Prediction of Study Period Students (Bachelor Degree) Muhammadiyah University of Sidoarjo Based on Decision Tree Method using C4.5 Algorithm

IOP Conf. Series: Journal of Physics: Conf. Series **1179** (2019) 012033 doi:10.1088/1742-6596/1179/1/012033

Prediction of Study Period Students (Bachelor Degree) Muhammadiyah University of Sidoarjo Based on Decision Tree Method using C4.5 Algorithm

Arif Senja Fitrani

Department of Informatics, Faculty of Engineering, Universitas Muhammadiyah Sidoarjo, Sidoarjo, Indonesia.

rohman.dijaya@umsida.ac.id

Abstract. *The implementation of the study of stratum 1 (one) students is taken for 4 (four) years at the Muhammadiyah University of Sidoarjo. There are 10 faculty consisting of 27 departments with multi discipline and characteristic of study program of differentiate. Where in the study period is reached within the period of at least 4 years or 8 semesters. Students complete their study periods step by step. During the study period of 8 semesters, students' time cannot run in a timely manner, so the graduation projections will be delayed and the study period will be more than 8 semesters. Many of the factors behind this condition both on the student's initial ability, finance, and achievement of GPA in each semester. To know the factors that cause the condition of the student is not on time in taking the study period underlying various conditions, then conducted the study prediction period of study by using method C 4.5. It is expected that the condition underlying delayed graduation student of Muhammadiyah University of Sidoarjo can be anticipated through this research. With this, between providers of higher education can control each other over the condition of graduation students. The prediction achievement of study period using data mining decision tree method C 4.5 can produce approach or factors that influence student graduation with degree percentage to 85%.*

1. Introduction

Muhammadiyah Sidoarjo University (UMSIDA) as one of the private higher education providers. In carrying out the educational function, having achievements towards students to be able to run the academic process to the maximum and the accuracy in the study period that is in accordance with SNPT in 2013 and this is also a commitment to the vision of UMSIDA namely UMSIDA to become a national level quality university in 2020. For faculty profile and UMSIDA study program, there are 10 faculties consisting of 27 study programs. *The implementation of the study of stratum 1 (one) students is taken for 4 (four) years at the Muhammadiyah University of Sidoarjo. There are 10 faculty consisting of 27 departments with multi discipline and characteristic of study program of differentiate. Where in the study period is reached within the period of at least 4 years or 8 semesters. Students complete their study periods step by step. During the study period of 8 semesters, students' time cannot*

run in a timely manner, so the graduation projections will be delayed and the study period will be more than 8 semesters[1]. *Many of the factors behind this condition both on the student's initial ability, finance, and achievement of GPA in each semester. To know the factors that cause the condition of the student is not on time in taking the study period underlying various conditions, then conducted the study prediction period of study by using method C 4.5.*

2. Related Work

2.1. Classification based on Decision Tree

Classification is the process of finding a model or function that explains or distinguishes a concept or class of data, with the aim of being able to estimate the class of an object whose label is unknown[2]. Classification is a learning function that maps (classifies) an element (item) data into one of several classes that have been defined. Input data for classification is a collection of records. Each record is known as an instance or instance, which is determined by a tuple (x, y), where x is a set of attributes and y is a specific attribute, which is expressed as a class label (also known as a target category or attribute). Some classification techniques used are decision tree classifier, rule-based classifier, neural-network, support vector machine, and naive bayes classifier[3][4][5]. Each technique uses a learning algorithm to identify the model that provides the most appropriate relationship between the set of attributes and the class label from the input data. The decision tree that is induced is not always the same in some experiments because of the sequence or how to select features as branch breakers[6]. There are many choices of algorithms to induce decision trees, such as Hunt, CART (C & RT), ID3, C4.5, SLIQ, SPRINT, QUEST, DTREG, THAID, CHAID, and so on. If you pay attention to the selection criteria of the breaker branch, the decision tree induction algorithm is one of the Entropy (impurity) criteria which is mainly used in the algorithm ID3, C4.5, and C5.0. It is based on a choice of solving points which maximizes gain information (maximum entropy reduction). A minimum value of zero when all data on the node is owned by one class, this implies the most informative.

