Indonesian Journal of Electrical Engineering and Computer Science
Vol. 12, No. 2, November 2018, pp. 570~576
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v12.i2.pp570-576

Data Mining Approach to Herbs Classification

Adillah Dayana Ahmad Dali, Nurul Aswa Omar, Aida Mustapha
Fakulti Sains Komputer dan Teknologi Maklumat, Universiti Tun Hussein Onn Malaysia

Article Info

Article history:

Received Apr 9, 2018
Revised May 20, 2018
Accepted Jul 11, 2018

Keywords:

Classification
Data mining
Herbs

ABSTRACT

Herbs are one of the high-value products in Malaysia. The term 'herbs' has more than one definition. It is also demanding by multiple manifolds. Herbs are used in many sectors nowadays. The ability to identify variety herbs in the market is quite hard without the intervention of human experts. Unfortunately, human experts are prone to error. Herbs classification is able to assist human experts and at the same time minimizing the intervention. This research performs identification and classification of herbs based on image capture ad variety of classification algorithms such as an Artificial Neural Network (ANN), K-Nearest Neighbors (IBK), Decision Table (DT) and M5P Tree algorithms. The selected algorithms are implemented and evaluated to their relative performance and IBK is found to produce the highest quality outputs.

Corresponding Author:

Nurul Aswa Omar,
Fakulti Sains Komputer dan Teknologi Maklumat,
Universiti Tun Hussein Onn Malaysia,
Parit Raja, 86400 Batu Pahat, Johor, Malaysia.
Email: nurulaswa@uthm.edu.my

1. INTRODUCTION

Malaysia is one of the leading exporter of herbs. The herbs industry was aiming to produce high-value products, totaling RM2.2 billion of the Gross National Income (GNI) as reported in The Malaysian Times newspaper (2013). Data shows that herbal product demand has multiplied manifolds. Herbal health foods have reached RM2,380 billion around 2001. Prior to that, from only RM2,093.8 million in 1980, drastic increase in the world herbal products are valued at RM950 billion in 1996. Currently, the trade value of the herb sector was expected to soar over RM2 trillion by the year of 2020. The value of threefold increase compared to the RM777 billion worth of trade was estimated in the herbs sector in 2009. On the local front, the ministry estimated the herb market to expand by 15 per cent a year from RM7 billion in 2010 to around RM29 billion by 2020.

Generally, it is known that herbs has contributed a lot in medicinal purposed from long time ago. All the facts, truths or principles of herbs has been passed down for periods of millenarian of years [1]. Herbs are nutritious as well as valuable plants. Truthfully, the biggest possible for new herb currently lies in the food sectors. It is being used in food preparation and not only that, it is also being widely used in medicine and cosmetic industry. Herbs are used by almost everyone nowadays either in the form of spices, herbs or daily food-based products. With the increasing use of herbs, there is an urgent need for the ability to identify variety herbs available in the market. Most herbs grow in the jungle and the way they identify is through the recognition of human experts.

It is very importance to automatically acknowledge the various type of herbs for herbs classification referred on their particular features due to short number of resources along with knowledgeable person. One way to identify herbs is through classification of the herbs. Computer Science has finally harnessed both the enormous storehouse of data and the vast computational power. Widely defined as Knowledge Discovery in

the databases, data mining is computerized or useful extraction of patterns. It represents the knowledge inevitably gathered in databases that resolves complication.

Data mining is about configuration of complication by examine and determine the data that already existed in the databases. It finds the important information hidden in large volumes of data [2]. Data mining is also described as the series of action of uncovering patterns in data. The possible use of data mining method defines that the approach in which a repository of data can be utilized may stretch far beyond what was perceive when the data was initially gathered. Lots of applications in machine learning to data mining as shown in the understanding. The important knowledge structures that are gained, the fundamental explanation, are at least as crucial, and frequently very much more substantial compare to the capability to accomplish well on new examples.

Lots of learning techniques look for structural definition of what is learned, definition that could be quite complicated and are commonly articulate as sets of rules. In the recent development of automated classification techniques, there has been a great deal of progress. From the combinations of artificial intelligence and statistical classification approaches, a significant number of new techniques have arisen. Machine learning and data mining have both fascinated reasonable interest in the classification algorithms of both in the research areas. A few external-memory algorithms [3-6] and parallel implementations [7], [8] have also been specified. It have been recommended with the purposed of boost up the implementation time also analyze on huge training sets.

