Evolutionary data mining design to visualize the Examination Timetabling data at a University: a first round development

J. Joshua Thomas, Dr. Ahamad Tajudin Khader

Abstract - Examination scheduling ("timetabling") at a University is a determined challenge. Allocating exams to stipulate "time slots" requires most advanced quantitative techniques. This study takes an alternate approach of applying the principles of data mining (DM) explicitly using undirected data mining, data preprocessing to get the patterns in data then understand the relationship between them. Combining the general structure and methods of EA and the needs that DM can discover the meaningful and interesting patterns through a defined number of generations or until craving a pattern quality is achieved. In this paper, data mining techniques are deployed to visualize the examination timetabling preprocessing data (ETPD). From the perspective of computer discipline, this paper surveys techniques used and contributions from the field such as evolutionary data mining representation through artificial intelligence, and visualization. This institutional research has conducted at a University dataset.

Keywords: Evolutionary data mining, information visualization, examination time tabling

1. INTRODUCTION

Evolutionary Algorithms (EAs) have previously proved their usefulness in examination timetabling problem in a variety of forms in last few years and, in real-world problems rather effectively. There are few assortment of evolutionary algorithms with applications have been developed. The Principal objective of an EA is either to find something, generally to solve a problem. Probably the evolutionary algorithms are stochastic optimization algorithms based on the mechanism of inborn genetics. These search techniques investigate combinatorial search spaces using simulated evolution. This work considers the examination timetabling which is mainly concerned with the assignment of examination to the time slots. For that, there is a great need to amalgamate EA and Data Mining technique to analyze the data pattern, information (data) contained in timetabling problems. Several of data representation in various educational institutions, universities, and colleges are used; as these data are the inputs to the EAs to produce an optimal solution as timetables. However the optimal solution might not be efficient in all the cases. The data set which are used by most of the institutions are in different blueprint and with diverse flavors. Consequently the need of data mining on these multi-variants is considered necessary. The work includes effectively mine the examination timetabling preprocessing data (ETPD) by using data mining techniques for instance inductive methods and discovery of classification to categorize and classify the data into significant visualization then spot the irregularity and regularity in it. A university examination data set has been taken into account with the few benchmark datasets from ftp://ftp.mie.utoronto.ca/pub/testprob.

2. DATA MINING AND EVOLUTIONARY ALGORITHMS

It is a radical nature to connect or interrelate in some way on these two broad fields together and make the EA process and Mining processing rational [1]. EA consists of certain steps that take place in order for a problem to be solved (ETP). It is necessary to probe a bit into the terminology in order to clarify and get a hold of these ideas.

A. *Evolutionary Process*

An extremely important aspect of an EA is the representation of a chromosome and eventually its genes. It must be said that there is no universally accepted proposal of a representation; however the types are bit strings, real numbers, permutation of elements, list of rules, program elements (GP) and through accepted data structure. Typically the representation of an individual must be adapted to the problem at hand. As EAs seem to handle quite a variety of problems, especially in the responsive area of computer databases and is of remarkable interest however one has to define the application well. To be more specific we could consider flat file data and DBMSs (Database Management Systems). Data has another active part besides more formal, the aspect of knowledge discovery, referred to as data mining. The paper contributes to the light discussion on EA problem formulation and the data mining is to visualize the closed data into interesting classification data pattern with meaningful results [3].

B. *Denotation the processes of Data Mining*

In the simple case we consider data files which contain enough data for effective data mining. In here, we can think about the questions on the data. "Is this data clearly giving me some idea?", "What data mining method is to be used to visualize the data and spot the relationship or regularities?"

The importance on these questions gives rise to more serious approach.

1) Undirected data mining: This is the simplest case of data mining. User who is just interested in getting any kind of patterns on the data may have a sort of relationships that may be assumed. However this category of data mining is most frequently used.

2) Data sampling: The most time-consuming task of the whole data mining process is usually the evaluation of the patterns collected. It would be nice to work with small data but still get the same valuable results. Sampling is the process of identifying representative parts of a data to use in data mining [8].

3) *Data preprocessing:* The suggestion of preprocessing is primarily based on the need to prepare the data before they are actually used in pattern extraction. There are no standard preprocessing practices however, rather frequently used ones such as, extraction of derived columns that quantities will accompany but not directly related to the data patterns.

