Cogn Comput (2011) 3:490–500
DOI 10.1007/s12559-010-9052-5

Developing Prognosis Tools to Identify Learning Difficulties in Children Using Machine Learning Technologies

Antonis Loizou · Yiannis Laouris

Received: 30 January 2010 / Accepted: 7 June 2010 / Published online: 25 June 2010

Abstract The Mental Attributes Profiling System was developed in 2002 (Laouris and Makris, Proceedings of multilingual & cross-cultural perspectives on Dyslexia, Omni Shoreham Hotel, Washington, D.C, 2002), to provide a multimodal evaluation of the learning potential and abilities of young children's brains. The method is based on the assessment of non-verbal abilities using video-like interfaces and was compared to more established methodologies in (Papadopoulos, Laouris, Makris, Proceedings of IDA 54th annual conference, San Diego, 2003), such as the Wechsler Intelligence Scale for Children (Watkins et al., Psychol Sch 34(4):309–319, 1997). To do so, various tests have been applied to a population of 134 children aged 7–12 years old. This paper addresses the issue of identifying a minimal set of variables that are able to accurately predict the learning abilities of a given child. The use of Machine Learning technologies to do this provides the advantage of making no prior assumptions about the nature of the data and eliminating natural bias associated with data processing carried out by humans. Kohonen's Self Organising Maps (Kohonen, Biol Cybern 43:59–69, 1982) algorithm is able to split a population into groups based on large and complex sets of observations. Once the population is split, the individual groups can then be probed for their defining characteristics providing insight into the rationale of the split. The characteristics identified form the basis of classification systems that are able to accurately predict which group an individual will belong to, using only a small subset of the tests available. The specifics of this methodology are detailed herein, and the resulting classification systems provide an effective tool to prognose the learning abilities of new subjects.

Keywords Automated prognosis tools · Learning difficulties · Machine Learning

A. Loizou (✉) · Y. Laouris
Cyprus Neuroscience and Technology Institute, Nicosia, Cyprus
e-mail: antonis@cnti.org.cy

Y. Laouris
e-mail: laouris@cnti.org.cy

Introduction

Learning is a complex process and although it is hard to quantify, research on difficulties in learning has led to an understanding on what kind of processes may be related with it. Learning difficulties are defined as referring to a number of disorders, which may affect the acquisition, organisation, retention, understanding or use of verbal or other information [5]. Such disorders may be diagnosed through an assessment of difficulties in reading, performance on verbal IQ and non-verbal IQ measures, as well as difficulties in oral and written speech.

However due to the large number of factors involved in identifying learning difficulties, the integration of the information provided by the various tests and measures in order to provide a reliable diagnosis is a non-trivial process. To that end, the work presented herein is carried out with the goal of automating this process through the use of established Machine Learning technologies. The solid mathematical background of these techniques, as well as their successful application in various fields provides confidence in their ability to discriminate between different types of subjects based on large and complex sets of observations. Moreover, their application can also provide valuable insight on which tests can be considered the most reliable in identifying learning difficulties.

Machine Learning techniques have also been previously applied in the area of analysing people's learning

performance. The algorithms used range from simple decision tree–based methods [6] to custom built algorithms designed for purpose [7]. The focus of these previous efforts seems to lie in one of two broad categories: the prediction of the learners performance [6, 7, 8, 9] and the identification of the most successful teaching strategies [10, 11, 12]. Typically, the data driving these systems is generated by the assessment methods of the institutions the learners belong to happen to be using. Instead, the work presented in this paper relies on data generated using well-established tests used to assess learning abilities in children. Another element of novelty of this work is that it simultaneously addresses the goals of both categories of identified previous research. That is the case since we have developed accurate prognosis tools for identifying learning difficulties by assessing the efficiency of indicators provided by the various tests, and combining the best ones in a Machine Learning framework.

An existing dataset [1, 2] of the scores achieved by 134 children in a variety of established tests designed to measure their learning abilities was used in order to develop the tools presented in this paper. Namely the tests used are the Mental Attributes Profiling System (MAPS) [15], the Wechsler Intelligence Scale for Children (WISC) [3], the rapid naming test developed by Wimmer et. al. [13], and Woodcock's Reading Mastery Test [14].

The work reported on in this paper rests on the assumptions that the 134 children who volunteered to participate in the tests are a representative sample of the population and that the four aforementioned tests sufficiently capture the learning abilities of the participating children. The underlying hypothesis of this work is that a group of weak learners can be identified based on the data and that membership to this group can be assessed using a much smaller set of tests. In showing that the hypothesis holds a set of automated tools which are able to identify children with learning difficulties based on a minimal set of tests is obtained. Moreover, their early identification facilitated by these tools is seen as an opportunity to further support the children in overcoming their learning difficulties.

The paper is organised as follows. First, the following section provides an overview of the various tests administered to the 134 children. Section 3 then gives a concise description of the Machine Learning technologies used in producing the experimental results presented in Sect. 4. Final remarks are provided along with the conclusions drawn in Sect. 5.

Dataset Description

The dataset consists of 134 children who took part in a study consisting of various tests designed to assess their learning abilities [15]. The participants came from three age groups: 44 were 7–8 years old, 44 participants were 9–10 years old, and 46 were 11–12 years old. Participants came from 16 regular elementary schools, equally sampled from urban, suburban, and rural public schools in Cyprus. The various tests administered in the context of that study along with the measurements recorded are provided in this section.

Mental Attributes Profiling System (MAPS)

The MAPS cognitive test is a battery of validated computer-based video-game type tests that assess the learning abilities of pre-elementary and elementary age school children.