A logical programming technique is proposed by duplicating the mechanism of the human brain, which is the objective of Artificial Neural Network (ANN). This technique simulates the main biological operations of the human brain utilizing a particular software. ANN is an algorithm that is able to perform human brain operations, composing decisions, creating results, produce conclusions referring to the existent information in case there are insufficient data, continually receiving, learning and remembering data in a computing environment [9].

At present, it is pretty complicated to apply machine vision to categorize herb, due to the substantial computational resources along the difficulties algorithms needed. The ANN can defeat few of the complication by extracting the features instantly as well as efficiently. ANN has arisen as the imitation of the biological nervous system. A mode of working of a computer by assimilated to the mode of working of a brain, neural network model was grown. An evaluation of a set of shape features [10].

2. RESEARCH METHODOLOGY

Research methodology is a process used to gather information and data. It is also known as a way of knowing the outcome of a specific problem and making decisions. Researchers construct their research by formulating and defining a research problem. A different criteria to determine the current research problems in methodology by researchers. In the methodology, it explains the way a problem is inspected and the reason for using a specific method and technique.

In the previous research, there are many types of classification methodologies mentioned such as ANN or Gabor-Wavelets. This section will brief about the chosen methodology used to determine that this project runs perfectly. A proper methodology plan is necessary to collect the required information and data. Figure 1 shows a classification framework for this research. It included four steps which are dataset acquisition, pre-processing, classification and performance analysis.

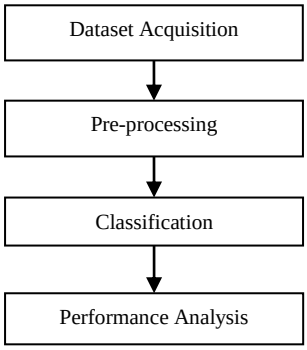

Figure 1. Classification framework

Datasets are prepared in first step where research data sources from machine learning repository and specific in leaf dataset. This dataset can be download from http://archive.ics.uci.edu/ml/datasets/Leaf.

Next step is pre-processing, where this step is to ensure the quality of the data result, the pre-processing data should be implemented. In this research, the input data is in the term of numerical. The pre-processing technique used to reduce the variation of herbs due to illumination factors. To improve the performance of herb classification, this process will help to enhance and normalize the herb dataset.

Based on the literature, four classification algorithm which are Artifical Neural Network [11], K-Nearest Neighbour [12], Decision Table and M5P Tree algorithms have been chosen for the experiments. These algorithms are chosen because they are the latest algorithms used in the literature. This algorithm is used in step three.

Last step is evaluation metric of the herbs dataset which correlation coefficient (CC), mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE) and root relative squared error (RRSE) on each experiment will best stated. For this purposed, Multilayer Perceptron Neural Network (MLP), K-Nearest Neighbours (IBK), Decision Table (DT) and M5P Tree algorithms will be compare based on those metrics. The result shown as following because the dataset is not a categorical dataset. It is a continuous dataset. To determine the best method for the performance for herbs classification, the performance of this algorithm will be recorded and tested. The table below shows the result of the tested algorithm. The next section will focus on the process each step in classification framework based on this research.

2.1 Dataset

A dataset is a group of data which is collected from a certain source such as the Internet. It is mostly noisy, incomplete and inconsistent. Dataset might consist of data for not just one but more members equivalent to the number of sequences. The concept dataset may also be used are more relatively, to refer to the data in a collection of closely related tables, equivalent to a particular experiments or event. Besides, it contains only aggregate data or often contains too much data to analyze which is lacking on the attribute's value.

The dataset lists values for each variable, which can be a number such as integers or real number. This research uses the Leaf dataset, available for download from http://archive.ics.uci.edu/ml/datasets/Leaf. This dataset includes 40 different plant species including herbs and the details of scientific names of each plant as well as the number of leaf specimen accessible by species are shown in Table 1. Species numbered from 1 until 15 and 22 until 36 exhibits simple leaves and species numbered from 16 to 21 and 37 to 40 have complex leaves. There are a total of 340 data. It contains 15 neurons, 1 input layer, a hidden layer contains of 23 neurons and 1 output layer.

