A Classification Model for Predicting Standard Levels of OTOP's Wood Handicraft Products by Using the K-Nearest Neighbor

Jittaporn Tarapitakwong
Asian Development College for Community Economy and Technology,
Chiang Mai Rajabhat University, Thailand
E-mail: nuyims@hotmail.com
Bungon Chartrungruang
Chiang Mai Rajabhat University, Thailand
bung_onc@yahoo.com
Samerkae Somhom
Chiang Mai University, Thailand
samerkae.s@cmu.ac.th
and Nuttiya Tantranont
Asian Development College for Community Economy and Technology,
Chiang Mai Rajabhat University, Thailand
nuttiya18@gmail.com

Abstract – **The aim of this research is to develop a classification model for predicting standard levels of OTOP's wood handicraft products by using the K-Nearest Neighbor (K-NN). To develop candidates of classification models, we used the analysis software Weka to apply the k-fold cross-validation method to our datasets. Then, the best K-NN from 22 product attributes were selected to be used for the models based on the Euclidean distances. The best classification model developed from the previous steps was able to predict standard levels of OTOP's wood handicraft products with high reliability. The model reported in this study can achieve accuracy, recall, and precision at the level of 88.34%, 88.30%, and 83.4%, respectively. Our result indicates that the model with the lowest Euclidean distance in the 3 aspects above can be efficiently used for predicting standard levels of product as specified by the OTOP project.**

Keywords – **K-Nearest Neighbor, Euclidean distance, Classification, OTOP product standards**

I. INTRODUCTION

In the present, computer technology plays an important role in people's livelihood. The technology improves convenience and speed of communication, education and research, and also reduces dependency on human experts. With the help of computer technology, researchers can easily employ the Case-based Reasoning approach to solve new problems based on experience and solutions of the past [1]. Researchers can also predict outcomes based on similarity of attributes as reported in several studies, such as prediction of VARK for developing rules of visual, aural, read/write and kinesthetic learning styles [2]. Teachers can develop learning resources and their management that are consistent with students' learning styles in order to improve motivation and comprehension of students toward learning. The study has been reported to use data mining for predicting scores of 3rd year middle school students and use the outcomes to suggest suitable high school programs for students [3]. This includes predicting and improving student scores by using the K-NN [4].

With the aims of improving capability and efficiency of the OTOP (One Tambon One Product) project, developing products and their market potential at home and abroad, the

Thai government launched the OTOP Product Champion (OPC) project that encourages entrepreneurs to improve quality of their products to meet the standards specified by the OTOP project (the 5-star level system). This move by the government can increase preparedness of products for markets at home and abroad, increase business income and improve self-reliance of Thai entrepreneurs [5]. According to the OTOP project, the products are registered into 6 categories; food, drink, fabric, clothes, handicrafts/decorations/ souvenir, and non-food herbal products. It has been reported that most of the OTOP products are registered in the category of handicrafts/ decorations/souvenir with wood handicraft products making up the largest fraction of this category [6]. However, a lot of products in the category were not ready for quality improvement in order to meet OTOP's standards [7]. The problems arise because the entrepreneurs lack necessary knowledge related to product attributes used in the evaluation and OTOP evaluation process; thus, making the entrepreneurs less prepared for product evaluation

The aim of this study is to develop a classification model for categorizing wood handicraft products into classes of standard levels specified by the OTOP project. The model will help potential entrepreneurs to evaluate their products before submitting application to the OTOP Product Champion (OPC) project. The model developed here is based on product attributes used by OTOP product evaluation committees to evaluate submitting products. In this study, products were initially classified into categories by using the K-NN Algorithm, and we later identified the best candidate model by using the Euclidean distance. These methods are known to be used for predicting learning styles as reported in [8][9][10]. The model reported in our study can be used for predicting standard levels of OTOP's wood handicraft products (1 to 5-star levels). This study will benefit potential entrepreneurs who want to develop their quality of their products to meet with product standards of the OTOP project.

