# Using Fourier coefficients in time series analysis for student performance prediction in blended learning environments

**Gamulin, Jasna; Gamulin, Ozren; Kermek, Dragutin**

*Source / Izvornik:* **Expert systems, 2016, 33, 189 - 200**

**Journal article, Published version**
**Rad u časopisu, Objavljena verzija rada (izdavačev PDF)**

https://doi.org/10.1111/exsy.12142

*Permanent link / Trajna poveznica:* https://urn.nsk.hr/urn:nbn:hr:211:534638

*Rights / Prava:* In copyright / Zaštićeno autorskim pravom.

*Download date / Datum preuzimanja:* **2024-03-29**

SVEUČILIŠTE U ZAGREBU
FAKULTET ORGANIZACIJE I INFORMATIKE
VARAŽDIN

*Repository / Repozitorij:*

Faculty of Organization and Informatics - Digital Repository

DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

Expert Systems



Article

DOI: 10.1111/exsy.12142

# Using Fourier coefficients in time series analysis for student performance prediction in blended learning environments

Jasna Gamulin,[1] Ozren Gamulin,[2]* and Dragutin Kermek[3]

*(1) University of Zagreb School of Medicine, Zagreb, Croatia*
*(2) Department of Physics and Biophysics, University of Zagreb School of Medicine, Zagreb, Croatia*
*E-mail: ozren@mef.hr*
*(3) Department of Computing and Technology, Department of Theoretical and Applied Foundations of Information Sciences, University of Zagreb Faculty of Organization and Informatics, Varaždin, Croatia*

**Abstract:** *In this work, it is shown that student access time series generated from Moodle log files contain information sufficient for successful prediction of student final results in blended learning courses. It is also shown that if time series is transformed into frequency domain, using discrete Fourier transforms (DFT), the information contained in it will be preserved. Hence, resulting periodogram and its DFT coefficients can be used for generating student performance models with the algorithms commonly used for that purposes. The amount of data extracted from log files, especially for lengthy courses, can be huge. Nevertheless, by using DFT, drastic compression of data is possible. It is experimentally shown, by means of several commonly used modelling algorithms, that if in average all but 5–10% of most intensive and most frequently used DFT coefficients are removed from datasets, the modelling with the remained data will result with the increase of the model accuracy. Resulting accuracy of the calculated models is in accordance with results for student performance models calculated for different dataset types reported in literature. The advantage of this approach is its applicability because the data are automatically collected in Moodle logs.*

***Keywords:*** educational data mining, student performance prediction, time series, frequency domain, discrete Fourier transforms

## 1. Introduction

E-learning is defined as the usage of any information technology or application for learning or learning support (Laurillard, 2004). More specifically, b-learning (blended learning, hybrid learning and mixed mode) is the form of learning environment where the traditional classroom teaching and face-to-face communication between teacher and students are blended with the computer-mediated interaction (Bubaš & Kermek, 2004). This type of learning is outspread in higher education in Croatia and worldwide, especially in Europe. According to mission statements of Croatian universities, the blended learning is the preferred way of applying information and communication technology in higher education.

Typical for blended learning is the usage of a form of learning management system (LMS) and/or a form of e-assessments during the course that provides numerous data on student behaviour patterns and results. These data stimulated the development of educational data mining (EDM). It is a relatively new research area pursuing the development of methods and computational approaches for exploring data originating from an educational context (Romero & Ventura, 2010).

In a log file (or access log), an LMS writes a record of student activities. In this paper, a student log denotes an activity report in Moodle (Moodle, 2014) for one student although the idea is generally applicable in any other virtual learning environment. In a student log, it can be seen what pages students accessed, the time and date they accessed them, the Internet protocol addresses they came from and their actions (view, add, update and delete). Each action is presented as one click. Words the *access* and the *click* are used interchangeably in this paper. In order to build the prediction model of student final exam result, the time pattern of student Moodle logs is analysed. Time series for each student is generated, containing the number and temporal order of a student accesses to a certain LMS course during the time when the course was taking place. The data from Moodle logs are pre-processed, and the discrete Fourier transforms (DFT) is calculated in order to reduce the amount of data needed for successful student performance prediction. The DFT is widely used in technology for signal processing and in economics for time series analysis. Nevertheless, the authors could not find the record that it has been used in EDM.

The approach proposed in this paper for student performance model building could be useful for understanding the learning process in the b-learning environment and for improving teaching techniques. It can be also used for administrative purposes in assigning student groups for the following courses, making teaching schedule for the programme, number of teachers needed and calculating the school's income from tuition fees. The advantage of

© 2015 Wiley Publishing Ltd

modelling with data from LMS log files is its applicability in many different cases because that data are automatically collected and accessible.

This paper is organized in the way that firstly the review of the up-to-date scientific research and motivation for this research is presented. Next, the underlying method of the study and the dataset used are described followed by analysis and discussion of results. At the end, the main conclusions, limitations of the study and guidelines for future research are presented.

## 2. Motivation and related work

One key application of EDM methods has been the improvement of student models (Baker & Yacef, 2009). A student model is providing data on a student who uses some kind of computer-based learning system in order to adjust that system according to the student. Model building process includes data pre-processing, parameter optimization and attributes selection steps. The student modelling and performance prediction in an e-learning and tutoring/adaptive learning systems (Kotsiantis *et al*., 2004; Lykourentzou *et al*., 2009; Cobo *et al*., 2011; Jovanović *et al*., 2012; Dorça *et al*., 2013; Lara *et al*., 2014) and moreover intelligent tutoring systems (Thai-Nghe *et al*., 2010) do differ from modelling in a b-learning environment (Dias & Diniz, 2013) because b-learning student models have less data available. The papers referring to b-learning environment are not so numerous and, expectably, do not rely on machine learning techniques so frequently (Delgado *et al*., 2006; Deneui & Dodge, 2006; Thai-Nghe *et al*., 2007; Dekkar *et al*., 2009; Divjak & Oreški, 2009).

In an e-learning environment, all the data on learning process are automatically stored in databases. In a b-learning environment, only a part of learning process is covered by automatically stored data. Examples of that data are access to some kind of learning material, general information about the organization of the class, self-evaluation tests, some forms of online assessments and sometimes forums posted on LMS. Part of the learning process taking place in classroom is generally not included in student modelling. Nevertheless, the blended learning as the emerging and dominant form of e-learning environment in Croatian higher education system deserves to be understood profoundly.

Most often used variables (attributes) for student modelling are the number of solved assignments, the number of quiz accesses, the number of quizzes passed, the number of quizzes failed, the number of forum posts, the number of forum posts read, the time used for assignments, the time used for quizzes and the time spent on forums (Jovanović *et al*., 2012) and so on. Divjak and Oreški (2009) used 30 attributes (sociological and student perception about their study) collected by a questionnaire to predict student performance using discriminant analysis. Usually, the class label for modelling is the final grade for the course (Romero *et al*., 2008; Lykourentzou *et al*., 2009; Jovanović *et al*., 2012). Dias and Diniz (2013) constructed the class label named *the quality of interaction* with an LMS. Dorça *et al*. (2013) are dynamically modelling to the probability distribution of student learning style according to the learning style inventory. The research is conducted as a simulation, not using the real dataset. Dekkar *et al*. (2009) are using as attributes the number of enrolled courses, their average grade, the number of the enrolled science courses, the number of the enrolled mathematic courses and the average grade for mathematical courses. As the class label, they are modelling using the drop-out rate.