Table 1. Detailed scientific name of each plant and the number of leaf specimen accessible by species

Scientific Name	#	Scientific Name	#
Quercus suber	12	Fraxinus sp.	10
Salix atrocinera	10	Primula vulgaris	12
Populus nigra	10	Erodium sp.	11
Alnus sp.	8	Bougainvillea sp.	13
Quercus robur	12	Arisarum vulgare	9
Crataegus monogyna	8	Euonymus japonicus	12
Ilex aquifolium	10	Ilex perado ssp. azorica	11
Nerium oleander	11	Magnolia soulangeana	12
Betula pubescens	14	Buxus sempervirens	12
Tilia tomentosa	13	Urtica dioica	12
Acer palmatum	16	Podocarpus sp.	11
Celtis sp	12	Acca sellowiana	11
Corylus avellana	13	Hydrangea sp.	11
Castanea sativa	12	Pseudosasa japonica	11
Populus alba	10	Magnolia grandiora	11
Acer negundo	10	Geranium sp.	10
Taxus bacatta	5	Aesculus californica	10
Papaver sp.	12	Chelidonium majus	10
Polypolium vulgare	13	Schinus terebinthifolius	10
Pinus sp.	12	Fragaria vesca	11

Figure 2 shows the visualization of the dataset. It is not the output of a classification model yet helps to visualized the dataset itself. It shows a matrix of two-dimensional scatter plots of every two of attributes.

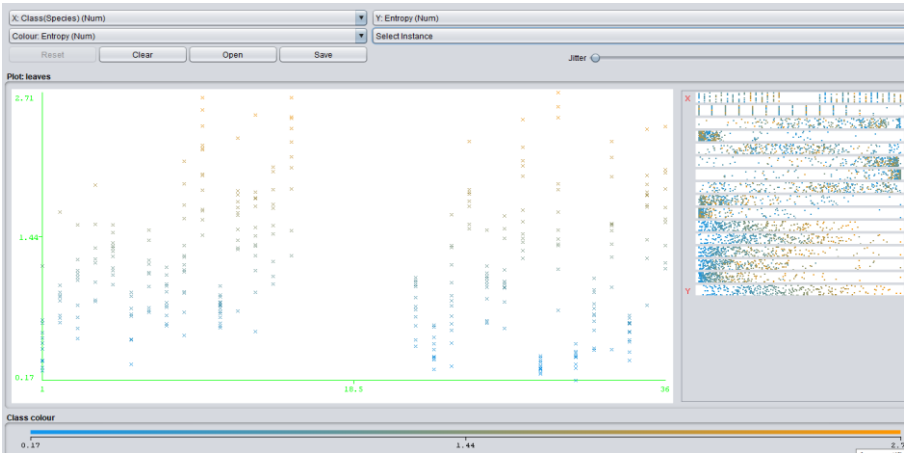

Figure 2. Visualization of dataset

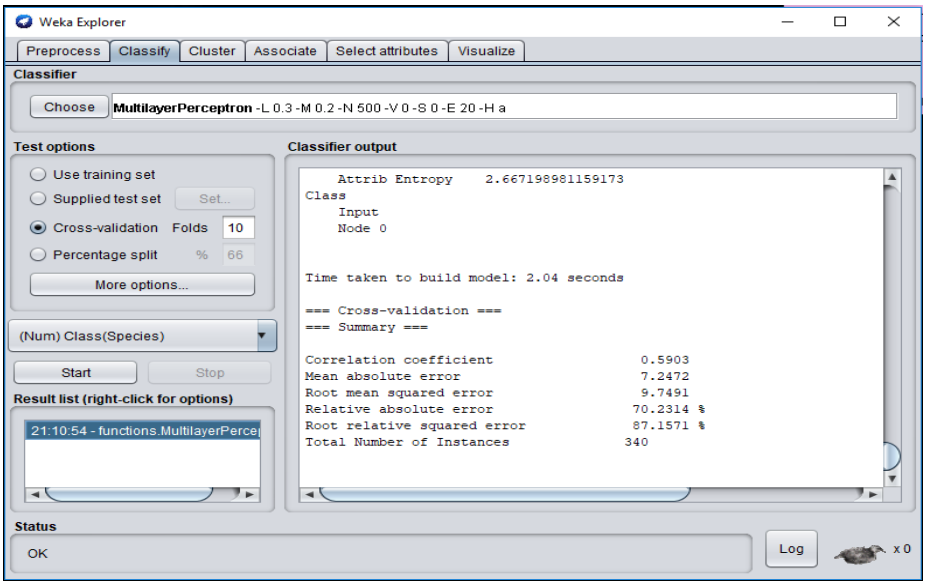

Figure 2. Summary of MLP algorithm

The dataset need to be inserted in the WEKA. After that choose classify, and choose an algorithm to use. As for an example in Figure 2, MLP algorithm has been chosen and the summary of the result has been stated.

