Using a TwoStep Clustering Technique for Academic Streaming

Ely Salwana Mat Surin, Norizan Mohd Yasin and Suraya Hamid

Abstract—There are two types of main academic streaming: the science stream and the arts stream in Malaysia. It is crucial to assign students to the correct stream, to avoid a distribution imbalance of students in both streams; as this will affect the planning human capital. There are few problems in determining streaming, which include academic performance of students in schools does not link substantially to the academic streaming process and the academic streaming are not determined systematically due to a lack of understanding of the important factors that influence the streaming process. It is therefore difficult to identify the real potential of individual students. In order to stream students systematically, a TwoStep clustering technique is proposed to allocate students to appropriate streams based on their academic performance data.

Keywords—clustering, academic streaming, data mining, human capital

I. INTRODUCTION

Schools have been using various methods to determine the streaming of students such as parental demands, homogenous and heterogeneous skills, and cluster grouping. Recently, clustering methods have been used extensively in education for the distribution of students according to their performance (Bian, 2010; Gentry, 1999; Trivedi, Pardos, Sárközy & Heffernan, 2011; Zimmermann & Raedt, 2009). Information technology (IT) can be employed as a tool for the streaming of students and has a direct and positive impact on planning and education (Delahunty, 2000).

The IT techniques that can be used for the streaming of students include clustering and data mining (Agarwal, Pandey & Tiwari, 2012; Romero & Ventura, 2013; Shovon, Islam & Haque, 2012). When large data is required for better decision making, then an appropriate method, such as data mining, can be employed to extricate the information from repositories. The purpose of data mining is to locate valuable information among large clusters of data (Baradwaj, 2011) by concentrating on the use of a few methods and processes to determine the data trends (Shirwaikar & Rajadhyax, 2012). In a clustering method in data mining, the student data is divided into natural groups and a useful summary is provided on the students’ learning progress (Hämäläinen, Kumpulainen & Mozgovoy, 2013). This study aims to i) investigate current practices applied in academic streaming for secondary school in Malaysia and the student’s academic performance system; ii) identify factors contribute to the academic streaming process and propose an academic streaming framework based on identified factors; and iii) identify and apply a suitable clustering model to distribute students based on their academic performance.

II. BACKGROUND

Prior Research

Student Learning Pathway in education refers to academic streaming for a particular student (Cooper, Coll, Bartko, Davis, & Chatman, 2006; Taylor, 2007). It can be described as direction or pathway of the student through their development in school that will support and prepare them to meet and exceed their expectations (Mittendorff, Jochems, Meijers, & den Brok, 2008). Determination of academic streaming is done based on the students’ academic performance for a specific subject.

The clustering method in data mining helps in the streaming of students, since potential students in a particular grade are placed together in one classroom (Gentry, 1999). This method is proposed to help in the streaming of students by placing the students in several groups by means of natural clustering according to their academic performance for the whole year. Clustering models are used to establish natural groups that provide an accurate portrayal of the learning pattern of students and also support the underlying teaching processes and learning targets (Hämäläinen et al., 2013).

III. THE STUDY

The study was done on 5 high schools involving 17 teachers as the interviewees and 465 pupils as the survey respondents. Currently, student’s academic performance and student's academic streaming are linked manually in order to identify suitable academic streams for each student. There is no existing system where student's academic performance and student's academic streaming are embedded electronically and link to the existing system that could help to identify suitable academic streams for the students. Data collection is done in three phases: the first phase involves interviewing the teachers to explore the problem, and detailed understanding of an academic streaming phenomenon in high school; second phase involves conducting a survey on 465 students to collect opinions and to discover problems faced analysis. Finally, the quantitative data collected from the survey is analyzed using the Partial Least Squares Approach - Structural Equation Modelling (PLS-SEM), and data collected via document analysis is analyzed using IBM SPSS Modeler® 15. The data analysis is shown below.
Table 1 shows an example how themes are constructed. Main themes are developed based on some initial codes, however others formed sub-themes and remaining others were thrown out because of redundancy or there were some reason such as not covered in the main process of academic streaming. A summary of findings from the interviews is shown in Table 2.

Table II
A SUMMARY OF FINDINGS FROM THE INTERVIEWS VALIDATED BY THE RESPONDENTS

<table>
<thead>
<tr>
<th>Factor influences academic streaming</th>
<th>T2</th>
<th>T3</th>
<th>T7</th>
<th>T11</th>
<th>T13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guide</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Teacher Advice</td>
<td>√</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Parental pressure</td>
<td>√</td>
<td>√</td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Self-concept</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Aptitudes</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Parent’s aspiration</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Interest</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Skills</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Exam results</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Friends</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
</tbody>
</table>

For validity purpose, the study needs to identify factors that should be measured. In this study there are 13 factors from 5 different groups of important factors need to be measured. Each factor considered is able to influence student's academic streaming process. Each factor has its own loading which are used to evaluate the factor. These loadings are based on the survey question. After the PLS algorithm implemented, the below diagram shown in Figure 1.
TwoStep technique combines the ability of the K-Means clustering to handle a very large dataset, and the ability of the Hierarchical clustering (HCA – Hierarchical Cluster Analysis) to give a visual presentation of the results (see Table 3, Table 4, Figure 2 and Figure 3).

The predictor importance is, sc4 (science subject for test no 4) because it is seen as an important predictor compared to other subjects. The three models used in the analysis in terms of cluster size, clustering quality, the importance of the cluster and ratio size of the cluster. The cluster quality for TwoStep is 0.413 which is close to fair quality, but the quality for K-means and Kohonen are nearly low which is 0.282 (K-means) and 0.273 (Kohonen). The scale used is between -1 and 1. Value of 1 indicates good quality, whereas, value of -1 indicates poor quality. A value of -1 would mean all cases are located in the cluster centers of some other cluster. Due to the number of clusters produced by the K-means and Kohonen, the ratio of cluster size is also high. This is because the calculation of the ratio of cluster size is made based on the largest cluster divided smallest cluster. The small ratio shows no significant
difference in the distribution of the clusters. Among the three models analyzed, TwoStep shows the smallest value of ratio of size which is 3.19. While for K-means and Kohonen the ratio of size is 7.67 and 14.00. The smaller size of ratio, indicating better quality of clusters produced.

IV. DISCUSSION

There are several considerations that take place when a student decides which academic stream they would prefer to follow. The choice is affected by influences that were described earlier, and involve the opinion of their school, peers, family expectations or traditions, and employment prospects. These factors are: education institution factor (guide, teacher advice), peers factor (friends), family factor (parental pressure, self-concept, and aptitudes), historical trends factor (parent’s aspiration) and employment market factor (interest, skills, exam results). The TwoStep model was deemed to be the most appropriate, as it combines the K-Means and Hierarchical Cluster Analysis models. Based on the analysis, it is apparent that TwoStep is the most important model that suit for the educational data of students’ performance with importance value is 0.807 compared to K-means (0.496) and Kohonen (0.478). This show that based on the student’s performance, TwoStep is the most suitable model to group students into 2 categories or academic streaming class which is: art or science.

V. CONCLUSION

The outcome of this study has resulted in a clearer understanding of the factors that are involved in processes to determine the streaming of students based on academic performance. As teachers are involved in handling academic streaming, their views are important. However, as the student will ultimately be affected by the decision process, understanding their angle is equally valid. As academic performance is the basis of the decision process, students to a great extent are masters of their destiny. There are, however, a range of influences that may be encountered by students when they have to reach a decision on which stream is more suitable for them. The TwoStep model was considered to be adequate for the purposes of this study on academic performance data.

REFERENCES