Classification and Data Analysis for Modeling Selected Colon Diseases

Anna Kasperczuk(1) and Agnieszka Dardzinska(2)

Bialystok University of Technology, Department of Biocybernetics and Biomedical Engineering

Wiejska 45a, 15-351 Bialystok, Poland

(1)a.kasperczuk@pb.edu.pl
(2)a.dardzinska@pb.edu.pl

ABSTRACT

Nowadays analyzing the vast amount of data, including medical ones, whose numbers are growing at a rapid rate, it becomes a very important task. Therefore, it is needed to find new tools, which will be helpful during the process of extracting knowledge, not only visible and obvious from data but also the hidden one. This is extremely important in medicine, where the users still do not know the causes of the emergence and treatment of many diseases. The emergence of methods that would help physicians to make proper decisions, to expand possibilities of proper diagnoses and contribute to development medicine, in a modern, automated way. In this paper, a method of constructing a classification model for ulcerative colitis, in which the process of formation is not fully known and understandable, is presented.

KEYWORDS

Data mining, classification, decision tree, medical database, ID3, J48, k-NN method, colon disease

1 INTRODUCTION

Today we observe huge development of science, but still the efficient analysis of stored data, including medical data, becomes an enormous challenge [1]. It generates the need for exploring new methods and tools helpful to extract interesting knowledge from the collected data. The rapid development of data generation, and processing technology creates the need for data collections analyzes. The clear answer to the need for advanced and automated data analysis, database storage and data warehouses becomes data mining technology. The main task is to discover non-trivial and previously unknown dependencies and patterns from data. One of the important issues related to data mining and knowledge discovery from databases is the induction of decision trees [2], [3], [4]. In this paper, we show an application of selected classification methods for modeling extracted knowledge from a gastrological information system [5]. Ulcerative colitis is a disease causing long-term inflammation of the colon, creating irritation or ulcers. This can lead to a debilitating abdominal pain and potentially life-threatening complications. Unfortunately, there is no study on ulcerative colitis. Specialists have to exclude many other diseases such as ischemic colitis, irritable bowel syndrome, Crohn's disease, and diverticulitis of colon cancer. Accurate physical examination and analysis of the history of previous illnesses help to narrow the scope of the study. Therefore it is extremely important to try to build a classification system that will help doctors work.

2 DATA MINING

Data mining is a relatively new discipline of learning. It seeks to understand the processes under investigation and the data they generate. While talking about data mining, references should be clearly made to the analysis of real, large sets of observations, tested to generate results which can be interesting from the user's expectations point of view [3]. Knowledge discovery and data mining can be treated as set of methods and approaches of analyzing observational data sets which generate unexpected relationships and summarize data in the original way, so that they can be both understandable and useful to their user [6]. One of the popular task of data
mining is classification process [7], [8], [9], [10].

2.1 Classification

Classification is the assignment of certain objects to the appropriate classes based on certain features of these objects. While dividing certain objects characterized by a variable (qualitative or quantitative) it is necessary to designate certain values of these variables as class limit values, creating a classification scheme.

The simplest classification scheme is a dichotomy, which is a simple division of objects into two classes, a class of objects having given feature and a class that does not have this feature. The example of such division is a partition of society into adults (here understood as people over the age of 18) and minors. Another selection is a fission into women and men.

The classification is based on finding a way of mapping data set to predefined classes. Based on the content of the database, a model (such as decision tree or logical rules) is built. It is then used either to classify new objects in the database or to deeper understanding of existing classes. For example, in the medical information systems classification rules for describing individual diseases can be extracted from knowledgebase, and then they can be applied automatically in diagnosed subsequent patient’s processes [4], [11].

2.2 Decision Trees

Decision tree models are the most common form of representation of knowledge discovered in data mining process by today's commercially available software. Decision tree can be treated as a form of description of classification knowledge [12].