Categorisation: The test presents an object on the lower part of the screen and invites the subjects to drag it in one of three squares that represent different "worlds" for which there is a match. The following "categories" were tested:

1. Objects of different colour to be placed in one of three squares of the same colour.
2. Geometrical shapes to be placed in squares containing other shapes of the same type.
3. A plant to be classed as vegetable, tree, or flower.
4. An animal to be placed in its suitable environment: sea, sky, or open fields.
5. Objects usually found in the home to be placed in the appropriate room (office, kitchen, or bathroom).

The software records whether the categorisation was correct, along with the time taken by the child to respond.

Lateral awareness: This test provides two types of measures. One, it evaluates the children's ability to make left–right discriminations on their own bodies. During the first part, the test shows a child 'sat' in the same orientation as the subject (i.e., the subject sees the back of the child on the screen) in front of two objects, one on the upper left, and the other on the upper right of the visual field. The subject is instructed to 'grab' an object by clicking on the left or right shoulder of the child displayed on the screen. The time taken to select an arm and whether or not the arm selected was the correct one are measured. The same procedure is repeated during the second part of the test, in which the orientation of the child on the screen is reversed, i.e., the child on the screen is facing the subject. The second type of measures are derived from Piaget's [16] tests to evaluate awareness of right–left relations outside the body.

Navigation: The navigation test consists of an 8x8 matrix of small pieces of cheese and a mouse.

The subject is verbally instructed to move the mouse in one of eight possible directions to 'eat' the corresponding piece of cheese. The software measures the number of correct responses and the number of trials carried out.

Sequencing: In this test, different objects or animals appear split in two, three, four, or five pieces, and the subject is requested to 'drag' the pieces and place them in the right order to complete the picture. The second part of the test presents pictures, which represent different stages of a temporal process. The subject is expected to put them in the correct chronological order. The test measured the time taken to complete each section of the test, which was comprised of six different types of exercises along with whether the subject assembled the image correctly or not.

Visual memory: A grid of cards is presented on the screen face down, and the subject is able to turn over pairs of cards. If the two cards feature the same picture, they would remain face up. Otherwise, they are turned back as they were, and the game continues until all cards are uncovered. The time taken to complete the test, and the number of cards turned over are recorded.

Visual discrimination: A group of three pictures with minor differences are presented to the subject, along with an additional picture which is identical to one of the pictures in the group. The subject is asked to select the matching picture from the group. The exercise is repeated four times, and the test records whether the correct picture has been selected along with the time taken to do so.

Auditory memory: The test was modelled using the digital phone metaphor. The subject is invited to dial a number. Two sets of two-digit numbers are followed by a three, four, five, and six digit number. It concludes by presenting a set of two seven-digit numbers. The test is terminated if the subject makes 3 consecutive errors. The number of correctly dialled sequences, as well as the number of correct digits for each number, is recorded.

Auditory discrimination: The main screen of the test features two human-like figures, who speak a word, one after the other. The subject is asked to decide whether the two words are the same or different by clicking on a ✔ or a ✘ symbol. Each word includes consonants which sound similar and are therefore confused by weak readers, especially by dyslexics in the Greek language. The following consonant combinations were tested: $\varphi - \beta$, $\delta - \vartheta$, $\zeta - \sigma$, $\chi - \gamma$, $\tau - \nu\tau$, $\kappa - \gamma\gamma/\gamma\kappa$, $\pi - \mu\pi$, $\tau\sigma - \sigma\tau$, $\gamma - \gamma\gamma/\gamma\kappa$ and $\xi - \kappa\sigma$. The test also evaluates the ability of the child to differentiate between the same letter combinations when they used in random strings of letters. The test keeps record of the time taken to respond and the correctness of each response.

For more details regarding the structure of the video-game interfaces used in the MAPS cognitive tests and the parameters measured, refer to [1, 15]. Moreover, the MAPS measurements are found to be accurate predictors of reading ability [17].

Wechsler Intelligence Scale for Children (WISC)

The Wechsler Intelligence Scale for Children (WISC) [18] is a measure for testing intelligence in children aged 6–16 years old. It is composed of ten core sub-tests and five supplemental ones through which verbal abilities and performance are assessed. The supplemental sub-tests are used to accommodate children in certain rare cases, or to make up for spoiled results which may occur from interruptions or other circumstances. None of the supplemental sub-tests have been administered to the subjects of this study, as that would give rise to different types of the data for children who completed some of the supplemental tests.

The ten core sub-tests are split into four categories: Verbal Comprehension (VCI), Perceptual Reasoning (PRI), Processing Speed (PSI), and Working Memory (WMI).

For the purposes of this study, only two VCI sub-tests were used from the WISC to assess verbal IQ in the participating children:

Vocabulary: The children are asked to describe the meaning of words presented to them.

Similarities: The children are asked to identify the relationship between two concepts.

In each case, the number of correct responses were recorded. The third and final VCI sub-test, **Comprehension**, has not been administered, as previous studies [19] have shown a large variance in the scores assigned by different judges to the same responses.

Moreover, the PRI, PSI, and WMI sub-tests have also not been administrated as they are very similar to certain MAPS tests, and the researchers [1] felt that their inclusion would increase the risk of children losing interest in completing highly similar tests causing them to underperform.

