Pruning nearest neighbor cluster trees

Samory Kpotufe, Ulrike von Luxburg

(Submitted on 3 May 2011 (v1), last revised 5 May 2011 (this version, v2))

Nearest neighbor (k-NN) graphs are widely used in machine learning and data mining applications, and our aim is to better understand what they reveal about the cluster structure of the unknown underlying distribution of points. Moreover, is it possible to identify spurious structures that might arise due to sampling variability?

Our first contribution is a statistical analysis that reveals how certain subgraphs of a k-NN graph form a consistent estimator of the cluster tree of the underlying distribution of points. Our second and perhaps most important contribution is the following finite sample guarantee. We carefully work out the tradeoff between aggressive and conservative pruning and are able to guarantee the removal of all spurious cluster structures at all levels of the tree while at the same time guaranteeing the recovery of salient clusters. This is the first such finite sample result in the context of clustering.

Subjects: **Machine Learning (stat.ML)**; Learning (cs.LG)

Cite as: **arXiv:1105.0540 [stat.ML]**

(or **arXiv:1105.0540v2 [stat.ML]** for this version)

Submission history

From: Samory Kpotufe [view email]

[v1] Tue, 3 May 2011 10:34:25 GMT (85kb)

[v2] Thu, 5 May 2011 14:13:49 GMT (85kb)

Which authors of this paper are endorsers?