Discussion Board 2

Week 5

Daniel Prusinski

What are the benefits of using statistical methods for discretization of the variable?

The pros for the discretization of variables are:

* In some cases the values reflect a more accurate representation of reality
  + ie 90 mph = speeding
  + ie 60 mph = normal
  + ie 0 mph = fully stopped
* Discretization can statistically strengthen a data set
* Easier to understand and explain
* The data is simplified into categories
* On the programming side, it can be quicker
* There are more techniques for analyzing discrete data

Given an item set, discretize the data, and use Weka and the Apriori method to simulate the Min-Apriori method. The item set I used was the soybean data:

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: soybean

Instances: 683

Attributes: 36

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 6

Size of set of large itemsets L(3): 2

Best rules found:

1. int-discolor=none 581 ==> sclerotia=absent 581 conf:(1)

2. mycelium=absent int-discolor=none 575 ==> sclerotia=absent 575 conf:(1)

3. leaves=abnorm sclerotia=absent 548 ==> mycelium=absent 547 conf:(1)

4. sclerotia=absent 625 ==> mycelium=absent 619 conf:(0.99)

5. int-discolor=none 581 ==> mycelium=absent 575 conf:(0.99)

6. int-discolor=none sclerotia=absent 581 ==> mycelium=absent 575 conf:(0.99)

7. int-discolor=none 581 ==> mycelium=absent sclerotia=absent 575 conf:(0.99)

8. leaf-malf=absent 554 ==> mycelium=absent 548 conf:(0.99)

9. mycelium=absent 639 ==> sclerotia=absent 619 conf:(0.97)

10. leaves=abnorm mycelium=absent 567 ==> sclerotia=absent 547 <conf:(0.96)>

This discretization is helpful when solidifying confidence intervals and showing the different buckets the data can be classified in.