3. EXAMINATION TIMETABLING

Examination timetabling problem is a well known optimization (NP-hard) problem undergone by schools, universities and other educational institutions. There are wide number of dissimilarity on the idea of exam timetabling with different establishments having different requirements and constraints. In most cases there will be a limited number of rooms in which examination should be placed in timeslot. In the same way, some educational establishments will have their sub- institutional courses whereby very few conflicts with exams from other departments.

In the most straightforward timetabling problem, each examination timetabling problem has a set $E = \{e_1 \ldots e_n\}$ of exams and a set $T = \{t_1 \ldots tm\}$ of timeslots into which all n exams must be scheduled. In this case this research considers an examination timetabling problem whilst in that there are 575 exams, number of student is 14589, number of enrollment is 58469, number of timeslot is 40 and number of seat per timeslot is 2500 with *Hard constraints* ($c_1 \ldots c_5$). The constrains are:

- Number of students that have exams in 2 consecutive slots
- Number of students that have exams in 3 consecutive slots
- Number of students that have exams in more than 3 consecutive slots
- Number of exams that are scheduled in inappropriate slots
- Number of exams that have not been scheduled

We will not attempt to survey the very wide range of approaches available in hand, as we are sure that better new approaches will be made available compared to the current measures. However, the problem is reasonably stable, and the data for the problem is known to keep on advancing in the quest to find an accurate solution. We have a range of clustering algorithms which possess certain common similarities and dissimilarity dimensions which we may attempt to represent visually, that will be discussed in subsequent session. We consider a University examination sample dataset (USM), a benchmark dataset (Carter-Georgetown University Law Center USA)

4. PROBLEM SPECIFICATION

The mining system [6] built should be able to extract "n" data patterns (rules) from a data file that contains examination timetabling data and its characteristics. The database/flat file has the preprocessing data, which is, going to be visualizing the assignment schedules. Therefore, the interested individual can be informed for possible relations between columns and rows and plan future schedules accordingly. For our consideration the widespread of data has been built upon a flat file however, in general it should be represented as an Relationship Model is shown in Fig 1.

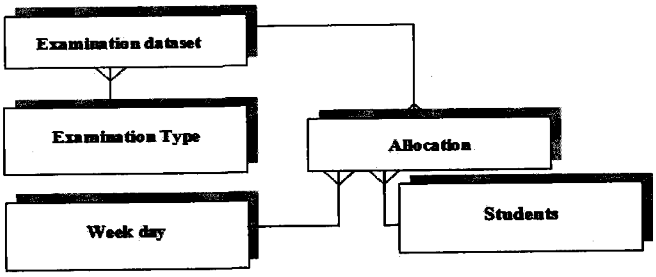


Fig. 1 Entity-Relationship Diagram, This model will be used for the data which is stored in excel, text files, Microsoft access files.

The most important attribute is the *examination*. This corresponds to an *enrollment* (e.g. 10, 20 etc,). There are *Examination types* (e.g. "Monday - Friday", with 3 exam periods) and *examination* is of a certain type. An *allocation is* spread for a certain *students* on a certain *weekday.*

5. THE SOLUTION

The problem has many aspects. It would be better to deal with one facet at a time to save curability. However the main concern on the solution is of the visualization part rather than the algorithmic measurement. Anyway here we explain the chromosome representation of the problem domain.

C. *Representation through data mining*

Before attempting to analyze the format of an EA, it is better to clarify the way mining will take place. As already mentioned, there are at least three types of data mining. At present, this paper deals with the first one, that is, undirected (pure) data mining using rattle package and a simple Genetic algorithms. This kind of mining allows us to extract explicit data patterns [9]. The explicit data patterns consist of three subsets E, S and P. E stands for Examination S for Slots and P for Prediction. Each subset refers to a portion of the dataset. In order to form a rule, they are connected by a suggestion. Separately form that we have novel visual representation to identify the data patterns.

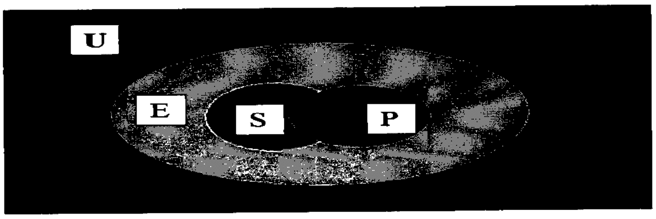


Fig-2 Undirected data mining a single suggestion

D. *Representation through chromosome arrangement*

The population is initialized using a model for each genome. It consists of various parts, those parts are columns and their relationships with each other are given below

1) Term: A term has the following form: "Variable"="Description" variable refers to a data variable.

