Contextual Spatial Outlier Detection with Metric Learning

Guanjie Zheng[1], Susan L. Brantley[2], Thomas Lauvaux[3], Zhenhui (Jessie) Li[1]

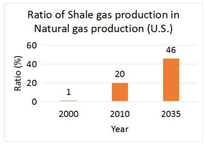
College of Information Sciences and Technology[1], Department of Geosciences[2], Department of Meteorology and Atmospheric Science[3]

Pennsylvania State University

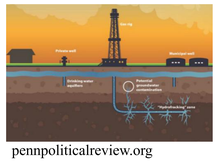
gjz5038@ist.psu.edu, jessieli@ist.psu.edu

INTRODUCTION

Improvements in high volume hydraulic fracturing, i.e., "fracking", which allows the development of shale gas, have changed the energy landscape. However, fracking can potentially contaminate environment. The biggest concern is the escape of methane into surface and ground waters. Collaborating with geoscientists, we aim at detecting the outlier water samples with high methane values, which could indicate gas leakage.



Data from [1]



pennpoliticalreview.org

Why not just measure methane? Samples with extreme methane concentration are often trivial outliers that are already known (i.e., due to a serious well leakage). We are more interested in detecting non-trivial, local outliers for unknown leakage.

A straightforward solution: Build a regression model using features to predict methane concentration. Use the prediction error as outlier score.

- Methane = f (distance to gas well) + g (land type) + h (time) + ...
- But, the contextual features may not be informative enough to learn a reliable regression model. Many determining factors are either unknown (e.g., underground geology) or not well documented (e.g., anthropogenic activities like coal mining, industrial waste, and old residential houses).

Our Solution: contextual spatial outlier detection, i.e., compare a sample with its contextual neighbors

PROBLEM DEFINITION

Problem 1 (Contextual Outlier Detection) *Given data set* $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n\}$*, where each data point* $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ *is composed of a contextual attribute vector* $\mathbf{x}_i \in \mathbb{R}^d$ *(including spatial coordinates) and a behavioral attribute value* $y_i \in \mathbb{R}$*. We wish to assign an outlier score* $S_i \in [0, 1]$ *to each data point* $\mathbf{z}_i$*, using* $\mathbf{x}_i$ *as contextual attributes and* y_i *as the behavioral attribute. Higher score indicates that this sample has a higher probability to be an outlier.*

Water Example: Outliers in groundwater with potential leakage

- Behavioral attribute: methane concentration in groundwater
- Context: sample location, distance to gas wells, distance to fault, …

House Example: Outliers in sold houses

- Behavioral attribute: house sold price
- Context: house location, square feet, year of built, #bedroom …

References

[1] Stevens. 2012. The 'shale gas revolution': Developments and changes. Chatham House (2012), 2–3.
[2] Weinberger and Tesauro. 2007. Metric Learning for Kernel Regression. In AISTATS. 612–619.
[3] Llewellyn et al. 2015. Evaluating a groundwater supply contamination incident attributed to Marcellus Shale gas development. PNAS112, 20 (2015), 6325–6330.
[4] Tobler. 1970. A computer movie simulating urban growth in the Detroit region. Economic geography 46, sup1 (1970), 234–240.

Acknowledgement. The work was funded by General Electric Fund for the Center for Collaborative Research on Intelligent Natural Gas Supply Systems and was supported in part by NSF awards #1639150, #1618448, #1652525, and #1544455. This work has also been funded by the U.S. Department of Energy National Energy Technology Laboratory (project DE-744FE0013590).

METHODS (MEtric Learning Outlier Detection Y)

Local Model: kNN Kernel Regression

Based on "near things are more related" [4] assumption, we apply **kNN kernel regression** to estimate the behavioral attribute value of one sample using its contextual neighbors [2].

$$\hat{y}_i = \frac{\sum_{j \in N_i} w_{ij} y_j}{\sum_{j \in N_i} w_{ij}} \qquad w_{ij} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right)$$

However, Euclidean distance in kNN regression will assign equal weights on different features.

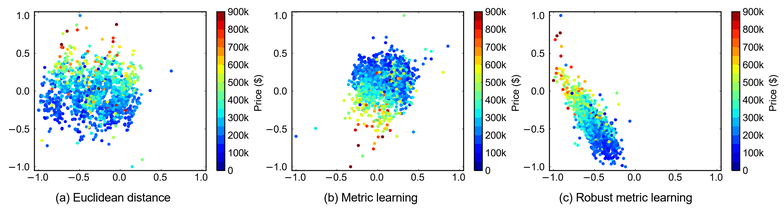
How can we combine distances from different features? (1) Features carry different meanings, e.g., when calculating the distance between two houses, how to combine spatial distance and the time difference in year of built? (2) Features have different scales, e.g., year of built has difference within 100, but square feet could have difference in the scale of 1000. (3) Some features may be not relevant.

Metric Learning for Contextual Neighborhood Discovery

Our Solution: We use Mahalanobis distance [2], which can learn the weight on different features automatically.

$$d_{ij}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}}^2 = \|(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)\|$$

The goal of metric learning for kernel regression is to learn the distance metric so that the regression error can be minimized.



(a) Euclidean distance (b) Metric learning (c) Robust metric learning

From left to right, distance metric is better modeled. Houses with different sold prices are more separated from each other, and houses with similar sold prices are closer to each other.

