REI502M - Introduction to Data Mining

Report on Project 1

Elías Snorrason October 20, 2019

Contents

1 Pre-processing

- The data set can be found in an arff format at https://github.com/ongxuanhong/Preprocessing-with-horse-colic-dataset/blob/master/horse-colic.arff
- If association rule mining is used for classification, the `outcome` attribute is set as the class attribute. The class values are renamed as:
  - lived
  - died
  - was euthanized
- The rectal temperature was discretized to throw out extreme values, using 8 bins.
- The pulse attribute was also discretized, this time into 8 bins with equal frequency.
- The respiratory rate was relabeled accordingly:
  - Normal: 8-10 bpm
  - Above Normal: 11-25 bpm
  - Fast: 26-35 bpm
  - Extreme: Over 35 bpm
- Total protein was discretized into 5 bins
  - $<$ 5.5 gms/dL
  - 5.5-6.4 gms/dL
  - 6.5-7.5 gms/dL
  - 7.6 - 10 gms/dL
  - $>$ 10 gms/dL
- Abdomcentesis total protein was split into 3 bins
  - 0- 1.5 gms/dL
  - 1.6 - 3 gms/dL
  - $>$ 3 gms/dL
- packed cell volume was split into 3 bins
  - $<$ 31 %
  - 31-50 %
  - $>$ 50 %
- nasogastric reflux pH was split into 3 bins
- Removed the Age attribute since it was so skewed

The following attributes were removed from the dataset:

- Tunni
- Type of lesion:
- Subtype of lesion:
- Pathology_cp_data:

Note: Using the `MergeManyValues` for numerical attribues...