IOP Conf. Series: Journal of Physics: Conf. Series **855** (2017) 012031 doi:10.1088/1742-6596/855/1/012031

Software Piracy Detection Model Using Ant Colony Optimization Algorithm

Nor Astiqah Omar[1], Zeti Zuryani Mohd Zakuan[2] and Rizauddin Saian[1]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia
[2]Faculty of Law, Universiti Teknologi MARA, Malaysia

E-mail: astiqah93lavenda@ymail.com, zeti@perlis.uitm.edu.my and rizauddin@perlis.uitm.edu.my

Abstract. Internet enables information to be accessible anytime and anywhere. This scenario creates an environment whereby information can be easily copied. Easy access to the internet is one of the factors which contribute towards piracy in Malaysia as well as the rest of the world. According to a survey conducted by Compliance Gap BSA Global Software Survey in 2013 on software piracy, found out that 43 percent of the software installed on PCs around the world was not properly licensed, the commercial value of the unlicensed installations worldwide was reported to be $62.7 billion. Piracy can happen anywhere including universities. Malaysia as well as other countries in the world is faced with issues of piracy committed by the students in universities. Piracy in universities concern about acts of stealing intellectual property. It can be in the form of software piracy, music piracy, movies piracy and piracy of intellectual materials such as books, articles and journals. This scenario affected the owner of intellectual property as their property is in jeopardy. This study has developed a classification model for detecting software piracy. The model was developed using a swarm intelligence algorithm called the Ant Colony Optimization algorithm. The data for training was collected by a study conducted in Universiti Teknologi MARA (Perlis). Experimental results show that the model detection accuracy rate is better as compared to J48 algorithm.

1. Introduction

University is a sacred place to nurture students so that they will become a better person and later can contribute to the society. However this is not the case nowadays. Malaysia as well as other countries in the world is faced with issues of piracy committed by the students in universities. Easy access to the internet is one of the factors which contributes towards piracy in Malaysia as well as the rest of the world. Piracy in universities has gained much attention due to recent high profile lawsuits faced by the universities [1]. The Software Alliance in 2013 has conducted Compliance Gap BSA Global Software Survey on software piracy. It was found that unlicensed software usage continued to be a major problem in 2013. According to the survey, 43 percent of the software installed on PCs around the world was not properly licensed and the commercial value of the unlicensed installations worldwide was $62.7 billion. Universities become concerned about the issue as they might face negative publicity and also they may face legal actions.

Piracy in the university can be in the form of software piracy, music piracy, movies piracy and piracy of intellectual materials such as books, articles and journals piracy. Students in university play a critical role in piracy due to the advances in peer-to-peer file sharing programs

which have made it easy for students to illegally download and share copyrighted content via campus network [2]. According to [1], piracy on campus has gained abundant attention due to recent high profile lawsuits against developers of file sharing software. Universities become concerned as they can face negative publicity along with the risk of large fines. Software piracy has been literally defined as "the unauthorized copying or distribution of copyrighted software" [3]. However, software piracy has no definite definition under the law. According to [2], software piracy is a form of copyright infringement which can happen in higher institutions where peers usually share information with each other. Thus, the act of software piracy can be dealt with under the provision of copyright infringement. For example, in Malaysia copyright works is governed by the Copyright Act 1987.

The Copyright Act 1987 provides protection for copyrightable works in Malaysia. Section 7(1) Copyright Act 1987 provides for works which are eligible for copyright which includes literary works; musical works; artistic works; films; sound recording; and broadcasts. Literary works according to Section 3 of the Copyright Act 1987 includes work in the form of computer program. Thus, software as a form of computer program is a copyrightable work and is protected under the Malaysian copyright law. These works according to Section 7(2) "shall be protected irrespective of their quality and the purpose for which they were created". This provision shows that the law emphasized on the importance of protecting one's work. Copyright infringement is defined under Section 36 Copyright Act 1987, which provides that copyright is infringed by a person when he does something with the work without the consent of the owner of the work. If software is pirated, the pirate according to Section 36 has infringed the copyright of the software. For any act of copyright infringement, Section 37(1) of the Copyright Act 1987 provides that it "shall be actionable at the suit of the owner of the copyright", which means that the owner of the copyright has the right to sue the infringer. The offences and punishments relating to copyright infringement are provided under Section 41 Copyright Act 1987 which is illustrated in table 1.

