Classification using Sum of Ranking Differences of Outlier Measures

Tony Lemos, John H. Kalivas
Department of Chemistry
Idaho State University
921 S. 8th Avenue, STOP 8023 Pocatello, ID 83209, USA
lemonot2@isu.edu, kalijohn@isu.edu

Abstract
A useful application in analytical chemistry is classifying unknown samples into classes. Single-class classification is a type of classification approach where only one well-defined class of interest. Outlier detection is useful for defining class membership for unknown samples, since outlier detection removes samples that are not represented by the sample class space. When using outlier detection, there are two problems: which outlier measure to use and the tuning parameter value for the chosen outlier measure. The proposed technique for single-class classification using outlier measures eliminates these problems. To avoid selecting any one particular outlier measure, multiple measures are evaluated by using sum of ranking differences (SRD). The method of SRD is used to evaluate multiple outlier measures to obtain a consensus in classifying a sample. In regards to training parameters, a parameter window is used to avoid doing more work, such as having a training set of samples to select a tuning parameter. Wavelength selection and using spectra from different instruments is used in conjunction with SRD to provide a robust characterization of the class of interest. Presented are results for the new classification approach on spectral food data sets.

Objectives

- Create a simple procedure to perform one-class classification
- Utilize multiple outlier measures to obtain a consensus in classifying a sample

Background

Definitions:
- Target class - Class of interest
- Non-target samples – Samples not belonging to the class of interest

Two types of classification techniques:
- Discriminant Classification
  - Samples are classified into more than one predefined class
  - One-class classification
  - Samples are classified into one predefined target class

Outlier Detection

- Outlier - A not-outlying observation, or outlier, is one that appears to deviate markedly from other members of the class in which it occurs
- Outlier detection is one-class classification
- Differentiating between data that appears normal (belonging to a class) and abnormal
- Difference: Application
  - Outlier detection – Which samples are not conforming to the normal behavior of similar samples
  - One-class classification – Is this sample behavior similar enough to the other samples to belong to their class?

Approach

- 17 outlier measures
- Comparison of columns (samples) across rows (merit)
- Determines a rank for each sample
- Sum of ranking differences (SRD)
- Comparison of ranks by random numbers (SRD)
- Determine the probability that the SRD sample rankings is not a random ranking
- Table 1: List of outlier measures used for the non-classification

Data sets

- Meat Mid-infrared (MIR)
  - 40 samples for each class
  - Process:
    - 5 samples from each class for validation
    - Maximum tuning parameter window: 24
    - 10 splits
- Strawberry puree MIR data
  - 351 strawberry samples
  - 632 non-strawberry (strawberry adulterated with other fruits) samples
  - Process:
    - 30 validation samples from each class
    - Outlier clean the target class
    - 10 splits
    - Stack wavelength regions

Results

- Strawberry puree MIR data
  - True positive (TP)
  - False positive (FP)
  - True negative (TN)
  - False negative (FN)
  - Sensitivity = TP / (TP + FN)
  - Specificity = TN / (TN + FP)
  - Accuracy = (TP + TN) / (TP + TN + FP + FN)

Meat Results

Target class - Turkey
- Figure 2: Spectra and the principle component (PC) plot (bottom) for each meat

Conclusion

- SRD is an effective one-class classification technique
- Generally increases in accuracy at higher windows
- Flexibility of SRD
- Outlier measures
- Instruments
- Preprocessing methods
- Tuning parameters
- Tuning parameter window
- Adjust sigma threshold

Acknowledgements
Work supported by the National Science Foundation under grant No. CHE-2004611 (co-directed by CDE and EE). Programs.com is gratefully acknowledged by the authors.