The Mining and Analysis of Data with Mixed Attribute Types

Ed Wakelam, Neil Davey, Yi Sun, Amanda Jefferies, Parimala Alva, Alex Hocking
School of Computer Science
University of Hertfordshire
Hatfield, UK
e-mail: {e.wakelam, n.davey, y.2.sun, a.l.jefferies, p.alva, a.hocking3}@herts.ac.uk

***Abstract*— Mining and analysis of large data sets has become a major contributor to the exploitation of Artificial Intelligence in a wide range of real life challenges, including education, business intelligence and research. In the field of education, the mining, extraction and exploitation of useful information and patterns from student data provides lecturers, trainers and organisations with the potential to tailor learning paths and materials to maximize teaching efficiency and to predict and influence student success rates. Progress in this important area of student data analytics can provide useful techniques for exploitation in the development of adaptive learning systems. Student data often includes a combination of nominal and numeric data. A large variety of techniques are available to analyse numeric data, however there are fewer techniques applicable to nominal data. In this paper, we summarise our progress in applying a combination of what we believe to be a novel technique to analyse nominal data by making a systematic comparison of data pairs, followed by numeric data analysis, providing the opportunity to focus on promising correlations for deeper analysis.**

Keywords-Data Mining; Educational Data Mining; Data Analytics; Numeric, Nominal Data Analysis; Dimensionality reduction; Knowledge Extraction.

I. INTRODUCTION

We are initially investigating the potential to apply Artificial Intelligence (AI) techniques to improve e-learning systems in both educational and business settings [1], In particular, we are focussing upon how learning systems can be designed to adapt to individual students during the learning activity. This adaptability would enable the e-learning system to monitor and adjust the teaching based upon a wide variety of analyses of the knowledge and performance of the student. In order to achieve this, we are investigating how student attributes may be analysed and deployed.

Our first steps have been to perform a variety of analyses on open source published student data [2] in order to identify factors which correlate with student performance [3]. Significant advances in the field of data mining [4] are providing opportunities for tools to be deployed in analysing education data [5]. There have also been continued developments in Machine Learning (ML), which aims to determine how to perform important tasks by generalizing from examples [6].

These results may then be used to improve the design of adaptive learning systems [7] using contemporary AI techniques.

In section II, we discuss each of the types of student features relevant to our research: Categorical, comprising Nominal and Ordinal, and Measurement (Quantitative). Section III introduces the open source student data set which we have used to explore applicable analysis techniques. In section IV, we describe our experimental analysis of this data, summarising our results in section V. Finally, we discuss our conclusions in section VI including further work already underway and recommendations for future work.

II. EXISTING DATA ANALYSIS TECHNIQUES

A. Categorical Data

- *Nominal Features*

Nominal data is data where the feature values are labels such as male/female or yes/no. There are a number of statistical techniques available to analyse nominal data sets, notably Chi-square and Cramer's V [8]. Each has its own limitations, for example, sensitivity to sample size and a stronger than justified evidence of correlations [9].

In the case of nominal data, it is not possible to compare attributes directly in order to search for correlations. However, we can compare the correspondence between groupings of attributes and we have explored the use of what we believe to be a novel technique to do so. In this case, we have chosen to compare correlations between pairs of attributes [10]. Future work is underway to apply alternative nominal data analysis techniques to our data in order to compare our results and to identify the strengths and weaknesses of our technique.

- *Ordinal Features*

Ordinal data is a type of categorical data in which order is important. The originators of our data set do not categorise any of the student data captured in their study as ordinal.

B. Measurement (Quantitative) Data

There are a variety of statistical techniques available to analyse quantitative (numeric) data sets. In this case we have selected to use Principal Components Analysis (PCA) to reduce the dimensionality of our data and Growing Neural Gas (GNG) to identify potentially interesting clusters of data. GNG [11] has been successfully used to identify clusters in data for many applications such as the analysis of Hubble Space Telescope images [12] and automatic landmark

extraction in images [13]. PCA and GNG have also been successfully combined for intrusion detection [14].

