A K-means Based Competitive Learning with Text Description Language Features for Practical Botanical Systematics

WEN-SEN LEE
Department of Computer Science and Information Engineering
De Lin Institute of Technology
No.1, Lane 380, Qingyun Rd., Tucheng City, Taipei County 236
TAIWAN
leehsu@dlit.edu.tw

Abstract: - The practical methodologies of stable pattern classification using artificial intelligence as advisory tools are researched here according to studies in the flowering plant genera Lithops N.E. Br. (Aizoaceae). In this paper, the use of K-means model is a practical generation of groups as a classifier for botanical taxa. In order to provide comparisons for this study of effective classification performance, the study here in the succulent plant genus Lithops involved the classification of 87 records that comprise about 35 species. It is demonstrated that the proposed system using artificial neural networks technique with statistical property of grouping method can achieve a classification rate of 88.57% separated records into 35 groups referred to the traditional plant taxonomic groups.

Key-Words: - K-means, botanical taxa, succulent plant, taxonomic

1 Introduction

Conventional expert systems require expert knowledge and experience for the development of identification and classification systems, In the 1960s, numerical or mathematical taxonomy began to develop [1]. Most original applications were widely used in the commerce and engineering areas rather than in biological taxonomy. In general, botanical classification needs to perform exactly or complex data collection, due to classification usually depends on descriptions, illustrations and specimens. People who write keys use their own expert knowledge to produce an expert system based on the definition of the classes. Using such traditional methods for botanical classification can take many months or years to get sufficient characters for classification keys. Beside, few studies have focused on exploiting existing clustering systems using modern computational methods. Artificial neural networks (ANNs) are less applied for botanical systematics. One of the computer-based clustering-oriented analysis with the modern adaptive resonance theory 2 (ART2) technique to date for the botanical classification was researched by [2]. ANNs is made up of many computational processing elements called neurons or nodes by [3] and [4]. The training of ANNs is carried out to associate correct output responses to particular input patterns. Once trained properly, an ANNs has the ability to generalize when similar [5], but not identical patterns are introduced to the network.

Unsupervised classification procedures are often based on some kind of clustering strategy, which forms groups of similar patterns. Such clustering techniques are very useful for pattern classification problems. K-means clustering [6] method used here is one of the most popular conventional computerised methods. This method is the unsupervised clustering algorithm for comparison with competitive learning. K-means clustering is an algorithm to classify or to group objects based on features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of Euclidean distance between data and the corresponding cluster centre. The K number of cluster must be determined in advance for the effective performance. Among all the proposed ANNs structures, the use of unsupervised neural networks to perform clustering in the succulent plant, genus Lithops, is a neural computing technique. The network is trained based on the competitive learning rule with the winner-take-all rule. Thus, an artificial intelligent classification technique for adaptively adjusting the number of groups was is presented as a good example of self-organisation. Furthermore, the ability of such networks to perform unsupervised learning is of particular interest recently. The case study presented here is of the genus Lithops. These are flowering succulent plants of...
horticultural. They consist of about 35 species of xerophytes that occur naturally only in Namibia and three provinces of South Africa, with four confirmed localities in Botswana. Therefore, this paper is the study to discover such competitive learning with optimised classification algorithm as a recommended clustering method for succulent plant. It would be highly attractive if these techniques could be applied to botanical systematics.

2 Proposed methods

2.1 Text data acquisition
It is valuable to describe the selection of characters and records [7], and provides ANNs technique which is the most suitable for botanical classification. This kind of plant has curious shape, similar to stones with marbled and spotted leaf surfaces. The model key generated from the characteristics of the plant such as colour, size and shape, is intended to be used to identify the species of the genus *Lithops*. In order to provide comparisons for the effective classification performance of ANNs, data were extracted from Cole’s [8] monograph with permission. The descriptions of the 87 taxa at subspecific and varietal rank were used as Operational Taxonomic Units (OTUs) [8]. This was because one aim was to test the hypothesis that existing data from text language descriptions of living plants could be successfully used for training and testing. Those selected were numeric and easily coded for proposed training data pre-processing. The characters selected have been treated as follow:

1). General features
   - Number of head: pairs of leaves on a plant with a single rootstock.
   - Face major axis (length): longest dimension of top.
   - Face minor axis (width): smallest dimension of top.