The papers engaged in LMS activity time pattern recognition are rather scarce. Cobo *et al*. (2011) are grouping students using clustering to the categories: active learner, lurker and shirker. They are based on student activities in online forums and represented as time series. Delgado *et al*. (2006) are modelling using student grades from the number of logs to Moodle segmented per month of accesses and from grades.

Lara *et al*. (2014) analyse the time pattern of Moodle logs in a strictly e-learning environment in order to build the model of student drop-out prediction based on the idea that the time pattern of Moodle activity is connected to student dropout. The historical model is built and then used for classification of new students with the goal to identify students who will probably drop out from the course. Classification was based on the time analysis of interaction with Moodle. Although the several categories of interaction with Moodle are constructed, the real data analysis shows that students predominantly use the action 'view'. Attributes used are the number of Moodle actions per week per a student, the number of days with action per week and the number of visualizations weekly for a student for the specific Moodle resource. They measure the Euclidian distance of each student attribute from the attribute of the model. The training set is 100 students, and validation set is 50 students. The data on four courses lasting for 20 weeks each are analysed. The approach provides better results than some standard machine learning methods.

Being the new research field, EDM has recently incorporated methods reserved for natural sciences. An example for this is a spectral learning approach to knowledge discovery (Falakmasir *et al*., 2013). In this work, the spectral algorithm is applied in order to improve knowledge tracing parameter-fitting time while maintaining the same prediction accuracy when using hidden Markov model for determining student knowledge of skills in adaptive educational systems and cognitive tutors. Warnakulasooriya and Galen (2012) introduced the notion of fractal dimension. They concluded that student answers in a tutor system considered as time series have characteristics of random walk or Brownian motion. Those examples encouraged the authors of this paper to apply Fourier transform to the time series of student LMS logs.


© 2015 Wiley Publishing Ltd

## 3. Methodology and dataset

### 3.1. Time series

Interest in time series data mining increased with the increased number of longitudinal databases. The prominent characteristic of such databases is the huge amount of data. The problem could be approached by combining classifiers in the way that algorithms try to forget irrelevant information instead of synthesizing all available information (Kotsiantis *et al*., 2010).

The problem could be also approached by compressing data without losing important information. The usual approach to data compressing is to cover the majority of variation of the series by a small number of data. DFT is a popular technique of transformation and compression in time series data mining (Bagnall & Janacek, 2005). Fourier transform is used to transform the time-domain data into the frequency domain and vice versa, and it is generally used in technology for signal analysis.

Time series are very data rich. For one-semester course lasting for 14 weeks, the number of data points in time series generated from Moodle logs per hour is approximately 2400 points per student. Having 300 students enrolled in the course, it means that we have a processing problem.

When we have Fourier series with signals of finite lengths, they could be represented by DFT (Chen & Chen, 2014). The improved method to calculate DFT is fast Fourier transform (FFT). FFT has identical results as DFT but requires less computation and has become one of the most frequently used algorithms in computer science, electrical engineering and time series analysis (Chen & Chen, 2014). Spectral analysis is not considered a conventional time-domain analysis. Spectral approach to time series data mining provides an insight into complex shape or autocorrelation structure (Bagnall & Janacek, 2005), and FFT is regarded to be better than autoregressive moving average (ARMA) models for problems where complex patterns in periodogram represent cyclic trends and autocorrelation structures. FFT is faster than fitting ARMA models and could be used to detect other forms of similarities that ARMA models cannot (Bagnall & Janacek, 2005).

Time series is generally a collection of observations $X_t$, each being recorded at time $t$. In our case, time series is discrete $t_i$ where $i = 0,1,2,3, \ldots, N$, where numbers stand for hours or days depending on how long in equal and equidistant periods $T$ the signal was measured. It practically means that $T = t_i - t_{i-1}$ is in fact the period of time, and we call it the accumulation time. Two accumulation times were used in this work: $T_{\text{hour}} = 1$ h and $T_{\text{day}} = 1$ day. It means that student clicks were counted within 1 and 24 h, respectively. As a result of these measurements, time series for each student was generated, describing the number and temporal order of a student clicks at a certain LMS course during the time when the course was taking place. Thereafter, the DFT was calculated from the generated time series. The normalized DFT of time series $x(n)$, $n = 0,1 \ldots N-1$ is a vector of complex numbers $X(f)$:

$$X\left(f_{k/N}\right) = \frac{1}{\sqrt{N}} \sum_{n}^{N-1} x(n) e^{-i2\pi kn/N},\ k = 0, 1 \ldots N-1$$

Time series in this work are real signals, and Fourier coefficients are symmetric around the middle one. Fourier transform represents linear combination of complex sinusoids

$$s_f(n) = \frac{e^{i2\pi kn/N}}{\sqrt{N}}$$

Therefore, the Fourier coefficients represent the amplitude of each of these sinusoids, after signal $x$ is projected on them. In order to simplify time series and capture most important periodicities hidden in those time series, the most dominant frequencies were used. By the most dominant frequencies, it is meant the ones that carry the most of signal energy. The way to identify those frequencies is to calculate periodogram by squaring magnitude of Fourier coefficients:

$$P\left(f_{\frac{k}{N}}\right) = \left\| X\left(f_{k/N}\right) \right\|^2, \quad k = 0, 1 \ldots \left[\frac{N-1}{2}\right]$$

Dominant frequencies are peaks in the periodogram and correspond to the coefficients with highest magnitude. From here on, when we refer to the best or largest Fourier coefficients, we mean the ones that have highest peak in the periodogram.

Frequency spectra of student Moodle logs for four different courses taking place at two faculties during one academic year are used for modelling student performance using state-of-the-art classification algorithms. The number of frequencies used to build the successful model is gradually changed in order to test how many Fourier coefficients are needed for successful modelling. By changing the time of logs accumulation (1 day or 1 h) and changing the number of categories, we can compare the classifier performance.

In technology, signals like recorded sounds or photographs can be improved by removing coefficients from the signal in frequency domain that can be associated with noise (Reis *et al*., 2009). According to this, we assume that we can reduce the number of DFT coefficients by removing unimportant ones from our dataset and in the same time preserve information about student success. Even more, removing unimportant information might improve the accuracy of modelling. Thus, DFT is used here as the way for improving and compressing (Vlachos *et al*., 2004) time series data. Important question is which Fourier coefficients contain relevant data useful for student success prediction modelling. Agrawal *et al*. (1995) used first $k$ Fourier coefficients (not in the EDM context) that are adequate for random walk time series. For the time series with strong periodicities, it is more appropriate to use the best $k$ Fourier coefficients because the most of the power is not in the low

© 2015 Wiley Publishing Ltd

frequencies only but is scattered at the whole spectrum (Vlachos *et al*., 2004).