Rapid Naming

The test is modelled after Wimmer et al. [13], consists of presenting the children with objects, which they are asked to name. The test is a two-stage process, with an increase in the degree of difficulty at the second stage. Two different adaptations of the test have been completed by the children:

Rapid naming of pictures: The subject is asked to name the objects depicted in a random sequence of 20 images, consisting of 5 different images that are each repeated four times. The images were presented on a single page,

with four lines of the same five objects ordered differently. The names of the objects presented in the first stage were words which start with the same single consonant cluster (e.g. $\kappa\alpha\pi\acute{\epsilon}\lambda o, \kappa\alpha\rho\acute{\epsilon}\kappa\lambda\alpha, \kappa\epsilon\rho\acute{\alpha}\sigma\iota$, $\kappa\alpha\rho\acute{\delta}\tau o, \kappa\lambda\epsilon\iota\delta\acute{\iota}$). The second stage consisted of objects whose names started with different consonant clusters (e.g. $\varphi\rho\acute{\alpha}ουλα, \pi\lambda\upsilon\nu\tau\acute{\eta}\rho\iota o, \sigma\kappa\acute{\upsilon}\lambda o\varsigma, \sigma\tau\alpha\upsilon\rho\acute{\delta}\varsigma, \mu\pi\alpha\nu\acute{\alpha}\nu\alpha$).

Rapid naming of letters: In each stage, the children were asked to name as fast as possible a random sequence of 20 letters appearing on a single page (5 different letters, each repeated four times). Only vowels were included in the first stage (α, η, ε, o, υ), while the second stage consisted of consonants which share similar characteristics and are usually confused by poor readers in Greek (π, τ, σ, δ, ϑ). The child had to say the name of the letter and not the sound that it makes, for an answer to be recorded as correct.

For each task, the time taken by the child to respond as well as the number of correct responses have been recorded.

Woodcock's Reading Mastery Test

The subjects' reading ability was assessed through two different tasks involving the reading of real words and pseudo-words. Both reading measures are Greek adaptations of Woodcocks Reading Mastery Test Revised [14] and have been used in previous studies [20, 21]. In both tests, the participants' score was the number of words read correctly within a minute.

Word identification: The test consists of 85 words forming a $2 \times 2 \times 2$ factorial design in terms of frequency (high/low), orthographic regularity (regular/exception), and length (bi-syllable/tri-syllable). Due to the absence of standard frequency counts in Greek, half of the words were sampled from the first and second grade language books, and the other half from third and fourth grade language books.

Word attack: The subjects were asked to read 45 pronounceable pseudo-words that were derived from real words after changing two or three letters (either by substituting them or using them backwards). The degree of difficulty was incrementally raised, as the test started with words consisting of two syllables, while the final words consisted of five.

Methodology

A description of the Machine Learning algorithms used to produce the results given in Sect. 4 is provided in this section. The descriptions are motivated by the aim to provide an intuition of the inner working of each algorithm, rather than to provide an exhaustive explanation of the specific details associated with each one. Thus, the algorithms are described in an as concise and clear a manner as possible, while the interested reader is directed to the referenced bibliography.

Principal Component Analysis (PCA)

PCA [22] is an extremely powerful method employed in analysing multivariate datasets. In mathematics, it is defined as an orthogonal linear transformation that projects the observations on a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

A useful analogy is the following: Imagine viewing a set of 2-D shapes from an angle perpendicular to the edge of the 2D surface. In this setting, it would be extremely difficult to differentiate between the various shapes, as they will appear as straight line segments. What PCA accomplishes is the identification of a new coordinate system whereby the viewing angle is shifted to be perpendicular to the face of the 2D surface allowing the full structure of the shapes to become visible.

With relation to real world datasets, PCA considers each repetition of an experiment as a point in a multi-dimensional space, with the number of dimensions equal to the number of observations recorded each time the experiment is carried out. In the context of this work, the experiment consists of administrating the tests described in the previous section. The experiment is repeated with each participating child, to obtain a point whose coordinates in the multi-dimensional space are given by the scores they receive in each test. The method proceeds by identifying a new set of variables equal in size to the set of original variables. Each principal component (or constructed variable) consists of a linear combination of all the original variables in such a way as to project the greatest differences between data points (in this case children) onto the first principal component, while the last few contain information that is highly similar across all data points. Moreover, PCA is theoretically proven to be the optimum linear transform for set of data, in terms of least square errors. As such the benefits that arise from the application of this method are 2-fold:

1. The data are projected onto a new set of coordinates that is *optimal* in discriminating between the data points.

2. The last few constructed variables can be safely ignored, since they contain information that is shared across the various data points. This results in a dataset that is smaller and easier to process.

In the interest of completion, we provide a short, formal description of the method. Given an $m \times n$ data matrix M, we can obtain a linear decomposition of the form:

$$M = U\Sigma V^T$$

The superscript ^{T} denotes the conjugate transpose of a matrix. Now, U is an $m \times m$ matrix whose columns are the *eigenvectors* of MM^T, and V an $n \times n$ matrix whose columns are the *eigenvectors* of M^TM. Finally, Σ is an $m \times n$ matrix of singular values (the square roots of the *eigenvalues* of MM^T), giving this type of decomposition the name *Singular Value Decomposition*. The projection Y of the original data matrix M obtained through PCA is then given by:

$$\begin{aligned} Y^T &= M^T U \\ &= V\Sigma \end{aligned}$$

Self Organising Maps (SOM)

Tuevo Kohonen's Self Organising Maps [4] algorithm is a particular type of artificial neural network, partly inspired by the way different types of sensory information are handled in separate parts of the cerebral cortex in the human brain. The version of the algorithm implemented in the context of this work uses a rectangular lattice of hexagonal neurons like the one shown in Fig. 1. This referred to as the *map*. The objective of the algorithm is to train different regions of the map to respond to different types of stimuli.

