TDDD41/732A75: Association Analysis -1

Goals

- Cluster a given dataset and use association analysis to describe the clusters obtained.

Procedure

- **Dataset**

  In this exercise, you will work with one of the most well-known datasets in the data mining literature, namely the Iris dataset. The dataset consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample, they are the length and the width of sepal and petal. You can find the dataset here.

  Open the iris.arff file. Since the association analysis in Weka (Apriori algorithm) cannot cope with continuous attributes, we should discretize the iris dataset before starting the mining process. Weka provides several filters to apply to the data. You can see them by pressing the **Choose** button. We are interested in the Discretize filter that you can find by selecting the directory Unsupervised first and then the directory Attribute. Now, click on the line that has appeared to the right of the Choose button to edit the properties of the filter. You can find a detailed description of the filter by pressing the **More** button. Select the attributes indices 1-4 (meaning that you do not want to discretize the 5th attribute as it is the class and thus already discrete) and select 3 bins (number of states of the discretized attributes). Press **OK**. Press **Apply**. Now you can edit the data again by pressing the **Edit** button and see that the data has actually been discretized.

- **Clustering**

  Appy SimpleKmeans clusterer to the data with 3 clusters (since we know there are 3 types of Iris flowers) and seed value 10. In the Cluster mode, select Classes to clusters evaluation to crosstabulate the clustering and class labeling. Ignore the class attribute.

- **Association analysis**

  To perform association analysis, click on the **Associate** tab. Press the **Choose** button to select the association algorithm (we recommend to use the Apriori algorithm). Click on the line that has appeared to the right of the Choose button to edit the properties of the algorithm. You can find a detailed description of the algorithm by pressing the **More** button. Set the desired properties and press **OK**. Click **Start**. Check the output on the right hand side of the screen. Note that after the conjunctions of attribute-value pairs on the right and left hand sides of each rule, there is a number. That number indicates the support of the determinant and of the determinant plus the consequent.

- **Visualization**

  By clicking on the **Visualize** tab, you can see the data crosstabulated for each pair of attributes. Set your visualization preferences with the bars at the bottom of the screen.

- **Describing clustering through association analysis**

  Now we use association analysis to assist you in describing the clusters found in the Iris dataset. The first thing to do is to create a new attribute that represents the cluster label assigned to each instance. For this purpose, click on the Preprocess tab and select the AddCluster filter (in the Attribute directory within the Unsupervised directory). Click on the line that has appeared to the right of the Choose button to edit the properties of the filter, e.g. which clustering algorithm to use, number of clusters (we recommend to use 3 clusters), ignored attributes (ignore the class), etc. Check the section on clustering above if in doubt. Press Apply. Check that a 6th attribute has been created with the clustering label. Now, run the association analysis as indicated in the corresponding section above. Find rules that are accurate and such that the antecedent does not contain the class attribute and the consequent only contains the cluster attribute. Find such rules for the 3 clusters. This should help you to describe the instances grouped in each cluster. Repeat the exercise above with a different combination of clustering algorithm, number of clusters and/or number of bins in the discretization filter, in order to see whether you get better or worse results.

Submission

Submit a report on your experiments and results as well as answers to all questions in the text. In addition to the experiment described in the text (3 clusters, 3 bins) your report should include discussion about at least two out of 3 possible varitions.

- Different clustering algorithm (avoid using MakeDensityBasedCluster wrapper in this case)
- Different number of clusters
- Different number of bins

When running experiments with different variations, make sure you reload the data and reapply the AddCluster filter. In each of the 3 (or 4) experiments you should identify at least one (possibly the best) association rule for each of the clusters (**Hint**: numOfRules parameter). In addition, try to explain the differences between results in different experiments as well as reasons for why these differences occured.

When explaining the experiments there is no need for explaining every step of the process (such as which buttons were clicked, etc). It is enough to say which algorithm was run and which arguements were used. In addition, do not copy-paste the full outputs from the tool. When presenting the results, present only those parts of the output which are relevant for what you are trying to explain. Avoid writing statements without motivation, such as "Results are better/worse." In other words, explain in what respect they are better/worse and why this happened.