This phenomenon requires attention from academics and researchers as the act of software piracy is detrimental to students. By having the law in place, students who commit software piracy might be facing legal consequences by the manufacturer seeking to thwart the theft of their commercial products. If the students are later found guilty, they might be dismissed from the university. Even though software piracy is common among students, it is actually a serious matter. The punishment imposed on the students might be able to act as a deterrence so that students will not do the same thing when they venture into the working environment. This study which involves predicting software piracy tendency among students, is important because after predicting their tendency, prevention can be done. Prevention is important for the students so that they are more sensitive to the likelihood of conviction and punishment which may act as a deterrent against software piracy.

Thus, this study has developed a classification model for detecting software piracy. The model was developed using an implementation of the Ant Colony Optimization (ACO) algorithm [4, 5] for data classification called the Ant-Miner [6, 7, 8].

2. Methodology

Three phases involved in developing the classification model as depicted by figure 1. The data pre-processing step will convert the data in an appropriate for Ant-Miner. In the second step, Ant-Miner trained the data to produce the classification model. The last step validates the classification model using predictive accuracy.

The first phase is the Data Pre-Procesing. The data was extracted from a survey conducted among the students at UiTM Perlis. Figure 2 shows an example of a question in the questionnaire. Each question was labelled by a question number in the most left column. The next column represents the question itself. Finally, the rest of the columns represents a number

Table 1. Offences and punishments relating to copyright infringement.

Offences	Punishment
Sales, hire, distribution, possession, exhibition, imports of any infringing copies	• First offence: a fine not less than RM2,000 and not more than RM 20,000 for each infringing copy or imprisonment for a term not exceeding 5 years or both. • Subsequent offence, a fine not less than RM4,000 and not more than RM 40,000 for each infringing copy or imprisonment for a term not exceeding 10 years or both.
Contrivance of infringing copies	• First offence, a fine not less than RM4,000 and not more than RM40,000 for each contrivance or imprisonment for a term not exceeding 10 years or both. • Subsequent offence, a fine not less than RM8,000 and not more than RM80,000 for each contrivance or imprisonment for a term not exceeding 20 years or both.
Circumvention, removal, alteration, distribution, import of the electronic works without authority	• First offence, a fine not more than RM250,000 or imprisonment not exceeding 5 years or both. • Subsequent offence, a fine not less than RM500,000 or imprisonment not exceeding 10 years or both.

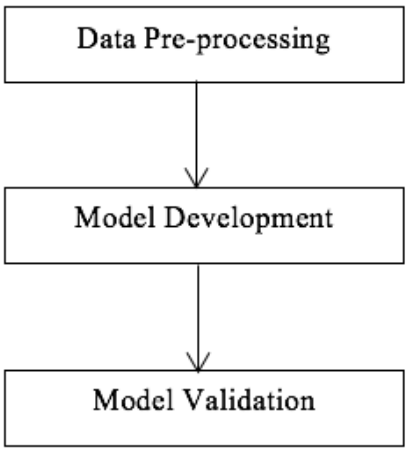

Figure 1. Research phases

of choices to be chosen by an individual which are strongly disagree (1), disagree (2), somewhat disagree (3), neither agree nor disagree (4), somewhat agree (5), agree (6) and strongly agree (7). The choices was divided into two parts. First, choices 1-4 reflects good attitude labeled as "Good" and second, choices 5-7 reflects bad attitude labeled as "Bad". Table 2 presents data

description of this study.

2	I have allowed people to copy digital contents I have purchased.	1	2	3	4	5	6	7

Figure 2. Example of Question in the Questionnaire.

Table 2. Data Description.

Attribute	Values
Gender	Male, Female
Attitude	Good, Bad
Income	High, Low
Ethics	Yes, No
Law	Yes, No
Perception	Will Piracy, Will Not Piracy

The second phase is the model development. The classification model was developed using ACO. ACO is a metaheuristic algorithm inspired by the foraging behaviour of ant colonies introduced in the early 1990s [9, 5]. The purpose of ACO is to find the best solution using a set of artificial ants that communicates indirectly using an item called the pheromone. The amount of pheromone on one of the trails used by the majority ants will increase as time passed by and will decrease on the less used trails. As a consequent, since ants tend to follow the pheromones, all or most of the ants will finally converge to the best trail, with the high density of pheromones, which happened to be the shortest trail from the nest to the food source. However, there are a small number of ants still using the longer branch. This is the effect of "path exploration", where the density of the pheromone will not bias some of the ants. ACO has also been applied to various fields such as the travelling salesman problem, sequential ordering, flow shop scheduling and the graph colouring problems.