2) Description: A description has of the following forms: "Value or "Numeric. A Value or Numeric may be set or initialized. Set refers the values remain unaltered whereby initialized means the opposite.

E. *The evaluation method*

How exactly the pattern going to be evaluated? What factor has much influence over the data? These are some of the interesting questions. One daring thing is to separate the variables of our data into functionally dependent and not dependent ones. A data pattern must be general and it must give the user a substantial and clear image of the insight data. A pattern must be accurate and precise. The measuring of accuracy and simplification shown below:

$$\text{correctness} = |E \cap S \cap P| / (|E \cap S \cap P| + |E \cap S \cap \overline{P}| \quad (1)$$

$$\text{generalization} = |E \cap S \cap P| / (|E \cap S \cap P| + |E \cap \overline{S} \cap P| \quad (2)$$

$$\text{Function} = |E \cap S \cap P| - (|E \cap C| \times |E \cap P|) / |S| \quad (3)$$

The evaluation function analyzed [8, 9] is simple and powerful.

F. EA Results

We run the experiments with the four benchmark data sets. They are in two different files that we consider here they are course. data, student. data. The 'course file' has two column of information for course versus enrollments and it is sorted in ascending order. The student data file contains courses which are organized in row basics. The Table below shows the best fitness, average fitness which has a pattern in the output. Let the fitness of the student data file has the lowest fitness of zero and the course data file has lower fitness in Uofulster course data set.

Bench mark dataset (file)	***Generation Number***	***Best Fitness***	***Average fitness***	***Standard deviation***
Georgecourse	1000	235.42	219.868	28.957
Georgestudent	1000	0.000	-1890.993	10.359
Nanyangcourse	1000	219989.831	206560.948	29682.673
Nanyangstudent	1000	**0.000**	**-93.492**	8.655
Swansea course	1000	918.147	858.231	119.498
Swansea student	1000	0.000	-1977.552	40.257
Uofulstercourse	1000	**8.978**	8.166	1.262
Uofulsterstudent	1000	**1711.608**	1601.452	222.896

Table 1 Simple Genetic Algorithm Results on few benchmark data to identify data patterns.

Table-2 illustrates the best member in the course data file which shows a pattern in the best member column on the 3 variable genome demonstrations.

Bench mark Data set (file)	***Generation Number***	***Number of variables***	***Best member***	***Best fitness***
Georgecourse	1000	3	2.000	235.642
Nanyangcourse	1000	3	2.000	219989.831
Swansea course	1000	3	2.000	918.147
Uofulstercourse	1000	3	4.999	8.978

Table-2 Simple Genetic Algorithm Results on course data file with best member patterns.

6. STATISTICAL VISUALIZATION

Clustering is one of the most popularly used data mining technique widely used in cross disciplines. The main perception of every researcher is to identify their data into significant groups among the data points, although this popularity of clustering is known for its theoretical properties. One of the main reasons is that it is very difficult to evaluate the quality of a partition of some given data set with other informal measures. A cluster should contain similar objects and should satisfy additional criteria [5]. In early days the discussion concentrated mainly on techniques, nowadays the whole process of clustering starting with the selection of cases and variables and ending with the validation of clusters become dominant.

A. Data preprocessing

The aim of this effort is to implement and evaluation of several commonly used clustering algorithms for examination timetabling data sets. The program has to fulfill the following requirements:

- Import of examination timetable datasets from flat files into the effective data mining system
- Knowledge visualization of the dataset. (Tree Maps, Parallel co-ordinates)
- Implementation of Hierarchical Clustering.
- 2- Dimensional visualization of the clustering results with exporting the results as images and data.

To show the usefulness of clustering programs all clustering algorithms will be evaluated by using an available dataset obtained from University Sains Malaysia (USM) and (Benchmark dataset)[4].

