APPLICATION OF DECISION TREE ALGORITHMS FOR DISCRIMINATING AMONG WOODY PLANT TAXA BASED ON THE POLLEN SEASON CHARACTERISTICS

Agnieszka Kubik-Komar1,*, Elżbieta Kubera1, Krystyna Piotrowska-Weryszko2 and Elżbieta Weryszko-Chmielewska3

1 Department of Applied Mathematics and Computer Science, University of Life Science in Lublin, 20-950 Lublin, Poland
2 Department of General Ecology, University of Life Science in Lublin, 20-950 Lublin, Poland
3 Department of Botany, University of Life Sciences in Lublin, 20-950 Lublin, Poland

*Corresponding author: agnieszka.kubik@up.lublin.pl

Abstract: The aim of this study was to verify which parameters of the atmospheric pollen season can distinguish between pollen types, the ranges of parameter values that delineate classes of taxa, and finally which taxa are similar to others within the domain of these parameter ranges. Decision tree algorithms were applied and the best tree was chosen to describe the rules of pollen classification. The study material consisted of airborne pollen grains of the following eight taxa: Alnus, Betula, Carpinus, Corylus, Cupressaceae, Fraxinus, Populus and Ulmus. Research was conducted in Lublin in eastern Poland during 2001-2013. The following six atmospheric pollen season parameters were analyzed: season start and end, duration, maximum daily pollen concentration, date of maximum pollen concentration, and the Seasonal Pollen Index (SPI). Four algorithms were used in data analysis and the J4.8 algorithm was chosen as the best for taxa classification, date of the end of season and the SPI value belonging to characteristics that served most to discriminate between pollen types. Based on the classification tree, the following four groups of taxa were identified: (i) Ulmus; (ii) Corylus, Alnus, Populus; (iii) Betula; and (iv) Carpinus, Fraxinus, Cupressaceae.

Key words: airborne pollen; characteristics of pollen season; decision tree; J4.8 algorithm.

INTRODUCTION

Allergic diseases are considered to be a significant global health problem. Recent research has revealed that in Poland over 45% of its inhabitants suffer from various allergies. These diseases are primarily prevalent among children and young people (Samoliński et al., 2007). The main sources of allergens include, among others, pollen grains (Holgate et al., 2001). Therefore, the investigation of pollen seasonal characteristics is a very important issue.

In the White Book on Allergy published by the World Allergy Organization (WAO), allergic diseases are considered to be a significant global health problem defined as “allergy epidemic” (Pawankar et al., 2011). Analysis of data from various epidemiological studies proves that the upward trend of cases of allergy persists and there is no prospect of a reduction in the prevalence of allergic diseases in the near future (Jackson, 2001; Marshall, 2004; Asher et al., 2006; D’Amato et al., 2007). The problems associated with allergies are expected to be even more severe due to the changes in climate and the environment. These changes affect the concentration of pollen grains and fungal spores in the air. The WAO’s mission is to spread knowledge about the health risks of allergic diseases and to improve this situation by integrated education and the promotion of research in this area as well as by taking actions designed to achieve effective prevention (Pawankar et al., 2011). The White Book on Allergy was presented to the European Parliament and the parliaments of the Member States, and allergic diseases were identified as a priority of the European Framework Program.
Due to the large number of allergic disease cases, monitoring the pollen content in the air is an important task not only because of its medical implications, but also for its economic and social implications. It is estimated that the economic cost of allergies (direct costs: expenditure on medications and health care services, and indirect costs: social costs of unemployment, social support, the loss of income tax revenue, lower productivity at work, etc.) is huge and amounts to a few hundred million Euros each year (Pawankar et al., 2011). Moreover, cross-reactions are observed between allergens of many plants, as for example in the Betulaceae family. This means that people allergic to birch pollen may also present symptoms of allergies to the pollen of hazel, alder and hornbeam (Rapiejko, 2008). In addition, polyvalent allergies (allergies caused by more than one factor) that affect more and more organs are reported more frequently (Kozłowska et al., 2007; Pawankar et al., 2011). This common risk of allergies has motivated us to attempt a classification of the most allergenic pollen types of woody plants based on the characteristics of the pollen season. In an earlier study conducted in Lublin, it was found that the spectrum of pollen grains was usually dominated by the pollen of woody plants and its average annual percentage was 58.4%. The highest concentrations of pollen were noted in April (33.3% of the average annual total). The pollen seasons of woody plants largely overlap, which can cause severe allergic symptoms due to several concurrent factors. It was found that birch pollen is the most frequent represented in the air of Lublin and reaches the highest percentage in the pollen spectrum (on average 23.6%) (Piotrowska-Weryszko and Weryszko-Chmielewska, 2014).

