Hybrid Genetic Algorithms Feature Selection and Decision Trees Classifier for Student's Computer Self-Efficacy

Wanphen Wirojcharoenwong[1], Montean Rattanasiriwongwut [2], and Monchai Tiantong[3]
King Mongkut's University of Technology North Bangkok, Thailand
[1]wi.wanphen@gmail.com
[2] montean@it.kmutnb.ac.th, www.it.kmutnb.ac.th
[3]monchai@kmutnb.ac.th, www.kmutnb.ac.th

Abstract - **Data mining techniques such as Decision Tree have been applied to the field of retail sale, e-commerce, banking etc. Data mining, the extraction of hidden relationship information from large dataset, is a powerful new technology with great potential to help business focus on the most important information in their data warehouses. Data mining techniques can learn normal and anomalous patterns from training data. Genetic algorithm can help selecting appropriate features and building optimum decision trees. In this paper, we propose a new hybrid mining approach in the design of an effective and appropriate to analyze student's computer self-efficacy model. The new proposed hybrid classification model is established base on a combination of genetic algorithm feature selection and decision tree (C4.5) result analysis using WEKA tool. The results show that a new hybrid classification model has even higher accuracy and lower complexity. The number of leaves and size of the constructed decision tree (i.e. complexity) are less, compared with decision tree models.**

Keywords - **Genetic Algorithms, Computer Self-Efficacy, Data Mining, Feature Selection, Decision Trees**

I. INTRODUCTION

The increased use of Computer and Internet technology is the main tool in education activities of students because students gather more information from internet and using digital library. Therefore, perceived computer self-efficacy among teachers and students play an import part in applying computer supported education and achieving its goal [4]. The paper is organized as follows: Section 2 cover a detailed confabulation on the related works done so far. Section 3 introduce the proposed Computer Self-efficacy, Decision Tree C4.5 and Genetic Algorithm Feature Selection techniques in construction of Decision Tree models for Student's Computer Self-Efficacy. Section 4 presents the experiment result and analysis from using the proposed method. Conclusion are discussed in Section 5.

II. RELATED WORK

In this section, we shall review the literature of Genetic Algorithm feature selection and Computer self-efficacy.

A. Genetic Algorithm

Min Chen and Ludwig, S.A. (2013) propose Fuzzy Decision Tree (FDT) classifier that is based on soft discretization by identifying the best "cut-point and applying a feature selection method that is based on the ideas of mutual information and genetic algorithms. The results show that FDT classifier obtains

in some cases higher values than other decision tree and fuzzy decision tree approaches based on measures such as true positive rate, false positive rate, precision and area under the curve [5].

Mohammad Khanbabaei and Mahmood Alborzi (2013) presents a new hybrid mining approach in the design of an effective and appropriate credit-scoring model. They are hybrid classification model is established based on a combination of clustering, feature selection, decision trees, and genetic algorithm techniques. The hybrid model choices and combines the best decision trees based on the optimality criteria. The results show the number of leaves and the size of the constructed decision tree are less, compared with other decision tree models [6].

B. Computer Self-Efficacy

Vehbi Celik and Etem Yesilyurt (2013). presents conducted in order to test the effect levels among the latent variables of attitude to technology, perceived computer self-efficacy, computer anxiety and the attitude toward doing computer supported education and these latent variables' ratios to each other. The participant group of the research consists of 471 pre-service teachers. Using exploratory factor analyses and the structural equation modeling techniques to analyses the data collected. The most significant finding of this study is that attitude to technology, perceived computer self-efficacy and computer anxiety are important predictors of teacher candidates' attitude toward using computer supported education [9].

Surej P John (2013) study is to identify the antecedents as well as the effects of computer self-efficacy on information systems acceptance and use. The study is conducted in the context of social networking sites adoption. 255 respondents from Bangkok, Thailand participated in this research. Structural equation modeling techniques have been employed to analyses the data collected. Results also show that Social factors do not play a major role in improving an individual's computer self-efficacy. Computer self-efficacy is found to be directly influencing perceived usefulness and indirectly influencing intention to use an information system [7].

III. INTRODUCTION TO COMPUTER SELF-EFFICACY, DECISION TREES C4.5 AND GENETIC ALRITHM FEATURE SELECTION

A. Computer Self-Efficacy

Computer self-efficacy was spread from the Social Cognitive Theory (Bandura, 1986) and has been derived from the broader construct of self-efficacy. Self-efficacy refers to person's capabilities in executing some tasks or facing challenges. Self-efficacy is thought to play a central role of behaviors individuals.