B. Classification: dissimilarity and similarity measures

Hierarchical methods have need of the dimension of a dissimilarity or similarity measure. As a result the distance measures and correlation coefficients are used to plan and distinguished the dissimilarity and similarity. We found that correlation coefficients and derived measures based on correlation coefficient have a smaller amount of benefit over the cluster cases. However, the distance measures and derived measures are apparently useful for the clustering variables. In divisive methods start with the assumption that all objects are part of a single cluster here in the dataset (2 x 575) variables which are department wise examination and timeslots. It computes a divisive hierarchical clustering of the dataset returning the object class *'diana'* [6].

The *'diana'-algorithm* starts with one large cluster containing all 575 observations then divided them until each cluster contains only a single observation. At each stage, the cluster with the largest diameter is selected. The metric string is used for calculating dissimilarities between observations is *"manhattan"* it leads to absolute differences on the dataset. If the *'diana' (**di**visive **ana**lysis* clustering) [6] metric is *'euclidean'* and the logical flag is TRUE then the dataset will be considered as dissimilarity matrix otherwise as a matrix of observations by two variables. Here the used technique is single linkage that requires nearest neighbor and it can be calculated by

$$d^{new}_{(p+q),i} = \min(d_{pi,} d_{qi}) res.s^{new}_{(p+q),i} = \max(s_{pi,} s_{qi}) \qquad (4)$$

Where $d^{new}_{(p+q),i}$ dissimilarity between the new cluster *(p+q)* with cluster *i*

$s^{new}_{(p+q),i}$ Similarity between the new cluster *(p + q)* with cluster *i*

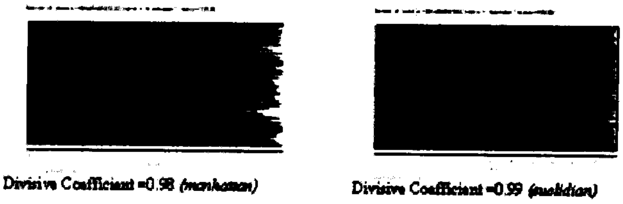


Fig-3: divisive method on the obtained dataset
3 (a) *manhattan* distance (b*) euclidean* distance.

The agglomeration schedule can be visualized as *dendrogram*. This method, helps to evaluate how many clusters are in the dataset, it has been determined on the basis of the *dendrogram*. The identification procedure is to mark the number of *'small'* hills (clusters) that combine objects at a low distance level. It has been visualized in Figure-4 illustrates which objects are combined in which step. In here, the used method is complete linkage it yields the agglomerative coefficient which measures the amount of clustering structure found from the obtained dataset. The equation for the re-calculation of dissimilarities and similarities are:

$$d^{new}_{(p+q),i} = \max(d_{pi,} d_{qi}) ress^{new}_{(p+q),i} = \min(s_{pi,} s_{qi}) \qquad (5)$$

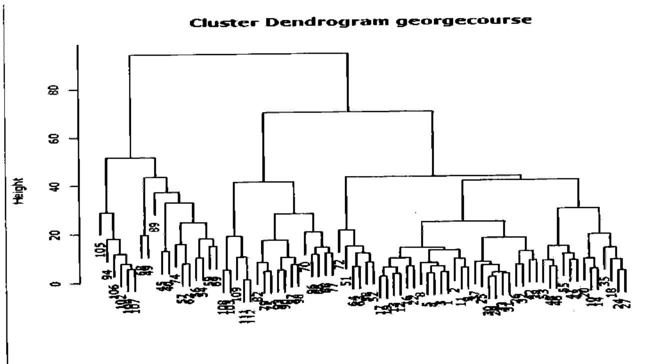


Fig-4: Agglomerative on the obtained dendrogram
(George Town University Course data set) refer Appendix

Complete linkage uses max function as results a very strong definition of the sameness of clusters are identified and it should be difference based on *'manhattan'* and *'euclidean'* distances. Complete linkage results in dilatation [5] and produce too many clusters in the *dendrogram* shows which objects are combined in which step. The below explanation is beneficial for better understanding.