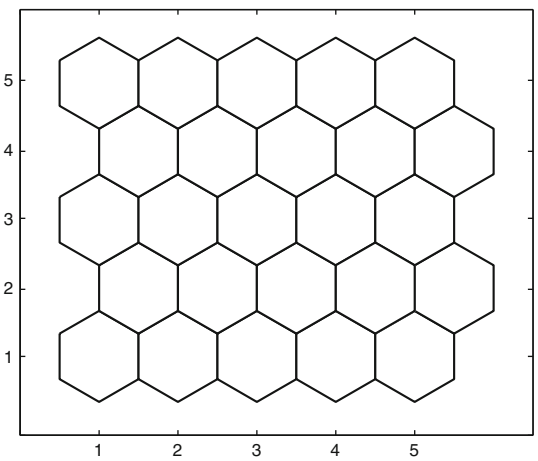

Fig. 1 A Self Organising Map is represented as a rectangular lattice of hexagonal neurons. Each hexagon represents a neuron, and shared edges represent connections between neurons

To do so, each neuron is associated with a *weight vector* of size equal to the number of variables recorded during an experiment. These weight vectors are initialised using small values randomly selected from a Gaussian distribution with 0 mean and 1 standard deviation. Each available data point is then presented to the map, which evolves in response in the following way. When a given data point (arising from experimental observations) is presented to the map, the neuron whose weight vector is the closest (in the multi-dimensional space) is first identified. This neuron is called the *Best Matching Unit (BMU)* with respect to that particular data point and subsequently, adapts its weight vector so that it moves even closer to the data point. Moreover, the adaptation is then propagated to other neurons on the map. The degree to which each neuron adapts depends on its distance from the BMU on the lattice which forms the map. This process of adaptation is called the *training phase* for a SOM and is described by the following equation for a particular neuron v:

$$W_v(t+1) = W_v(t) + \theta(v,t)\alpha(t)(D(t) - W_v(t))$$

In our case, each data point is presented in random order to a 5×5 map. The number of training cycles is set empirically by running the algorithm a large number of times and identifying when the BMU for each case ceases to change in subsequent training cycles. With respect to the dataset used in this study, it was found that the BMUs corresponding to each datapoint remain the same after 2000 training cycles. As such, the variable t keeps track of time, increasing by one each time a data point is presented to the map. $W_v(t)$ is then the weight vector of neuron v at time t, while $D(t)$ encodes for the datum presented to the map at time t. Thus, $(D(t) - W_v(t))$ gives the distance (in multi-dimensional space) between the datum $D(t)$ and (the weight vector of) neuron v. Therefore, if this quantity is added in full to the weight vector of neuron v, W_v, it will cause its displacement to coincide with $D(t)$. However, this displacement is dampened through the following two functions:

The neighbourhood function, $\theta(v,t)$: This function describes the intuition that the amount to which a neuron adapts in response to a given data point should depend on its distance from the BMU on the map. Neurons that are close to the BMU should adapt more, while others that reside in more distant areas of the map less. In our work, the neighbourhood function is a Gaussian distribution centred at the BMU. Initially, the standard deviation of this distribution is equal to the size of the map, so that all neurons will be affected by each datapoint. As t grows large, the standard deviation is reduced until it reaches 0 at the final cycle, where only the BMU adapts in response to an input.

The learning rate, $\alpha(t)$: In order for the map to converge to a stable state after the training phase, the neurons will have to adapt less in the latter cycles than at the beginning. There is a wide range of functions available to encapsulate this, and in the context of this work, we selected a linear function of the form:

$$\alpha(t) = \alpha(0)\frac{T - t + 1}{T}$$

where T is the total number of cycles, and $\alpha(0)$ is the initial learning rate.

As described above, both functions describe quantities that reduce as the training of the map proceeds. At the beginning, when the neighbourhood is broad and the learning factor large, the self-organising takes place on the global scale. When the neighbourhood has shrunk to just a couple of neurons and the learning factor becomes small, the neuron weights are converging to local estimates.

The outcome of this process is a map segmented into different areas, each trained to respond to data of a different type. Borders between the various areas are represented by neighbouring neurons with significantly different weight vectors. By presenting the data to the map one last time, and without altering the structure of the map, we can record which area of the map responds to each data point and in this way split the dataset into groups. This is called the *mapping phase* of the SOM.

This process of identifying groups is beneficial in that it eliminates any *a priori* assumptions on the nature of the groups, and the characteristics that define them. As such the groups are entirely emergent from the collected data, and independent of any classification bias typically associated with data processing carried out by humans. Once the process has been completed, one can then probe the different groups to discover the specific ways in which they significantly differ.

Bayesian Classification

The aim of the work presented in this paper is to develop effective prognosis tools to use in the early identification of learning difficulties in children. Thus, the task is ultimately one of classification, with the aim of using the least number of parameters possible. Bayesian classification provides a powerful tool to do this [23], and this section will provide a brief overview of the method's inner workings.

