Proposal About Rank-by-feature [3]

Yunxin Sun

1 Preprocess

1.1 Objective

Distinguish a column between categorical and numerical

1.2 Methods

Go through all the data items:

1. If any of the data items happens to be NaN string(”object” or ”bool” in pandas), set it to be categorical.

2. Or consider the following factors [2]

- Contain head zero
- unique value makes up less than 30 percent
- All strings are equal-length.

If any of the two above conditions hold true, set the column to be categorical, otherwise to be numerical.

2 Feature Ranking

2.1 Categorical

Compute variance for frequency of a dimension [5]
The less the variance is, the better. If the variances are equal, then the bigger the unique value, the higher the score.
Some non-legal dimensions exist, refer to readme for details.

2.2 Cluster

Ranking Criterion:
Conditional Entropy [1]

2.3 Outlier

Ranking Criterion:

2.3.1 LOF

LOF [4]

$$reachability - distance_k(A, B) = max\{k - distance(B), d(A, B)\}$$

$$lrd_k(A) = 1/(\frac{\sum_{B \in N_k(A)} reachability - distance_k(A, B)}{|N_k(A)|})$$

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} lrd(B)}{|N_k(A)|} / lrd(A)$$

Choose of k:
$k = 20$ if number of data samples is greater than 20 else half of number of data samples

2.3.2 Number of Outliers

An item of value d is considered as a outlier if $d > (Q3+1.5*IQR)$ or $d < (Q1-1.5*IQR)$.[3]

2.4 Association

Ranking Criterion:
Pearson's correlation coefficient(r) of two certain columns

$$r(s) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

References

[1] Guo D, Gahegan M, Peuquet D, et al. Breaking Down Dimensionality: Effective and Efficient Feature Selection for High-Dimensional Clustering[C]//Workshop on Clustering High-Dimensional Data. 2003.

[2] https://datascience.stackexchange.com/questions/9892/how-can-i-dynamically-distinguish-between-categorical-data-and-numerical-data

[3] Seo J, Shneiderman B. A rank-by-feature framework for interactive exploration of multidimensional data[J]. Information visualization, 2005, 4(2): 96-113.

[4] Breunig M M, Kriegel H P, Ng R T, et al. LOF: identifying density-based local outliers[C] ACM sigmod record. ACM, 2000, 29(2): 93-104.

[5] Filippova D, Shneiderman B. Interactive exploration of multivariate categorical data: exploiting ranking criteria to reveal patterns and outliers[J]. Human-Computer Interaction Lab, University of Maryland, Technical Report# HCIL-2009-38, 2009.