III. PORTUGUESE STUDENT DATA SET

In order to investigate the predictive accuracy of student achievement data was taken from a set of students from a Portuguese study [15]. This data consists of information taken from two Portuguese secondary schools and each student has 33 attributes. The data includes three labels: first period grade, second period grade and final grade. The subjects are Mathematics (395 students) and Portuguese Language (649 students) and the data was collected during the 2005-2006 academic year. The attributes comprise 16 numeric (including the labels: first period, second period and final performance grades) and 17 nominal (Tables I and II).

TABLE I. EXAMPLES OF THE NUMERIC ATTRIBUTES

Identifier	Description
Age	Student's age (numeric: from 15 to 22)
Absences	Number of school absences (numeric: from 0 to 93)
Studytime	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

TABLE II. EXAMPLES OF THE NOMINAL ATTRIBUTES

Identifier	Description
Gender	Student's gender (binary: "F" - female or "M" - male)
Mjob	Mother's job (nominal: "teacher", "health" care related, civil "services" (e.g., admin or police), "at_home" or "other")
Romantic	With a romantic relationship (binary: yes or no)

For consistency we have adopted the original attribute types as used in the Portuguese study, although there are a small number of the attributes defined as numeric which could be considered as ordinal.

IV. EXPERIMENTAL ANALYSIS

A. Analysis of Nominal Data

Our method is to compare the correspondence between pairs of our nominal data attributes. To illustrate, the technique, here is a worked example of a data set of 4 students, each with 2 nominal attributes (Table III).

TABLE III. EXAMPLE DATA SET

Student	Attribute 1 (a1)	Attribute 2 (a2)
s1	p	x
s2	p	y
s3	q	z
s4	p	y

After setting a counter to zero we compare every possible pairing of student attribute values in the attribute 1 column of Table III with the corresponding pair in the attribute 2 column. If the selected pair from attribute 1 have the same value and the corresponding pair from attribute 2 also have the same value then we increment the counter by 1. Similarly if they both have different values then we increment the counter by 1. Otherwise, we decrement the counter by 1 (see Table IV).

So, for example, looking at step 1 below, the values of attribute 1 are both “p” (i.e., the same), whereas the values of attribute 2 are “x” and “y” (i.e., different), so we decrement the counter by 1. However, looking at step 2, the values of attribute 1 are “p” and “q” (different), and the values of attribute 2 are “x” and “z” (different), so we increment the counter by 1.

TABLE IV. STEP BY STEP PROCESS

Step	Student pairing	a1	a2	Score	Cumulative counter
1	(s1 s2)	(p p)	(x y)	-1	-1
2	(s1 s3)	(p q)	(x z)	+1	0
3	(s1 s4)	(p p)	(x y)	-1	-1
4	(s2 s3)	(p q)	(y z)	+1	0
5	(s2 s4)	(p p)	(y y)	+1	1
6	(s3 s4)	(q p)	(z y)	+1	2

We repeat this process for all combinations of attribute values and the resultant counter totals are used to populate a correlation matrix. This is done by inserting the counter total into the correlation matrix cell which corresponds to the respective attribute. Obviously, each attribute fully correlates with itself resulting in identical values across the matrix diagonal. We normalise our resulting matrix by dividing all entries by this value to keep all correlation matrix values between -1 and +1 (see Table V).

TABLE V. NORMALISED CORRELATION MATRIX FOR ILLUSTRATIVE EXAMPLE 1

	a1	a2
a1	1	$^1/_3$
a2	$^1/_3$	1

Positive values represent positive correlations between the respective attributes, negative values represent negative correlations and the magnitude of the value represents the strength of the correlation.

For example, where there are a high proportion of student pairs where the corresponding attributes, such as Mother’s job and gender are correspondingly the same or different this will result in a relatively higher correlation value (for example, 1/3 in Table V) between the two attributes.

For each attribute, we evaluate its correlation with all other attributes and find the mean value over all these correlations. As a first indicator of interesting attributes, particular attention was paid to those correlations where the magnitude of the mean value was high in comparison to the mean values of other attributes. Those correlations where the

magnitude was above the mean for that attribute then provided additional correlations for consideration.

