Attribute Discretization and Selection

Clustering

NIKOLA MILIKIĆ
nikola.milikic@fon.bg.ac.rs

UROŠ KRČADINAC
uros@krcadinac.com
Naive Bayes Features

- Intended primarily for the work with nominal attributes
- In case of numeric attributes
  - Use the probability distribution of attributes (Normal distribution is default) for probability estimation for the each attribute
  - Discretize the attribute’s values
Attribute Discretization

Discretization is the process of transformation numeric data into nominal data, by putting the numeric values into distinct groups, which length is fixed.

Common approaches:

- Unsupervised:
  - Equal-width binning
  - Equal-frequency binning
- Supervised – classes are taken into account
**Equal-Width Binning**

*Equal-width binning* divides the scope of possible values into $N$ subscopes (bins) of the same width:

\[
\text{width} = \frac{(\text{max value} - \text{min value})}{N}
\]

Example: If the scope of the values is between 0 and 100, we should create 5 subscopes (bins) in the following manner:

Width = \((100 - 0) / 5 = 20\)

Subscopes (bins): [0-20], (20-40], (40-60], (60-80], (80-100]

Usually, the first and the final subscope (bin) are being expended in order to include possible values outside the original scope.
Equal-frequency binning (or equal-height binning) divides the scope of possible values into $N$ subscopes where each subscope (bin) carries the same number of instances:

Example: We want to put the following values in 5 subscopes (bins):

$5, 7, 12, 35, 65, 82, 84, 88, 90, 95$

So, each subscope will have 2 instances:

$5, 7, | 12, 35, | 65, 82, | 84, 88, | 90, 95$
Discretization in Weka

We apply certain *Filters* to attributes we want to discretize.

*Preprocess* tab

Option: *Choose -> Filter*

`filters/unsupervised/attribute Discretize`. 

FishersIrisDataset.arff
Discretization in Weka

Equal-width binning is the default option.

- **attributeIndices** – the *first-last* value means that we are discretizing all values. We can also name the attribute numbers.

- **bins** – the desired number of scopes (bins)

- **useEqualFrequency** – *false* by default; *true* if we use Equal Frequency binning
Discretization in Weka

Applying the filter

The resulting subscopes (bins)
Data, before and after discretization

Before

After
Attribute Selection

Attribute Selection (or Feature Selection) is the process of choosing a subset of relevant attributes that will be used during the further analysis.

It is being applied in cases where the dataset contains attributes which are redundant and/or irrelevant.

- Redundant attributes are the ones that do not provide more information than the attributes we already have in our dataset.
- Irrelevant attributes are the ones that are useless in the context of the current analysis.
Attribute Selection Advantages

Excessive attributes can degrade the performance of the model.

Advantages:

• Advances the readability of the model (because now the model contains only the relevant attributes)
• Shortens the training time
• Generalization power is higher because it lowers the possibility of overfitting

If the problem is well-known, the best way to select attribute is to do it manually. However, automated approaches also give good results.
Approaches to Attribute Selection

Two approaches:

• *Filter* method – use the approximation based on the general features of the data.

• *Wrapper* method – attribute subsets are being evaluated by using the machine learning algorithm, applied to the dataset. The name *Wrapper* comes from the fact that the algorithm is wrapped within the process of selection. The chosen subset of attributes is the one for which the algorithm gives the best results.
Attribute Selection Example

census90-income.arff
Attribute Selection Example

We want to apply the selection of attributes
Attribute Selection Example

ClassifierSubsetEval is our choice for the evaluator
Attribute Selection Example

NaiveBayes classifier
We need to discretize the numeric attributes.
Attribute Selection Example

As the search method we choose the BestFirst.
Attribute Selection Example

Filter is set and can be applied.
Attribute Selection Example

The number of attributes is reduced to 7
Clustering

*Clustering* belongs to a group of techniques of unsupervised learning. It enables grouping instances into groups, where we know which are the possible groups *in advance*.

These groups are called *clusters*.

As the result of clustering each instance is being added a *new attribute* – the cluster to which it belongs. The clustering is said to be successful if the final clusters make sense, if they could be given meaningful names.
K-Means algorithm in Weka

FishersIrisDataset.arff
Choosing the clustering algorithm

We choose the SimpleKMeans algorithm.
Parameter settings

**numClusters** – the number of desired clusters; we set it to 3 because we have 3 kinds

**displayStdDevs** – if *true*, the standard deviation will be displayed
Running the Clustering

Clustering over the imported data

We ignore the Species attribute
Results of Clustering

Centroids of each cluster and their standard deviations

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Full Data</th>
<th>Cluster#</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(150)</td>
</tr>
<tr>
<td>Sepal Length</td>
<td>5.8433</td>
<td>5.8885</td>
</tr>
<tr>
<td>Sepal Width</td>
<td>3.0573</td>
<td>2.7377</td>
</tr>
<tr>
<td>Petal Length</td>
<td>3.758</td>
<td>4.3967</td>
</tr>
<tr>
<td>Petal Width</td>
<td>1.1993</td>
<td>1.418</td>
</tr>
</tbody>
</table>

Number of instances in each cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>61</td>
<td>41%</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>33%</td>
</tr>
<tr>
<td>2</td>
<td>39</td>
<td>26%</td>
</tr>
</tbody>
</table>
Evaluation of Results

Select the attribute which we want to compare the results with.

Which classes are in which clusters

Names of classes which are given to clusters
Visualization of Clusters

- Right click
  - Visual representation of clusters
Was clustering successful?

*Within cluster sum of squared error* gives us the assessment of quality.

It is being counted as the sum of square differences between the value of the attribute of each instance and the value of the centroid of the given attribute.
How to figure out the number of clusters?

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55.6</td>
</tr>
<tr>
<td>2</td>
<td>12.1</td>
</tr>
<tr>
<td>3</td>
<td>7.0</td>
</tr>
<tr>
<td>4</td>
<td>5.5</td>
</tr>
<tr>
<td>5</td>
<td>5.0</td>
</tr>
<tr>
<td>6</td>
<td>4.8</td>
</tr>
<tr>
<td>7</td>
<td>4.7</td>
</tr>
<tr>
<td>8</td>
<td>4.2</td>
</tr>
<tr>
<td>9</td>
<td>4.1</td>
</tr>
<tr>
<td>10</td>
<td>3.6</td>
</tr>
<tr>
<td>20</td>
<td>1.7</td>
</tr>
<tr>
<td>50</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Little difference in error
Using Clusters for Classification

AddCluster – our choice of the filter

Setting no class
Using Clusters for Classification

We choose the **SimpleKMeans** as the clustering algorithm.

In terms of clustering, we ignore the attribute 5 (Species).
Using Clusters for Classification

After the filter is being applied (Apply) we add the new attribute by the name of cluster.
Using Clusters for Classification

Optional: this attribute can be removed before we create a classification model.
Using Clusters for Classification

We use the NaiveBayes classifier.

We do the classification according to the cluster attribute.

The confusion matrix:
Thank you notes

Weka Tutorials and Assignments @ The Technology Forge

• Link: http://www.technologyforge.net/WekaTutorials/

A survey for you, to judge us :)  

http://goo.gl/cqdp3I
Any questions?

NIKOLA MILIKIĆ
nikola.milikic@fon.bg.ac.rs

UROŠ KRČADINAC
uros@krcadinac.com