Compared to the other tree classification methods, decisions can be made very quickly. The primary advantage of using decision trees is a clear and fast representation of knowledge, the ability to use multidimensional data and the use of large data sets. In addition, the accuracy of this method is comparable with the accuracy of other classification methods. On the other hand, the disadvantage of this method is the high sensitivity to the missing values of attributes, as there are no open assumptions about the full availability of information gathered in the database [2]. That is why it is extremely important to prepare data properly before proceeding to the analysis. We can use ERID algorithm which helps to extract knowledge from incomplete information systems [7], [8], [12].

Classification trees are used to determine the affiliation of objects to the dependent variables qualitative classes. This can be done by measuring one or more predictive variables. The classification tree represents the process of dividing a set of objects into several homogeneous classes. The division is based on the value of the feature of the objects, the list corresponds to the classes to which the objects belong, while the edges of the tree represent the values of the attributes from which the division was made [13], [14].

Tree nodes are described by the attributes of the explored relationship. The tree borders specify all possible values for selected attribute. Tree leaves are values of a class attribute. Classification is done by viewing the tree from root to the last leaf through all edges described by attribute values [2], [11], [13], [14].

The algorithm for creating a decision tree can be written as follows [2]:

Step 1: For a given set of objects, make sure that they belong to the same class (if they belong - finish the procedure if they are not - consider all possible divisions of a given set into the most uniform subsets).

Step 2: Evaluate the quality of each of these subsets according to the previously accepted criterion and select the best one.

Step 3: Divide the set of objects as you like.

Step 4: The steps to perform for each of the subsets.

For the purpose of this paper, we use modified C4.5 algorithm. The C4.5 algorithm is one of two most popular algorithms used in practice. This algorithm is actually an extension of the ID3 algorithm. In this method, we work on incomplete information system, where using
the containment relation we build a new dataset, which is more complete then the primary one [2], [10]. The C4.5 algorithm recursively passes through all the nodes by selecting a possible division as long as further subdivision are possible. For qualitative variables, this algorithm by definition creates separate branches for each value of the qualitative attribute. This may be due to the greater branching of the tree than it is desirable, as some values may be rare and naturally associated with other ones [2]. The general idea of the tree induction algorithm using modified C4.5 algorithm is as follows:

Step 1: The tree starts with a single node representing the entire training set.

Step 2: If all the examples belong to one decision class, then the examined node becomes a leaf and is labeled with that decision.

Step 3: Otherwise, the algorithm uses the measure of entropy (heuristic function) as the heuristics for selecting the attribute that best shares the set of training examples.

Step 4: For each test result one branch is created and training examples are appropriately separated into new nodes (subtrees).

Step 5: The algorithm continues in a recursive manner for whole set of examples.

Step 6: The algorithm ends when the stop criterion is reached.

Because of the high sensitivity of the algorithm for missing data, the k-Nearest Neighbor’s method (k-NN) can be used. Looking for the most similar solution, we get the 1-NN algorithm, but sometimes it is good to look for the most similar solutions and take what we have usually done in the past. This means that we are looking for the most similar solutions, then we count how many times we have chosen a specific method and we choose the most popular solution [15]. Similarity is easily judged by distance calculation. The shorter the distance the more similar the case [16]. The most common distances are as follow [12], [15], [16]:

1. Euclidean:

\[ D(x, y)^2 = \sum_{i=1}^{n} (x_i - y_i)^2 \]  \hspace{1cm} (1)

2. Manhattan:

\[ D(x, y) = \sum_{i=1}^{n} |x_i - y_i| \]  \hspace{1cm} (2)

3. Chebyshev:

\[ D(x, y) = \max_{i=1:n} (|x_i - y_i|) \]  \hspace{1cm} (3)

In order to improve the performance of the k-NN algorithm, the commonly used technique is standardization or normalization of data. Its use causes all dimensions for which the distance is calculated to have equal significance. Otherwise, one dimension would be dominated by other dimensions.

Standardization consists in bringing a situation where the average value of a particular feature is 0 and the standard deviation is equal to 1.

\[ x_j(i) = \frac{x_j(i) - \text{mean}(x_j)}{\text{std}(x_j)} \]  \hspace{1cm} (4)

Where:

\( i \) – another vector index;

\( j \) – index of feature (variable);

\( \text{mean}(x_j) \) - average value of variable \( j \)

\( \text{std}(x_j) \) - standard deviation of variable \( j \).