We applied the technique to each of the Mathematics and Portuguese language data sets in turn. For each data set, we were then able to identify those pairs of attributes that were most strongly correlated – whether positively or negatively. This enabled us to consider the potential influences on student behaviours.

We were also able to compare the correlations in the Mathematics data set with those in the Portuguese language data set.

Using the correlation matrix generated by this technique we then produced corresponding PC1 v PC2 scatter plots for each of our Mathematics and Portuguese Language student data sets in order to visualize potential clusters for future analysis and comparison with any clusters identified in our numeric data. In order to visualize and more easily identify potential clusters we produced a PCA scatter plot for each of the four final grade intervals (using final grades 0-5, 6-10, 11-15, 16-20 as our labels) for each student data set.

B. Analysis of Measurement Data

After normalisation of the Mathematics and Portuguese Language student numeric data sets, respectively (by subtracting the mean and dividing by the standard deviation) we performed a linear Principal Component Analysis (PCA), plotting each of the leading three principle components, PC1 v PC2, PC2 v PC3, PC1 v PC3. In each Figure, the amount of variance accounted for by the respective principal components is reported. For example, in Figure 1 PC1 and PC2 account for 26% of the total information in the data.

In each case a visual inspection suggested possible clusters. In order to try and identify these clusters we applied GNG, with key parameters set to 50 training runs and a maximum of 200 nodes. This technique [16] identified a small number of clusters and their respective centroids as well as allowing us to identify the actual students in each cluster.

V. RESULTS

We are looking to identify interesting correlations in our student data attributes, providing the opportunity to focus on promising correlations for deeper analysis.

A. Nominal data

- *Mathematics students*

The top and bottom three cross-correlating attributes ranked by highest and lowest mean value are shown in Tables VI and VII respectively.

TABLE VI. HIGHEST MEAN VALUE MATHEMATICS STUDENT ATTRIBUTES

Attribute	Mean value
Higher education wish	0.23
School	0.19
Parent cohabitation	0.18

TABLE VII. LOWEST MEAN VALUE MATHEMATICS STUDENT ATTRIBUTES

Attribute	Mean value
Paid tutor	0.008
Gender	0.006
Extra-curricular activity	0.003

Our results show potential correlations may exist between the student's wish to take Higher Education and other nominal attributes - the school attended and parent cohabitation status, followed by receipt of extra educational support, Mother's job, access to the internet, the reason for choice of school and nursery school attendance.

Mother's job also shows potential correlations with other factors, including the wish for higher education, parent cohabitation, school attended, educational support and choice of school.

Paid extra tuition does not correlate strongly with other factors, even parent's jobs, which we might have expected. This is also true for students receiving educational support from within the family. However, future analyses may show that such extra tuition correlates with student performance measured by their grades.

Internet access also shows potential correlations with a number of factors, including the wish for higher education, school attended, parent cohabitation, address, the level of educational support by the school and Mother's job.

Factors which show very low correlations with others are the level of extra-curricular activities, whether the student was male or female and paid tutoring, followed by romantic relationships, Father's job, and family size.

- *Portuguese Language students*

The top and bottom three cross-correlating attributes ranked by highest and lowest mean value are shown in Tables VIII and IX respectively.

TABLE VIII. HIGHEST MEAN VALUE PORTUGUESE LANGUAGE STUDENT ATTRIBUTES

Attribute	Mean value
Paid tutor	0.20
Higher Education wish	0.18
Parent cohabitation	0.16

TABLE IX. LOWEST MEAN VALUE PORTUGUESE LANGUAGE STUDENT ATTRIBUTES

Attribute	Mean value
Family education support	0.02
Gender	0.01
Extra-curricular activity	0.003

Our results show potential correlations may exist between paid tutoring, the student's wish to take higher education and parent cohabitation followed by educational support and Mother's job.

Paid extra tuition shows potential correlations with a number of other factors including the level of educational support, the wish for higher education, parent cohabitation, and Mother's job. This is also true for extra educational support provided by the school, correlating with the use of paid tutors, parent cohabitation, and Mother's job.