Decision tree algorithms were applied in order to establish the most discriminative parameters of pollen seasons and to describe the rules of pollen classification. A tree represents the process of dividing a set of objects into homogeneous classes. This division is based on a set of attributes (parameters of the atmospheric pollen season). A tree consists of the root (entire sample), nodes where a decision is made by checking the condition (test), the branch leading to the next floor below the node (node child), and finally the leaf – a class to which the observation is assigned (Maimon and Rokach, 2010). This analysis, as a non-parametric method, does not require information about data distribution, is resistant to outliers and the nature of the functional relationship between the classifiers and the objects need not be specified/known (Rokach and Maimon, 2008).

A short introduction to decision tree methodology and its application in biological sciences can be found in Kingsford and Salzberg (2008) and Geurts et al. (2009). Decision trees can be applied to both classification and regression problems. For example, Csépe et al. (2014) used the regression versions of two algorithms used in this paper (J4.8 and REPTree) for predicting daily values of Ambrosia pollen concentrations and alert levels 1-7 days ahead for Szeged (Hungary) and Lyon (France). In this paper, we have used the classification versions of these algorithms to reach the set out aims.

**MATERIALS AND METHODS**

**Biological data**

Pollen monitoring was performed in Lublin (eastern Poland) during 2001-2013. The sampling site was located close to the city center (22°32’25” E and 51°14’37” N; 197 m a.s.l.). Pollen data were recorded using a Hirst-type volumetric trap (Lanzoni VPPS 2000). The sampler was placed on the flat roof of the University of Life Sciences building in Lublin, at a height of 18 m above ground level. Pollen grains were identified in 4 horizontal traverses of the slide. Daily average pollen counts were expressed as the number of pollen grains per cubic meter of air (P/m³) (Mandrioli et al., 1998). This particular research methodology and equipment are recommended by the European Aeroallergen Network and the International Association for Aerobiology. The following atmospheric pollen season (APS) parameters were analyzed: season start (S_START) and end (S_END), duration (number of days – S_DUR), maximum daily pollen concentration (peak value – S_PEAK) expressed as a number of pollen grains/m³, date of maximum pollen concentration (S_PEAK_DATE),
and Seasonal Pollen Index (SPI) (the sum of pollen grains during the pollen season). Pollen seasons were calculated using the 95% method, in which the start of the season was defined as the date when 2.5% of the seasonal cumulative pollen count was trapped and the end of the season when the cumulative pollen count reached 97.5% (Andersen, 1991; Myszkowska et al., 2011). The following taxa were taken into account: *Alnus* – alder, *Betula* – birch, *Carpinus* – hornbeam, *Corylus* – hazel, *Cupressaceae* – cypress family, *Fraxinus* – ash, *Populus* – poplar, and *Ulmus* – elm.

**Statistical analysis**

Statistical analysis of the data was performed using the Data Mining module of STATISTICA 10 software (StatSoft Inc., 2011) and WEKA open source tool for data mining tasks (Hall et al., 2009). As part of the data mining process, the decision tree method is usually applied to large sets of data in order to create a model and check its performance. In our research, the dataset was relatively small (8 taxa x 13 years), which is why different algorithms were tried in order to obtain the best classification. The measure of the quality of each applied tree was the percentage of correctly classified observations from both the cross-validation method, which is used when the amount of data is limited (Ozer, 2008), and reused observations from the training set as test data. According to Rokach and Maimon (2008), the cross-validation estimate of the generalization error is the overall number of misclassifications divided by the number of examples in the data, which is one minus classification accuracy (percentage of correctly classified instances). This is why these two measures (number of misclassifications and number of correctly classified instances) can be used interchangeably. In the cross-validation method, we divided the dataset into 13 random parts in which the classes are represented in approximately the same proportions as in the full dataset. Each part is held out in turn and the algorithm is trained on the 12 remaining parts. Then, the classification accuracy is calculated on the holdout set. Finally, the averaged value of the 13 accuracies is calculated.

