Identifying Groups among Binary and Ordinal Phenotypic Antimicrobial Resistance Data Using Cluster Analysis Techniques

Alali, W.Q. [1], Scott, H.M.[1], and Gold, D.L.[2]
[1]Department of Veterinary Integrative Biosciences; [2]Department of Statistics; Texas A&M University, College Station, Texas 77843, USA.

Abstract

Describing multivariate phenotypic antimicrobial resistance (AR) data has been of critical importance in many recent studies, especially given the inherent high variability among these types of data. In order to assess various cluster analysis techniques, we have applied hierarchical algorithms and iterative partitioning methods on binary (i.e., classified as resistant or susceptible) and ordinal (i.e., minimum inhibitory concentration [MIC] values) AR data sets. The AR data were obtained from an ongoing longitudinal study (2004-2007) assessing the potential for transmission of AR among integrated livestock and human populations. Our example data sets represent 504 *Escherichia coli* colonies isolated from human wastewater, then phenotypically characterized as to resistance to 15 antibiotics. Ward's minimum variance with squared Euclidean distance was the only method to describe the clustering in our binary data at a solution of 15 clusters with $R^2 = 0.83$. No single unique phenotype found in one cluster existed in others. 46% of the isolates (233/504) were pan-susceptible and grouped in one cluster. The 57 different multi-resistance AR patterns fell in different clusters suggesting certain AR genes are linked in the bacterial isolates. None of the cluster algorithms was able to group the ordinal MIC data. The association between the obtained clusters and the risk factors (including host-occupational exposure to animal agriculture) in our AR study will be assessed.

Introduction

The development, selection and spread of antimicrobial-resistant bacteria in human and veterinary medicine is a major concern worldwide. AR phenotypes are the *in vitro* resistant characteristics of bacterial isolates against the action of one or more antimicrobial agents. These characteristics often are assessed by measuring the MIC using broth microdilution methods or by disk diffusion methods. These values are usually further interpreted into susceptible or resistant phenotypes according to breakpoints that are determined by a variety of committees (e.g., NCCLS). AR resistance profiles for a bacterial isolate can range from pansusceptible to resistant to multiple antibiotics.

Finding AR phenotypic groups with similar properties in hundreds of bacterial isolates is of great significance to understand resistance patterns. Visualizing and consequently analyzing large multinomial AR phenotypic data sets with high variability is difficult. Therefore, being better able to classify AR phenotypes would help our understanding of the relationship among the resistance clusters and the risk factors in a study, and also would aid in the interpretation of seemingly unique versus more common 'garden variety' AR phenotypic patterns.

Cluster analysis has been applied in different branches of science and recently was used in an AR phenotypic data analysis (Berge et al., 2003). Those authors applied several cluster analysis techniques on continuous (i.e., interval) AR phenotypic data. To our knowledge, there are no papers describing the performance of cluster analysis techniques for use with binary or ordinal AR data.

Objective

The objective of this study was to assess the relative utility and appropriateness of a variety of cluster analysis techniques for AR phenotypic data; that is, identifying clusters with similar resistance patterns in order to better describe AR phenotypes among bacterial isolates.

Methods

A subset of AR phenotypic data from our ongoing longitudinal antimicrobial resistance transmission study were used to assess a variety of cluster analysis techniques. The example data set included 504 *E. coli* isolates that were obtained from human wastewater samples and were tested for antimicrobial susceptibility. A more complete description of our study population and the antimicrobial susceptibility test procedures may be found in Scott et al.(2005). The data were in two forms: 1) ordinal scale – which represent a variety of MIC cutpoints from a restricted range of dilutions, and 2) binary scale (i.e., susceptible or resistant) which represent the interpretation of the MIC values for 15 antibiotics. Squared Euclidean distance for dissimilarity was used for both types of data. Six hierarchical agglomerative clustering methods (single linkage, complete linkage, average linkage, centroid linkage, median linkage, and Ward's minimum variance) and one iterative partitioning method (k-means) were applied to *E. coli* phenotypes to determine their cluster memberships. The MIC values were used in their original form, standardized with a mean of 1 and a standard deviation of 1, and log transformed. The cluster analysis was performed using both SAS® and SPSS® statistical packages.

Cluster analyses usually provide more than one cluster solution. The final number of clusters in SAS® was determined using the squared multiple correlation (R^2), the cubic clustering criterion (CCC), the Pseudo-F, and the pseudo-T^2 (PST2) statistic.

Results

Using our example data set, we were unable to obtain meaningful cluster distributions that best describe both types of data – ordinal or binary – using the six hierarchical agglomerative and k-means in SPSS. Similar resistance patterns (phenotypes) that were found in one cluster also existed in one or more other clusters, which indicated lack of homogenous clusters. In SAS®, cluster algorithms (hierarchical agglomerative and k-means) with squared Euclidian distance were unable to cluster MIC values into meaningful clusters with or without standardization or transformation.

The six hierarchical methods with a squared Euclidean distance were compared with cluster solutions of 10, 15, and 22 to find the find the cluster fit that best described the binary data. The method with highest R^2 that explained the highest variability and met our objective was Ward's minimum variance at 15 cluster solution and with $R^2 = 0.83$. Other hierarchical methods were not able to identify homogenous clusters. K-means produced 2 large clusters, and 13 small clusters using the cluster solution (n=15) that was suggested by Ward's minimum variance.