2.2. J48 Algorithm

The J48 algorithm is one class classifier in Weka's data mining application that implements the C4.5 algorithm. In building a model in the form of a decision tree, the C4.5 algorithm uses the information gain theory approach. C4.5 algorithm has advantages because it can produce a model in the form of a tree. The model produced by the C4.5 Algorithm (J48 algorithm in WEKA) is produced in the training process from the training data in the form of a decision tree. In C4.5 algorithm, the selection of attributes that will be processed using information gain[7]. If we choose an attribute to break objects in several classes, we must select the attribute that produces the greatest information gain.The information gain size is used to select the test attribute on each node in the tree. This size is used to select attributes or nodes in a tree. Attributes with the highest information gain value will be selected as the parent for the next node. Before calculating the gain, the entropy value must be calculated first. The information gain size is used to select the test attribute on each node in the tree. This size is used to select attributes or nodes in a tree. Attributes with the highest information gain value will be selected as the parent for the next node. Before calculating the gain, the entropy value must be calculated first.

2.3. Decision Tree Concept

Decision tree is a tree that is used as a reasoning procedure to get answers to the problems that are included. The tree formed is not always a binary tree. If all the features in the set use 2 types of categorical values, the tree shape obtained is in the form of a binary tree. If the feature contains more than 2 types of categorical values or uses a numeric type, then the tree shape obtained is usually not a binary tree[8]. Flexibility makes this method attractive, especially because it provides the advantage of visualizing suggestions (in the form of a descision tree) that can make the predictive procedure observable [9]. Descision tree is widely used to resolve decision-making cases such as in the fields of

medicine (diagnosis of patient disease), computer science (data structure), psychology and (decision making theory).

3. Research Methodology

3.1. Requirement Analysis System

In this sub-chapter will be explained related to what is needed to build the system. The stages starting from input, preprocessing, process, output as describes bellow.

3.2. Input Specification

In this study sampling is a series of data taken from the New Student Admissions Unit (PMB) and the Academic Administrative Bureau (BAA) of the Muhammadiyah University of Sidoarjo. The system only processes data from undergraduate (S1) students, 2010-2011 academic year. The number of training data is 791 students from 11 majors.

3.3. Pre-processing

Preprocessing steps to carry out the classification process are needed. Preprocessing the test and training data set (training). This process needs to be closer to accuracy and confined to the classification process. From the characteristics of the data provided, at least also pay attention to the expected process and output, so that the selection of features must be appropriate.

There are 17 variables from student data for test and training data sets. There are 6 (six) for categorical features for Gender, Department, Class, School and Job. Furthermore, 9 (nine) for the nomic features are the TPA Value variable, Semester 1 GPA up to Semester 8 GPA. And 2 (two) variables are the NIM student ID and Name. Feature values use categorical and nomic types, the details are as follows :

1. Gender features included in training data included in categorical properties include {Male, Female}.
2. Department features are used because there are types of Departments that are exact, educational, social and religious in the UMSIDA. But not all departments are used for training tests, from 27 majors that are only used by 11 majors with a large number of students. Features majors with categorical types include {Communication Science (ikom), Psychology (psi), Early Childhood Education Teacher Education (PGPAUD), Industrial Engineering (IND), State Administration (an), Electrical Engineering (ELEK), Mechanical Engineering (MSN), Management Economics (management), Accounting Economics (ak), Primary School Teacher Education (PGD), Informatics Engineering (INF)}.
3. The class system is determined in effective students to start lectures which are divided into morning classes and evening classes. In categorical features include {morning (P) and evening (S)}.
4. Before joining UMSIDA, prospective students go through the Academic Potential Test (TPA) process, where the character of the assessment is nomicic. Then the numerical multi splitting feature is divided into {<= 40>, <= 60>, <= 70>}.
5. In the series before joining UMSIDA, filling in personal data as well as the origin of the Vocational High School. The categorical features include {SMA, SMAS, MA, SMK}.
6. Taking into account the demographics of Sidoarjo and its surroundings are industrial areas and SME centers, the Work Status in categorical features includes {Not Working (TB) Working (B)}.
7. The results of the assessment of learning achievement at the end of the study program in each semester are expressed by the Grade Point Average (GPA). GPA retrieval starts from semester 1 to 8. The numerical features include {<= 2.5>, <= 3>, <= 3.5>}.

For the use of datasets, several variables are successfully obtained from student data, with various characteristics found in each student. The total data of class of 2010-2011 students used was 791

consisting of 507 students who were graduated for 8 semesters and 284 students were declared to have not graduated exactly for 8 semesters.