Computer self-efficacy (CSE) refer to a person's belief about his or her ability to use a computer to perform a computing task successfully. (Compeau & Higgins, 1995). Marakas, Yi and John (1998) propose that CSE affect not only a person's belief about his or her ability to use a computer to perform a computing task but also his or her intention to ward future use of computers.

B. Decision Tree C4.5

The decision tree C4.5 is a classification method for machine learning techniques and data mining. It is targeted at supervised learning [10]. One attractive method involves the construction of a decision tree, a collection of decision nodes, connected by branches, extending downward from the root node until terminating in leaf nodes. Beginning at the root node, which by convention is placed at the top of the decision tree diagram, attributes are tested at the decision nodes, with each possible outcome resulting in a branch. Each branch then leads either to another decision node or to a terminating leaf node [8] show in Fig. 1.

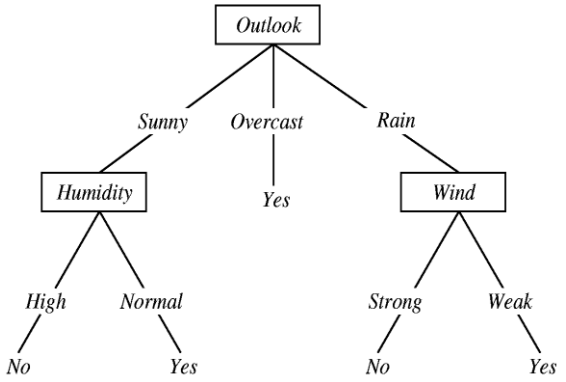

Fig 1. A Decision Tree for the Concept Play Tennis [8]

C4.5 algorithm was proposed in 1993, developed by J. Ross Quinlan. C4.5 algorithm is an extension of Quinlan's earlier ID3 algorithm. C4.5 is not one algorithm but rather a suite of algorithms. The generic description of how C4.5 works shown in Fig. 2. [10] All tree introduction methods begin with a root node that represents the entire, given dataset and recursively split the data into smaller subset by testing for a given attribute at each node. The subtrees denote the partitions of the original dataset that satisfy specified attribute value tests. This process typically continues until the subsets are "pure," that is all instances in the subset fall in the same class, at which time the tree growing is terminated [10].

Algorithm 1.1 C4.5(D)
Input: an attribute-valued dataset D
1: Tree = {}
2: **if** D is "pure" OR other stopping criteria met **then**
3: terminate
4: **end if**
5: **for all** attribute $a \in D$ **do**
6: Compute information-theoretic criteria if we split on a
7: **end for**
8: a_{best} = Best attribute according to above computed criteria
9: Tree = Create a decision node that tests a_{best} in the root
10: D_v = Induced sub-datasets from D based on a_{best}
11: **for all** D_v **do**
12: $\text{Tree}_v = \text{C4.5}(D_v)$
13: Attach Tree_v to the corresponding branch of Tree
14: **end for**
15: **return** Tree

Fig 2. Decision Tree C4.5 Algorithm [10]

C. Genetic Algorithm Feature Selection

Genetic Algorithms (GAs) is a heuristic search algorithm that mimics the processes by natural selection operates and has been successfully applied in many search optimization and apply them to solve business and machine learning problems.

John Holland developed the Genetic Algorithm in 1970. GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution factors as mate selection, reproduction, mutation, and crossover of genetic information [2], [4]. GA belong to the larger class of evolutionary algorithms (EA).

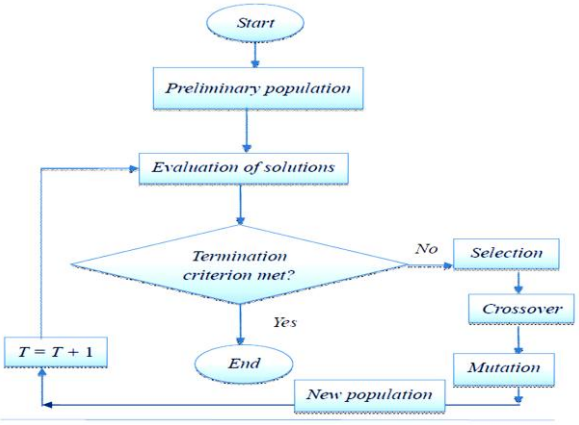

Fig 3. Simple Genetic Algorithm [4]

IV. EXPERIMENT RESULT AND ANALYSIS

The main purpose of this study the GA and Decision Tree C4.5 hybrid could produce a better classification than the current best performer of Decision Tree C4.5 alone.