### 3.2. Experimental data

In this work, four courses are analysed. All courses are taught at the University of Zagreb. The first-year and the fourth-year Physics course are taught at the School of Medicine, and the next two courses are taught at the Faculty of Organization and Informatics. All courses are taught as blended learning courses, meaning that the traditional classroom teaching is supplemented by learning material and general info posted on Moodle platform. Nevertheless, each of those courses is also specific in a certain way. The Medical Physics 1 course (MP1) in the first study year has a large number of enrolled students with the great variety of learning material types published on the Moodle course web page. Nevertheless, the teaching approach is dominantly classical with lectures, seminars and laboratory exercises performed in classroom. Students attend classes during 3.5 months. The second course at the School of Medicine (Physics of Medical Diagnostics, PMD) is taught at the fourth study year and has also large number of enrolled students, but it lasts for only 1 week with smaller amount and diversity of teaching materials available online. Teaching approach is again dominantly classical. At the Faculty of Organization and Informatics, we processed the data on the undergraduate course Web-design and Programming (WEB) and the graduate course Advanced Web-technologies and Services (AWT). Those courses are enrolled by a smaller number of students than courses at the School of Medicine, 80 and 64, respectively, but the number of Moodle accesses is in total comparable with the courses at the School of Medicine. This could be explained by the fact that students enrolled at the Faculty of Organization and Informatics are intrinsically oriented towards information and communication technology usage. Such a diversity of courses allows checking for applicability of the proposed way of student modelling. The data on all the courses and generated initial datasets are listed in Table 1. From each initial dataset, approximately 30 reduced datasets were generated by changing the number of DFT coefficients used. The process of forming the initial datasets and reducing the number of DFT coefficients needed for modelling is described in Section 3.3.

For all courses, students receive grades from 1 to 5. Exceptionally, the PMD course is graded by pass and fail only. In order to compare results for all courses, we discretized the dataset into two categories, Good and Poor. In Good category, we placed all the students with grades 3, 4 and 5, and in Poor category, all the students with grades 1 and 2. In case of PMD, we already have students divided into two categories. Good (G) are the students who passed the exam at the first exam term, and Poor (P) are the students who failed. The three-category datasets [grades 5,4 → category Very Good (VG); grade 3 → category Standard (S); grades 2,1 → category Poor (P)] were generated for AWT and WEB courses because they have similar teaching approach and the number of students, and it was possible to compare modelling results also between two-category and three-category datasets. The number of respondents within categories is displayed in Table 1, and the distribution of categories is apparently quite balanced.

### 3.3. Pre-processing

Generally, the data accumulated in the Moodle database are not in the form suitable for data mining. Thus, in majority of papers dealing with some sort of data mining, a kind of pre-processing is performed. In this paper, time series for each student is extracted from the Moodle log table, containing the number and temporal order of a student accesses to a certain LMS course during the duration of the course. From those time series, DFT are calculated and used for prediction of student final performance.

**Table 1:** *The overview of the examined courses and generated initial datasets*

| *Name of the course* | *Dataset label* | *Number of students in 2012–2013* | *Duration of the course (days)* | *Approximate number of clicks per course* | *Number of class label categories* | *Number of respondents within categories G/P and VG/S/P* | *Accumulation time* |
|---|---|---|---|---|---|---|---|
| Medical Physics 1 | MP1 $2C_{day}$ | 300 | 100 | 100000 | 2 | 165/135 | $T_{day}$ |
| Physics of Medical Diagnostics | PMD $2C_{hour}$ | 270 | 5 | 22000 | 2 | 175/94 | $T_{hour}$ |
| Advanced Web-technologies and Services | AWT $2C_{hour}$ | 64 | 120 | 100000 | 2 | 36/28 | $T_{hour}$ |
| | AWT $2C_{day}$ | 64 | 120 | 100000 | 2 | 36/28 | $T_{day}$ |
| | AWT $3C_{hour}$ | 64 | 120 | 100000 | 3 | 16/20/28 | $T_{hour}$ |
| | AWT $3C_{day}$ | 64 | 120 | 100000 | 3 | 16/20/28 | $T_{day}$ |
| Web-design and Programming | WEB $2C_{hour}$ | 80 | 120 | 100000 | 2 | 41/39 | $T_{hour}$ |
| | WEB $2C_{day}$ | 80 | 120 | 100000 | 2 | 41/39 | $T_{day}$ |
| | WEB $3C_{hour}$ | 80 | 120 | 100000 | 3 | 15/26/39 | $T_{hour}$ |
| | WEB $3C_{day}$ | 80 | 120 | 100000 | 3 | 15/26/39 | $T_{day}$ |

G/P, good/poor; VG/S/P, very good/standard/poor.


© 2015 Wiley Publishing Ltd

More precisely, the data in the Moodle log table contain student access data to the course resources (Figure 1(a)). In order to generate time series of student accesses to the course resources, accumulation time of access ($T$) is defined. $T$ is the period of time in which all accesses or clicks made by one student are added together. In this paper, two accumulation times are used, $T_{\text{hour}}$ and $T_{\text{day}}$, as described in Section 3.1. Time series for a single student is generated in a way that starting date and time of the course are defined as 00:00 on the day when the course starts. For each $T$ from the starting time to the end of the course, the number of clicks for each observed student is counted. As a result, time series