3.4. Pre-processing

In understanding a table form with attributes and records, a decision tree is obtained from it. Where the attribute states a parameter that is made for the criteria of decision tree formation.
In general, C4.5 algorithm in the stage of determining the decision tree can be passed by:

1. Starting from the root node.
2. For all features, calculate the entropy value for all samples (on nodes).
3. Select the maximum or highest gain information.
4. Use this feature as a node breakdown into a branch.
5. Recursively do each branch created by repeating steps 2 to 4 until all data in each node only gives one class label. Nodes that cannot be broken down are leaves that become the decision.

In determining the attributes that are positioned as roots by looking for the highest gain value of some of the attributes. From the highest attributes the formula shown in equation 3.1 can be used;

$$\text{Gain (S,A)} = \text{Entropy (S)} - \sum_{i=1}^{n} \frac{|si|}{|S|} * Entropy\ (Si) \quad \text{....(3.1)}$$

Where S is a set of cases, A is an attribute, n is the number of partition attributes A, whereas | Si | is the number of cases on the I and | S | partitions is the number of cases in S.
Then in the calculation process entropy value is used the formula shown in equation 3.2;

$$\text{Entropy } (s) = \sum_{i=1}^{n} - pi * \log_2 pi \text{.....(3.2)}$$

Where s is a set of case, A is a feature, n is the number of partitions s, while pi is proportion of the to s.
From the process of calculating the decision tree for GPA semester 2 as the root node with the attributes <= 3 and> 3, as in Figure 1. From the final decision tree calculation process for school features as an internal node, as in Figure 2.