Mother's job shows potential correlation with the use of paid tutoring, educational support, parent cohabitation and attendance at a nursery school.

Internet access only correlated modestly with other factors for Portuguese language students.

Factors which show very low correlations with others are the level of extra-curricular activities, student gender and family educational support, followed by romantic interest, guardian, Father's job and school attended.

- *Comparisons between Mathematics and Portuguese Language analysis results*

The wish to take higher education shows potential correlation with Mother's job, cohabitation status and receipt of extra educational support for both sets of students.

In both cases Mother's job correlates with other factors. In contrast, Father's job, along with romantic relationships and extra-curricular activities shows very low correlations with other factors in both sets.

Additional educational support provided by the school also shows potential correlation with a number of other factors in both sets.

In comparison with Portuguese language students, paid extra tuition in the case of Mathematics students does not correlate strongly with other factors.

Interestingly, gender, considered to be an influential factor, does not correlate well with other attributes in either set.

In the case of Mathematics students, internet access shows potential correlations with a number of factors, such as the wish to take further education, school attended, and parent cohabitation. However, in the case of Portuguese Language students, internet access shows only modest correlations.

- *Principal Component Analysis*

As described in section 1, above, a PCA projection will allow visualization of multi-dimensional data in a two dimensional representation. For each data set the initial PCA plot including all final grades proved too challenging to visualize and so we produced four plots, one for each of the four final grade intervals. We have included one example from each data set. Principle component analysis of our Mathematics and Portuguese Language student data shows no evidence of potential clustering.

For example, a PC1 v PC2 nominal data plot of Mathematics students' achieving final grades of between 11 and 15 (Figure 1).

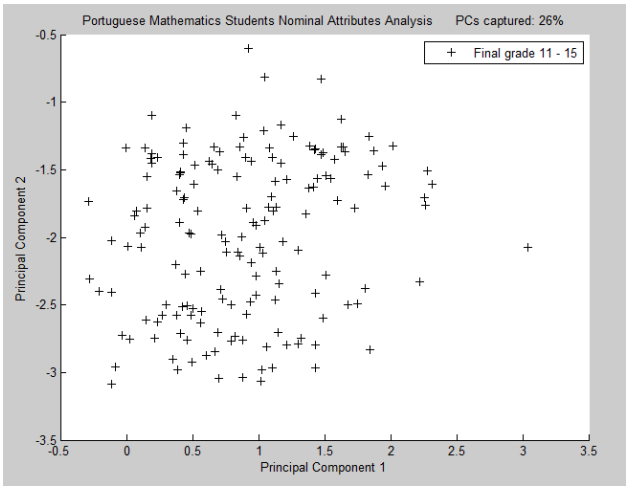

Figure 1. Mathematics nominal data PC1 v PC2 Final Grades 11-15

A further example shows a PC1 v PC2 nominal data plot of Portuguese language students' achieving grades of between 11 and 15 (Figure 2). This data plot appears to exhibit a lower boundary delineation which we believe to be a result of a predominance of very narrow variances in the attribute values in this particular data set.

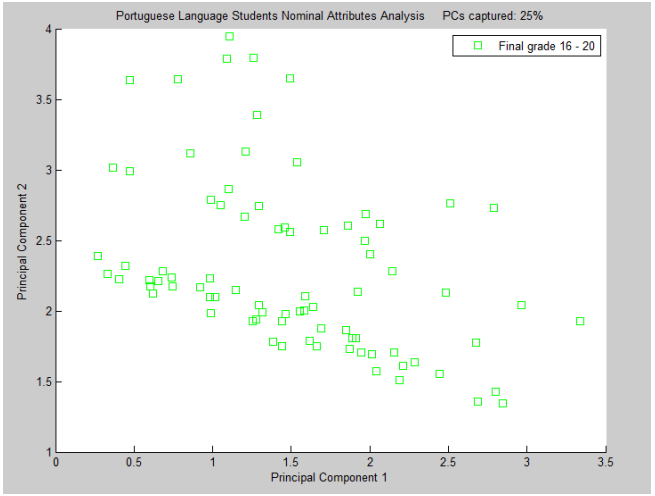

Figure 2. Portuguese Lang nominal data PC1 v PC2 Final Grades 16-20

B. *Measurement data*

- Mathematics students

GNG identified modest clustering in each of the PC1, PC2, PC3 comparisons For example, in Figure 3 we can see that three clusters have been identified. The centroids are shown in red and in each case the students in each cluster are identified in order to for look for potential correlations with the results of our nominal data analysis.