In formal terms, the probability model for a classifier is a conditional model of the form:

$$P(C|V_1, \ldots, V_n)$$

where the dependent variable C encodes the class of a particular object, and the feature variables $V_1 \ldots V_n$ are recorded through experimental observation. Bayes' theorem states that:

$$P(C|V_1, \ldots, V_n) = \frac{P(C)P(V_1, \ldots, V_n|C)}{P(V_1, \ldots, V_n)}$$

The various terms of the equation above are defined as follows:

$$posterior = \frac{prior \times likelihood}{evidence}$$

- The posterior probability, $P(C|V_1,\ldots, V_n)$, is the probability that the objects belong to class C, given that the variables $V_1 \ldots V_n$ have been observed.
- The likelihood, $P(V_1,\ldots, V_n|C)$, gives the probability of an object which is known to belong to class C to exhibit the observed features $V_1,\ldots, V_n$.
- The evidence, $P(V_1,\ldots, V_n)$, is the probability of observing an object that exhibits the observed features $V_1,\ldots, V_n$, regardless of the class it belongs to.

For practical purposes, the denominator of the fraction (the probability of the evidence) can be safely ignored. This is so as it remains constant for each individual object, and the classification occurs by identifying the class with the largest probability to have generated the object.

Now, the numerator of the fraction is equal to the joint probability model $P(C, V_1,\ldots, V_n)$. By repeatedly applying the definition of conditional probability, we obtain:

$$\begin{aligned} P(C, V_1, \ldots, V_n) &= P(C)P(V_1, \ldots, V_n|C) \\ &= P(C)P(V_1|C)P(V_2, \ldots, V_n|C, V_1) \\ &= P(C)P(V_1|C)P(V_2|C, V_1) \\ &\quad P(V_3, \ldots, V_n|C, V_1, V_2) \\ &\quad \vdots \\ &= P(C)P(V_1|C)P(V_2|C, V_1) \\ &\quad P(V_3|C, V_1, V_2) \cdots P(V_n|C, V_1, \ldots, V_{n-1}) \end{aligned}$$

Typically, Bayesian classifiers are designed under a strong independence assumption stating that each observed variable is independent of any other. Although this assumption is incorrect in most cases (the value of one variable depends on the values taken by others), it simplifies the model greatly:

$$P(C, V_1, \ldots, V_n) = P(C)\prod_{i=1}^{n} P(V_i|C)$$

Bayesian classifiers that use this assumption are commonly called 'naive' Bayes classifiers, as the assumption is most likely incorrect. However in our dataset, 226 variables are used to describe the

performance of each child. This number is small in computational terms, and the assumption needs not be made.

Summary

To summarise, the Machine Learning framework used in this study comprises of three key steps:

1. Utilising PCA to reduce the number of variables contained in the original dataset to a smaller set of combined variables in the projected dataset while retaining most of the information contained within it. The projected dataset makes the application of the SOM unsupervised clustering algorithm both easier and more effective. It becomes easier as the smaller number of variables reduces the computational requirements of the algorithm, and more effective as the first principal components (which form the new set of variables) encode the largest differences between subjects.
2. Applying the SOM algorithm to identify clusters in the population of subjects. The identified clusters however do little in determining which of the original variables to assign a subject into a cluster, since they are computed based on the first principal components.
3. Using the clusters identified by the SOM as class labels for the development and evaluation of Bayesian classifiers using the original dataset, and thus enabling the identification of the tests (giving rise to the original variables) that can be assessed in order to assign subjects to clusters.

Experimental Results

This section provides a detailed description of the results obtained by applying the techniques presented in the previous section to our dataset of 134 children. A total of 226 variables have been recorded for each experimental subject. Principal Component Analysis has been applied to the dataset, to obtain an optimal projection of the data in terms of discriminatory clarity. The 226 variables have thus been replaced by the first 66 principal components, to obtain a model that accounts for 91% of the variance in the original dataset. That is to say that using PCA, we have been able to reduce the number of variables to less than a third of those recorder while retaining 91% of the information contained in the original dataset. Noting that some of the variance in the original dataset must be attributed to noise (measurement errors, unrecorded changes in experimental conditions, etc.), we consider the projected dataset based on the first 66 principal components as a very accurate model for the original dataset.

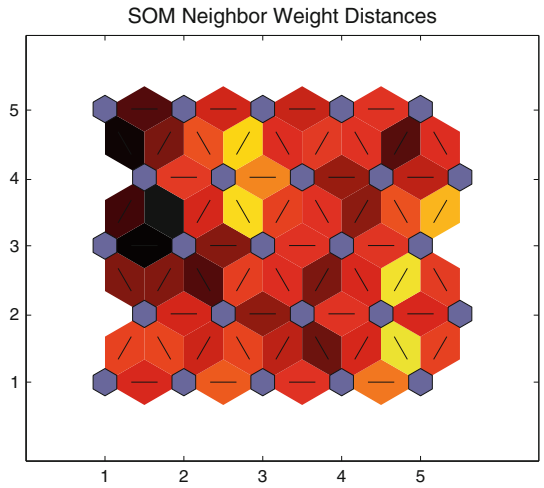

Fig. 2 Graphical representation of the SOM after 2000 cycles of training. Neurons are represented by the *small blue hexagons*. The distances between the weight vectors associated with each neuron are colour coded such that *light colours* represent smaller distances, while dark regions encode for larger ones. As such, sequences of dark coloured connections between neurons are interpreted as borders on the map

The projected data has subsequently been used as the input to the Self Organising Map algorithm for 2000 training cycles. Figure 2 gives a visual representation of the SOM's structure after the training phase has been completed. The figure shows that the self organisation has resulted in a "**U**" shaped border represented by the dark red and black connections between neurons on the map. As such, the experimental subjects are split in two broad categories by the map: one inside the "**U**", and one to its right side. Moreover, we can observe a weaker border inside the category represented by neurons on the right side of the map at coordinates {5 , 2}—this is the rightmost connection between the neurons on the second row from the bottom. As such this category can be split into two more specific ones.