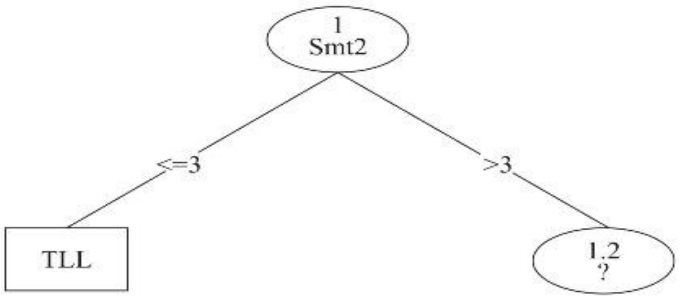

Figure 1. Decision Tree Node

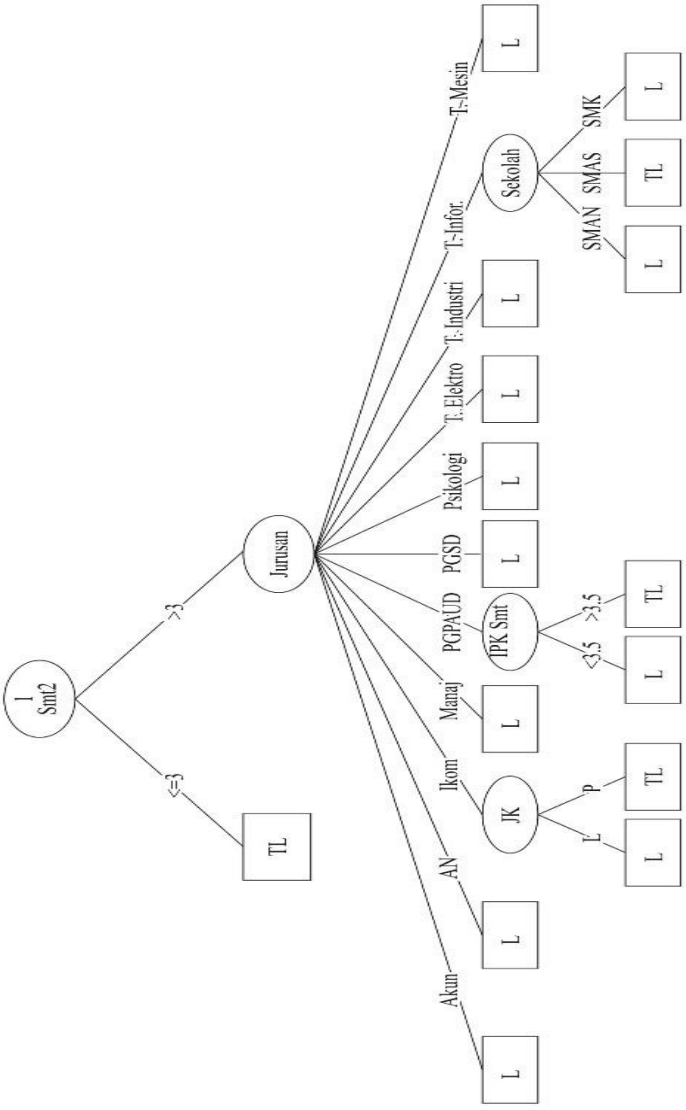

Figure 2. Decision Node Calculation Result

4. Determination of Training Set, Classify Decision Tree J48

By starting to open the WEKA program, proceed to open Explore, select the Explore button. Interface explore is already open, "Open file" image 5 to direct a training dataset that has been prepared with the datasetTra.arff file name, if successful it will display the attributes and amount of student data with a total of 791 students. And also accompanied by visualize according to the attributes of each image 6. Specifying the "Classify" tab, select "Choose" eka weka → classifyers → trees → J48 to determine the method to be used. Where the method for this research is Decision Tree C4.5 in this case the WEKA program is J48. From some Test options, the option "Use training set".

4.1. Training Set Process, Classify Decision Tree J48

If the steps are correct starting from preprocessing and classify, then do the process to get the Training set. Do "Start" to get the training set of figure 5.5. The part that results from the processing of this J48

method is the "Result list". Where is the Result list with id "00:04:48 - trees.J48" which will be saved as a model. Explanation of information from processing results in the "output classifier" with the text type includes Run information, J48 pruned tree, Evaluation on training set, Detailed Accuracy Class, and Confusion Matrix.

4.2. Save Model Output Training Set, Classify Decision Tree J48

The training set results will be taken to the Log interface, output Classifier and Result list. The log will record user activity. In the Classifier output, gradually the output in the form of text is described. And on the Result list is a marking on the process carried out from the method used. Right-click on "00:04:48 - trees.J48" on the "Result list" to save from the model that has been generated .

4.3. Data Test Stages, Classify Decision Tree J48

Repeating from the initial stage, starting from selecting Explore, opening the same Data.arff dataset as before. Then open the Classify tab, at this stage the difference is without specifying the method. In "Test options", select "Sumpplied test set" to open the data test file. Test data files have been prepared in advance, namely datasetTest.arff as in. The data will then be tested or predicted from accuracy by relying on training data that has been saved by the results. The attribute related to this data is the same as attribute training data. There is a description of the condition of data tests related to attributes as many as 15. Activating on "More options", "Output prediction" with status "Output additional attributes" is filled with "text".Next on the Result list, call back the model that was previously saved with the file name ModelDataTra.model. if it is successful, do the steps to "evaluate the model on current test set", which is to reprocess the new status of the test data. In a sense, the results of the stored test data are called again to appear in the Classifier output then the prediction process from the test data which has also been called.

4.4. Result

The prediction results for the graduation rate for 2 (two) data testing there are differences. Vocational testing data with a number of 278. From the testing of the application of testing data with different attribute models above, the level of difference or the percentage difference is 1%. Table 2.