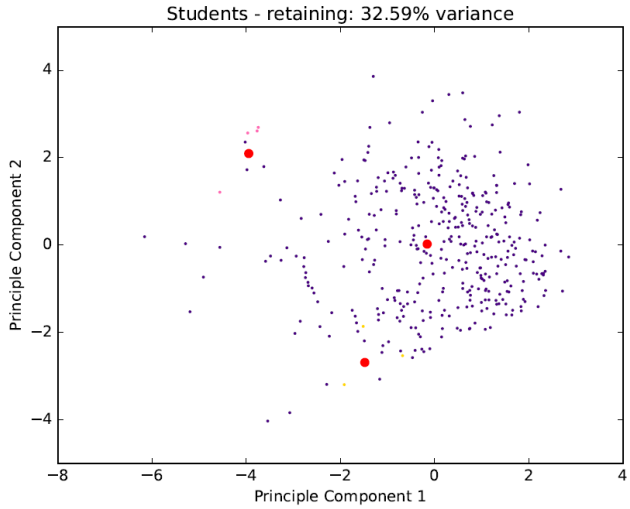

Figure 3. Mathematics students numeric data PC1 v PC2 scatter plot

- Portuguese Language students

GNG did not identify useful clustering in either of the PC1, PC2, PC3 comparisons. In all cases only one cluster was identified, for example, in Figure 4. As above, the centroids are shown in red.

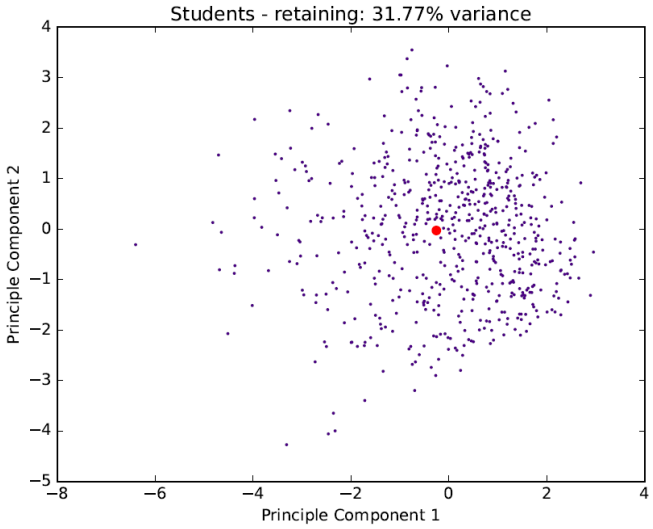

Figure 4. Portuguese Lang. students numeric data PC1 v PC2 scatter plot

We repeated the GNG analysis, adjusting the key parameters, increasing the number of training runs from 50 to 100 and maximum nodes from 200 to 600. However, this did not result in improvement. Further work is underway to identify alternative techniques to identify potential clustering in the Portuguese Language student numeric data, such as Curvilinear Component Analysis (CCA).

VI. CONCLUSION AND NEXT STEPS

In this paper, we have taken the first steps in exploring a mixed attribute type (numeric and nominal) data set provided by real student data with the objective of identifying useful potential correlations between attributes.

We have applied a novel approach to the analysis of the nominal data, comparing the correspondence between pairs of nominal attributes.

We then investigated if the analysis would identify interesting information in the data set, which to some extent it did. Our PCA plot of the Mathematics nominal data showed no evidence of clustering. Further work is underway to apply a non-linear visualization method in order to investigate potential clustering.