2.2 Pre-processing

To ensure the quality of the data result, the pre-processing data should be implemented. In this research First, select subset of available data. Then, pre-process data which organize the selected data. Last but not least, transform the data that ready for machine learning.

2.3 Classification Algorithm

Predicting a new data happens by tree modelling of data which is the use of classification [2]. In WEKA, the algorithms chosen include the Bayesian classifiers, trees, rules, functions, lazy classifiers, and a final miscellaneous category. Only certain algorithms in WEKA are capable to perform regression or support predicting continuous variable. The following algorithms are used because the dataset contains continuous class variable. In this research study, we describe such an approach.

a) Artificial Neural Network: It could be composed as mathematical equations in a rationally natural way. Multilayer Perceptron is a neural network that trains that apply back-propagation. Besides, it is a precise predictor for the underlying classification difficulty.

b) K-Nearest Neighbor: The training instances is stored by the lazy learners. It only do real work when the classification time come. IBK classifiers that being used is the IBK which apply the identical distance metric. The number of nearest neighbors (default k=1) could also be described particularly or might as well determined automatically utilizing leave-one-out cross validation, subject to an upper limit given by the particular value.
c) Decision Table: Explaining the outcome has an easy way which is to make it the equivalent as input in machine learning. It creates a decision table majority classifiers.
d) M5P: It is a model tree learner that capable to build logistic model trees. M5P unite a common decision tree along the probability of linear regression functions at the nodes. In every leaf of regression model, the M5P regression model are competent with a linear regression model.

2.4 Evaluation Metric

Because of the numerical nature of the dataset, the primary quality measure proposed by the error rate is no longer suitable. Errors are not easily present or absent; they come in variety sizes. To figure out the outcome of numerical prediction, a few of alternative methods can be applied. The equation of the evaluation is shown in the following where:

n = the number of error
p = predicted values
a = the actual values

a) Correlation coefficient (CC) measure how strong a relationship is between two data. Correlation coefficient calculated using the following equation:

$$\frac{S_{PA}}{\sqrt{S_P S_A}}$$

where: $S_{PA} = \frac{\sum_i (p_1 - \bar{p})^2 (a_1 - \bar{a})}{n-1}$, $S_P = \frac{\sum_i (a_1 - \bar{a})^2}{n-1}$ and $S_A = \frac{\sum_i (a_1 - \bar{a})^2}{n-1}$

The equation return value between -1 and 1. The value 1 is considered as strong positive relationship meanwhile value -1 is strongly negative and the value 0 has no relationship. If the result shows a greater number than value 1 and less than -1, a mistake has been made.

b) Mean absolute error (MAE) is an average the magnitude of the individual errors without taking account of their sign. The equation for the MAE is as the following:

$$\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n}$$

c) Root Mean Squared Error (RMSE) measures the differences between values. It is a standard deviation of the residuals (prediction errors). The residuals are a measure of how far the regression line data/attribute points are. RMSE is calculated using the equation as stated in the following:

$$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$$

d) Relative Absolute Error (RAE) is relative to a simple predictor, the average of the actual value. The error is just the total absolute error instead of the total squared error. RAE takes the total absolute error and normalize it by dividing by the total absolute error of the simple predictor. It is calculated using the following equation:

$$\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \bar{a}| + \cdots + |a_n - \bar{a}|}$$

e) Root Relative Squared Error (RRSE) is a relative to what it would have been if a simple predictor had been used. The predictor is the average of the actual values. The equation use is as follows:

$$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \cdots + (a_n - \bar{a})^2}}$$

3. RESULTS AND DISCUSSIONS

The result obtains is stated in CC, MAE, RMSE, RAE and RRSE. It does not calculate the accuracy because dataset values is in continuous number instead of categorical or nominal values. Table 2 shows the comparison result of the selected algorithms while Figure 3 illustrates the performance.

Table 2. Comparison result of selected algorithms

Algorithm	CC	MAE	RMSE	RAE	RRSE	Time (Seconds)
MLP	0.59	7.25	9.75	70.23	87.15	2.37
IBK	0.61	5.06	9.90	49.00	88.43	0.01
DT	0.43	8.13	10.18	78.8	91.02	0.42
M5P	0.53	7.97	9.50	77.32	84.97	0.82

