f(x)dx = \int_R \left[f_C(x) + f_P(X)\right] dx = p(C) + p(P) = 1$, obtained by integrating (1).

Bayes' theorem states that the conditional probability of a subject belonging to the group P after observing $x \in A = (a - \varepsilon; a + \varepsilon)$, where ε is the radius of a small region around the actual observation a, can be written as

** Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf.

*Correspondence to: Jussi Mattila, VTT Technical Research Center of Finland, Tekniikankatu 1, FIN-33101, Tampere, Finland. Tel.: +358 40 592 7979; Fax: +358 20 722 3499; E-mail: jussi.mattila@vtt.fi.

ISSN 1387-2877/11/$27.50

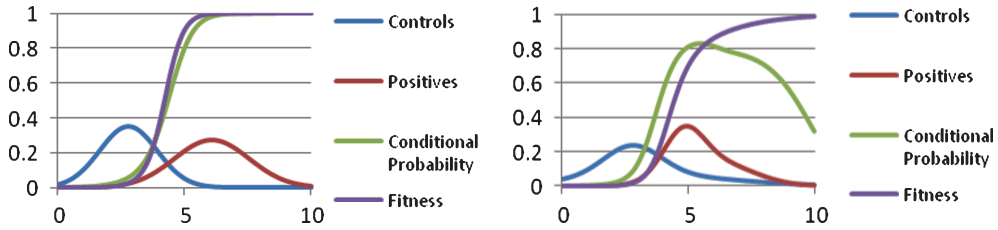

Fig. 2. Conditional probability and *fitness* computed with two distributions of control and positive cases. On the left, synthetic data with no outlier observations produces monotonously increasing curves with both methods. On the right, data with wide distribution tails causes drastic change to conditional probability behavior, rendering it sub-optimal for human interpretation.

$$p(P|x \in A) = \frac{p(x \in A|P)p(P)}{p(x \in A)} = \frac{p(x \in A|P)p(P)}{p(x \in A|P)p(P) + p(x \in A|C)p(C)} \to \frac{f_P(x)}{f_P(x) + f_C(x)}, \quad (2)$$

when $\varepsilon \to 0$. While estimating the probability densities from empirical data, one needs to smooth the estimates e.g. by using a sufficiently wide kernel estimate, or to use a large enough ε to compensate for the measurement noise and errors caused by the finite number of samples. If ε is chosen to be too large, $p(P|x \in A)$ approaches the a priori fraction of the positive cases $p(P)/(p(P) + p(C))$, and equation (2) loses its predictive power. The values $p(P)$ and $p(C)$, also called the "a priori" probabilities, are needed when applying the Bayes' rule. This can be a great asset, but could also be regarded as a drawback when used or interpreted incorrectly. In addition, distributions of the form (2) have also some inconvenient properties, which can make their interpretation difficult, as shown later in Fig. 2.

Instead of using conditional probability (2), let us introduce a *fitness* function *Fit(a)*, which increases monotonously. In a sense, *fitness* describes the location of the subject with value a relative to distributions f_C and f_P. Let us first define the left and right integrals for f_C and f_P,

$$L_P(a) := \int_{-\infty}^{a} f_P(x)\,dx \quad \text{and}$$
$$R_C(a) := \int_{a}^{\infty} f_C(x)\,dx, \quad (3)$$

which are also illustrated in Fig. 1. For completeness, $R_P(a)$ and $L_C(a)$ are defined in an analogous manner. If one consider value a as the clinical threshold for classification between the controls C and positives P, one can construct a new boolean classifier, described in Table 1.

Table 1
Classification performance when using the value a as the threshold to discriminate between controls C and positives P

	False Negatives	True Positives
P	$L_P(a)$	$R_P(a)$
	True Negatives	False Positives
C	$L_C(a)$	$R_C(a)$
	$x \le a$	$x > a$

Table 1 shows that $P(x \le a) = L_C(a) + L_P(a)$ and $P(x > a) = R_C(a) + R_P(a)$ for the columns and $P(P) = L_P(a) + R_P(a)$ and $P(C) = L_C(a) + R_C(a)$ for the rows. In particular, the fraction of rejection errors (false negatives) from all the errors (both false negative and false positive) can be written as

$$Fit(a)^*: = \frac{FN(a)}{FN(a) + FP(a)} = \frac{L_P(a)}{L_P(a) + R_C(a)}, \quad (4)$$

where the abbreviations *FN* and *FP* refer to false negatives and positives, i.e. the counts of incorrectly classified instances. It is obvious from equation (4) that $Fit(a)* \in [0, 1]$ and one can intuitively expect that $Fit(a)^*$ increases along with increasing values of a, which is proved by differentiating (4):

$$\frac{d}{da} Fit(a)^* = \frac{\frac{d}{da}L_P(a)R_C(a) - L_P(a)\frac{d}{da}R_C(a)}{[L_P(a) + R_C(a)]^2} = \frac{f_P(a)R_C(a) + L_P(a)f_C(a)}{[L_P(a) + R_C(a)]^2} \ge 0 \quad \text{for each } a, \quad (5)$$