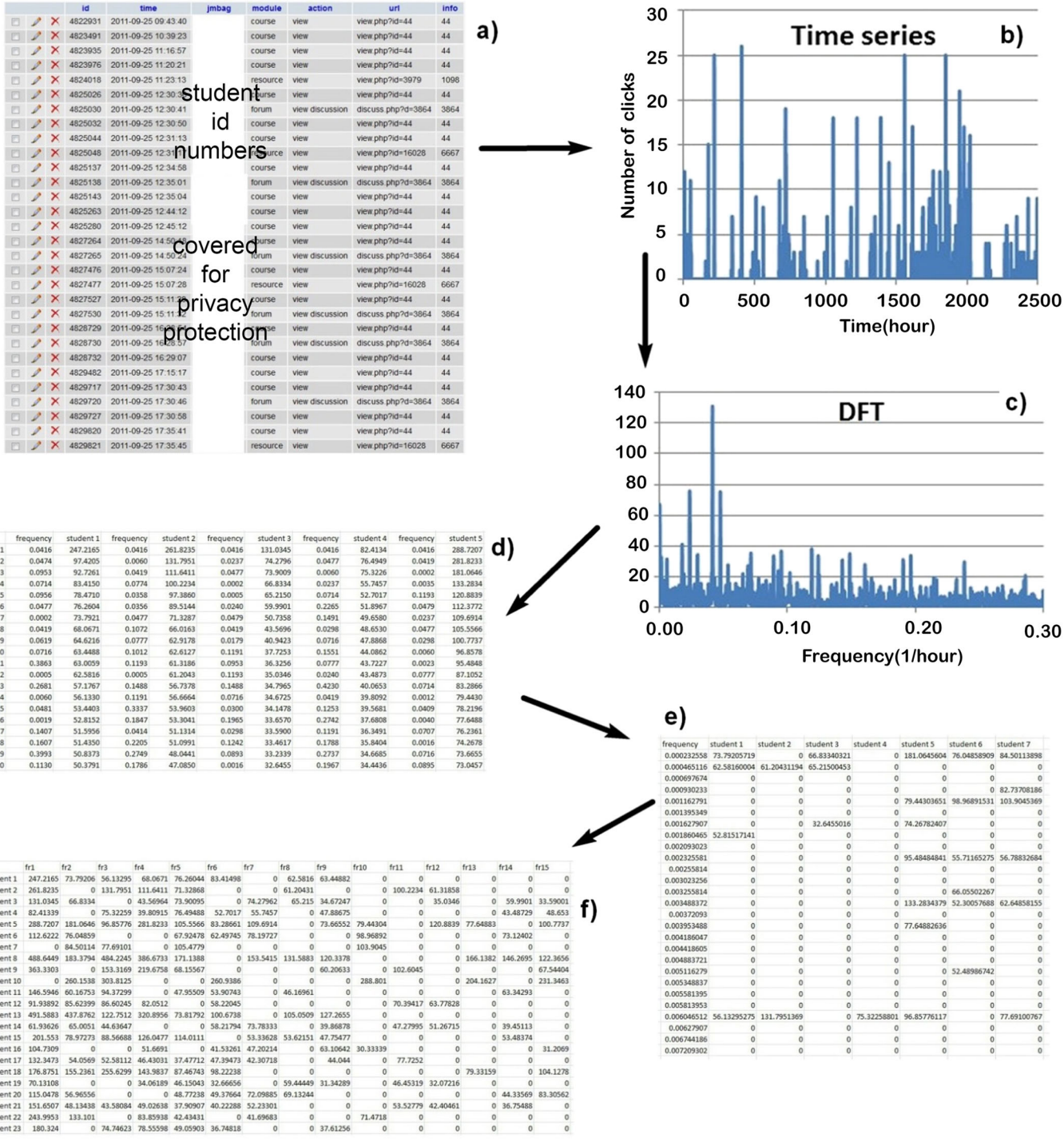



**Figure 1:** *The data pre-processing. (a) The Moodle database table containing data on all students and all accesses to a certain course; (b) student access time series generated for each student; (c) the sample periodogram, that is, discrete Fourier transforms coefficients for one student; (d) the list of 20 most intensive discrete Fourier transforms coefficients for each student; (e) the list of common coefficients appearing in all student sets; and (f) resulting tables for each observed course.*

© 2015 Wiley Publishing Ltd

of student accesses is formed, and a sample for one randomly chosen student is presented in Figure 1(b). In a case of $T_{hour}$, the number of points in the time series is approximately 2400 points for one-semester course, and in a case of $T_{day}$, the number of points in the time series is approximately 90 points. Sequences like this are created for each student enrolled in the courses processed in this paper. The third step of data pre-processing is calculation of DFT for all generated time series. Result is the periodogram. The number of DFT coefficients in a periodogram equals to the number of points in the original time series. The sample periodogram for one student is presented in Figure 1(c). From all periodograms, 20 most intensive coefficients for each student are chosen (Figure 1 (d)). Those sets of 20 coefficients are not identical for all students. The new list is formed containing coefficients that appear in all sets (Figure 1(e)). According to that list, the new table is created with coefficients in the first column and their intensities for each student in other columns. If the coefficients from the first column exist in the 20 coefficients set for a particular student, its intensity is copied in the coefficient row of the corresponding student column, and if not, the cell is filled with zero. In the last column, the number of non-zero coefficients is counted. The table is then sorted according to the last column from higher to lower values. Resulting tables for each observed course have 60 to 600 coefficients depending on accumulation time and the number of students (Figure 1(f)). This is how 10 initial datasets were formed (Table 1). From each initial dataset, the actual modelling datasets were generated by gradually reducing the number of coefficients from maximum number to only two or three remaining coefficients. The reduction was made by removing the least represented coefficients first and then gradually removing more common coefficients until only two or three most common DFT coefficients remain. This is the way how approximately 30 modelling datasets were generated from each initial dataset. It means that altogether approximately 300 models were formed per modelling algorithm.

### 3.4. Algorithms and software

The log data were transformed into time series and then into periodograms with small scripts written in MATLAB (Mathworks, Natick, MA, USA). In order to obtain the preliminary information about differences between student time series and to compare it with periodograms, the principal component analysis (PCA) was performed by MATLAB add-on PLS Toolbox (Eigenvector Research, Inc., Wenatchee, WA, USA). When PCA showed that the difference between student categories is present in observed datasets, student performance modelling was performed using algorithms listed in Table 2. According to literature, those algorithms are usually used for prediction models in EDM. For model calculations, the RAPIDMINER 5.3 (Rapidminer GmBH, 2014) software was used. In all models, the split validation with stratified sampling was performed, meaning that one set of data was used for model building (70%) and the other for validation of the model (30%). During the model calculation, attributes were optimized and selected with genetic algorithm because it was shown (Oreški *et al.*, 2012; Gamulin *et al.*, 2014) that the model accuracy was much higher when optimization was applied.

We used several algorithms frequently described in literature in order to show that all of those algorithms give accuracy higher than random guessing and comparable with other modelling results. Also, all the applied algorithms give out higher accuracy once we reduce the number of coefficients in the way described in Section 3.3. We used the default parameters in RAPIDMINER (Rapid Miner, 2014) for the applied algorithms. The authors believe that it is important that all the algorithms gave out similar results in order to imply the general trend that should be investigated in detail later on in a subsequent research. In order to avoid overfitting, the cross-validation was carried out for all models. In addition, the experiment itself in which four modelling algorithms were applied and compared having similar modelling results showed that overfitting was not important problem in this case.

Applying the selected four algorithms to the described models was very time demanding because approximately 1080 calculations were carried out. For some models, it took several hours to be computed. That is why for most of the models only onefold validation was performed on the subset of 30% of data. Tenfold validations were performed as a check on a few randomly chosen modelling algorithms and datasets. The results of those check-ups showed the several per cent lower accuracy than for onefold validation. Nevertheless, the general tendency was the same. Because it was not the point in maximizing accuracy, rather

**Table 2:** *Algorithms used in this experiment*

| *Algorithm* | *Abbreviation* | *Label* | *Reference* |
|---|---|---|---|
| Naïve Bayes | NB | Polynominal | Kotsiantis *et al.*, 2004; Jovanović *et al.*, 2012 |
| Artificial neuron networks | ANN | Polynominal | Kotsiantis *et al.*, 2004; Delgado *et al.*, 2006; Paliwal & Kumar, 2009; Jadric *et al.*, 2010; Jovanović *et al.*, 2012; Oreški *et al.*, 2012, |
| Support vector machines | SVM | Binominal | Kotsiantis *et al.*, 2004; Lykourentzou *et al.*, 2009; Thai-Nghe *et al.*, 2009; Lara *et al.*, 2014 |
| *K*-nearest neighbour | kNN | Polynominal | Kotsiantis *et al.*, 2004 |


© 2015 Wiley Publishing Ltd

investigating the behaviour of the model in dependence of the number of the coefficients used, we performed only onefold validation for most of the models.

The aim of this paper is to show that student logs time series contain information sufficient for building student performance prediction model. Another aim is to show that by transforming time data into the frequency domain (DFT), the information is still preserved and that it is also preserved when the number of DFT coefficients is reduced to a smaller number of coefficients. The reduction of the dataset allows faster computing. It is also assumed that the distribution of DFT coefficients can be helpful for recognition of the student behaviour pattern on LMS and subsequently for the improvement of LMS content. The part with recognition of student behaviour pattern is not elaborated here and should be investigated in detail in a future research.

