Riffle: an **R** package for Nonmetric Clustering

Geoffrey B. Matthews and Robin A. Matthews
Western Washington University
Bellingham, WA, USA

February 17, 2006

We present here an **R** package for Riffle, a nonmetric clustering technique [2]. This is a algorithm for clustering (unsupervised learning) that does not rely on a similarity measure for multivariate data, and uses only nonparametric (order) statistics. It is suitable for mixed nominal, ordinal, *etc.* attributes, as are often found in environmental data analysis.

The current implementation of Riffle in **R** has a number of improvements over the original implementation [2], utilizing a marginalized expectation maximization (EM) approach to speed up the search. This has advantages in avoiding local minima, as well. Also, a technique for creating useful seed clusterings has been developed (rather than completely random initial clusters, as in [2]), substantially speeding up the clustering and making the final cluster less susceptible to noise. Procedures are also provided to use the resulting clustering to find an optimal subset of the attributes, and to create a naive Bayes classifier.

Riffle has been used successfully in a variety of clustering tasks, and we have found it to be a useful, intuitive technique for graduate and undergraduate students in environmental sciences.

Although nonmetric clustering is over a decade old, it is not widely known. A recent authoritative survey [1],pp. 541-542, does not include a discussion of them, and laments, “How do we treat vectors whose components have a mixture of nominal, ordinal, interval and ratio scales? Ultimately, there are rarely clear methodological answers to these questions. ... We have given examples of some alternatives that have proved to be useful. Beyond that we can do little more than alert the unwary to these pitfalls of clustering.” It is hoped that this **R** package will promote a more widespread use of nonmetric clustering in situations where it is appropriate.

References

[1] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification, Second Edition.* John Wiley & Sons, Inc., New York, NY, 2001.

[2] Geoffrey Matthews and James Hearne. Clustering without a metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):175–184, 1991.