II. LITERATURE REVIEW

A. Classification

Classification is the method of categorizing information into classes based on similarity of attributes of information. The method is commonly used in data mining [11]. The classification algorithm works by checking both predictor and target variables for their similarity and predictive relationships between the variables. For example, being elderly women may highly correlate to having high income. The associated dataset with known classes initially supply to the model is the training set. The model will later classify new information with unknown classes into classes based on predictive relationships obtained from the training set. For example, a 63 year-old female professor may be classified to a group of persons who have high income [9].

B. K-Nearest Neighbor

The K-Nearest Neighbor (K-NN) is the method of classification that uses instance-based learning for predicting values [9]. The algorithm employs distance function to determine similarity of information. One of the simplest distance functions is the Euclidean distance [12]. To perform K-NN, the following steps have to be done;

(1) Define K of the nearest neighbors of an instance to be considered in the algorithm

(2) Calculate the Euclidean distance of each instance

(3) Select K neighboring attributes that have the lowest Euclidean distance as shown in the Fig. 1.

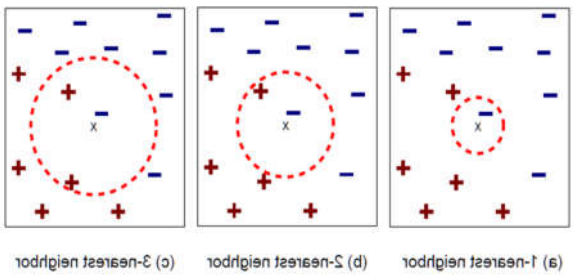

Fig. 1 The K-NN of an instance [9]

C. Cross-validation

Cross Validation is the method used for estimating error and reliability of a model. Generally, a dataset is initially re-sampled and then a testing set is classified into classes. Validation is made by comparing results of the testing set with a training set [13]. Cross validation generally has its applications in the fields of Network architecture and Classification. In this study, we used the k-fold cross-validation [12], which the sample is randomly classified into k sub-samples. Then, one of the k sub-samples is selected as the testing set, while other sub-samples left are used as training sets. These processes are repeated by changing the testing set to one another sub-samples until every sub-sample is used.

III. METHODOLOGY

A. Data preparation

Initially, we collected information from 163 wood handicraft products that have been registered to the OTOP Product Champion (OPC) project and have been classified into the 5-star system of the OTOP project. A total of 22 product attributes used in this study are shown in the Table 1. The products were annotated according to their standard levels as shown in Fig. 2. Information of the products was prepared in a format of *.CSV file due to its compatibility with the analysis software Weka.

Table I
Product attributes that are used by the OTOP project for classifying products into the 5-star system

Attribute	Variable	Attribute	Variable
Sources of materials	A1	Income in comparison to the previous year	A12
Increasing of production capacity	A2	Brand continuity	A13
Environmental friendly manufacturing process	A3	Background of a product	A14
Potential of large-scale production	A4	Integration of local identities	A15
Product development in 1 year	A5	Meticulousness in production	A16
Packaging development	A6	Design in general/Harmony in visual design	A17
Packaging	A7	Important Product Attributes	A18
Time required to start a business	A8	Quality of materials	A19
Participation of local communities	A9	Uniqueness of design	A20
Accounting	A10	Utility	A21
Main distribution channels	A11	International market potential	A22

ID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	Class
1	3	3	2	2	3	3	3	3	3	2	3	2	5	6	8	10	7	3	5	3	5	1	4
2	3	3	3	3	3	3	3	3	3	2	2	3	5	4	8	5	3	3	3	1	1	1	2
3	3	2	2	1	2	2	2	1	1	1	1	1	2	2	5	5	3	3	3	1	1	1	1
4	2	3	2	3	3	3	3	3	3	3	3	3	5	6	8	5	3	5	5	3	1	3	3
5	3	3	3	3	3	3	2	3	2	2	3	2	5	6	8	5	3	3	3	3	1	3	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Fig. 2 Example of product attributes prepared in a format of *.CSV

B. Data classification

This step requires classification of products into classes based on the training set. The dataset is initially classified into 2 categories; 1) Training set and 2) Testing set. The training set is used to explore predictive relationships of input data (attributes) and outcomes (1 to 5-star classes), while testing set is used for predicting their unknown outcomes (classes) shown in the Fig. 3.