The first ACO implementation for data classification is called the Ant-Miner. Ant-Miner follows a divide-and-conquer approach to discover classification rules called the rule induction. Rule induction is a technique to extract set of rules from a data set [10, 11]. Each rule is in the form of IF <term1 AND term2 AND ... > THEN <class>. A rule contains two parts: the rule ancedent and the class. The rule antecedent is a combinations of one or many terms. Each term is a triple <attribute = value>, such as <Attitude = Good>. The rule consequent specifies the class predicted.

First, an ant starts with an empty rule. The ant will add one term at a time to its current partial rule. The rule addition will stop if adding any term to the rule would make the rule cover less than a pre-defined minimum number of cases, or all attributes have already been used by the ant for the rule antecendent.

Ant-Miner uses a heuristic measure as evaluation measure to fill in the antecedent part of the rule, by selecting the best term to be included into the partial rule. The heuristic measure (2) is the normalization of entropy measures between terms (1). The algorithm selects one best rule from a set of discovered rules, based on a quality measure using some fitness function.

$$H\left(W|A_i = V_{ij}\right) = -\sum_{w=1}^{k} P\left(w|A_i = V_{ij}\right) \cdot \log_2 P\left(w|A_i = V_{ij}\right) \tag{1}$$

where W is the class attribute, k is the number of classes and $P\left(w|A_i = V_{ij}\right)$ is the number of tuples for $A_i = V_{ij}$ with class w.

$$\eta_{ij} = \frac{\log_2 k - H\left(W|A_i = V_{ij}\right)}{\sum_{i=1}^{a} x_i \cdot \sum_{j=1}^{b_i} \left(\log_2 k - H\left(W|A_i = V_{ij}\right)\right)} \tag{2}$$

where a is the total number of attributes, x_i is set to one if atttribute A_i was not yet used by the current ant or to zero otherwise, and b_i is number of values for attribute A_i.

Ant-Miner has an exploration behaviour and exploitation behaviour. The exploration behaviour is contributed by a value called the pheromone level. Ant-miner uses a probability (3) that is proportional to the product of heuristic value and pheromone level for that term, to add terms to a rule. Dorigo in his book, called this transition rule, random proportional transition rule [4]. Detail algorithms are described in [7].

$$P_{ij} = \frac{\eta_i j \times \tau_{ij}}{\sum_{i=1}^{a} x_i \cdot \sum_{j=1}^{b_i} \left(\eta_{ij} \cdot \tau_{ij}\right)} \tag{3}$$

where τ_{ij} is the current amount of pheromone for term $A_i = V_{ij}$.

Table 3 shows an example of entropy, heuristic and probability calculation for each terms.

Table 3. An example of entropy, heuristic and probability calculation.

Terms	Entropy	Heuristic	Probability
gender=male	0.937	0.097	0.100
gender=female	0.943	0.087	0.101
attitude=good	0.936	0.099	0.100
attitude=bad	0.950	0.077	0.102
income=high	0.890	0.168	0.095
income=low	0.968	0.049	0.104
ethics=yes	0.931	0.106	0.100
ethics=no	0.946	0.082	0.101
law=yes	0.884	0.179	0.095
law=no	0.964	0.056	0.103

In this study, the number of ants was set to 100, the maximum uncovered cases = 10, rule for convergence = 10 and number of iterations = 100 are constant use but minimum case per rule will change. The second column in table 5 shows the result of accuracy for minimum case per rule from 6 to 10 with 10 folds cross validation. The percentage shows that the interval of accuracy in between 64.38% to 64.94% at every different minimum case per rule. Thus, these results show more transparent in a graph which is illustrated in figure 5.