In this study, we assessed the analytic power of various decision trees. The following algorithms were applied: (i) Classification and Regression Trees (CRT); (ii) Chi-square Automatic Interaction Detection (CHAID); (iii) J4.8 based on Iterative Dichotomiser (ID3) and (iv) Reduced Error Pruning Tree (REPTree).

CRT (Breiman et al., 1984) is a data mining method for optimal partitioning of a data set. As a result, the partitioning can be represented graphically as a binary decision tree. In the case of classification, a measure of the diversity of $k$ classes is the Gini index:

$$G(S) = 1 - \sum_{i=1}^{k} p_i^2$$  \hspace{1cm} (1)

where $p_i$ denotes the probability that the observation is from class $C_i$ (estimated as a proportion of the number of observations in $C_i$ to a number of observations in training set $S$ - $n/n$). The values of the Gini index are between 0 and 1, and 0 occurs when all the data in the node are from the same category (Soman et al., 2006). Thus, in the splitting criteria the attribute with the minimum value of this index is chosen for classification in the node to minimize impurity of subsets.

CHAID is the AID (Automatic Interaction Detection) algorithm using the $p$-value of the chi-square test for the contingency table, which summarizes the response variable (dependent) and the independent one.

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{t} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$  \hspace{1cm} (2)

where $m_{ij} = p_i n_{S_j}$, $S_j$,..., $S_t$ are subsets of the training set, $p_i$ denotes the probability that the observation is from class $C_i$, $n_{S_j}$ is the number of observations in $S_j$ and $n$ is the number of observations from $C_i$ class that appeared in $S_j$ subset. CHAID was originally proposed by Kass (1980). The algorithm only accepts nominal or ordinal categorical predictors. When predictors are continuous, they are transformed into or-
dinal predictors. This algorithm forms a hierarchical classification tree with multiple branches unlike CRT, where all splits are binary (Kim and Loh, 2001).

J4.8 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool, while C4.5 is an extension of ID3 (Quinlan, 1993). It uses gain ratio as splitting criteria, which is defined as:

$$Gr_a(S) = \frac{G_a(S)}{E_a(S)},$$

where $E_a(S)$ is entropy (impurity function) of training set $S$ splitting by an $a$ attribute (feature) and $G_a(S)$ is information gain calculated according to the following formulas:

$$E_a(S) = \sum_{j=1}^{t} \frac{n_{S_j}}{n} E(S_j),$$

$$G_a(S) = E(S) - E_a(S)$$

with $E(S) = -\sum_{i=1}^{k} p_i \log_2 p_i$

All designations are explained above (Eq. 1, 2). For each attribute, the gain ratio is calculated and the

---

**J4.8 Tree growing procedure**

Start with the root. Do until node is a leaf:

Stop criterion: all instances belong to one class OR not enough instances to split (in our case <10)

Leaf: yes, no

for each attribute except class attribute:

Sort training instances according to a given attribute values

Compute values of infoGain for all possible split indices. Find maximum infoGain and get GainRatio for the best split. Set the split point to the greatest value in all data that is smaller or equal to an average of two attribute values just before and after split index for the best split

Find best attribute to split on - attribute with the highest GainRatio and infoGain higher than average infoGain computed for all attributes

Split data in the node into two subsets according to chosen attribute and split point. New subsets constitute new nodes. Repeat tree growing procedure for both nodes

**Pruning procedure**

Starting from nodes just above the leaves and going up to the root of the tree. Examine a subtree in each node

Compute error (number of incorrectly classified instances) if the subtree would be substituted by largest branch (with higher number of instances)

Compute error if the subtree would be substituted by a leaf

Compute error for the whole subtree (sum errors for all leaves)

Check which error would be the smallest and modify the tree if it is a better choice

---

Fig. 1. The schema of the growing and pruning algorithm of J4.8 tree.
one with the maximum value is chosen. In Fig.1, the scheme of the J4.8 algorithm is presented. In addition to growth of the tree, the error-based pruning procedure is shown. Pruning methods are applied after the tree grows to avoid the over-fitting to the training set. The tree is cut back into a smaller tree by removing sub-branches if the error rate after pruning is less than for the unpruned tree. This error rate is estimated as the upper bound of the statistical confidence interval for proportions, which are the numbers of misclassifications (Maimon and Rokach, 2010).

