Data Mining, Lecture 7: Outlier Analysis II

S. Nõmm

[1] Department of Computer Science, Tallinn University of Technology

March 14, 2016

Probabilistic Models

Mixture-based generative model assumes that the data were generated from a mixture of k distributions with the probability distributions $\mathcal{G}_{\infty} \ldots \mathcal{G}_{\|}$ based on the following process:

1. Select a mixture component with prior probability α_i, where $i \in \{1, \ldots, k\}$. Let rth component is selected.
2. Generate data point from $\mathcal{G}_{\rangle}$.

Denote generative model as $\mathcal{M}$

Probabilistic Models

- The probability density function of the data point $\bar{X}_j$ being generated by the model is :

$$f(\bar{X}_j|\mathcal{M}) = \sum_{j=1}^{k} \alpha_i f^i(\bar{X}_j)$$

- For data set $\mathcal{D}$ containing n data points the probability density of the data set being generated by model $\mathcal{M}$ is

$$f(\mathcal{D}|\mathcal{M}) = \prod_{j=1}^{n} f(\bar{X}_i|\mathcal{M})$$

- The log-likelihood fit $\mathcal{L}(\mathcal{D}|\mathcal{M})$ of the data set $\mathcal{D}$ with respect to $\mathcal{M}$ is

$$\mathcal{L}(\mathcal{D}|\mathcal{M}) = \log\left(\prod_{j=1}^{n} f(\bar{X}_j|\mathcal{M})\right) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{k} \alpha_i f^i(\bar{X}_j)\right)$$

Clustering for Outlier Detection

- The detection of outliers as a side-product of clustering methods is, however, not an appropriate approach because clustering algorithms are not optimized for outlier detection.
- A simple way of defining the outlier score of a data point is to first cluster the data set and then use the raw distance of the data point to its closest cluster centroid.
- A simple way of defining the outlier score of a data point is to first cluster the data set and then use the raw distance of the data point to its closest cluster centroid.
- Clustering methods are based on global analysis. Therefore, small, closely related groups of data points will not form their own clusters in most cases.
- The major problem with clustering algorithms is that they are sometimes not able to properly distinguish between a data point that is ambient noise and a data point that is a truly isolated anomaly.

Distance-Based Outlier Detection

- The distance-based outlier score of an object O is its distance to its k-th nearest neighbor.
- The distance-based outlier score of an object O the average distance to the k-nearest neighbors.
- Pruning methods are used only for the case where the top-r ranked outliers need to be returned, and the outlier scores of the remaining data points are irrelevant (can be used only for the binary-decision version).
- Local Distance Correction Methods (Local Outlier Factor (LOF) method).
- Histogram- and Grid-Based Techniques

Kernel Density Estimation

- Kernel density estimation methods are similar to histogram techniques in terms of building density profiles.
- The major difference is that a smoother version of the density profile is constructed (a continuous estimate of the density is generated at a given point).
- The value of the density at a given point is estimated as the sum of the smoothed values of so called kernel functions associated with each point in the data set.
- Each discrete point $\bar{X}_i$ in the data set is replaced by a continuous function $K_h(\cdot)$ that peaks at $\bar{X}_i$ and has a variance determined by the smoothing parameter h. Example: is the Gaussian kernel with width h.

$$K_h(\bar{X} - \bar{X}_i) = \left(\frac{1}{\sqrt{2\pi}h}\right)^d \cdot e^{\dfrac{-||\bar{X} - \bar{X}_i||^2}{(2h^2)}}$$

Kernel Density Estimation

- Each discrete point $\bar{X}_i$ in the data set is replaced by a continuous function $K_h(\cdot)$ that peaks at $\bar{X}_i$ and has a variance determined by the smoothing parameter h. Example: is the Gaussian kernel with width h.

$$K_h(\bar{X} - \bar{X}_i) = \left(\frac{1}{\sqrt{2\pi}h}\right)^d \cdot e^{\dfrac{-||\bar{X} - \bar{X}_i||^2}{(2h^2)}}$$

- The kernel estimation $f(\bar{X})$ based on n data points of dimensionality d, and kernel function $K_h(\cdot)$ is defined as follows:

$$f(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\bar{X} - \bar{X}_i).$$

- The estimation error is defined by the kernel width h, which is chosen in a data-driven manner.

Information-Theoretic Models

- Measure the increase in model size required to describe the data as concisely as possible.
- consider two strings:
  1. ABABABABABABABABABABABABABABABABAB
  2. ABABACABABABABABABABABABABABABABAB

  the first one is the 17 reapeatings of AB but the second one contains one element which makes descriptive model more complicated.
- In general, outliers increase the length of the description in terms of these condensed components to achieve the same level of approximation. For example, a data set with outliers will require a larger number of mixture parameters, clusters, or frequent patterns to achieve the same level of approximation. Therefore, in information-theoretic methods, the components of these summary models are loosely referred to as code books.

Outlier Validity

Receiver Operating Characteristic

- Outlier detection algorithms are typically evaluated with the use of external measures where the known outlier labels from a synthetic data set or the rare class labels from a real data set are used as the ground-truth.
- In outlier detection models, a threshold t is typically used on the outlier score to generate a binary label.
- For any given threshold t on the outlier score, the declared outlier set is denoted by $\mathcal{S}(t)$.
- Let $\mathcal{G}$ is the true set of outliers.
- True positive rate (frequently referred and *recall*) is

$$\frac{|\mathcal{S}(t) \cap \mathcal{G}|}{|\mathcal{G}|}$$

- False positive rate (frequently referred and *recall*) is

$$\frac{|\mathcal{S}(t) \setminus \mathcal{G}|}{|\mathcal{D} \setminus \mathcal{G}|}$$

Outlier Detection with Categorical Data

- Probabilistic Models
- Clustering and Distance-Based Methods
- Binary and Set-Valued Data

Other special cases

- High-Dimensional Outlier Detection
- Outlier Ensembles