Outlier Score with Local Confidence

(Global) Outlier Score = difference between predicted value and actual value

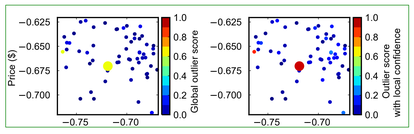
$$S_i^G = |y_i - \hat{y}_i|$$

Sample A in a neighborhood with non-outliers vs. Sample B in a neighborhood with all outliers

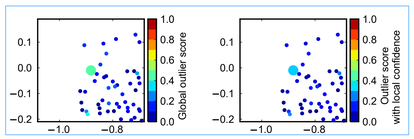
→ Sample A is more likely to be an outlier

Our Solution: Outlier score with local confidence to address the heterogeneity inside regions.

$$C_i = \frac{1}{\sum_{j \in N_i} S_j^G} \qquad S_i^L = S_i^G \times C_i = \frac{S_i^G}{\sum_{j \in N_i} S_j^G}$$



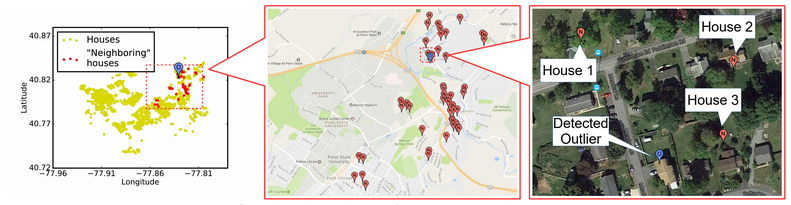
House 1 Sold price: $715K
Global outlier score: 0.63 (high)
Neighbors: low global outlier scores → homogeneous neighborhood
Local outlier score: 0.93 (high) → outlier



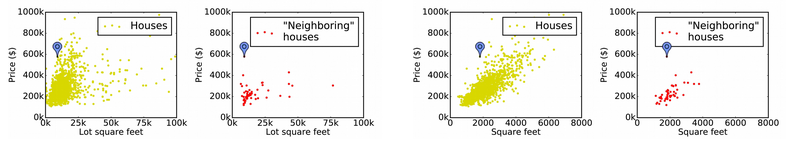
House 2 Sold price: $949K
Global outlier score: 0.45 (high)
Neighbors: high global outlier scores → heterogeneous neighborhood (house prices range from $15K to $90K)
Local outlier score: 0.32 (low) → non outlier

EXPERIMENTS

Case Study on Zillow House Dataset



(a) The detected outlier and "Neighboring" houses on the map



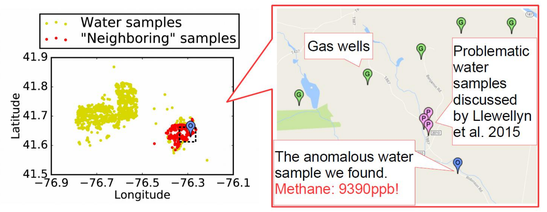
(b) Sold price w.r.t. Lot square feet (c) Sold price w.r.t. Square feet

House	Square feet	# Bedrooms	# Bathrooms	Lot square feet	Year built	Year sold	Price ($)	Zestimate ($)	Address
Detected outlier	1,801	5	2.5	9,148	1955	2016	580,000	234,056	120 Meadow Ln State College, PA 16801
House 1	1,778	3	2	8,712	1943	2016	120,000	192,430	1707 Puddintown Rd State College, PA 16801
House 2	1,840	4	2.5	11,326	1950	2014	181,000	222,883	1800 Puddintown Rd State College, PA 16801
House 3	2,818	5	3	11,761	1952	2014	215,000	329,950	310 Bottorf Dr State College, PA 16801

(d) Three houses that are very similar to the detected outlier

The detected outlier is similar to its neighbors in contextual attributes such as spatial location, square feet and lot square feet, but its sold price is much higher. In addition, Zestimate value provided by Zillow also suggests that this is an outlier ($234,056 Zestimate vs. $580,000 sold price).

Case Study on Water Dataset



(a) Outlier and its "Neighbors" (b) The outlier we detected and three known problematic water samples

Methane value of the detected outlier (blue balloon) ranks 66/1645 (top 4.0%) globally, and ranks 8/300 (top 2.6%) in neighbors (red dots). The detected outlier is only 1km downstream from a site where we know that methane leaked into three homes (pink balloons) [3]. This outlier location could also be affected by the upstream gas wells.

Gas well Outlier we detected Discovered problem in reference

Quantitative Evaluation

Methods \ Datasets	Zillow	Water	Air	El Nino	Hydro
LOF	0.159	0.071	0.024	0.616	0.422
CAD	0.354	0.110	**0.244**	0.439	0.146
ROCOD.CART	0.422	0.121	0.208	0.595	0.769
ROCOD.RIDGE	0.403	0.118	0.104	0.333	0.611
LR	0.389	0.114	0.057	0.285	0.770
XGBOOST	**0.477**	0.119	0.083	**0.780**	**0.935**
ZS	0.206	0.122	0.219	0.622	0.234
SOD	0.167	0.054	0.034	0.292	0.487
GLS-SOD	0.188	**0.142**	0.208	0.619	0.254
MELODY	**0.687 (+44%)**	**0.182 (+28%)**	**0.716 (+193%)**	**0.970 (+24%)**	**0.965 (+3.2%)**

AUC comparison of different algorithms

Learn more about our work: https://faculty.ist.psu.edu/jessieli/