The final phase figure 1 is the model validation. This study calculate the predictive accuracy of the classification model to know the dataset quality using k-fold cross validation with $k = 10$[12]. The accuracy was calculated using formula (4).

$$Accuracy = \left(\frac{predicted - actual}{actual}\right) \times 100\% \quad (4)$$

Finally, this study makes a comparison between Ant-Miner and J48 based on the percentage of accuracy. J48 is an implementation of C4.5 algorithm [13] in Weka software [14]. Classification model that has a higher percentage of accuracy is better.

3. Findings and Discussion

This study tested the effect of changing the value of min. cases per rule parameter. First, this study has a look at the simplicity of the discovered rules. The simplicity of discovered rules is measured by the number of discovered rules and the average number of terms in a rule.

Table 4 shows the number of rules and the number of terms for different values of min. cases per rule. The lower the number of rules and terms, the better the classification model is. The number of rules decreased as the value of min. cases per rule is decreased. The lowest number of rules is 5.7 when min. cases per rule is set to 10 as depicted by figure 3. The lowest number of rules is 6.4 when min. cases per rule is set to 7 as depicted by figure 4.

Table 4. Predictive accuracy, number of rules and number of terms for different values of min. cases per rule.

Min. Cases per rule	Number of rules	Number of terms
6	6	7.6
7	5.5	6.4
8	5.8	6.7
9	5.9	7
10	5.7	6.5

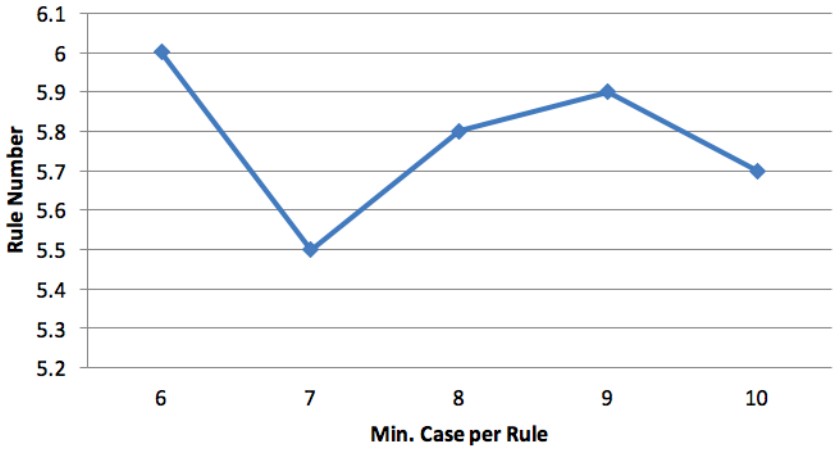

Figure 3. Number of rules for different values of min case per rule.

Table 5 shows the predictive accuracy. The higher the value of the accuracy, the better the classification model is. According to figure 3, the highest accuracy is 64.94% when the min. cases

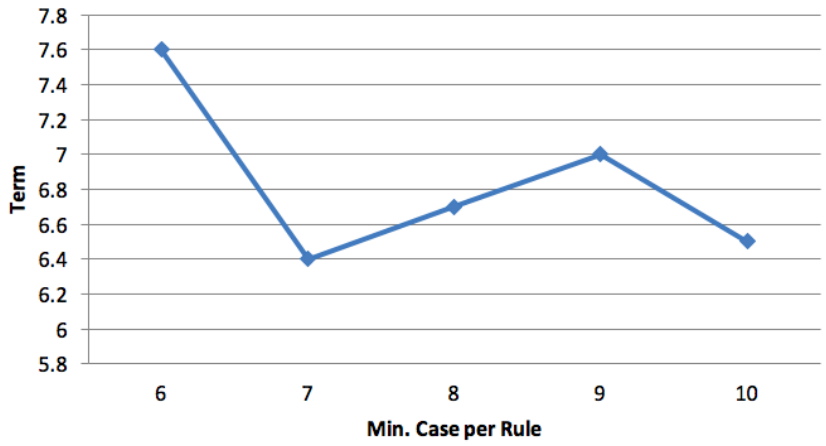

Figure 4. Number of terms for different values of min case per rule.

Table 5. Predictive accuracy for different values of min. cases per rule.