A. Dataset

The data were drawn from students at five University in Thailand. Over two semesters, 2025 usable questionnaires were collected in the face-to-face classes. It has to be cleaned up during the pre-processing stage by eliminating incomplete data and remove of noise or outliers and handling missing data. The data 1436 pass on to the next stage of data mining. The attribute are as follows:

No.	Attribute	Class
1	sex	2
2	age	3
3	education background	3
4	major	4
5	faculty	4
6	year	numeric
7	gpa	6
8	Computer owner	2
9	Experience in Computer	4

10	Experience in Internet	4
11	frequency	4
12	Job1	6
13	Job 2	6
14	Job 3	6
15	place	3
16	average	3
17	decision	4

B. Evaluation Methods

We used WEKA 3.7.13 for feature selection and classification. The feature extractor will produce numerical outputs in the form of Attribute Related File Format (ARFF) files. The Decision Tree models were performed based on 10-fold cross-validation technique.

In order to compare the performance of the Decision Tree C4.5 without feature selection and Decision Tree C4.5 with GA-based feature selection method. The description Decision Tree C4.5 without feature selection of run information is show in Fig. 4.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     WekaExcel
Instances:    1436
Attributes:   17
              sex
              age
              edu
              major
              course
              year
              gpa
              owner
              exp_com
              exp_net
              frequency
              job1
              job2
              job3
              place
              average
              decision
Test mode:    10-fold cross-validation

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         789               54.9443 %
Incorrectly Classified Instances       647               45.0557 %
Kappa statistic                          0.1894
Mean absolute error                      0.2612
Root mean squared error                  0.4092
Relative absolute error                 84.1067 %
Root relative squared error            103.8923 %
Coverage of cases (0.95 level)          85.585  %
Mean rel. region size (0.95 level)      62.5174 %
Total Number of Instances             1436

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.209    0.042    0.291      0.209    0.243      0.195    0.655     0.171     computer
                 0.832    0.630    0.606      0.832    0.701      0.229    0.635     0.632     database
                 0.269    0.055    0.405      0.269    0.323      0.257    0.700     0.287     internet
                 0.203    0.098    0.425      0.203    0.275      0.139    0.580     0.366     network
Weighted Avg.    0.549    0.374    0.510      0.549    0.507      0.206    0.630     0.484

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
  23  62   6  19 |   a = computer
  30 642  30  70 |   b = database
   7 106  47  15 |   c = internet
  19 250  33  77 |   d = network
```

Fig 4. Run Information C4.5 without Feature Selection

In Fig. 5, description Decision Tree C4.5 with GA-based feature selection method (GATree). The chosen crossing factor was 0.6, the mutation probability 0.033, the population 20 and the number of generations 20. There are two evaluation methods to evaluate the predict performance of the new proposed hybrid Student's Computer Self-Efficacy model and decision tree used for comparison in this paper: 1) Percentage of the correctly classified instances, 2) Number of leaves of the tree, and 3) Size of the tree.

V. EXPERIMENTAL RESULTS

We used descriptive machine learning to obtain experimental results. Table I elaborates characteristics of the decision tree constructed without feature selection in the Student's Computer Self-Efficacy dataset.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     completedata922_predicted-weka.filters.unsupervised.attribute.Remove-R10-11
Instances:    1436
Attributes:   10
              Sex
              Age
              Education
              Major
              GPA
              Owner
              Fre
              Job
              Place
              Decsion
Test mode:    10-fold cross-validation

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1382               96.2396 %
Incorrectly Classified Instances        54                3.7604 %
Kappa statistic                          0.9281
Mean absolute error                      0.0258
Root mean squared error                  0.126
Relative absolute error                  9.8614 %
Root relative squared error             34.5434 %
Coverage of cases (0.95 level)          98.8858 %
Mean rel. region size (0.95 level)      29.7702 %
Total Number of Instances             1436

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.731    0.007    0.655      0.731    0.691      0.686    0.918     0.719     internet
                 0.977    0.034    0.959      0.977    0.968      0.941    0.992     0.990     network
                 0.965    0.023    0.978      0.965    0.972      0.942    0.989     0.987     database
                 0.500    0.001    0.857      0.500    0.632      0.653    0.908     0.634     computer
Weighted Avg.    0.962    0.028    0.963      0.962    0.962      0.934    0.989     0.981

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
  19   4   3   0 |   a = internet
   4 632  11   0 |   b = network
   4  21 725   1 |   c = database
   2   2   2   6 |   d = computer
```