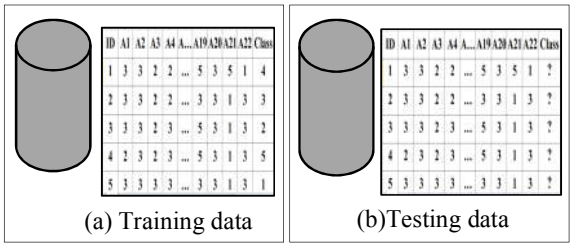

Fig. 3 Classification of the training set and the testing set

C. Developing the classification models

We used the K-NN to develop classification models that are able to classify candidate products into 5 classes, according to the 5-star system of the OTOP project. The algorithm is widely used for predicting values based on the Instance-based Learning method [8][9][10]. In this study, the method was used to compare similarity of OTOP products by calculating the Euclidean distances between the testing set and the training set of the product attributes. The Euclidean distances of each product attribute were sorted in ascending order and k attributes with the lowest Euclidean distances were selected as predictors for classification models. A low Euclidean distance shows high predictive relationship between variables. The equations used in developing the models are shown in the Fig. 4 as follows [14].

$$\mathrm{dist}(x_i, x_j) = \sqrt{\sum_{k=1}^{n} \left(x_{i,k} - x_{j,k}\right)^2}$$

Where,

$\mathrm{dist}(x_i, x_j)$ is the Euclidean distance between x_i and x_j

n is the total number of product attributes

$x_{i,k}$ is the k^{th} attribute of an instance x_i

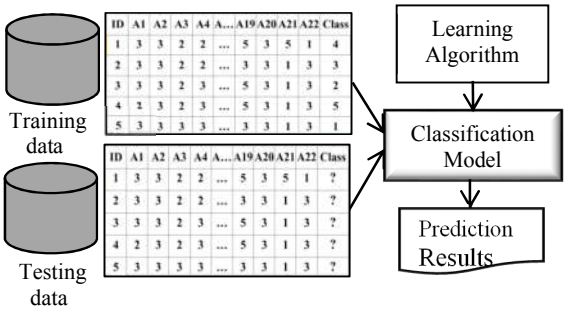

Fig.4 Developing classification models for predicting standard levels of OTOP products [15]

D. Validation

To validate the reliability of the classification models, we re-classified products into classes by using the K-NN. The products were separated into 2 categories; 1) Training set and 2) Testing set. In this study, we used the k-fold cross-validation as a sampling method [12]. The method required us to separate the data set into k sub-samples. Then, we selected one of the sub-samples to be the training set, while other sub-samples became the testing set. The processes were repeated until every sub-sample was used. Finally, we calculated accuracy, recall and precision of the classification models in order to find the model with the lowest Euclidean distance [9] as follows;

Accuracy is the measurement of correctness as shown in the following equation;

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Where,

TP denotes the number of true positives
TN denotes the number of true negatives
FP denotes the number of false positives
FN denotes the number of false negatives

Recall is the sensitivity of prediction which is the fraction of the number of label assigned correctly to the total number of labels assigned, as shown in the equation below;

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

Precision is the measurement of degree of the model's correctness as shown in the following equation;

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})}$$

IV. RESULTS OF THE RESEARCH

As shown in the research methodology, we developed the classification models for classifying OTOP's wood handicraft products by using the K-NN in the analysis software Weka as shown in the Fig. 5. The software used the Euclidean distance to calculate similarity of data sets. We studied the models with the k-fold cross-validation at k = 5, 10, 15, 20 and 25 folds, respectively. Validation of the classification models is shown in the Table 2., while the results of classification is shown in the Fig. 6.