Using Ward's method, there was no single resistance pattern (phenotype) found in one cluster that existed in others. 46% of the isolates (233/504) were pansusceptible and grouped in cluster A (Table 1). Singly resistant phenotypes for ampicillin and tetracycline were found in cluster D and E, respectively, accounting for 7% (36/504) of the isolates. Doubly resistant patterns existed in 28% (139/504) of the isolates and appeared in clusters B, G, and H with resistance to cephalothin with ceftiofur, or ampicillin, or tetracycline respectively, and in cluster C nalidixic acid with tetracycline. At least one of the 4 antibiotics: ampicillin, cephalothin, nalidixic acid, and tetracycline were found in one or more of the multiple resistance clusters (B-O) (Table 1).

Discussion

There appears to be no "best" cluster technique that works for all types of data. Therefore, several cluster techniques should be applied to recover the best cluster structure in a data set. Furthermore, commercially available statistical packages (e.g., SAS and SPSS) may produce different cluster solutions, possibly due to differences in their clustering algorithms. In our paper, as well in another

AR paper (Berge et al., 2003), Ward's minimum variance with squared Euclidian distance produced well-separated homogenous clusters. Milligan (1981) has supported, in his Monte Carlo tests of cluster analysis review, that Ward's minimum is the best method to produce well separated clusters.

Table1 AR binary phenotypic clusters using Ward's method. Table shows 15 clusters and the number of isolates per cluster. Grey-shaded boxes show the number of resistant isolates to each antibiotic.

	Antibiotics[a]															No. of
Cluster	AUG	AMP	FOX	TIO	CRO	CEP	CHL	CIP	GEN	KAN	NAL	STR	SUL	TET	SXT	isolates
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	233
B	0	0	0	1	0	100	0	0	0	0	0	0	0	0	0	100
C	0	0	0	0	0	0	0	0	0	0	15	0	0	1	0	15
D	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	16
E	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	20
F	0	5	0	0	0	4	0	0	0	0	1	0	12	0	1	12
G	0	12	0	0	0	12	0	0	0	0	0	0	0	0	0	12
H	0	0	0	0	0	12	0	0	0	0	0	0	0	12	0	12
I	0	2	0	0	0	9	0	0	0	0	9	0	0	0	0	9
J	15	8	0	0	0	15	0	0	0	0	2	0	2	1	0	15
K	0	16	0	0	0	12	0	0	0	1	1	8	0	16	0	16
L	0	0	0	0	0	2	0	1	0	3	0	11	4	10	0	11
M	1	14	0	0	0	13	0	0	0	1	0	11	18	18	15	18
N	1	9	0	0	0	7	5	0	1	0	2	10	9	2	9	10
O	2	5	0	0	0	4	2	4	3	2	4	4	5	3	5	5

[a] amoxicillin/clavulanic acid (AUG), ampicillin (AMP), cefoxitin (FOX), ceftiofur (TIO), ceftriaxone (CRO), cephalothin (CEP), chloramphenicol (CHL), ciprofloxacin (CIP), gentamicin (GEN), kanamycin (KAN), nalidixic acid (NAL), streptomycin (STR), sulfamethoxazole (SUL), tetracycline (TET), trimethoprim/ sulfamethoxazole (SXT).

The cluster algorithms we assessed were unable to group MIC values (i.e. ordinal data) into well-separated clusters. We believe that the nature of these ordinal data, based on 2-fold dilutions, caused instability in cluster algorithms which resulted in non-homogenous clusters that do not represent the data properly.

Our data set contained 57 different phenotypic resistance patterns. These patterns ranged from singly resistant through multiple resistance to as many as 12 antibiotics. Resistance genes were suggested to be genetically linked in bacterial isolates. Bacterial exposure to an antibiotic may cause the organism to express resistance to multiple antibiotics that could be genetically related. Also, resistance may persist due to the genetic linkage of several AR genes, providing for their continued existence even as antibiotic selection pressures change. The clusters we obtained represent 'groupings' among the resistance phenotypes that might best explain the AR genetic linkage in the bacterial isolates.

This study represents an early application of cluster techniques for non-continuous AR data. Papers describing AR studies often underutilize MIC values in their analysis, often due to the difficulty arising from the nature of these data. Thus, the question of how to group MIC data into meaningful clusters is still a problem open for solutions.

References

Berge, A.C., Atwill, E.R., Sischo, W.M. (2003) Assessing antibiotic resistance in fecal *Escherichia coli* in young calves using cluster analysis techniques. *Preventive Veterinary Medicine*, 61(2):91-102.

Scott, H.M., Campbell, L.D., Harvey, R.B., Bischoff, K.M., Alali, W.Q., Barling, K.S., and Anderson, R.C. (2005) Patterns of antimicrobial resistance among commensal *Escherichia coli* isolated from integrated multi-site housing and worker cohorts of humans and swine. *Foodborne Pathogen and Disease*, 2(1): 24-37.

Milligan, G.W. (1981). A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research*, 16: 397-407.