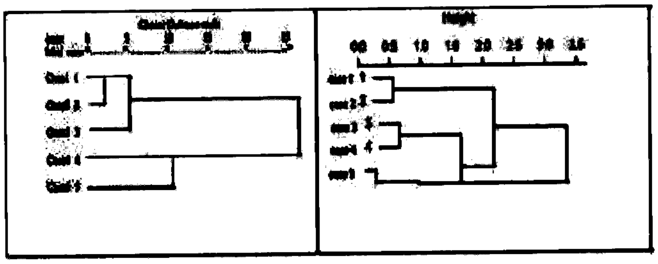


Fig-5: Dendrogram 4(a) Explanation of case, 4(b) Dataset-Y

In figure-5(a) the object (case) 1 and 2 are combined in step 1. In the step 2 objects 3 is merged into the cluster built by object 1 and 2. In step 3 object 4 and 5 are combined. On the top a cluster distance measuring scale is labeled; with respect to the 5(a), figure-4(b) is being measured and hence the distance is calculated in the form of height mentioned as 'small' hills. In another case the agglomeration levels are considered to identify the clusters, in the method of sharp increase (dissimilarity are clustered) or decrease (similarity are clustered) in figure-5(a) 4-5 a significant increase between step 3 and 4 hence, the best cut is step 3 instead of step 4 and it increased widely. Therefore, step 3 corresponds to the solution with two clusters in the same way; Figure-5(b) and rest of the data components are measured then the dissimilarity matrix has tabulated the higher value indicates the high dissimilarity.

DataSet	Objects	Matrix Dissimilarities		Min		Max		Mean		Median	
		manhattan	euclidean	manhattan	euclidean	manhattan	euclidean	manhattan	euclidean	manhattan	euclidean
DataSetS01022A	73	2628	2628	0.05495	1.000	7.037400	77.878	2.635400	30.911	1.747200	29.530
DataSetS01022B	35	595	595	0.22876	1.4142	7.15240	43.9320	2.82860	19.0890	2.76540	18.9740
DataSetS01022C	44	946	946	0.090909	1.000	6.508800	54.489	2.604200	24.462	2.579200	25.060
DataSetS01022D	5	10	10	1.6779	4.2426	5.8671	33.0150	3.1869	18.24	2.9392	17.1300
DataSetS01022F	22	231	231	0.18182	1.000	6.70470	36.8920	2.72520	16.4590	2.74340	16.1250
DataSetS01022H	67	2211	2211	0.16368	1.4142	7.05810	70.2140	2.73550	28.6070	2.61200	27.2950
DataSetS01022I	36	630	630	0.11111	1.000	6.50190	42.202	2.74740	19.429	2.64060	19.000
DataSetS01022K	22	231	231	0.27657	1.4142	6.95270	42.0480	2.79880	17.9770	2.56850	16.7630
DataSetS01022L	34	561	561	0.11765	1.000	6.48430	43.909	2.74340	19.529	2.62820	19.313
DataSetS01022M	28	378	378	0.14286	1.000	7.05180	41.6290	2.77620	16.6700	2.44430	15.0170
DataSetS01022P	54	1431	1431	0.074074	1.000	6.767000	58.613	2.683400	24.779	2.640600	24.187
DataSetS01022R	49	1176	1176	0.081667	1.000	6.867400	54.406	2.745800	23.824	2.645100	23.420
DataSetS01022S	44	946	946	0.090909	1.000	70249400	50.912	2.773200	21.723	2.644300	21.024
DataSetS01022U	14	91	91	0.57579	1.4142	5.93510	37.4830	2.82220	17.0400	2.94720	15.8110
DataSetS01022Y	15	105	105	0.26786	1.000	6.18720	36.0560	2.79150	15.7210	2.65490	14.1420
DataSetS01022Z	32	496	496	0.25000	2.000	6.7496	41.012	2.7153	18.353	2.7003	18.868
DataSetS01022	575	Variar		0.006957		3.576000		0.247700		0.152700	

Table 3: Dissimilarity matrix for the obtained dataset (USM)

7. INFORMATION VISUALIZATION

It is all about taking data and turning into useful information. It is a technique which is completely useful, and widely used. The massive collection of number (data) is difficult to identify with, and needs a lot of thinking to parse the data to get a grip on it. One way to improve this is to make this easier for the user to understand is if the computer processes the information for you by changing the raw figures into percentages. The goal of information visualization is to make the huge data coherent, present the data in several levels of details and tell stories about the data.

A. Advanced visualization

There are techniques available in Information visualization to render a complex hierarchical dataset in a relational view. It is capable of plotting many data points and keeping it visually consistent. This session will visualize the obtained dataset using TreeMaps which will be discussed as next sub-division.

