Viet Nam National University, Ho Chi Minh City

University of Science

Faculty of Information Technology

TRẦN ĐẠI CHÍ - 18127070
PHAN TẤN ĐẠT - 18127078

LAB 02

|Topic|

Classification and Clustering

|Lecture|

M.Sc Lê Ngọc Thành

Nguyễn Ngọc Đức

Dương Nguyễn Thái Bảo

Thành phố Hồ Chí Minh - 2020

Thankfulness

Our group would like to send our most sincere thanks and deep gratitude to the lecturers for creating conditions for the group to learn and complete the lab. And our group also would like to thank the lecturers for their enthusiastic guidance and help for the group to successfully complete this lab.

In the process of implementation, it's difficult to avoid some mistakes, I hope lecturers can ignore and give suggestions so that the group can learn from experience for the next projects.

We sincerely thank you!

1. Which classification method typically has the best result? In two datasets “preprocess_default.arff” and “preprocess_reduce_noise.arff”, the percentage of classification when using J48 always give us the best result and it just deviates a little from using “NaiveBayesSimple”.
2. Which method does not work well and why? As we can see from “Result.xlsx” file, Id3 not work well in experiment A because our dataset contains both numeric and nominal attributes and we just only use our dataset with Id3 after filtering in experiment B and C but it just useful with evaluation strategy “use training set” and almost gain very low percentage when using “cross-validation” and “percentage split”. So if our dataset is continuous, we should use another classification algorithms instead of Id3.
3. Why should we use the discretized version of the data set instead of the original one? Because discretized transforms the continuous values of the variable to discrete ones (an instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes, it’s simply binning, skip the class attribute if set). Therefore, it can help us improve significantly the classification performance and we also use Id3 with our dataset in this case but if we keep default no filter and change all attributes of our dataset from numeric to nominal, the percentage of correctly classification will increase significantly.
4. Which evaluation strategy tends to overestimate the accuracy and why? In experiment A, B and C, evaluation strategy “use training set” always tends to overestimate the accuracy because “use training set” is useful when we have all the data and we’re interested in creating a descriptive than a predictive model and our current dataset was collected on random samples of three different species of hawks: red-tailed, sharp-shinned, cooper’s hawks and we just use this dataset for analyzing purpose, so when we use “training set”, it will prepare our model on the entire training dataset, then evaluate the model on the same dataset that can achieve a perfect score.
5. Which evaluation strategy tends to underestimate the accuracy and why? In experiment A, B and C, evaluation strategy “percentage split” always tends to underestimate the accuracy because it split our dataset into a training and a testing partitions with ratio 66%, 34% and can give

us a very quick estimate of performance but it just preferable only when we have a large dataset, but our current dataset is a subset of the original dataset and it’s not large enough to use this evaluation strategy usefully.