### 3.5. *Principal component analysis*

In order to explore the intuitive assumption that the information about student performance could be acquired from the time pattern of student Moodle logs, the preliminary research was performed. The PCA was applied to the 1-year data of student grades at the final exam for a b-learning course and the sequences of student access to the course at LMS. The analysed time series show that the Moodle logs time series do show the separation of the classes on student performance.

## 4. Results and discussion

### 4.1. *Preliminary results*

The PCA is applied to the student access time series data collected for one of the observed courses (AWT graduate course) as independent variables and the final exam grade as the dependent variable. The grades achieved at the final exam are discretized into three categories as described in Section 3.2. The result of the PCA for the AWT course is presented in Figure 2(a), where the separation of time series according to the student performance can be observed.

Additional information such as periodicity of a student accesses can be extracted if transformation from the time to the frequency domain is accomplished using DFT (Bagnall & Janacek, 2005). It is expected that information stored in time series will be preserved, and in addition, hidden information about student access pattern will be revealed. To check the assumption that the information will be still preserved also in frequency domain, the PCA of frequency spectra periodogram produced by DFT is carried out, and the PCA results are presented in Figure 2(b).

Results of the PCA on access time series (Figure 2(a)) and frequency spectra (Figure 2(b)) are similar, indicating that information about student final success is preserved. This result encouraged authors to take the next step and check the possibility of modelling using compressed data in frequency domain.

### 4.2. *Modelling using compressed data in frequency domain*

The difference between access data in the time and in the frequency domain is presented in Figure 3, where time series and periodograms generated using DFT for two randomly chosen students are presented. In Figure 3(a), the access time series are shown, and the student access pattern seems to be completely random. On the contrary, in Figure 3(b) are presented the DFT periodograms obtained from time series presented on upper figure, and the common prominent coefficients can be noticed. Only the first half of the periodogram is shown in Figure 3(b) in order to emphasize the common parts (coefficients). We assume that

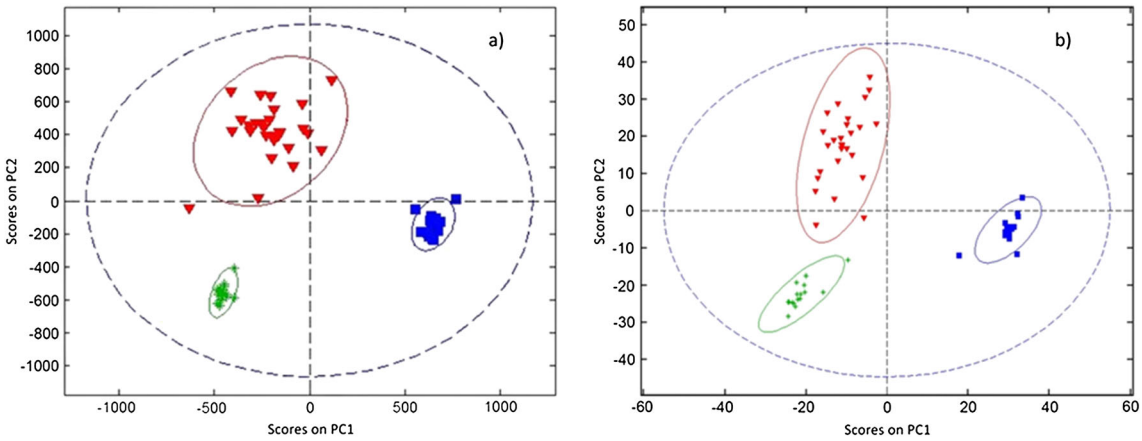



**Figure 2:** *The principal component analysis results for students enrolled in the Advanced Web-technologies and Services course and student performance categories: very good (green stars), standard (blue squares) and poor (red triangles). Ellipses depicted in the figure define confidence limits, within which 95% of the data are allocated. (a) The principal component analysis results on access time series and (b) on frequency spectra.*

© 2015 Wiley Publishing Ltd

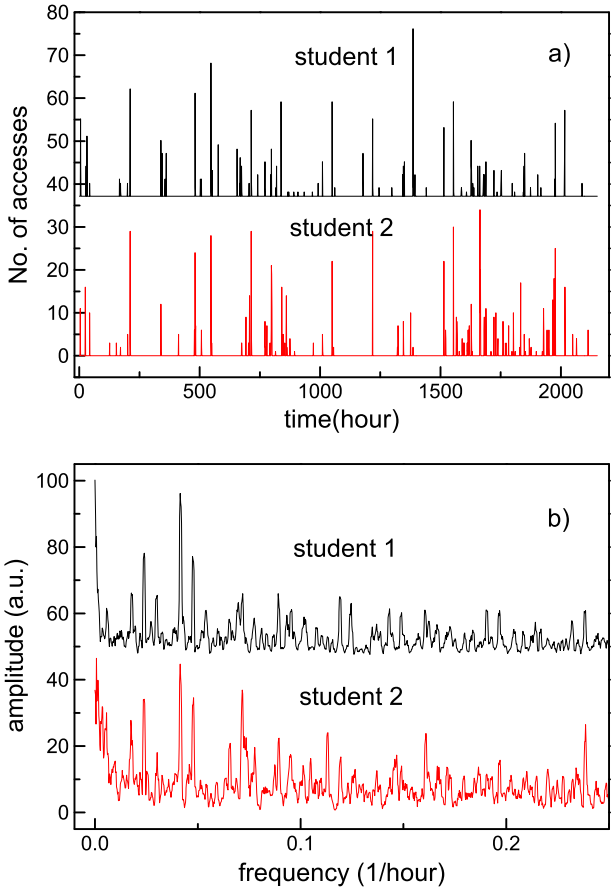



**Figure 3:** *(a) Student access time series for the $AWT_{hour}$ course for two randomly chosen students; (b) discrete Fourier transforms periodograms calculated from student access time series for the same two students. Because of better visibility, only lower half of periodogram is presented.*

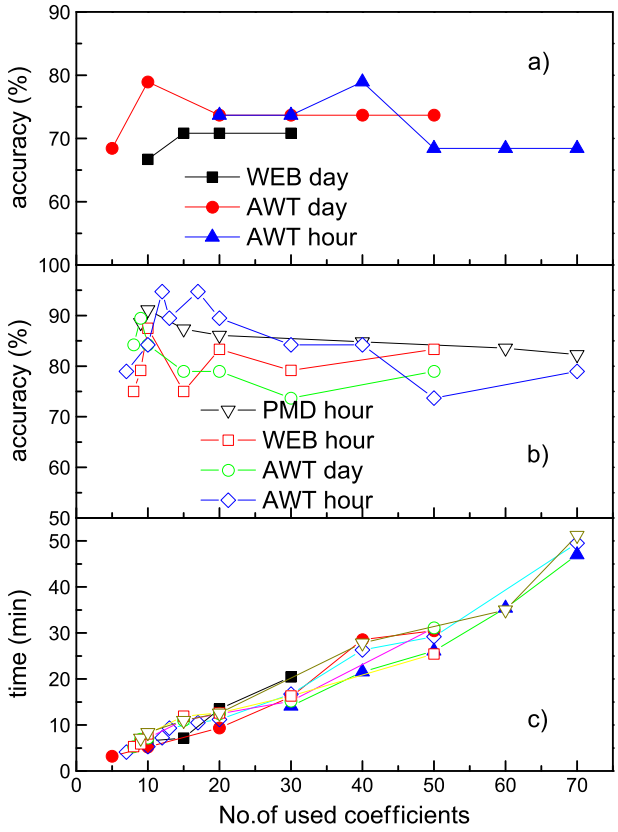



**Figure 4:** *Student performance models generated using artificial neuron networks. (a) Model accuracy dependence on number of coefficients used for datasets discretized into three categories; (b) for datasets discretized into two categories; and (c) time spent for model calculation (min) in dependence on number of used coefficients for the different datasets.*

information useful for student performance modelling is stored in these prominent coefficients.