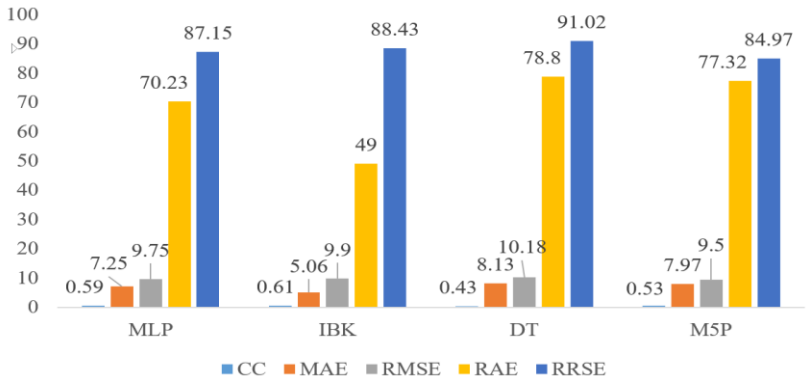

Figure 3. Comparative results across all classification algorithms

As stated in the table above, IBK gave the best result amongst those algorithms that were tested. For the CC, IBK has the result of 0.61 compared to MLP, DT and M5P which are 0.59, 0.43 and 0.53. As for the time taken for each algorithms to produce results, the KNN only took 0.01 second to produce. As for the other algorithms, DT took 0.42 seconds and M5P took 0.82 seconds meanwhile the longest time taken for algorithms to produces results is 2.37 which is MLP.

4. CONCLUSION

This research has accomplished the main objective of evaluating crucial features for herbs classification. It recognizes the most applicable algorithms for the achievement of herbs classification. All those algorithms were run and tested in WEKA tools. The comparison of all algorithms has been made and stated in order to find which algorithms gives the most excellent result for the herbs. This classification algorithm is very easy to implement in the classification tool such as WEKA. Besides, it is a flexible feature.

ACKNOWLEDGEMENT

We would like to say thank you to Universiti Tun Hussein Onn Malaysia (UTHM) and Office for Research, Innovation, Commercialization and Consultancy Management (ORICC), UTHM for kindly proving us with the internal funding (Vot E15501).

REFERENCES

[1] Brown, D. The Royal Horticultural Society – *Encyclopedia of Herbs and Their Uses*. Dorling Kindersley, London.1995.

[2] Bhargava, G. Sharma, R. Bhargava, M. Mathuria, "*Decision Tree Analysis on J48 Algorithm for Data Mining*," International Journal of Advanced Research in Computer Science and Software Engineering, vol 3, no 6, pp. 1114-1119. 2013.
[3] Alsabti, S. Ranka, V. Singh, "*CLOUDS: Classification for Large Out-of-Core Datasets*", Proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 2-8, 1998.
[4] Gehrke, V. Ganti, R. Ramakrishnan, W.-Y. Loh, "*BOAT-Optimistic Decision Tree Construction*", Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 169-180, 1999.
[5] Gehrke, R. Ramakrishnan, V. Ganti, "*Rain Forest A Framework for Fast Decision Tree Construction of Large Datasets*", Data Mining and Knowledge Discovery, pp. 127-162, July 2000.
[6] Mehta, R. Agrawal, J. Rissanen, "*SLIQ: A Fast Scalable Classifier for Data Mining*", Proc. 1996 International Conf. Extending Database Technology, pp. 18-32, 1996.
[7] Joshi, G. Karypis, V. Kumar, "*ScalParC: A New Scalable and Efficient Parallel Classification Algorithm for Mining Large Datasets*", Proc. 1998 Int'l Parallel Processing Symp. and Symp. Parallel and Distributed Processing, pp. 573-579, 1998.
[8] Srivastava, E.-H.(Sam) Han, V. Kumar, V. Singh, "*Parallel Formulations of Decision-Tree Classification Algorithms*", Data Mining and Knowledge Discovery, vol. 3, no. 3, pp. 237-261, Sept. 1999.
[9] Lim, W.Y. Loh, Y.S. Shih, "*A Comparison of Prediction Accuracy Complexity and Training Time of Thirty-Tree Old and New Classification Algorithms*", Machine Learning, vol. 40, no. 3, pp. 203-228, Sept. 2000.
[10] Silva, A.R. Marcal, R. M. A. da Silva, "*Evaluation of featurees for leaf discrimination*," International Conference on Image Analysis and Recognition
[11] A. Yasar, I. Saritas, M. A. Sahman, A. O. Dundar, "*Classification of Leaf Type Using Artificial Neural Network*," International Journal of Intelligent Systems and Applications in Engineering, 2015.
[12] D. S. Guru, Y. H. Sharath, S. Manjunath, "*Texture Features and KNN in Classification of Flower Images*," IJCA Special Issues on "Recent Trends in Image Processing and Pattern Recognition, 2010.