B. Tree Maps

The Tree Map is a six line algorithm [11] has represented the information into a 2-dimensional rectangular map which provides visual representation of the dataset. In TreeMap everything is a node, and has a number of children. The parameters to the node are respectively, the Examination (department wise) or Slots, the size of the node and the color coded to the node. The size of the node is purely arbitrary and it computes the render size for the node based on this size and the total of all sizes at that level in the node tree.

Figure-6 illustrates the examination timetabling dataset variables (data points) are mapped into the rectangular area to represent 2 x 575 matrix of observations. This space- filling approach [11] highlights the insight of data into multiple regions. In the Figure-3 "DataSetS01022" is a top level node, containing two child nodes "Examination" and "Slots" as the dataset splits alphabetically [A..z] In addition it has been visualized with different values as parameters (60F). To implement TreeMap a .NET TreeMap control has been identified and used. The display shows the range of department wise examination and slot (data points) which build up the insight of the data into global picture. Suppose the user want to know different hierarchy it can be spot by the color combination. The main goal which we try to relate is to show the relationship between the hierarchical clustering with the TreeMap visualization which is not really applied and explained at this stage.

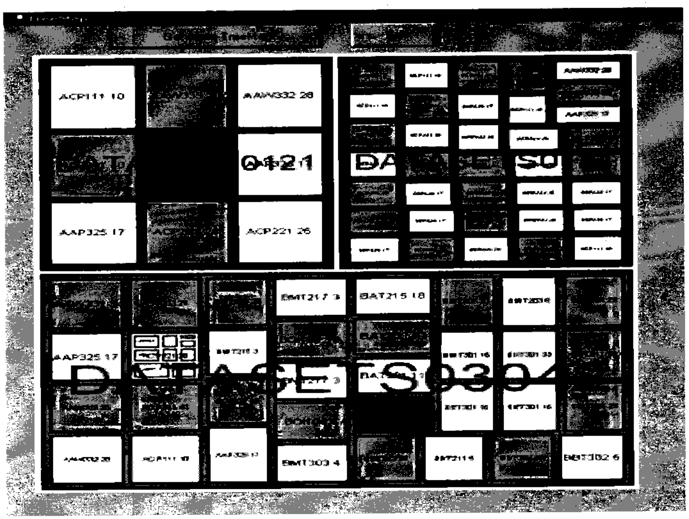


Fig -6: Information Visualization- Tree Maps

D. Parallel coordinates

Parallel coordinate plots to explore multivariate data. In fig-7 we explore the two variables data set which contains examination, timeslot and course, enrollment. The first variables are categorical and the second variables are continuous. Here we use brush mode, and make sure the focus in the parallel coordinate plots is on the first variable and highlight the cases of interest. Investigating on the outlier in the exam region 31 examinations are allocated to timeslots and 8 courses are enrolled by students Fig-7. By examining this case we change the focus of the plots to be in brush mode and brush the points.

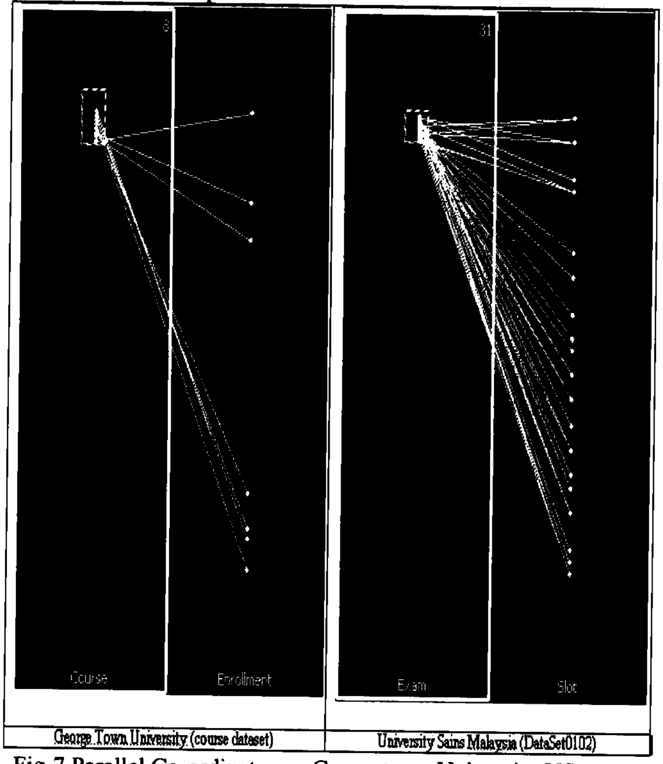


Fig-7 Parallel Co-ordinates on Georgetown University US, University Sains Malaysia Datasets.