In this research, potentially interesting periods are spread over wide part of frequency spectrum (Figure 3(b)). Therefore, the approach using best $k$ Fourier coefficients was chosen.

To check the possibility of modelling using compressed data in frequency domain, we built student performance models based on DFT calculated from the student access time series to the course web pages. In the first step, models are generated using artificial neural network (ANN) with genetic algorithm for attributes selection optimization using seven initial datasets selected from Table 1 ($AWT\,3C_{hour}$, $AWT\,3C_{day}$, $WEB\,3C_{day}$, $AWT\,2C_{hour}$, $AWT\,2C_{day}$, $PMD\,2C_{hour}$ and $WEB\,2C_{hour}$). The accuracy criterion is used to measure the number of positive and negative instances correctly classified in proportion to all other instances. Results are presented in Figure 4.

In Figure 4(a) and 4(b), the model accuracy is presented in dependence on number of used DFT coefficients. For all the used datasets, the model accuracy is increasing when the number of coefficients is increasing. After an optimal number of coefficients are reached, the model accuracy is at its maximum and then starts to decrease. The behaviour of model accuracies presented in Figure 4(a) and 4(b) is in accordance with the assumption that a certain number of DFT coefficients can be neglected because they do not bear information about performance. Removing those coefficients helps improving the accuracy of the model to the certain level. If the number of coefficients drops down below some optimal value, the accuracy of the model is decreasing. Likewise, if the number of DFT coefficients exceeds the optimal value, the accuracy of the model is also decreasing. This result confirms the assumption that Fourier coefficients that rarely appear in the student 20 most intensive Fourier coefficients set have low or even negative influence on accuracy of the student performance model. The optimal number is different for different datasets (Figure 4(a) and 4(b)). We assume that Fourier coefficients removed from datasets are behaving as noise that is covering important information about students. Expectedly, at the same time as the number of coefficients used for modelling is decreasing, the time used for model calculation is also decreasing. Reducing calculation time can be very important because some modelling algorithms require


© 2015 Wiley Publishing Ltd

computing time that could last for several hours and longer for a large number of coefficients and students. As an example, the time used for some calculation of ANN models with all possible coefficients lasts for more than 5 h using a standard desktop computer. Figure 4(c) represents the dependence of time required for an ANN model calculation on number of used Fourier coefficients.

In Table 3, the results of modelling with four datasets with three categories are listed. In the second column, the modelling algorithm is denoted; in the third column, the maximum number of coefficients in the dataset is listed; in the fourth column, the maximum achieved model accuracy is listed; and in the last column is the number of coefficients used to achieve the maximum accuracy listed in the previous column. It should be noted that if maximum accuracy appears for more than once, the maximum accuracy listed in the table is the result achieved with the smallest possible number of coefficients.

Similar table is carried out for the two-category datasets (Table 4). The maximum accuracies in both tables are achieved with the number of coefficients ranging from 5 to 30. There are few exceptions where the number of coefficients exceeds 30, and it is still significantly lower than the total (maximum) number of coefficients. Those results are confirming the assumption that reducing the number of coefficients in periodogram to the smaller number of most commonly (frequently) used coefficients will not degrade information about observed student and on the other side, according to Figure 4(c), will enable time-efficient model calculation. Model accuracies for all used algorithms listed in Tables 3 and 4 are comparable with accuracies found in literature sources (Kotsiantis *et al*., 2004; Delgado *et al*., 2006; Thai-Nghe *et al*., 2007; Romero *et al*., 2008; Lykourentzou *et al*., 2009; Kotsiantis *et al*., 2010; Jovanović *et al*., 2012; Gamulin *et al*., 2014; Lara *et al*., 2014) confirming that idea of using Fourier coefficients for student success prediction is plausible. Small differences between courses can be ascribed to the different numbers of students, student type (graduate, undergraduate, technology oriented or not), quantity and quality of resources posted at course page and so on. In future, the influence of those parameters to the success and accuracy of the model should be investigated.

**Table 3:** *The overview of accuracy and the required number of DFT coefficients for three-category models, $T_{hour}$ and $T_{day}$*

| *Dataset* | *Method* | *Maximum number of coefficient* | *Maximum accuracy (%)* | *Required number of DFT coefficient* |
|---|---|---|---|---|
| AWT | NB | 89 | 84 | 30 |
| $3C_{day}$ | kNN | 89 | 89 | 35 |
| | ANN | 89 | 70 | 10 |
| AWT | NB | 336 | 79 | 13 |
| $3C_{hour}$ | kNN | 336 | 79 | 11 |
| | ANN | 336 | 75 | 15 |
| WEB | NB | 80 | 75 | 12 |
| $3C_{day}$ | kNN | 80 | 83 | 20 |
| | ANN | 80 | 79 | 10 |
| WEB | NB | 409 | 79 | 7 |
| $3C_{hour}$ | kNN | 409 | 83 | 20 |
| | ANN | 409 | 78 | 40 |

DFT, discrete Fourier transforms; NB, naïve Bayes; kNN, *K*-nearest neighbour; ANN, artificial neuron networks.

**Table 4:** *The overview of accuracy and the required number of DFT coefficients for two-category models, $T_{hour}$ and $T_{day}$*

| *Dataset* | *Method* | *Maximum number of coefficient* | *Maximum accuracy (%)* | *Required number of DFT coefficient* |
|---|---|---|---|---|
| MP1 $2C_{day}$ | NB | 60 | 79 | 13 |
| | kNN | 60 | 76 | 13 |
| | SVM | 60 | 78 | 25 |
| | ANN | 60 | 78 | 40 |
| PMD $2C_{hour}$ | NB | 600 | 90 | 7 |
| | kNN | 600 | 92 | 15 |
| | SVM | 600 | 92 | 5 |
| | ANN | 600 | 91 | 11 |
| AWT $2C_{day}$ | NB | 89 | 95 | 40 |
| | kNN | 89 | 95 | 6 |
| | SVM | 89 | 89 | 7 |
| | ANN | 89 | 95 | 25 |
| AWT $2C_{hour}$ | NB | 336 | 89 | 12 |
| | kNN | 336 | 89 | 7 |
| | SVM | 336 | 89 | 17 |
| | ANN | 336 | 95 | 12 |
| WEB $2C_{day}$ | NB | 80 | 87 | 30 |
| | kNN | 80 | 84 | 10 |
| | SVM | 80 | 87 | 11 |
| | ANN | 80 | 83 | 11 |
| WEB $2C_{hour}$ | NB | 409 | 88 | 9 |
| | kNN | 409 | 88 | 8 |
| | SVM | 409 | 88 | 20 |
| | ANN | 409 | 88 | 10 |

DFT, discrete Fourier transforms; NB, naïve Bayes; kNN, *K*-nearest neighbour; SVM, support vector machines; ANN, artificial neuron networks.

