LOCI: Fast Outlier Detection Using the Local Correlation Integral
Summary

Brandon Sim

October 12, 2014

1 Benefits

1. LOCI provides an automatic, data-dictated cutoff to determine whether a point is an outlier (i.e. no hyperparameters forcing users to pick cut-offs)
2. LOCI is quickly computable (compared to previous best methods) and approximate LOCI is practically linear in time.

2 Intuition

1. Introduce the multi-granularity deviation factor (MDEF)
2. Propose a method which selects a point as an outlier if its MDEF value deviates significantly (more than 3 σ) from local averages

3 MDEF

Let the r-neighborhood of an object p_i be the set of objects within distance r of p_i. Then, intuitively, the MDEF at radius r for a point p_i is the relative deviation of its local neighborhood density from the average local neighborhood density in its r-neighborhood. So, an object with neighborhood density that matches the average local neighborhood density will have MDEF 0; outliers will have MDEFs far from 0.
This is defined formally as

$$MDEF(p_i, r, \alpha) = 1 - \frac{n(p_i, \alpha r)}{\hat{n}(p_i, \alpha, r)} \tag{1}$$

Here, $n(p_i, \alpha r)$ is the number of αr-neighbors of p_i; that is, the number of points $p \in \mathbb{P}$ such that $d(p_i, p) \leq \alpha r$, including p_i itself such that $n(p_i, \alpha r) > 0$ strictly.

Also, $\hat{n}(p_i, \alpha, r)$ is the average of $n(p, \alpha r)$ over the set of r-neighbors of p_i; that is,

$$\hat{n}(p_i, \alpha, r) = \frac{\sum_{p \in \mathcal{N}(p_i, r)} n(p, \alpha r)}{n(p_i, r)} \tag{2}$$

Also, define

$$\sigma_{MDEF}(p_i, r, \alpha) = \frac{\sigma_{\hat{n}}(p_i, r, \alpha)}{\hat{n}(p_i, r, \alpha)} \tag{3}$$

where

$$\sigma_{\hat{n}}(p_i, r, \alpha) = \sqrt{\frac{\sum_{p \in \mathcal{N}(p_i, r)} (n(p, \alpha r) - \hat{n}(p_i, r, \alpha))^2}{n(p_i, r)}} \tag{4}$$

4 LOCI algorithm

For each $p_i \in \mathbb{P}$, compute $MDEF(p_i, r, \alpha)$ and $\sigma_{MDEF}(p_i, r, \alpha)$. If $MDEF > 3\sigma_{MDEF}$, flag p_i as an outlier. If for any $r_{\min} \leq r \leq r_{\max}$ a point p_i is flagged as an outlier via the aforementioned mechanism, then we consider that point to be an outlier.

These cutoffs can be determined on a per-problem basis, but in general we use the following. We set $r_{\max} \approx \alpha^{-1} R_{\mathbb{P}}$ and $r_{\min}$ such that we have $\hat{n}_{\min} = 20$ neighbors.