8. CONCLUSION

This preliminary development of this work symbolizes the significance of EAs and data mining are important for the timetable designer to understand the data prior to the algorithmic simulations. Even though it needs more experimentation we intend to continue this research using the configuration mentioned in this work. There are several questions to be answered. This paper is the first approach to the time tabling area which might have new directions for the researchers.

9. REFERENCES

[1] Freitas, Alex A, *Data Mining and Knowledge Discovery with Evolutionary Algorithms.* New York Springer-Verlag 2002, pp. 31-45

[2] Wai-Ho Au, Keith C. C Chan, and Xin Yao, Fellow, IEEE. "*A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction", IEEE* Transactions on Evolutionary computation, Vol. 7, No. 6, December 2003.

[3] Wayne Smith, "*Applying Data Mining to Scheduling Courses at a University"* ,Communications of the Association for Information Systems Vol. 16, 463-474, 2005

[4] Muhammad Rozi Malim, Ahamad Tajudin Khader, Adli Mustafa *"Artificial Immune Algorithms for University Timetabling"*,E.K. Burke, H. Rudova (Eds): *PATAT* 2006, pp. 234-245.

[5] A.K Jain, M.N. Murty and P.J. Flynn. "*Data Clustering A Review*". ACM Computing Surveys, Vol. 31, No.3, September 1999.

[6] R Development Core Team (2006). "*R A language and environment for statistical computing"*. R Foundation for Statistical Computing, Vienna Austria. ISBN 3-900051-070-0, URL: http://www.r-project.org

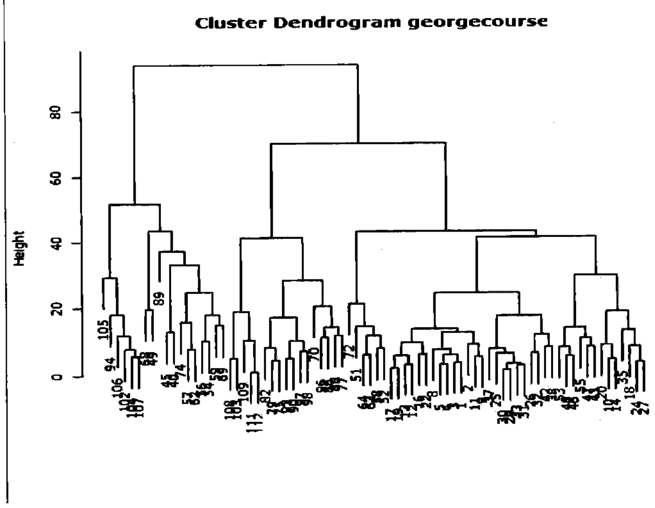
[7] Po Shun Ngan, Man Leung Wong, Wai Lam,, *"Medical Data Mining using evolutionary computation"*, ELSEVIER Artificial Intelligence in Medicine 16 (1999) pp 73-96.

[8] U.M Fayyad, and G. Piatesky-Shapiro, 1996. "*Advances in knowledge discovery and data mining, AAAI Press.*

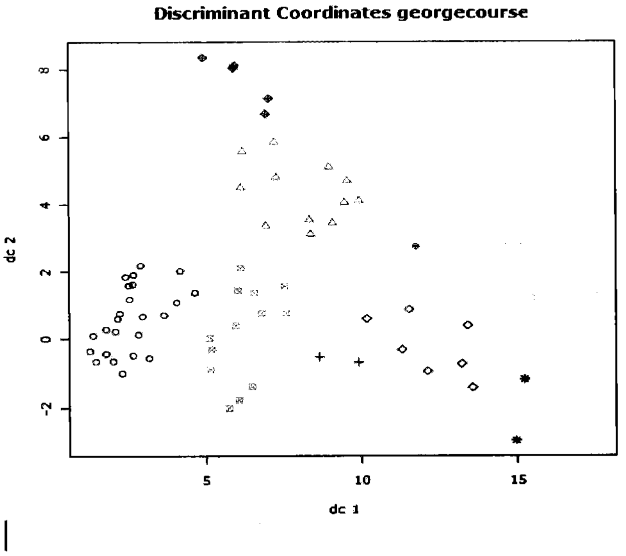
[9] I. W. Flockhart, N.J. Radcliffe, 1995. "*GA-MINER: Algorithms,* Final Report, the University of Edinburgh.