The observations made above are confirmed by the results of the mapping phase of the SOM shown in Fig. 3. During this phase, the available data is presented one last time to the map, recording the neuron which responds to each subject instead of making changes to the structure of the map. It can be seen from the figure that three neurons respond to the vast majority of experimental subjects. Moreover, the borders identified above are placed in such a way as to separate the three neurons, indicating significant differences between the children each neuron responds to. The bottom right neuron (at coordinates {5 , 1}) on the map responds to 49 children and is separated by the weak border at {5 , 2} (identified above, see Fig. 2) by the neuron at

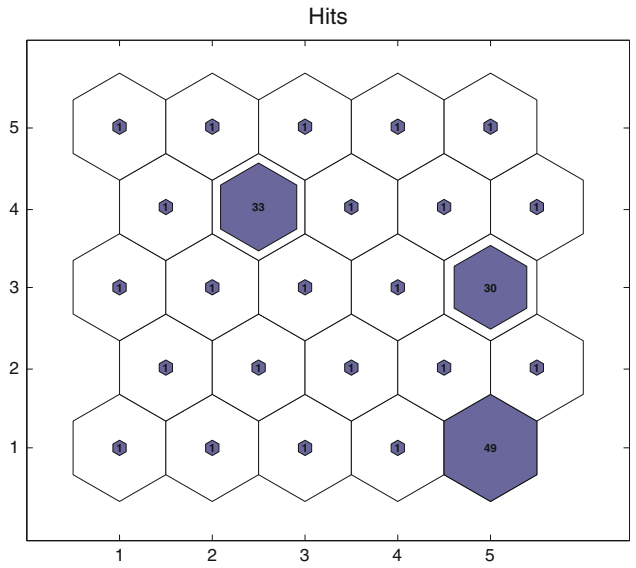

Fig. 3 Results from the mapping phase of the SOM. The number of experimental subjects each neuron responds to are displayed within the hexagon representing the neuron

{5 , 3} which responds to a further 30 cases. Throughout the remainder of this paper we will refer to the subjects each of these neurons responds to as *Cluster 1* and *Cluster 2*, respectively. *Cluster 3* will refer to the category which resides inside the "**U**" shaped border, i.e. the 33 cases to which the neuron at {2.5 , 4} responds to.

Once the children contained within each cluster have been identified, conventional statistical analysis tools can be used to identify significant differences between them, with respect to the original set of variables. The two-sample *t*-test is an appropriate statistic to evaluate whether two samples originate from the same population. Therefore, two one-sided two-sample *t*-tests (one left and one right) at the 99% level were carried out for each variable and each cluster, to assess whether the members of a cluster are significantly different from the rest of the population with respect to each variable. The two samples were assumed to have unequal variance, and the results of this analysis are summarised in Table 1.

A visual inspection of the table quickly reveals that *Cluster 2* performs significantly better than the rest of the

Table 1 Summary of results obtained using two-sample one-sided *t*-tests with unequal variance at the 99%level for each variable

		Cluster 1	Cluster 2	Cluster 3
Mental Attributes Profiling System (MAPS)	Age	+	+	−
	Grade		+	−
	Categorisation (time)		+	
	Categorisation (correct answers)		+	
	Lateral awareness (same orientation—time)	+	+	−
	Lateral awareness (same orientation—correct answers)	+	+	−
	Lateral awareness (different orientation—time)	+	+	−
	Lateral awareness (different orientation—correct answers)			−
	Navigation (correct answers)		+	−
	Sequencing (time)	+	+	−
	Sequencing (moves made)		+	−
	Visual memory (time)		+	−
	Visual discrimination (time)	+	+	−
	Visual discrimination (correct answers)	+	+	−
	Auditory memory (correct answers)		+	−
	Auditory memory (3 digits)	+	+	−
	Auditory memory (4–6 digits)		+	−
	Auditory discrimination (time)	+	+	−
	Auditory discrimination (correct answers)	+	+	
WISC-III	Vocabulary		+	−
	Similarities		+	−
Rapid naming (time)	Rapid naming		+	−
	Rapid naming (correct answers)	+	+	−
Woodcock	Word identification and word attack		+	−

Variables are grouped together using the tests from which they were recorded. A '+' sign indicates a significantly better performance of the members of a cluster with respect to the rest of the population on a given test, while the '−' indicates significantly worse performance

Table 2 The 12 variables with the largest discriminatory power with respect to the three clusters identified by the Self Organising Map algorithm

	Variable name	Classification accuracy
1	Acoustic memory total correct	72.32%
2	Navigation total correct	71.43%
3	Word identification	66.96%
4	Rapid naming average time (pictures)	65.18%
5	Word attack	65.18%
6	Grade	64.29%
7	Age	64.29%
8	Sequencing average time	64.29%
9	Auditory memory (5 digits)	64.29%
10	Rapid naming average time (letters)	63.39%
11	Auditory discrimination average time	63.39%
12	Lateral awareness (same orientation) average time	59.82%

sample in all but one test, while at the same time *Cluster 3* is the weakest, performing significantly worse than the other experimental subjects in all but three tests. *Cluster 1* resides in between the two clusters, as its results in 12 of the 22 tests do not significantly differ from the rest of the population. Members of this cluster do however perform significantly better than the remaining experimental subjects with respect to the other 10 tests.

Moreover, it is important to note that members of both *Cluster 1* and *Cluster 2* are both significantly older than the rest of the population which appears in the dataset. In addition, members of *Cluster 2*—the strongest cluster—are also in a significantly higher grade. The question of whether these two variables, Age and Grade, are the reason for the observed differences in performance is addressed below.