Example of accuracy dependence on the number of coefficients used for AWT course dataset with two categories and support vector machine algorithm is presented in Figure 5. Comparison of two accumulation times $T_{\text{hour}}$ and $T_{\text{day}}$ for the first 100 coefficients is presented in the inset of Figure 5. It shows that along with increasing number of coefficients, the accuracy of the calculated model is increasing until some optimal number of coefficients is reached. In the specific case, optimal number is approximately 20. After that point, further increase of the used number of coefficients causes the decrease of model accuracy.

The accuracy change is not a monotonic function of the number of used coefficients as can be observed in Figure 5. There is oscillation of model accuracy with the number of used coefficients. That oscillation is prescribed to the importance diversity of coefficients used for model calculation. Genetic algorithm optimization shows that from the set of coefficients used in model calculation, some coefficients are used, while some other are not. In this experiment, coefficients are reduced by their position in the

© 2015 Wiley Publishing Ltd

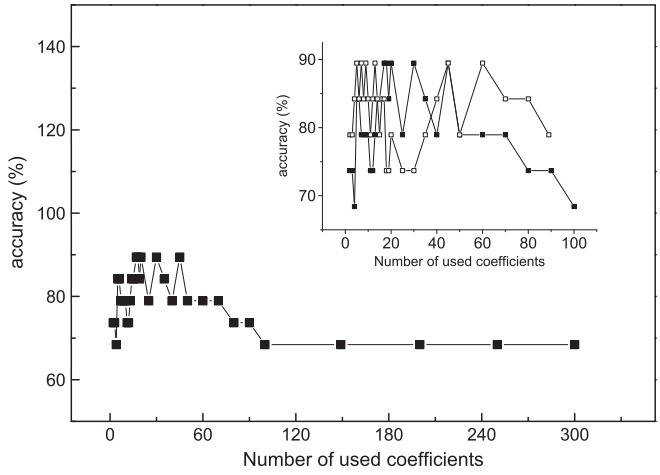



**Figure 5:** *Accuracy in dependence on the number of used coefficients for support vector machines algorithm and $AWT2C_{hour}$ course dataset with two categories when $T_{hour}$ is selected. In the inset of Figure 5, the comparison of accuracy with $T_{hour}$ (empty squares) and $T_{day}$ (full squares) is presented.*

dataset, from the most uncommon to the most common coefficients. During the elimination process, the importance of a certain coefficient is not considered, rather the intensity and the frequency of appearance of coefficients. Thus, it is possible that in some sets of coefficients, the higher number of unimportant coefficients is used than in some other set of coefficients, resulting in local decrease of model accuracy and oscillation of the accuracy presented in Figure 5. In a future investigation, an attempt to recognize and use important coefficients will be carried out.

In the inset of Figure 5, it is presented that in spite of the result oscillation depicted in both graphs, similar behaviour of model accuracy for different accumulation times can be observed. Additional check of these assumptions is performed by Wilcoxon matched pairs test. The test confirmed that the accuracy results on $T_{\text{hour}}$ and $T_{\text{day}}$ datasets show no significant difference when calculated for a certain course and using the same algorithm.

Comparison of the results for the two used accumulation times $T_{\text{hour}}$ and $T_{\text{day}}$ in Tables 3 and 4 and the results of Wilcoxon matched pairs test shows no significant difference between them indicating that the accumulation time does not influence modelling result. This is an important result because the choice of a certain accumulation time is not completely free. According to experience, if density of clicks per student is too low, the longer accumulation time is necessary to obtain meaningful periodogram. For instance, in a case of MP1 course, 300 students clicked approximately 100000 times during the semester, and the density of clicks per hour and per student proved to be too low to produce the useful periodogram. In a case of MP1 course and accumulation time 1 hour, all coefficients have similar intensity, and it was impossible to properly extract the 20 most intensive coefficients for each student. When extending accumulation time to 24 h, the periodogram with high-intensity coefficients suitable for modelling was obtained. In contrast, AWT course had approximately 100000 clicks too, but for 64 enrolled students only. The number of clicks per student and per hour was high enough to calculate Fourier transform that can be used for prediction modelling.

## 5. Conclusion

In this work, the authors show that student access time series generated from Moodle log files contain information sufficient for successful prediction of student final results in blended learning courses. It is also shown that if time series is transformed into frequency domain using DFT, the information contained in time series will be preserved, and resulting periodogram and its coefficients can be used for generating student performance models with the algorithms commonly used for that purposes. For lengthy courses, the number of coefficients can be huge. Nevertheless, the compression of data is possible. If, in average, all but 5–10% of the most intensive and the most commonly (frequently) used coefficients are removed from datasets, the modelling with the remaining data will result with the increase of the model accuracy. Compression of data is also decreasing model calculation time, allowing fast calculation on standard computers. The advantage of this approach is its applicability because the data are automatically collected in Moodle logs. Access time series can be produced using different accumulation times, depending on the course length and student access density. It is shown that selection of specific accumulation time is not influencing the final accuracy of the model for the most of the used algorithms.

It is true that different courses require different student online behaviours. One can see that also by comparing the DFT spectra, middle spectra of different courses are different, and the distribution of individual frequencies is different for two different courses. Nevertheless, the DFT spectra of similar students within one course are similar. In a future research, the student behaviour pattern should be connected to frequencies in DFT spectra. On the other hand, the experimental data indicate that it is possible to predict the student performance based on historical data within one course. It means that students with similar performance have similar LMS behaviour pattern within one course.

This paper pointed out some general tendencies, and it is up to the further research to check in more details the influence of accumulation time on final model accuracy. Also, other classifier performance measures aside from accuracy should also be considered (precision, sensitivity and specificity) as well as stricter reliability and validity


© 2015 Wiley Publishing Ltd

tests. Certain coefficients detected in observed periodograms can be connected with some events in course curriculum such as tests, laboratory exercises, seminars and midterm exams. In order to understand the student behaviour, it is important to connect as many coefficients as possible with real course events, and it is also the direction for the future work.

## References


AGRAWAL, R., K.I. LIN, H.S. SAWHNEY and K. SHIM (1995) Fast similarity search in the presence of noise, scaling, and translation in time-series databases, in *Proceedings of the 21th International Conference on Very Large Data Bases*, Zurich, Switzerland, Morgan Kaufman, 490–501.

BAGNALL, A. and G. JANACEK (2005) Clustering time series with clipped data, *Machine Learning*, **58**, 151–178.

BAKER, R. and K. YACEF (2009) The state of educational data mining in 2009: a review and future visions, *Journal of Educational Data Mining*, **1**, 3–17.

BUBAŠ, G. and D. KERMEK (2004) The prospects for blended learning in Croatian academic institutions, in *Proceedings of the 6th CARNet Users Conference CUC 2004*, Zagreb, Croatia.

CHEN, M.Y. and B.T. CHEN (2014) Online fuzzy time series analysis based on entropy discretization and a fast Fourier transform, *Applied Soft Computing*, **14**, 156–166.

COBO, G., D. GARCÍA-SOLÓRZANO, E. SANTAMARÍA, J.A. MORÁN, J. MELENCHÓN and C. MONZO (2011) Modeling students activity in online discussion forums: a strategy based on time series and agglomerative hierarchical clustering, in *Proceedings of the Fourth International Conference on Educational Data Mining Eindhoven*, The Netherlands, 253–258.