Table 1 The output claassivier outputs Decision Tree C4.5

Inst	JK	Jurusan	Kelas	TPA	Sekolah	Kerja	IPK Semester								Status	Predicted	Error	Probaility	Distrib	Hasil
							1	2	3	4	5	6	7	8						
1	P	Akun	P	24	SMAS	TB	3.5	3.2	3.2	3.26	3.23	3.29	3.28	3.27	L	1:L	+	*1	0	B
2	P	Akun	P	51	SMAS	TB	3.5	3.2	3.2	3.26	3.23	3.29	3.28	3.27	L	1:L	+	*1	0	B
3	P	Akun	S	60	SMK	B	3.52	3.4	3.41	3.32	3.3	3.27	3.3	3.29	L	2:TL	+	0.143	*0.857	S
4	P	Akun	S	48	SMK	B	3.56	3.42	3.37	3.33	3.29	3.28	3.31	3.3	L	2:TL	+	0.143	*0.857	S
5	L	Akun	S	52	MA	TB	0	3.36	3.22	3.36	3.31	3.29	3.31	3.3	L	2:TL	+	0.2	*0.8	S
6	P	Akun	S	56	SMAS	B	3.48	3.35	3.37	3.33	3.29	3.3	3.31	3.32	L	1:L	+	*0.667	0.333	B
7	P	Akun	P	40	SMK	TB	3.25	3.26	3.27	3.3	3.25	3.3	3.32	3.31	L	1:L	+	*1	0	B
8	P	Akun	S	40	SMAN	B	3.42	3.36	3.36	3.35	3.34	3.3	3.32	3.29	L	1:L	+	*1	0	B
9	P	Akun	P	40	SMK	TB	3.4	3.38	3.4	3.4	3.34	3.38	3.33	3.32	L	1:L	+	*1	0	B
10	P	Akun	P	48	SMAS	TB	3.48	3.34	3.4	3.33	3.32	3.32	3.33	3.32	L	1:L	+	*1	0	B
11	P	Akun	S	44	SMK	TB	3.5	3.45	3.48	3.4	3.33	3.34	3.34	3.33	L	2:TL	+	0.2	*0.8	S
12	P	Akun	P	48	SMK	TB	3.46	3.45	3.41	3.37	3.33	3.35	3.35	3.33	L	1:L	+	*1	0	B
13	P	Akun	P	48	SMAN	TB	3.38	3.49	3.47	3.42	3.37	3.36	3.35	3.33	L	1:L	+	*1	0	B
14	P	Akun	S	48	SMK	TB	3.6	3.49	3.48	3.4	3.33	3.33	3.35	3.34	L	2:TL	+	0.2	*0.8	S
15	L	Akun	P	56	MA	TB	3.5	3.55	3.44	3.43	3.39	3.39	3.37	3.36	L	1:L	+	*1	0	B
16	P	Akun	S	48	SMK	B	3.81	3.65	3.6	3.54	3.44	3.45	3.44	3.41	L	2:TL	+	0.143	*0.857	S
17	L	Akun	S	24	SMAS	B	3.65	3.63	3.58	3.59	3.5	3.47	3.46	3.45	L	1:L	+	*0.667	0.333	B
18	P	Akun	P	44	SMAS	TB	3.5	3.51	3.52	3.46	3.43	3.47	3.46	3.46	L	1:L	+	*1	0	B
19	P	Akun	S	59	SMK	B	3.88	3.74	3.7	3.65	3.5	3.5	3.48	3.45	L	2:TL	+	0.143	*0.857	S
20	P	Akun	S	67	SMK	B	3.48	3.51	3.54	3.53	3.45	3.45	3.48	3.49	L	2:TL	+	0.143	*0.857	S
21	P	Akun	S	56	SMK	B	3.69	3.6	3.6	3.58	3.54	3.52	3.49	3.47	L	2:TL	+	0.143	*0.857	S
22	P	Akun	P	64	SMAN	TB	3.52	3.53	3.5	3.49	3.48	3.5	3.49	3.48	L	1:L	+	*1	0	B
23	P	Akun	S	51	SMAN	TB	3.79	3.62	3.63	3.63	3.54	3.5	3.5	3.48	L	2:TL	+	0.2	*0.8	S
24	P	Akun	P	40	SMAN	TB	3.44	3.52	3.52	3.53	3.47	3.48	3.51	3.48	L	1:L	+	*1	0	B
25	P	Akun	P	44	SMAS	TB	3.75	3.63	3.63	3.61	3.56	3.55	3.53	3.51	L	1:L	+	*1	0	B
26	P	Akun	P	76	SMAN	TB	3.5	3.53	3.52	3.57	3.52	3.51	3.53	3.51	L	1:L	+	*1	0	B
27	P	Akun	P	56	SMAS	TB	3.48	3.59	3.56	3.58	3.6	3.6	3.59	3.56	L	1:L	+	*1	0	B
28	L	Akun	P	64	SMAS	TB	3.73	3.78	3.71	3.68	3.64	3.65	3.65	3.63	L	1:L	+	*1	0	B
29	P	Akun	P	44	SMAN	TB	3.67	3.64	3.62	3.63	3.63	3.65	3.66	3.65	L	1:L	+	*1	0	B
30	L	Akun	S	55	SMAN	TB	3.96	3.86	3.82	3.79	3.75	3.75	3.73	3.68	L	2:TL	+	0.2	*0.8	S
31	P	Akun	S	64	MA	B	2.71	2.34	2.61	2.54	2.47	2.5	2.4	2.32	TL	2:TL	+	0.118	*0.882	B
32	L	Akun	S	32	SMAS	B	2.94	3.11	3.02	2.99	2.93	2.73	2.62	2.57	TL	2:TL	+	0.118	*0.882	B
33	L	Akun	S	40	SMAN	B	1.9	2.12	2.33	2.5	2.6	2.66	2.71	2.73	TL	2:TL	+	0.118	*0.882	B
34	P	Akun	S	44	SMK	TB	3	2.91	2.9	2.86	2.83	2.71	2.72	2.67	TL	2:TL	+	0.118	*0.882	B
35	P	Akun	S	36	SMAN	B	2.63	2.86	2.9	2.89	2.88	2.7	2.74	2.72	TL	2:TL	+	0.118	*0.882	B
36	P	Akun	S	44	SMAS	TB	2.94	2.99	2.96	2.96	2.91	2.75	2.78	2.76	TL	2:TL	+	0.118	*0.882	B
37	P	Akun	S	39	SMK	TB	2.94	2.76	2.86	2.88	2.86	2.79	2.79	2.75	TL	2:TL	+	0.118	*0.882	B
38	P	Akun	S	44	SMAS	B	2.51	2.76	2.78	2.82	2.85	2.79	2.8	2.71	TL	2:TL	+	0.118	*0.882	B
39	P	Akun	P	48	SMAS	TB	3	2.9	2.79	2.86	2.84	2.79	2.8	2.74	TL	2:TL	+	0.118	*0.882	B
40	L	Akun	S	32	SMAN	TB	3.19	2.98	2.9	2.85	2.84	2.79	2.82	2.77	TL	2:TL	+	0.118	*0.882	B