2). Shape Features
   - Body profile: shape of the leaf body from the side view.
   - Curvature of lobe (from the top): the state of curve.
   - Cleft of lobe divergence.
   - Facial shape (from the top).
   - Fissure depth: measured from the base of the fissure.

3). Facial Features
   - Surface texture: texture of top surface.
   - Margins (visual).
   - Window: a translucent or transparent area on the top of the leaves.
   - Channel: broad or narrow areas between the islands and the margins.
   - Island: raised area above the level of the window.
   - Rubrications: short or long line, connected into a network, and/or dashes, hooks or dots.
   - Dusky dots: dots present or absent on the surface of leaves top.
   - Tanniniferous Idioblast (TI) (longitudinal section): absence or presence in the side part of the plant body.

4). Flower and Fruit Features
   - Flower colour.
   - Flower diameter: measured when flowers are fully open.
   - Seed capsules: number of fruit segments.
   - Seed colour.
   - Surface texture of seed.

2.2 Text data processing
Character is a different qualitative features which is simply a statement, and quantitative feature which is a number or measurement [9]. The more successful training the classification of genus *Lithops*, it is usually necessary to concern character values which are useful and weighted. For this purpose, there are three considered factors here for the effective classification performance. In the first considered factor, there are four different states for the character of surface texture. Thus, the ‘smooth’ state is numbered to be the value of 1, ‘smooth-slightly rugose’ is numbered to be the value of 2, ‘slightly rugose’ is numbered to be the value of 3, and ‘distinctly rugose or bullate’ is numbered to be the value of 4. For the training record of the *L.*
*gracilidelineata* v. *gracilidelineata*, the state has been described as ‘usually distinctly rugose, occasionally smooth’ by Cole. In these records, the network was trained with the state of ‘distinctly rugose’ which has been weighted more precisely by the statistical distribution of [10]. That is the second considered factor for the training records. In the third considered factor, for the character of ‘Merosity’ of *L. aucampiae* var. *aucampiae*, the state has been described as ‘mostly 6-merous (89%), otherwise 7-merous (7%) or 5- or 8-merous (4%)’. For these training records, the network was trained with the state of mostly 6-merous (89%) in the value of 6. The ideal way to present a quantitative character is to state its frequency of different measurement with a parameter. This is by giving the restrictions of the ranges as the modal percentage described (as 89%).

### 2.3 Methodology for proposed system

In the first instance, the purpose of data pre-processing [11] is to ensure that the inputs with larger absolute values are given the same importance as the inputs that have smaller magnitudes. Therefore, an attempt of data transformation is made to train a simple network for an adaptive network search. Here the method uses a technique based on the generalized learning rule. The transfer function used in the input and desired output layers involved a critical value of 0.9 that corresponds to the species of the current input record unless the desired output is 0.1[12][13]. In addition, the input vectors were normalised in the range between 0.1 and 0.9 to prevent any initial weighting of characters [12][13], given by

\[
R_{\text{new}} = \frac{(R - R_{\text{min}})(R_{\text{max}} - R_{\text{min}})}{D_{\text{max}} - D_{\text{min}}} (1)
\]

where \(R_{\text{min}}\) and \(R_{\text{max}}\) are the minimum and maximum value of one character over all records; \(D_{\text{min}}\) and \(D_{\text{max}}\) are the minimum and maximum of the desired output value. The competitive learning is based on the network of winner-take-all rule. Input nodes compete based on some specified criteria and winner is stated to classify input patterns. K-means is one of the unsupervised learning algorithms that solve the well known clustering problem. In this study, K-means algorithm was used for practical generation of groups. With respect to unsupervised learning, the use of K-means clustering to perform classification in the genus *Lithops* is a cluster analysis.