Reduced Error Pruning Tree (REPTree) is a fast decision tree algorithm based on the principle of calculating the information gain with entropy and reducing the error arising from variance (Ozer, 2008).

RESULTS AND DISCUSSION

Four decision trees were created and verified on the basis of the taxon classification error. The J4.8 algorithm was the best in both cross-validation (73% of correctly classified observations) and classification where the training and test sets were equal (87.5%) (Table 1). Based on these results, this tree was chosen for the analysis.

Table 1. Percentage of correctly classified instances for each applied algorithm in the case of cross validation (CV13) and training set used as test set.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CV13</th>
<th>test = training set</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRT</td>
<td>71%</td>
<td>84.6%</td>
</tr>
<tr>
<td>J4.8</td>
<td>73%</td>
<td>87.5%</td>
</tr>
<tr>
<td>CHAID</td>
<td>52%</td>
<td>70%</td>
</tr>
<tr>
<td>REPTree</td>
<td>62.5%</td>
<td>80.8%</td>
</tr>
</tbody>
</table>

The limit values of the atmospheric pollen season parameters dividing the set of observations into subsets as well as the number of leaf instances with the number of incorrectly classified observations are presented in the graph of the J4.8 classification tree (Fig. 2). Three of the six parameters analyzed were used in the tree construction process (Fig. 2). These features played the main role in taxon identification, especially the end of season and the sum of pollen grains. This tree has a hierarchical structure and this means that the discriminative power of S_END is higher than SPI and S_PEAK_DATE. Taking into consideration

Fig. 2. J4.8 tree trained on taxa data set. In parentheses: the number of leaf instances/the number of misclassified instances for test set=training set.
only the value of the most distinguishing attribute (S_END), the taxa can be divided into two groups – the first one with the end of the season before the 119th day of the year and the second for which the end of the pollen season is later. The results of the tree classification according to the S_END parameter are confirmed by the mean values for the end of the season for the studied taxa shown in Fig. 3.

Based on the results of the J4.8 tree, four groups of taxa were identified. Two consisted of a single taxon and are characterized by extreme values of the studied parameters: *Ulmus* (early end of season and low pollen count) and *Betula* (late end of season and high pollen count). The other two groups contain several taxa and are as follows: *Corylus, Populus, Alnus* with the value of S_END not higher than 119 and SPI value above 640, and a group consisting of *Carpinus, Fraxinus* and *Cupressaceae* with S_END higher than 119 and SPI less than 3146. The maximum pollen concentration date was also included in the tree structure. Its values separate *Populus* (higher values of S_PEAK_DATE) from *Alnus* and *Corylus* (S_PEAK_DATE less than 93 and 96 respectively).

Table 2. Confusion matrix for cross validation (CV13) and the training set used as a test set. The studied taxa are denoted by letters a-h. Correctly classified observations are on the main diagonal and misclassified ones are off-diagonal. Each row represents the classification of a given taxon and the column contains the numbers of observations classified as a given taxon.

<table>
<thead>
<tr>
<th>Classified as</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a=Corylus</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b=Alnus</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c=Cupressaceae</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d=Populus</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>e=Ulmus</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>f=Fraxinus</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g=Betula</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>h=Carpinus</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

Fig. 3. Box-plot (mean and standard deviation) of S_END parameter for taxa divided into left and right branch of the J4.8 tree by grey vertical dashed line. Horizontal lines represent the limit values of this parameter for nodes splits separately for both branches.

Fig. 4. Box-plot (mean and standard deviation) of SPI parameter for taxa divided into left and right branch of the J4.8 tree by grey vertical dashed line. Horizontal lines represent the limit values of this parameter for nodes splits separately for both branches.
The matrices (Table 2) allowed us to check the quality of the tree by showing the correctness of each taxon classification, but they also show which pollen was frequently confused with other analyzed taxa because of their mutual similarity based on the seasonal parameters studied. It is easily seen that the percentage of correctly classified instances was underestimated mainly by *Carpinus*, which was often confused with *Fraxinus* and *Cupressaceae*. These taxa were often misclassified, probably because of the similar values for the end of the season or SPI. The *Fraxinus* pollen season ended on average on May 4, and for *Carpinus* on May 5. In Lublin during 2001-2013, the average SPI values for *Fraxinus* and *Cupressaceae* were similar and amounted to 1624 and 1576 pollen grains, respectively. The sum of *Carpinus* pollen grains was a little bit lower than the sum of *Fraxinus* and *Cupressaceae* pollen.