[10] T.Back, 1996. *"Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms.* Oxford, Univ. Press

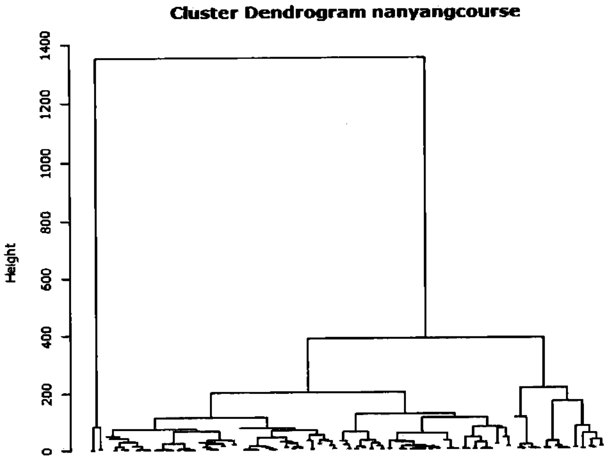
APPENDIX



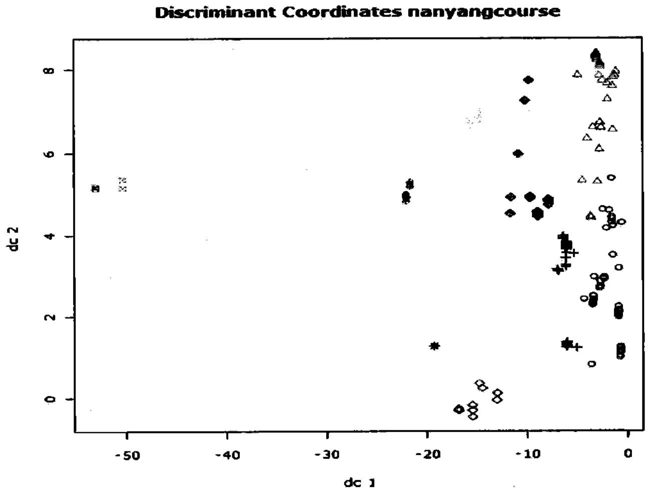
A (1) George Town University US Course data file



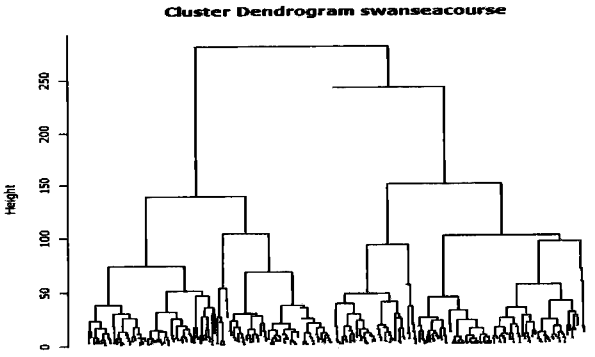
A(1) George Town UniversityUS Course data file discriminant



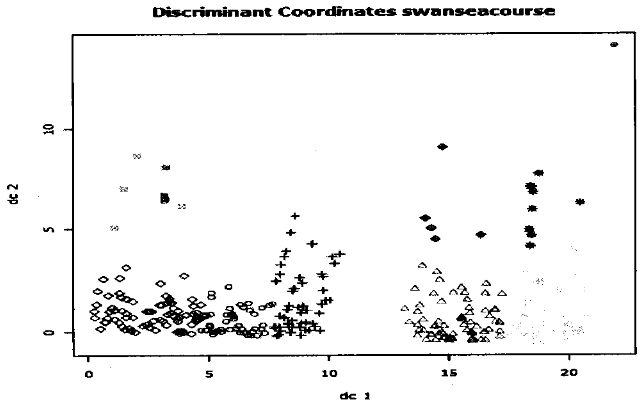
A(2) Nanyang University Course data file



A(2) Nanyang Univeristy Course data file discriminant



A(3) Swansea University Course data file



A(3) Swansea University Course data file discriminant