DEKKAR, W.G., M. PECHENIZKY and M.J. VLEESHOUWERS (2009) Predicting students drop out: a case study, in *Proceedings f the 2nd International Conference on Educational Data Mining, 2009*, Cordoba, Spain.

DELGADO, M., E. GIBAJA, M.C. PEGALAJAR and O. PÉREZ (2006) Predicting students' marks from moodle logs using neural network models, in *Proc. International Conference on Current Developments in Technology Assisted Education*, Sevilla, Spain, 586–590.

DENEUI, D.L. and T.L. DODGE (2006) Asynchronous learning networks and student outcomes: the utility of online learning components in hybrid courses, *Journal of Instructional Psychology*, **33**, 256–259.

DIAS, S.B. and J.A. DINIZ (2013) FuzzyQoI model: a fuzzy logic-based modelling of users' quality of interaction with a learning management system under blended learning, *Computers & Education*, **69**, 38–59.

DIVJAK, B. and D. OREŠKI (2009) Prediction of academic performance using discriminant analysis, in *Proceedings of the 31st Int. Conf. on Information Technology Interfaces (ITI 2009)*, Cavtat, Croatia.

DORÇA, F.A., L.V. LIMA, M.A. FERNANDES and C.R. LOPES (2013) Comparing strategies for modeling students learning styles through reinforcement learning in adaptive and intelligent educational systems: an experimental analysis, *Expert Systems with Applications*, **40**, 2092–2101.

FALAKMASIR, M.H., Z.A. PRADOS, G.J. GORDON and P. BRUSILOVSKY (2013) A spectral learning approach to knowledge tracing, in *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*, Memphis, TN, USA, 28–34.

GAMULIN, J., O. GAMULIN and D. KERMEK (2014) Comparing classification models in the final exam performance prediction, in *Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2014/CE)*, Opatija, Croatia, 781–786.

JADRIC, M., Z. GARACA and M. CUKUSIC (2010) Student dropout analysis with application of data mining methods, *Management*, **15**, 31–46.

JOVANOVIĆ, M., M. VUKIČEVIĆ, M. MILOVANOVIĆ and M. MINOVIĆ (2012) Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study, *International Journal of Computational Intelligence Systems*, **5**, 597–610.

KOTSIANTIS, S., C. PIERRAKEAS and P. PINTELAS (2004) Predicting students' performance in distance learning using machine learning techniques, *Applied Artificial Intelligence*, **18**, 411–426.

KOTSIANTIS, S., K. PATRIARCHEAS and M. XENOS (2010) A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education, *Knowledge-based Systems*, **23**, 529–535.

LARA, J.A., D. LIZCANO, M.A. MARTÍNEZ, J. PAZOS and T. RIERA (2014) A system for knowledge discovery in e-learning environments within the European higher education area – application to student data from Open University of Madrid, *UDIMA, Computers & Education*, **72**, 23–36.

LAURILLARD, D (2004) E-learning in higher education. In Ashwin, P. (ed.), *From Changing Higher Education*, Routledge Falmer, London, UK.

LYKOURENTZOU, I., I. GIANNOUKOS, V. NIKOLOPOULOS, G. MPARDIS and V. LOUMOS (2009) Dropout prediction in e-learning courses through the combination of machine learning techniques, *Computers and Education*, **53**, 950–965.

Moodle. Available at https://moodle.org/ (accessed July 2014).

OREŠKI, S., D. OREŠKI and G. OREŠKI (2012) Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment, *Expert Systems with Applications*, **39**, 12605–12617.

PALIWAL, M. and U.A. KUMAR (2009) A study of academic performance of business school graduates using neural network and statistical techniques, *Expert Systems with Applications*, **36**, 7865–7872.

Rapid Miner. Available at http://rapidminer.com/ (accessed July 2014).

REIS, M.S., P.M. SARAVIA and B.R. BAKSHI (2009) Denoising and signal-to-noise ratio enhancement: wavelet transform and Fourier transform, *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, **2**, 25–55.

ROMERO, C., S. VENTURA and E. GARCÍA (2008) Data mining in course management systems: moodle case study and tutorial, *Computers & Education*, **51**, 368–384.

ROMERO, C. and S. VENTURA (2010) Educational data mining: a review of the state-of-the-art, *IEEE Transaction on System, Man, and Cybernetics, Part C Applications and Reviews*, **40**, 601–618.

THAI-NGHE, N., P. JANECEK and P. HADDAWY (2007) A comparative analysis of techniques for predicting academic performance, in *Proceedings of 37th IEEE Frontiers in Education Conference (FIE'07)*, Milwaukee, USA, IEEE Xplore, T2G7–T2G12.

THAI-NGHE, N., A. BUSCHE and L. SCHMIDT-THIEME (2009) Improving academic performance prediction by dealing with class imbalance, in Proceeding of 9th IEEE International Conference on Intelligent Systems Design and Applications (ISDA'09), Pisa, Italy, IEEE Computer Society, 878–883.


© 2015 Wiley Publishing Ltd

Thai-Nghe, N., L. Drumond, A. Krohn-Grimberghe and L. Schmidt-Thieme (2010) Recommender system for predicting student performance, *Procedia Computer Science*, **1**, 2811–2819.

Vlachos, M., C. Meet and Z. Vagena (2004) Identifying similarities, periodicities and bursts for online search queries, in *Proceedings of the 23rd ACM SIGMOD International Conference on Management of Data*, Paris, France.

Warnakulasooriya, R., and W. Galen (2012) Categorizing students' response patterns using concept of fractal dimension, in *Proceedings of the 5th International Conference on Educational Data Mining*, Chania, Greece, 214–215.

# The authors

## Jasna Gamulin

Jasna Gamulin is a PhD student at the University of Zagreb Faculty of Organization and Informatics, working at the University of Zagreb School of Medicine, Center for International Cooperation. Aside from her work with international students at Medical Studies in English program, she has published eight scientific papers in international conference proceedings. Her areas of interest include educational data mining with prediction modelling and improving hybrid learning in higher education using web-based formative assessments as well as knowledge management.

## Ozren Gamulin,

Ozren Gamulin was born in Zagreb, Croatia. He received his PhD in Solid State Physics from the University of Zagreb, Croatia. He is author and co-author of more than 50 papers in international journals, conference proceedings and books. More than 30 published papers were cited in Current Contents database. Published papers are in field of semiconductor physics, biophysics, Raman spectroscopy, Fourier transform infrared spectroscopy and application of data mining and statistics in spectroscopy signal processing. Currently, he is an assistant professor and department head of the Department of Physics and Biophysics in the School of Medicine at University of Zagreb.

## Dragutin Kermek

Dragutin Kermek received his PhD in Information Sciences from University of Zagreb, Croatia. He has published over 50 research papers in various international journals, books and conferences. He served at the University of Zagreb Faculty of Organization and Informatics as Vice Dean for Academic affairs in three terms from academic year 2005/2006 to 2010/2011. Currently, he is the Full Professor in the Department of Theoretical and Applied Foundations of Information Sciences in the University of Zagreb Faculty of Organization and Informatics. His research interests include web engineering, design patterns, and e-learning.


© 2015 Wiley Publishing Ltd