In order to assess the discriminatory power of each variable, 226 Bayesian classifiers have been developed, each based on a single variable. The 22 experimental subjects that have not been assigned to one of the clusters, shown in Fig. 3 and which form the borders of Fig. 2, are not considered in the process of developing classifiers. The remaining 112 children were randomly divided into eight groups. Seven groups are of size 15, while the last one contains only 7 children. Each classifier then uses the subjects contained in seven of the groups to calculate the required probabilities discussed in Sect. 3.3. Its accuracy is then measured by using it to classify the members of the remaining group. The process is repeated eight times, using a different group to measure classification accuracy. Table 2 provides the 12 variables with the highest average accuracy over the eight trials.

With respect to the concerns voiced above, regarding the fact that members of the stronger clusters also appear to be older and, in the case of *Cluster 2*, in a higher grade at school, we find that Age and Grade receive only a joint 6th place in the ranking provided by Table 2. The total number of correct responses to the Acoustic Memory test gives rise to the best single variable classifier, with 72.32% accuracy. In addition, six of the top ten variables (with the exception of Age and Grade) are ones recorded through the MAPS test. The remaining four variables were recorded using the Rapid Naming and Woodcock's Reading Mastery tests.

Furthermore, classifiers based on specific combinations of variables have also been developed. The most accurate classifier built using only MAPS variables reaches a predictive accuracy of 92.86% using a total of 17 variables: the individual times taken to complete each Auditory Discrimination test, the total number of correct responses to the Auditory Memory test, and the individual times taken to complete each Sequencing test.

In contrast, the most accurate classifier that can be built using up to five variables predicts the correct cluster for 94.64% of the cases. The variables used to develop this classifier were as follows: the total number of correct responses to the Auditory Memory test, the total number of correct responses to the Navigation test, the Word Identification score, the Word Attack score, and the Rapid Naming of Pictures score.

Discussion and Conclusions

The experimental results obtained in the previous section provide evidence that the four testing systems—MAPS, WISC-III, Rapid Naming, and Woodcock's Reading Mastery Test—identify the same intrinsic quality. This is so since through the projection of the entire data on an optimally informative coordinate system (via PCA), and the use of the SOM algorithm to split the participating children into emergent groups effectively separates children that achieve significantly better scores in all four tests

from those that perform significantly worse. Moreover, the algorithm identifies a third group whose scores in 10 of the 22 individual tests are significantly higher than the rest of the children participating in the experiment, while no significant difference has been found in their results for the remaining 12. The presence of this intermediate group is important, as it allows the identification of potential learning difficulties to be made at finer levels of detail.

In addition, the results suggest that 94.64% classification accuracy can be achieved through the combination of only four tests:

- Auditory memory (MAPS [1])
- Navigation (MAPS, [1])
- Word Identification and Word Attack (Woodcock reading mastery tests [14])
- Rapid naming of pictures (Wimmer et.al. [13])

To this end, we propose that only these four tests can be used in cases where the potential of a child suffering from learning difficulties needs to be preliminary assessed using a quick and concise series of tests. Such cases may include situations where large numbers of children will undergo a preliminary assessment to facilitate the early identification of problematic learners.

In an ideal setting, the learning abilities of a child would develop by improving at a steady pace the older he/she gets. Indeed, it was found that the members of the strongest group (*Cluster 2*) were significantly older and in a higher grade at school. However, the fact that Age and Grade were not found to be the most powerful discriminators shows that there are weak learners (*below–average*) in the group studied herein: older children who perform badly in the tests administered to them, and who are consequently grouped together by the SOM together with younger, and thus naturally expected to perform worse, ones. Conversely, as Age and Grade are not found to be perfect discriminators, younger children who achieve high scores are placed by the algorithm in the same cluster as older children who can even be in a higher grade. These children can be characterised as having *above–average* learning abilities. Moreover, the results are compatible with the findings of [15], where the authors were not always able to demonstrate a developmental effect of the MAPS tests.

The objective of this work has been to develop effective tools to facilitate the early identification of learning difficulties. To that end, we have been able to develop classifiers that can categorise children in the three groups identified, based on small subsets of the variables recorded in administrating the various tests while at the same time reaching levels of accuracy higher than 90%. These can in turn be used to effectively assess a child's learning abilities in the following way. First, the group a child is expected to belong to is identified using his/her age and current grade at school. Subsequently, the variables required by the more accurate classifiers are measured (by administrating a subset of the tests), and the group the child should belong to is reassessed. If the two processes identify the same group, the child can be expected to possess *average* learning abilities. However, if the classification process identifies a different group than the one expected based on the child's age and grade, one of the following prognoses can be made:

Below-average performance: When a child is significantly older and attends a higher grade at school, but at the same time is placed by the classifier at the weakest cluster, this is a clear indication that the child may suffer from learning difficulties.

Above-average performance: This is the opposite situation to the one described above. A younger child who attends a lower grade at school has performed so well in the various tests that it has been assigned to the strongest cluster by the classifier. Such children can be considered to possess particularly strong learning abilities.

Mildly below-average performance: This prognosis is appropriate when a child who is significantly older and in a higher grade at school is classified in *Cluster 1*. The prognosis identifies a small degree of underperformance, which can perhaps be attributed to less severe factors than professing that a child suffers from learning difficulties. A similar prognosis can be attributed to children who may be significantly older but do not attend a higher grade than those contained in the weakest cluster. When such children (who would be expected to belong to *Cluster 1*) are classified in the weakest cluster, their performance can also be characterised as mildly below–average.