Normalization consists in bringing about a situation where the value of the variable belongs to the range [0,1]. Normalization is expressed by the formula:

\[ x_j(i) = \frac{x_j(i) - \min(x_j)}{\max(x_j) - \min(x_j)} \]  \hspace{1cm} (5)

Where:

\( i \) – another vector index;

\( j \) – index of feature (variable);

\( \max(x_j) \) – maximum value of variable \( j \);

\( \min(x_j) \) – minimum value of variable \( j \).

The description of k-NN algorithm, consisting of teaching and testing modules [12] which looks as follows:

Teaching
1. Make alternatives: standardize / normalize / leave data as they are.
2. Remember the entire training set.

**Testing**

1. Make standardization / normalization / leave the data that are (testing).
2. Count the distance between the test vectors and all the vectors of the training set.
3. Sort the distances from the largest to the smallest.
4. See the k-label of the closest vectors to the test vector. Make a histogram of the frequency of each label from "k-nearest" (how many and whose labels were among the nearest ones).
5. Assign the most common label as a test vector label.
6. If there was an impasse (two classes had the same number of votes) solve the problem randomly.

**3 EXPERIMENT**

Medical data can be hollowed out to derive rules that combine diagnoses with disease symptoms. These rules can be used to automatically classify (detect a disease) of new, previously undiagnosed patients based only on their symptoms. They can also be used to find the hidden relationship between the medical condition and predicates affecting the medical condition described. Medical databases contain patient information, among others. Registered during a doctor's visit or hospital stay, diagnostic test results.

The dataset contains clinical information about 152 patients affected by ulcerative colitis. Patients are characterized by 117 attributes, and classified into two groups: patients with ulcerative colitis (class ‘0’) and patients with other diseases of the digestive system, which is not a coexisting disease for the disease under examination (class ‘1’). Our goal was to build classification model which help to reclassify patients to group of not healthy persons.

First, we use selection methods to obtain the set of attributes that are most strongly associated with the classifier (dependent). Selected attributes are following: age, smoke_now, blood_feces, nu_blood_feces, eosinophils, AIAT, sodium, potassium, stricture_crohn_disease. The number of quantitative predictors is five and the number of quality predicates is four. Then we use k-NN method to complete missing data.

To find the best classifier we should pay attention to the following parameters we receive in output [17]:
- TP Rate - rate of the instances correctly classified as a given class;
- FP Rate - rate of the instances falsely classified as a given class;
- Precision - proportion of instances that are truly of a class divided by the total instances classified as that classes;
- Recall - proportion of instances classified as a given class divided by the actual total in that class;
- F-Measure - indicator of quality of the model;
- ROC Area - the accuracy of the test depends on how well the test separates the group being tested into those with and without the disease in question;
- Kappa Statistic - a measure of conformity between the proposed allocation instance of the class and the actual, which is about the overall accuracy of the model;
- Number of correctly classified instances.

<table>
<thead>
<tr>
<th>Class</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP Rate</td>
<td>0.895</td>
<td>0.778</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0.212</td>
<td>0.105</td>
</tr>
<tr>
<td>Precision</td>
<td>0.846</td>
<td>0.852</td>
</tr>
<tr>
<td>Recall</td>
<td>0.895</td>
<td>0.788</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.87</td>
<td>0.819</td>
</tr>
</tbody>
</table>

Table 1 shows a set of quality measures J48 model. For the classification attribute characterized by ulcerative colitis, the true positive value has high value (90%), which is a satisfactory result. The false positive rate is lower (21%), which may indicate a sufficient quality of the generated model.

The F-Measure measure, which estimates the overall quality of the model, has a high value.
(87%) and indicates a properly constructed medical process.

**Table 2.** Statistics for J48 model

<table>
<thead>
<tr>
<th>Factor</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly classified instances</td>
<td>84.87%</td>
</tr>
<tr>
<td>Incorrectly classified instances</td>
<td>15.13%</td>
</tr>
<tr>
<td>Kappa Statistic</td>
<td>0.69</td>
</tr>
</tbody>
</table>

The percentage of correctly classified attributes by the decision tree is equal to 84.87%. It is a correct result and indicates the good quality of the decision tree generated. The Kappa stats index is relatively low, which means that there are close to 15% of the observations with which the randomizer failed to cope.