The training procedure is a simple and easy way to classify the training data set. The certain number of clusters fixed the known groups. The first step is to define K centers, one data for each cluster. These centers should be placed in a critical way because different place produces different result. Since all the places of the centers are moving with the new members, the current data must be adjusted every time. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When all training data found the nearest center, the first step is completed and an early grouping is done. Then the network must re-calculate for the K new centers from the previous step. After we have these k new centers, a new binding has to be done between the same data set points and the nearest new center. As a result of this loop we may notice that the K centers change their location step by step until no more changes are done. Finally, this algorithm aims at minimizing an objective function [13], given by

\[
J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_{i}^{(j)} - c_{j} \|^2
\]

where \(\| x_{i}^{(j)} - c_{j} \|\) is a chosen distance measure between a data point \(x_{i}^{(j)}\) and the cluster centre \(c_{j}\), is an indicator of the distance of the \(n\) data points from their respective cluster centers. Each character with a real continuous value was used for one input and grouping the idealized training data into each center of K cluster with the minimum Euclidean from the centre of each of the cluster. The input vectors were normalised to avoid initial character weighting and these normalised vectors were fed to the network for training. Due to the sensitivity of the K-means to presentation order of the input patterns, the input vectors were presented in random order. In order to provide an objective method for the effective clustering performance, the statistical properties of grouping method was used for the final results produced.

### 3 Experimental results

For this study, there are 35 species of succulent xerophytes. Each of these possesses 21 continuous characters that are measured with the calculation of information statistics. The experiment was an attempt to group all 87 records into the clusters of genus *Lithops*. The K-means method was used to perform botanical classification. Each character with a real continuous value was used for one input and grouping the idealized training data into each center of K cluster with the minimum Euclidean from the centre of each of the cluster. The input vectors
were normalised to avoid initial character weighting and these normalised vectors were fed to the network for training. Due to the sensitivity of the K-means to presentation order of the input patterns, the input vectors were presented in random order. In order to provide an objective method for the effective clustering performance, the statistical properties of grouping method was used for the final results produced.

3.1 Two tests performed

3.1.1 Test 1: K-means algorithm by the lowest Euclidean distance from multiple runs.
The results are shown by the k-means algorithm on presentation of the test set at the point where no new records are moved to other groups. The initialized random seed was randomly chosen at 321. There are 100 results were produced from 100 runs. Each run produced several numbers of results that comprise a different number of members for 35 groups. It is due to the training data presentation of order are randomised chosen. The single best run was then selected with the lowest Euclidean distance using random presentation of order. Since the aim is the classification of the genus *Lithops*, the effective performance was then examined according to Fearn’s [14] and Cole’s (2005) classification systems.

3.1.2 Test 2: K-means and statistical properties of grouping (SPG) method from multiple runs.
The results are also shown by the k-means algorithm on presentation of the test set at the point where no new records are moved to other groups. In order to provide a comparison with the statistical properties of grouping (SPG) method [12], it was decided to extract 5 results with lowest mean Euclidean distance from 100 multi runs. Thus, there are 5 different results were produced for the purpose of using statistical method. The cluster analysis using the SPG method was also examined according to Fearn’s (1981) and Cole's (2005) classification systems.

3.2 Statistical significance tests

Table 1 shows mean distance between individual data records for those species that have infraspecific variants. There are twenty of the thirty-five species considered here. In Table 2, the result shows the statistical significance tests performed on the Mean Node Distance results by the K-means algorithm with Euclidean and SPG method. In Table 2a, the value compared is the overall mean node distance between infraspecific variants (subspecies and varieties). It is not necessary to first carry out an F-test, since test use the same ordered set of random seeds for the tests; a paired t-test can be performed. The null hypothesis for the t-test is that there is not a significant difference between the results of the two tests. The 2-tailed probability of their being a significant difference is returned by the paired t-test as shown in Table 2b. The hypothesis is not refuted, there being no significant difference at the 5% level.