Min. Cases per rule	Accuracy (%)
6	64.38
7	64.88
8	64.94
9	64.90
10	64.70

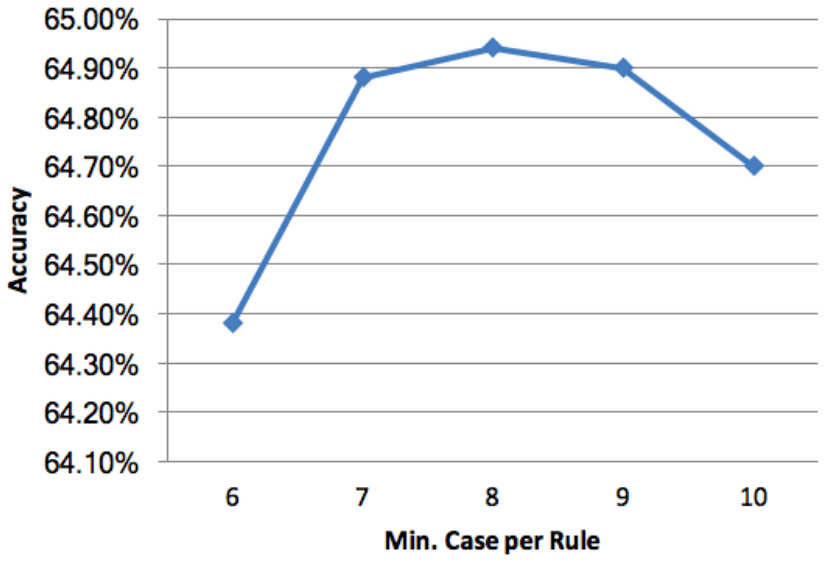

Figure 5. Accuracy for different values of min case per rule.

per rule is set to the value of 8. However, the accuracy starts to decrease when the parameter value was increased.

The results comparing the predictive accuracy of Ant-Miner and J48 are reported in table 6. As shown in table 6, the accuracy of classification model by J48 is 64.10% while the accuracy of classification model by Ant-Miner is 64.94%. Ant-Miner discovered rules with a better predictive accuracy than J48.

Table 6. Predictive accuracy (%) of Ant-Miner and J48.

J48	Ant-Miner
64.10	64.94

4. Conclusion

The main objective of this study is to develop a classification model for detecting software piracy tendency using a variant of ACO algorithm called the Ant-Miner. It is found that the Ant-Miner produced a more accurate classification model for detecting software piracy tendency as compared to J48.

Acknowledgments

The authors wish to thank Universiti Teknologi MARA for funding this study under DKCP (600-UiTMPs/PJIM&A/ST/DKCP (02/2012)).

References

[1] Chiang E P and Assane D 2008 *The Journal of Socio-Economics* **37** 1371–1380
[2] Gerlach J H, Kuo F Y B and Lin C S 2009 *Journal of the American Society for Information Science and Technology* **60** 1687–1701
[3] BSA 2014 The compliance gap: Bsa global software survey. Tech. rep. BSA — The Software Alliance
[4] Dorigo M and Sttzle T 2004 *Ant colony optimization* (the MIT Press) ISBN 0-262-04219-3
[5] Dorigo M, Maniezzo V and Colorni A 1996 *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* **26** 29–41
[6] Parpinelli R S, Lopes H S and Freitas A A 2001 An ant colony based system for data mining: applications to medical data *Proceedings of the genetic and evolutionary computation conference (GECCO-2001)* pp 791–797
[7] Parpinelli R S, Lopes H S and Freitas A A 2002 *IEEE transactions on evolutionary computation* **6** 321–332
[8] Parpinelli R S, Lopes H S and Freitas A A 2002 *Data mining: A heuristic approach* 191–208
[9] Dorigo M, Maniezzo V and Colorni A 1991 Positive feedback as a search strategy Tech. Rep. 91-016 Dipartimento di Elettronica, Politecnico di Milano Milan, Italy
[10] Han J, Pei J and Kamber M 2011 *Data mining: concepts and techniques* (Elsevier)
[11] Tan P N, Steinbach M and Kumar V 2006 *Introduction to data mining* (Pearson Addison Wesley Boston)
[12] Witten I H, Frank E and J Pal C 2016 *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)* (San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.) ISBN 0128042915
[13] Quinlan J R 2014 *C4.5: Programs for machine learning* (Elsevier)
[14] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P and Witten I H 2009 *ACM SIGKDD explorations newsletter* **11** 10–18