Generated decision tree J48:

```
blood_feces = N
  | AIAT <= 13: 1 (22.0)
  | AIAT > 13
  |   | stricture_Crohn_disease = 0
  |   |   | sodium <= 133: 1 (4.0)
  |   |   | sodium > 133
  |   |   | AIAT <= 24
  |   |   |     | age <= 20: 0 (6.0)
  |   |   |     | age > 20
  |   |   |     |   | AIAT <= 19
  |   |   |     |     | sodium <= 138 |
  |   |   |     |     |     | eosinophils <= 1.3: 1 (10.0)
  |   |   |     |     |     | eosinophils > 1.3
  |   |   |     |     |     |     | sodium <= 137: 0 (2.0)
  |   |   |     |     |     |     | sodium > 137
  |   |   |     |     |     |     | age <= 30: 0 (4.0)
  |   |   |     |     |     |     | age > 30: 1 (6.0)
  |   |   |     |     |     |     | sodium > 138: 0 (4.0)
  |   |   |     |     |     |     | AIAT > 19: 1 (10.0)
  |   |   |     |     |     |     | AIAT > 24
  |   |   |     |     |     |     | age <= 19: 1 (2.0)
  |   |   |     |     |     |     | age > 19: 0 (20.0)
  |   |   |     | stricture_Crohn_disease = 1: 1 (2.0)
  |   | stricture_Crohn_disease = 2: 0 (0.0)
  |   | stricture_Crohn_disease = 3: 0 (0.0)

blood_feces = Y
  | stricture_Crohn_disease = 0
  | age <= 35
```

The ROC curve is one of the ways to visualize the quality of the classification, showing the correlation between the TPR (True Positive Rate) and FPR (False Positive Rate). The graph is more convex, the classifier is better [18].

**Table 3.** Confusion matrix

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Real</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class_positive</td>
</tr>
<tr>
<td>Class_positive</td>
<td>TP</td>
</tr>
<tr>
<td>Class_negative</td>
<td>FP</td>
</tr>
</tbody>
</table>

\[
TPR (\text{\textit{sensitivity}}) = \frac{TP}{TP + FN} \quad (6)
\]

\[
FPR (1 - \text{\textit{specificity}}) = \frac{FP}{TN + FP} \quad (7)
\]

How does the ROC curve appear:

- We calculate the value of the decision function.
- We test the classifier for different alpha thresholds. Let us recall, alpha is the threshold of the estimated probability above which the observation is classified into one category (Class_positive) and below which - to the second category (Class_negative).
- From each classification carried out at the established alpha threshold, we obtain a pair (TPR, FPR), which is a single point of the ROC curve.
- Each classification, carried out at the established alpha threshold, corresponds to a certain error matrix.

The quality of classification using the ROC curve can be estimated by calculating the area...
under the curve (AUC). The larger value of the AUC indicates better model: AUC = 1 (ideal classifier), AUC = 0.5 (random classifier), AUC <0.5 (invalid classifier, worse than random) [18], [19].

Table 4. Confusion matrix for J48 algorithm

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>Classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>9</td>
<td>a = 0</td>
</tr>
<tr>
<td>14</td>
<td>32</td>
<td>b = 1</td>
</tr>
</tbody>
</table>

In our experiment we received ROC Curve which indicates very good classifier. The AUC for all model is equal to 0.92.

5 CONCLUSIONS

Classification methods are very useful in modern medicine. They are very helpful in finding new symptoms and patients treatment methods. In this paper, we built classification model for dependent variable. It becomes important to find the symptoms that affect whether the patient is ill or not. In this work, we use J48 method to a classification task. The decision tree algorithm shows what attributes have the greatest impact on ulcerative colitis or are the most linked to it.

Research was performed as a part of projects MB/WM/8/2016 and financed with use of funds for science of MNiSW.

REFERENCES