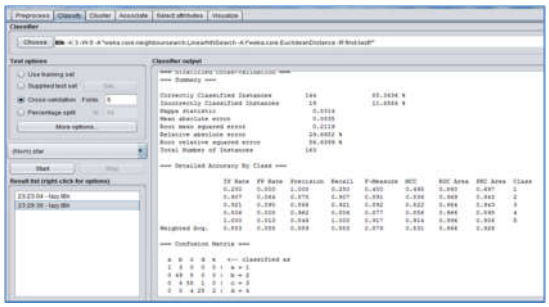

Fig. 5 Developing classification models in the analysis software Weka

Table II
Validation of the classification models used for classifying potential products into OTOP's 5-star system

Cross validation	K-Nearest Neighbor	Recall	Precision	Accuracy
5	3	88.30%	88.90%	88.34%
	5	83.40%	81.80%	83.44%
	7	79.80%	78.40%	79.76%
10	3	88.40%	87.70%	87.73%
	5	82.20%	83.40%	83.44%
	7	80.00%	81.00%	80.98%
15	3	86.40%	85.90%	85.89%
	5	80.60%	82.20%	82.21%
	7	80.50%	81.60%	81.60%
20	3	87.10%	86.50%	86.50%
	5	81.40%	82.80%	82.82%
	7	79.80%	81.00%	80.98%
25	3	84.00%	83.40%	83.44%
	5	80.00%	81.60%	81.60%
	7	77.20%	78.50%	78.53%

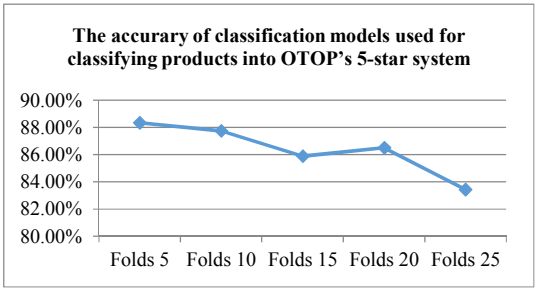

Fig. 6 Results of classification at K-Nearest Neighbor = 3

Comparison of the highest values of recall, precision and accuracy for each fold of the k-fold cross-validation used for classifying products into OTOP's 5-star system is shown in the Fig. 7.

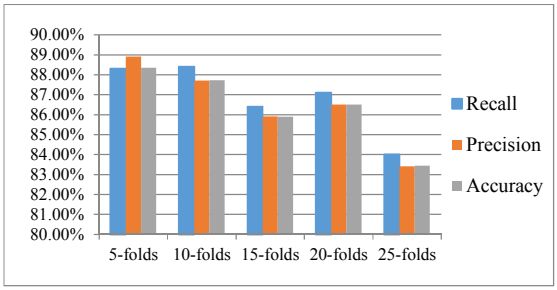

Fig. 7 The highest vales of recall, precision and accuracy

V. CONCLUSIONS

The aim of this research is to develop a classification model for predicting standard levels of OTOP's wood handicraft products by using the K-NN. We identified the best candidate model by using the Euclidean distance. The classification model was used to classify OTOP products in our study into 5 categories (from 1 to 5-star levels) based on 22 attributes as shown in the Table 1. We validated the reliability of the candidate models by using the k-fold cross-validation at 5, 10, 15, 20, 25 folds, and K-NN at k equals to 3, 5 and 7. We found that the model with k-fold cross-validation at 5 folds and K-NN at k equals to 3 yields the best prediction at the accuracy of 87.73%. The second best candidate is the model with k-fold cross-validation at 20 folds and K-NN at k equals to 3 yields the second-best prediction at the accuracy of 86.50%. The third best candidate is the model

with k-fold cross-validation at 15 folds and K-NN at k equals to 3 yields the third-best prediction at the accuracy of 85.89%. Finally, the lowest reliable candidate is the model with the k-fold cross-validation at 25 folds and the K-NN at k equal to 3 yields the best prediction at the accuracy of 83.44%. Our results are consistent with the study of Pansumret *et. al.*[12] who were able to identify the best candidate of classification models by using the k-fold cross-validation approach, and Somkantha *et. al.* [4] who were able to predict students' scores by using the K-NN approach and found that the K-NN yielded better results of prediction than the Bayes approach. In conclusion, the results of our study can help potential entrepreneurs to evaluate their wood handicraft products according to OTOP's 5-star system based on the product attributes before submitting for application to the OTOP project.