Table 2 Types of Data Testing Test Results

No	Attribute	Training	Testing	Suitable	Not suitable
1	All Attributes	513	278	200 (72%)	78 (28%)
2	Department Attributes	513	40	29 (73%)	11 (27%)

5. Conclusion

Many of the factors behind this condition both on the student's initial ability, finance, and achievement of GPA in each semester. To know the factors that cause the condition of the student is not on time in taking the study period underlying various conditions, then conducted the study prediction period of study by using method C 4.5. It is expected that the condition underlying delayed graduation student of Muhammadiyah University of Sidoarjo can be anticipated through this research. With this, between providers of higher education can control each other over the condition of graduation students. The prediction achievement of study period using data mining decision tree method C 4.5 can produce approach or factors that influence student graduation with degree percentage to 85%.

References

[1] Marín-Díaz V, López-Pérez M and Sampedro-Requena B E 2017 Personal Learning Environment within the Lecture Room: A Contribution from the Halls of Childhood Education Degree *Procedia - Soc. Behav. Sci.* **237** 360–4

[2] Dutta M K, Sengar N, Kamble N, Banerjee K, Minhas N and Sarkar B 2016 Image processing based technique for classification of fish quality after cypermethrine exposure *LWT - Food Sci. Technol.* **68** 408–17

[3] Herrera-Semenets V, Andrés Pérez-García O, Hernández-León R, van den Berg J and Doerr C 2018 A data reduction strategy and its application on scan and backscatter detection using rule-

based classifiers *Expert Syst. Appl.* **95** 272–9
[4] Pawening R E 2015 Classification of Textile Image using Support Vector Machine with Textural Feature *International Conference on Information, Communication Technology and System(ICTS)*
[5] Koivo H N 2008 NEURAL NETWORKS : Basics using MATLAB Neural Network Toolbox By 1–59
[6] Saettler A, Laber E and de A. Mello Pereira F 2017 Decision tree classification with bounded number of errors *Inf. Process. Lett.* **127** 27–31
[7] Panigrahi R and Borah S 2018 Rank Allocation to J48 Group of Decision Tree Classifiers using Binary and Multiclass Intrusion Detection Datasets *Procedia Comput. Sci.* **132** 323–32
[8] González M P, Lorés J and Granollers A 2008 Enhancing usability testing through datamining techniques: A novel approach to detecting usability problem patterns for a context of use *Inf. Softw. Technol.* **50** 547–68
[9] Baxter R, Hastings N, Law A and Glass E J . 2008 *Datamining: Concept, Models and Techniques* vol 39