We then applied numeric data analysis techniques to identify clustering and potential correlations in our numeric attributes identifying some potentially interesting patterns.

In the case of our Mathematics student data using Principle Component Analysis followed by the GNG technique we were able to identify some clustering of the data, however the corresponding analysis of our Portuguese Language student data did not identify useful clusters.

Further work is underway to analyse and make comparisons between the numeric and nominal data sets to identify correlations, and subsequently to use these analyses to develop methods to predict student performance.

From the educational perspective, this would then allow us to perform follow up analyses on the extent to which different attributes can influence student achievement.

Future work includes the application of alternative nominal data analysis techniques to our nominal student data in order to compare the results and evaluate the advantages and disadvantages of these techniques in comparison with those of the technique deployed.

The novel nominal data analysis technique may provide a useful additional tool in the analysis of nominal data. We have shared the technique and corresponding MATLAB code with colleague researchers to gain further feedback on its usage and ideas on how to increase the sophistication of the method. Please contact us for a copy of the code.

REFERENCES

[1] E. Wakelam, A. Jefferies, N. Davey and Y. Sun, “The potential for using artificial intelligence techniques to improve e-Learning systems”, 2015.

[2] P. Cortez and A. Silva, “Using data mining to predict secondary school student performance”, in A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008), Porto, Portugal, April, 2008, pp. 5-12.https://archive.ics.uci.edu/ml/datasets/Student+Performance#

[3] V. Ramesh, P. Parkavi and K. Ramar, "Predicting student performance: a statistical and data mining approach". International journal of computer applications 63, no. 8, 2013, pp 0975 – 8887.

[4] P. Bhalchandra et al, "Prognostication of student’s performance: An hierarchical clustering strategy for educational dataset." In Computational Intelligence in Data Mining—Volume 1, Springer India, 2016, pp. 149-157.

[5] D. Fatima, S. Fatima, “A survey on research work in educational data mining.” IOSR Journal of Computer Engineering (IOSR-JCE), 17, 2015.

[6] T. Hastie, R. Tibshirani, J. Friedman and J. Franklin, “The elements of statistical learning: data mining, inference and prediction.” The Mathematical Intelligencer. doi:10.1007/BF02985802, 2005.

[7] D. Clow, "An overview of learning analytics." *Teaching in Higher Education*, 2013, pp. 683-695.

[8] A. Agresti, "Categorical data analysis." Vol. 996. New York: John Wiley & Sons, 1996.

[9] P. M. Bentler,, and D. G. Bonett, "Significance tests and goodness of fit in the analysis of covariance structures." Psychological bulletin, 88(3), 588, 1980.

[10] P. Ashrafi, "Predicting the absorption rate of chemicals through mammalian skin using Machine Learning algorithms." (Ph.D. unpublished). University of Hertfordshire, 2016.

[11] B. Fritzke, "A growing neural gas network learns topologies." Advances in neural information processing systems 7, 1995, pp. 625-632.

[12] A. Hocking, J. Geach, Y. Sun, N. Davey, N. Hine, "Unsupervised image analysis & galaxy categorisation in multi-wavelength Hubble space telescope images", Proceedings of the ECMLPKDD 2015 Doctoral Consortium (ECML 2015), 2015, pp. 105-114.

[13] E. Fatemizadeh, C. Lucas and H. Soltanian-Zadeh, "Automatic landmark extraction from image data using modified growing neural gas network." Information Technology in Biomedicine, IEEE Transactions on 7, no. 2, 77-85, 2003.

[14] G. Liu, and X. Wang, "An integrated intrusion detection system by using multiple neural networks." IEEE Conference on Cybernetics and Intelligent Systems, 2008, pp. 22-27.

[15] P. Cortez, and A. Silva, "Using data mining to predict secondary school student performance." In the Proceedings of 5th Annual Future Business Technology Conference, 2008, pp. 5–12.

[16] A. Parimala, "Using machine learning and computer simulations to analyse neuronal activity in the cerebellar nuclei during absence epilepsy." (Ph.D. unpublished). University of Hertfordshire, 2015.