Table 1. Distance between infraspecific taxa – comparison between Euclidean and SPG method for classification performance

<table>
<thead>
<tr>
<th>Species Name</th>
<th>No. of varieties or subspecies per species</th>
<th>Mean Node Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>K-means</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Euclidean</td>
</tr>
<tr>
<td>aucampiae</td>
<td>4</td>
<td>0.33</td>
</tr>
<tr>
<td>bromfieldii</td>
<td>4</td>
<td>0.42</td>
</tr>
<tr>
<td>comptonii</td>
<td>2</td>
<td>0.37</td>
</tr>
<tr>
<td>dinteri</td>
<td>4</td>
<td>0.47</td>
</tr>
<tr>
<td>divergens</td>
<td>2</td>
<td>0.40</td>
</tr>
<tr>
<td>fulviceps</td>
<td>2</td>
<td>0.50</td>
</tr>
<tr>
<td>gesineae</td>
<td>2</td>
<td>0.65</td>
</tr>
<tr>
<td>gracilidelineata</td>
<td>3</td>
<td>0.44</td>
</tr>
<tr>
<td>hallii</td>
<td>2</td>
<td>0.29</td>
</tr>
<tr>
<td>hookeri</td>
<td>7</td>
<td>0.33</td>
</tr>
<tr>
<td>julii</td>
<td>4</td>
<td>0.63</td>
</tr>
</tbody>
</table>
Table 2. Statistical Test Results comparing Euclidean with SPG method

<table>
<thead>
<tr>
<th></th>
<th>Node Distance</th>
<th>Overall Mean</th>
<th>paired t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Euclidean</td>
<td>SPG</td>
</tr>
<tr>
<td>Mean Values</td>
<td></td>
<td>0.76</td>
<td>0.79</td>
</tr>
</tbody>
</table>

2a

Table 3. *Lithops*: 21 Characters, evaluation of effective performance

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Number of Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cole</td>
<td>22</td>
</tr>
<tr>
<td>Fearn</td>
<td>27</td>
</tr>
<tr>
<td>Cole &amp; Fearn</td>
<td>31</td>
</tr>
<tr>
<td>Performance (Cole system)</td>
<td>62.86%</td>
</tr>
<tr>
<td>Performance(Fearn system)</td>
<td>77.14%</td>
</tr>
<tr>
<td>Performance (Cole &amp; Fearn)</td>
<td>88.57%</td>
</tr>
</tbody>
</table>

4 Discussions

The results were referred to Fearn’s and Cole’s classification system. Table 3 shows that K-means method achieved the classification accuracy according to Fearn’s and Cole’s classification system. If all members of an group were all in the same Fearn’s and Cole’s system then an accuracy of 100% was recorded [12], given by

\[
\text{Accuracy} \% = \frac{\text{Specimen}_{\text{cor}}}{\text{Specimen}_{\text{tol}}} \times 100\%
\]  

where Specimen_{cor} is number of specimen classified correctly, and Specimen_{tol} is total number of specimen in Fearn’s and Cole’s system. Therefore, Table 3 shows 35 species of xerophytes for the genus *Lithops* that there are 22 groups referred to Cole’s classification system; there are 27 groups referred to Fearn’s classification system.

There are 31 groups are cooperated with Cole and Fearn’s classification system. Since the K-means technique is based on competitive learning with the winner-take-all rule in the competitive learning systems, a winner-take-all over all records is also evaluated. This is given a value of 100% if a correct classification has the highest percentage classification, or shared between joint winners. They are shown of the total samples of the test species that are classified to the corresponding species. It can be clearly seen that the statistical property of groupings method has effective performance of 62.86% that respect to Cole’s system and the rate of 77.14% that respect to Fearn’s system. If the Fearn’s classification system can be considered as part of Cole’s
5. Conclusions

There are many advantages of ANNs that have made them to be useful in various research fields related to play important roles in computational biology [15] [16]. The grouping method has been described that the K-means has implemented an effective performance that estimated an optimized number of groups and matched some existing species concepts in genus Lithops. The results presented here demonstrate that the K-means network learned quickly, and easily finds particularly clusters (groups) when producing a small number of groups. Separation was achieved based on the major groups of species according to Fearn’s (1981) and Cole’s system with the mean lowest Euclidean distance methods. This is capable of taking new data and grouping it according to clusters that were discovered by analyzing the training dataset. This study has also demonstrated that choosing text description language features for training and testing K-means can successfully classify the genus Lithops according to the output units created. The technique involves the extraction and processing of text description. Certainly, this study presented here is a recommended method to enable taxonomist to build systems using neural networks for practical botanical systematics.

Acknowledgements

I would like to thank the “Lithops – flowering stones” editor, Lino Di Martino, for the permission to use the genus Lithops data from Cole’s (2005) monograph.

References