Contact: Erich Huang
INSTITUTION: National Cancer Institute
Primary Category: AS Informatics
Secondary Category: AS Image Management, Image Analysis and Workflow

A novel statistical method for lossless compression of diagnostic imaging features

E Huang, PhD, Rockville, MD; J B Freymann, BS; J Kirby; C Jaffe, MD; A E Flanders, MD (erich.huang@nih.gov)

BACKGROUND

To classify image diagnostic content researchers often use itemized scales to record diagnostic features (e.g. BI-RAD scores). Categorization lists with fewer response elements are cognitively more concise, lead to better inter-reader agreement, and offer advantages in statistical analysis but at the risk of information loss. We present a novel Paired Category Collapsing process to successively compress imaging feature categories while retaining the information contained in the original categories.

EVALUATION

We applied this technique to a study of 75 glioblastomas collected for the Cancer Genome Atlas (TCGA) initiative analyzed by a team of six neuroradiologists who reviewed the MR images classifying each according to 24 categorical imaging features describing tumor morphology. Beginning with the original feature categories, we searched through all possible level scale pairs for that whose aggregation results in the least information loss, namely the pair such that within each category the distributions of the other features are most similar. We then aggregated the selected pair of categories and repeated the process until any further aggregation resulted in substantial information loss. We compared the overall survival prediction accuracy of multivariate Cox regression models using the original categories to that of models using the compressed categories to assess whether diagnostic feature information was indeed lost during the compression.

DISCUSSION

We significantly reduced the numbers of response elements in several feature categories, including tumor location, deep white matter invasion, and distribution, where the number of categories declined from 6 to 4, 8 to 5, and 4 to 2 respectively. We found no evidence of differences in the prognostic accuracy of the original feature categories versus the compressed ones (p-value 0.21), which suggests our compression of the feature categories is indeed lossless.

CONCLUSION

This novel statistical approach reduces the complexity of observational imaging feature data without information loss.