Fig 5. Run Information C4.5 with GA-Based Feature Selection

In the next table (Table II), we used characteristics of the decision tree constructed by the new proposed hybrid classification model in the Student's Computer Self-Efficacy dataset. Using 9 attributes for predictive.

TABLE I
CHARACTERISTICS OF C4.5 CONSTRUCTED WITHOUT FEATURE SELECTION IN THE STUDENT'S COMPUTER SELF-EFFICACY DATASET.

Total number of Instances	***Correctly Classified Instances***	***Percentage of the correctly classified instances***	***Number of Leaves***	***Size of the tree***
1436	789	54.9443 %	285	391

TABLE II
CHARACTERISTICS OF C4.5 CONSTRUCTED BY THE NEW PROPOSED HYBRID CLASSIFICATION MODEL IN THE STUDENT'S COMPUTER SELF-EFFICACY DATASET.

Total number of Instances	*Correctly Classified Instances*	*Percentage of the correctly classified instances*	*Number of Leaves*	*Size of the tree*
1436	1382	96.2396%	108	142

VI. CONCLUSIONS

This study has proposed a new hybrid classification model for designing the Student's Computer Self-Efficacy. C4.5 constructed without feature selection, it was a large tree containing more leaves. Number of leaves and size of the tree in the decision tree (i.e. complexity) of the new propose hybrid classification model in this paper were lower than of decision trees. This shows that the decision tree of the new propose hybrid classification model has even higher accuracy and lower complexity.

REFERENCES

(Arranged in the order of citation in the same fashion as the case of Footnotes.)

[1] Compeau, D.R. and Higgins, C.A. (1995). "Computer Self-Efficacy: Development of a Measure and Initial Test". MIS Quarterly, 19, (2), pp. 189-211.

[2] Daniel, T.L. and Chantal, D.L. (2015). "Data Mining and Predictive Analytics". 2nd Ed. John Wiley & Sons., USA.

[3] Marakas, G.M., Yi, M.Y., and Johnson, R.D. (1998). "The multilevel and multifaceted character of computer self-efficacy: To-ward clarification of the construct and an integrative framework for research". Information Systems Research, 9(2), pp. 126-163.

[4] Mitchell, M. (2002). "An Introduction to Genetic Algorithms second edition". MIT Press, Cambridge, Mass.

[5] Min, C. and Simone, A.L. (2013). "Fuzzy decision tree using soft discretization and a genetic algorithm based feature selection method". Proceedings of Nature and Biologically Inspired Computing (NaBIC). pp. 238-244.

[6] Khanbabaei, M. and Alborzi, M. (2013). "The use of Genetic Algorithm, Clustering and Feature Selection Techniques in Construcion of Decision Tree techniques in construction of Decision Tree model for Credit Scoring". International Journal of Managing Information Technology (IJMIT) Vol.5, No.4, November 2013. pp. 13-31.

[7] Surej, P.J. (2013). "Influence of Computer Self-Efficacy on Information Technology Adoption". International Journal of Information Technology, Vol.19, No.1., pp. 1-13.

[8] Tom, M.M. (1997). "Decision Tree Learning". in Machine Learning. The McGraw-Hill Companies, Inc., pp. 52-78.

[9] Vehbi, C. and Etem, Y. (2013). "Attitudes to technology, perceived computer self-efficacy and computer anxiety as predictors of computer supported education". International Journal Computers & Education. Vol.60 Issue 1. Elsevier Science Ltd. Oxford, UK, UK pp. 148-158.

[10] Xindong, W. and Vipin, K. (2009). "The Top Ten Algorithms in Data Mining (CRC Data Mining and Knowledge Discovery Series) 1st Edition". Taylor & Francis Group, LLC.,Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business.