Mildly above-average performance: When children that are significantly younger and in a lower grade perform well enough in the various tests administered to be classified in *Cluster 1*, instead of the weakest cluster as would be expected, a tendency to perform better in terms of cognitive abilities can be identified. This however is a weak observation, as the preceding one.

Based on the observations made above, this paper concludes that effective prognoses can indeed be made with respect to learning difficulties, using a small number of tests while at the same time being reasonably reliable through achieving over 90% predictive accuracy with respect to the sample studied here. However, it must be noted that diagnoses for learning difficulties in children should be made with great care, as they cannot always be assumed to be beneficial for the children. With respect to this, we must state that the technology developed and described in this article is not intended to be used as the

single means of diagnosis. Rather it is intended as tool to improve the ability of caregivers to diagnose learning difficulties early, and not to replace but to complement existing tools and methods. Moreover, the reader is reminded that classification using Bayesian methods occurs by identifying the class which has the greater possibility of having generated a particular observation. As such, those cases where a class is assigned with only marginal differences in the posterior probabilities computed for each class should be treated with more care and in-depth assessment before a reliable diagnosis is made.

Finally, we would like to note that the identification of particular types of learning difficulties and their correspondence with individual tests is the subject of ongoing work. We anticipate that the application of similar techniques and methods to those presented herein will prove valuable in doing so.

Acknowledgements The authors would like to thank Tatjana Taraszow, Lawrence Kalogreades and Elena Aristodemou for their valuable comments and support in developing this article.

References

1. Laouris Y, Makris P. M.A.P.S. mental attributes profiling system computerized battery for dyslexia assessment. Proceedings of multilingual & cross-cultural perspectives on Dyslexia, Omni Shoreham Hotel, Washington, D.C.; 2002.
2. Papadopoulos TC, Laouris Y, Makris P. The validation of a computerised cognitive battery of tests for the diagnosis of children with dyslexia. Proceedings of IDA 54th annual conference, San Diego; 2003.
3. Watkins MW, Kush J, Glutting JJ. Discriminant and predictive validity of the WISC–III ACID profile among children with learning disabilities. Psychol Sch. 1997;34(4):309–19.
4. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern. 1982;43:59–69.
5. Lerner JW (2000) Learning disabilities: theories, diagnosis, and teaching strategies. Houghton Mifflin, Boston.
6. Dekker GW, Pechenizkiy M, Vleeshouwers JM. Predicting students drop out: a case study. Proceedings of 2nd international conference on Educational Data Mining, Cordoba, Spain; 2009. p. 41–50.
7. Nugent R, Ayers E, Dean N. Conditional subspace clustering of skill mastery: identifying skills that separate students. Proceedings of 2nd international conference on Educational Data Mining, Cordoba, Spain; 2009. p. 101–10.
8. Romero C, Ventura S, Espejo PG, Hervas C. Data mining algorithms to classify students. Proceedings of 1st international conference on Educational Data Mining, Montreal, Canada; 2008. p. 8–17.
9. Gong Y, Rai D, Beck J, Heffernan N. Does self-discipline impact students knowledge and learning? Proceedings of 2nd international conference on Educational Data Mining, Cordoba, Spain; 2009. p. 61–70.
10. Mavrikis M. Data-driven modelling of students' interactions in an ILE. Proceedings of 1st international conference on Educational Data Mining, Montreal, Canada; 2008. p. 87–96.
11. Feng M, Heffernan N, Beck J, Koedinger K. Can we predict which groups of questions students will learn from? Proceedings of 1st international conference on Educational Data Mining, Montreal, Canada; 2008. p. 218-225.
12. Cho K. Machine classification of peer comments in physics. Proceedings of 1st International Conference on Educational Data Mining, Montreal, Canada; 2008. p. 192–196.
13. Wimmer H, Mayringer H, Landerl K. The double deficit hypothesis and difficulties in learning to read a regular orthography. J Educ Psychol. 2002;92:668–80.
14. Woodcock RW. Woodcock reading mastery tests Revised NU: examiner's manual. American Guidance Service, Circle Pines; 1998.
15. Laouris Y, Papadopoullos T, Makris P. Validation of MAPS in 16 schools; computer-based battery of 8 mental attributes tests (in preparation).
16. Piaget J. Judgement and reasoning of the child. Paul Kegan, London; 1928.
17. Aristodemou E, Taraszow T, Laouris Y, Papadopoulos T, Makris P. Prediction of reading performance using the MAPS (Mental Attributes Profiling System) multimodal interactive ICT application (in preparation).
18. Wechsler D. The Wechsler Intelligence Scale for Children. Psychological Corp, New York; 1949.
19. Brannigan GG, Rosenberg LA, Loprete LJ, Calnen T. Scoring of WISC-R comprehension, similarities, and vocabulary responses by experienced and inexperienced judges. Psychology in the schools 14, 430. Wiley, New York; 1977.
20. Papadopoulos TC. Phonological and cognitive correlates of word-reading acquisition under two different instructional approaches. Eur J Psychol Educ. 2001;26:549–68
21. Papadopoulos TC, Charalambous A, Kanari A, Loizou M. Kindergarten intervention for dyslexia: the PREP remediation in Greek. Eur J Psychol Educ. 2004;19:79–105
22. Pearson K. On lines and planes of closest fit to systems of points in space. Philos Mag. 1901;2(6):559–72
23. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn. 1997;29: 103–37