

Abstract – We present a cluster-based sampling and ensemble method to learn from large, imbalanced data set for bleeding detection in CE videos. Our method selects training examples randomly according to the data distributions derived from clustering. Multiple training sets are created such that data balance is restored. The sampling probability is proportional to the cluster distribution, and within each cluster the probability of a sample being selected is proportional to the distance to the center of the cluster. Classifiers are evaluated to compute performance-based weights and the prediction is made by aggregating decisions from the ensemble. Experiments were conducted using 8 annotated full-length videos. The cluster-based sampling provides training examples that preserve the innate data distribution with much less number of instances. Our experiments demonstrate that ensemble coupled with cluster-driven sampling achieves superior sensitivity and very competitive specificity. The one way ANOVA analysis reveals that our method greatly outperforms conventional SVM method.

Key Words – Capsule Endoscopy, Classification, Clustering, Video Analysis.