Similarly to Lublin, in Poznań, Szczecin, Rzeszów and Sosnowiec the pollen seasons of *Carpinus* and *Fraxinus* also ended at approximately the same time (Weryszko-Chmielewska, 2006). Similar values of this parameter for the taxa in question were also found in Germany (Melgar et al., 2012). As regards the sum of *Fraxinus* and *Cupressaceae* pollen, different centers reported large differences resulting from the floristic diversity of the area. In southern Spain, Italy and Germany, significantly more *Cupressaceae* pollen was recorded compared to that of *Fraxinus* (Giner et al., 2002; Rizzi-Longo et al., 2007; Melgar et al., 2012). Different results were obtained in Rzeszów and Sosnowiec (Poland), where significantly more *Fraxinus* pollen occurred than *Cupressaceae* pollen (Weryszko-Chmielewska, 2006). In the case of *Alnus* and *Populus*, in some years a similar date for the completion of the season was found, which seems to explain why these two taxa were sometimes confused by the algorithm. Similar results regarding the end of the *Alnus* and *Populus* pollen seasons were obtained by Emberlin et al. (1990) in London. Among the studied types of pollen, *Ulmus* and *Carpinus* belong to taxa with the lowest values of SPI, which explains the error in the matrices (Table 2). In other Polish regions, the concentrations of airborne pollen of *Ulmus* and *Carpinus* usually do not reach high values either (Weryszko-Chmielewska, 2006).

In addition, three taxa have the greatest stability within the end of the pollen seasons, *Fraxinus*, *Cupressaceae* and *Carpinus*, while the greatest variations of this feature are presented by the two earliest flowering taxa, *Alnus* and *Corylus*. These data show that due to the high variability of these two previously mentioned taxa, the permanent prediction of the seasonal length is required. Among the tested types of pollen, *Betula* and *Alnus* have the highest variations of SPI.

This study confirms the results of previous aerobiological studies but also provides some new data. Decision trees as insensitive to outliers are a useful tool for the analysis of pollen data, allowing for the simultaneous analysis of many taxa for which it is possible to compare the most discriminating parameters in a single graph. The best distinction between different classes of taxa was achieved using the J4.8 tree application. This tree allowed us to describe various taxa with characteristics of season along with their limits. Determination of these limits could have an important practical significance for people that are allergic, as it can inform them about the end of exposure to allergens in the environment. Limit values can also be used for comparative analysis with other regions. S_END and SPI are the main APS features distinguishing classes of taxa. Taking into account these parameters, the following groups of taxa were obtained: (i) a group characterized by an early end to the season and a limited total value of airborne pollen – *Ulmus*; (ii) a group in which the end of the season does not exceed the 119th day of the year and the SPI is above 640 – *Corylus*, *Populus*, and *Alnus*; (iii) a group characterized by a later end to the season and a high total concentration of airborne pollen – *Betula*; (iv) a group with the value of the S_END parameter above 119, while the SPI is below 3146 – *Carpinus*, *Fraxinus*, and *Cupressaceae*. Analysis of these two parameters also allowed us to identify three taxa that exhibited the highest yearly variabilities (*Alnus*, *Betula*, and *Corylus*).
Acknowledgments: This research was supported by Poland’s Ministry of Science and Higher Education as part of the statutory activities of the University of Life Sciences in Lublin.

Authors’ contributions: Agnieszka Kubik-Komar provided the idea of the research, analysis performance, execution of calculations and figures in STATISTICA, processing of results, conclusions, and text editing; Elżbieta Kubera performed data analysis in WEKA, execution of calculations and figures, assistance in the processing of results and drawing conclusions; Krystyna Piotrowska-Weryszko implemented the aerobiological research, database development, biological literature review, discussion of results, preparation of a part of the figures, text editing; Elżbieta Weryszko-Chmielewska implemented the aerobiological research and discussed the research results. All authors read and approved the final manuscript.

Conflict of interest disclosure: The authors declare no conflict of interest.

REFERENCES