For further studies, we would like to compare the reliability of other approaches used in model validation, such as Naïve Bayes, C4.5 and Rule base, in order to identify the best approach of validation. These kinds of studies will help to develop a system for predicting standard levels of product into categories the 5-star system specified by the OTOP project. The studies can also provide applications for classification and standardization of other types of products.

ACKNOWLEDGMENT

We would like to thank the Department of Community Development and members of the OTOP product evaluation committee for giving interviews and providing necessary information for the research. We also would like to thank OTOP entrepreneurs who provided their products to be used in the research and also participated in the study.

REFERENCES

[1] N. Wareeprasirt and N. Lumdee, *Artificial Intelligence*. Bangkok: KTP com&consult, 2009.

[2] O. Pantho and M. Tiantong, "Using Decision Tree C4.5 Algorithm for Predicting VARK Learning Styles**,"** *International Journal of the Computer, the Internet and Management,* vol. 24, pp. 58-63, 2016.

[3] S. Vilailuck, V. Jaroenpuntaruk, & D. Wichadakul, "Utilizing Data Mining Techniques to Forecast Student Academic Achievement of Kasetsart University Laboratory School Kamphaeng Saen Campus Educational Research and Development Center," *Journal, Science and Technology Silpakorn University*, vol. 2, pp. 1-17 , 2015.

[4] K. Somkantha, W. Kultangwattana, T. Hassago, and J. Raodchompoo, "Online Student Forecast System by Using K-Nearest Neighbor. *Proceedings of The 3rd International Conference on Knowledge and Smart Technologies*," 1(1), BuraphaUniversity. Chonburi, vol. 1, 2011.

[5] Community development, department, *Guidelines and criteria for the OTOP Product Champion in Thailand in* ๒๕๕๕, 1st ed., Bangkok, Thailand : BTS Press, 2012.

[6] Community development, department. (2013). *Summary of Operating Results for the project of registered manufacturer and entrepreneurs of OTOP in 2555.* Bangkok : Chatra(Big Ideas Come To Life).

[7] S. Phloprakarn. (2014) The results of OTOP product Champion in 2555. [Online]. Available: http://www.thaitambon.com/OPC2555/560301_5.xls

[8] P. Ponchob and S. Nitsuwat, "Recommender System for Notebook Purchasing using Content-Based Filtering," *Proceedings of the 6TH National Conference on Computing and Information Technology*, vol. 182, pp. 595-600, 2010.

[9] Sinsomboonthong, S.(2015). *Data Mining*. Bangkok : Chamchuree products.

[10] S. Sivilai, and C. S. Namahoot, "Applying Case-based Reasoning to Recommend Appropriate Food for Inpatients," *Proceedings of the 8 th Naresuan Research Conference*. vol. 2, pp. 1257-1265, 2012.

[11] W. Wirojcharoenwong and B. Eamthanakul, "Comparative Study of lassification Properties Between Decision Tree, Rule-base and K-Nearest Neighbor," *Proceedings of the Eighth National Conference on Computing and Information Technology*, vol. 8, pp. 554-560, 2012.

[12] Y. Pansumret, J. Phuboon-ob, and W. Pongsiri, "On Comparison of Data Mining Algorithmsfor Analysis of Factors Affecting the Academic Performance of Students," *Journal of Science and Technology Mahasarakham University*, vol. 9, pp. 281-289, 2013.

[13] N. Chirawichitchai, *Data Mining Techniques for Automatical Disease Analysis*, Bangkok, Thailand: Suan Sunandha Rajabhat University, 2010.

[14] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed., Waltham, USA: Elsevier, 2012.

[15] E. Pacharawongsakda, (2016) Introduction to Data Mining and Big Data Analytics. [Online]. Available: http://dataminingtrend.com/2014/data-mining-techniques/ensemble-model/