In the special case $L_P(a) = R_C(a) = 0$ where (5) is not defined, the result can be interpolated from closest values of a where (5) is defined. Finally, to eliminate the influence caused by varying proportions of $p(P)$ and $p(C)$ between different populations, the normalized *fitness* value is defined as

$$Fit(a) := \frac{L_P(a)/p(P)}{L_P(a)/p(P) + R_C(a)/p(C)} \quad (6)$$

Derivation of the *fitness* function can be conducted in an analogous manner if populations are interchanged, resulting in a monotonously decreasing function. In addition, alternate formulations of *fitness* functions to account for non-continuous variables, such as nominal and ordinal variables, can also be derived easily by counting these values as point masses while computing the integrals. The resulting *fitness* values obtained by evaluating (6) are in many situations close to the conditional probabilities (2) but *fitness* behaves in a more intuitive manner with real-life empirical distributions, as demonstrated in Fig. 2. For example, it is known that atrophy decreases the size of hippocampus in Alzheimer's disease; the smaller the size of hippocampus the higher the DSI value should be, i.e., the function should be monotonous. However, if the number of cases in the training set is small and conditional probabilities are used, posterior probability can decrease even while the hippocampus volume is decreasing.

It merits restating that the number of instances in either class of the training set does not bias *fitness* (6), which makes it robust against disparity between the numbers of class instances. Since the *fitness* values are not intended to be used solely as a machine learning classifier but accompanied with visual analysis tools, this choice offers more intuitive ratings for measured values. Additional information related to the class probabilities, i.e., disease incidence and prevalence, should be presented to clinicians via the graphical user interface.

Derivation of the composite Disease State Index

In clinical practice, multiple variables must be considered simultaneously. Combining results from several *fitness* functions would allow evaluation of large quantities of heterogeneous patient data at once. Due to its simplicity and interpretability, the weighted arithmetic mean is employed for combining *fitness* values. Let us define the composite Disease State Index (DSI) as

$$DSI(a_1, a_2, \ldots, a_n) := \frac{\sum_{i=1}^{n} \mathrm{w}_i Fit(a_i)}{\mathrm{w}_1 + \mathrm{w}_2 + \cdots + \mathrm{w}_n}, \quad (7)$$

where $[a_1, a_2, \ldots, a_n]$ are the data measured from the subject and $\mathrm{w} = \mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_n$ are the non-negative weights for each of the variables according to their *relevance*. *Relevance* is a parameter quantifying a variable's ability to differentiate classes C and P. To compute the *relevance* of the ith variable, the classification accuracy is estimated by applying the *fitness* function to the training data itself:

$$Acc(i) = \frac{|C_T : Fit(a_i) < \frac{1}{2}| + |P_T : Fit(a_i) > \frac{1}{2}|}{|C_T| + |P_T|} \quad (8)$$

where C_T and P_T are the corresponding training sets for the controls and positives (with the i:th variable present) and $\frac{1}{2}$ is the classifier threshold value for a. Now, *relevance* of a variable is formally defined as

$$Rel(i) := \max\left\{0, \left(Acc(i) - \frac{1}{2}\right) * 2\right\}. \quad (9)$$

If the *relevance* is zero, it discriminates the classes as poorly as a random label. A *relevance* of one indicates that the variable is capable of fully discriminating between training classes C and P, thus being an excellent candidate for estimating the disease state. Substituting w_i in (7) with (9) yields

$$DSI(a_1, a_2, \ldots, a_n) := \frac{\sum_{i=1}^{n} Rel(i) Fit(a_i)}{\sum_{i=1}^{n} Rel(i)}. \quad (10)$$

It is clear from (10) that like *fitness*, composite $DSI(a_1, a_2, \ldots, a_n) \in [0, 1]$. It must be emphasized that DSI cannot be considered as the probability of having the disease. Instead, it is a score that increases with the probability of having the disease, taking into account the assumption that having abnormally high (or low) values is worse than being inside the normal range. Thus, DSI is defined as a value derived from a series of observed facts that describes the rank of patient data relative to control and positive cases.