

Outlier Detection Methods for mixed-type and large-scale data like Census

Frantisek Hajnovic Data Scientist, Big Data Team frantisek.hajnovic@ons.gov.uk

Alessandra Sozzi Data Scientist, Big Data Team <u>alessandra.sozzi@ons.gov.uk</u>

Content

- Background
- Challenges
- Algorithms
- Wrap up

Background

Background

- Outlier detection (OD) refers to the problem of finding patterns in data that do not conform to expected <u>normal behaviour</u>.
- Outliers in census have in the past arose due to various reasons
- A mechanism for pointing in the right direction will: save time, improve quality and minimise the risk of serious errors identifying them earlier.

Errors in the questionnaire by the respondent

Scanning errors

Coding errors

Imputation errors

Unknown unknowns: errors we don't know yet (new online Census)

> Purely normal anomalies, worth flagging to ensure it's no error

Errors introduced at later steps of the Census processing



Example n.1



NAME: Paul AGE: 92 yo Working Hrs/Week: 40

Example n.2



NAME: Larry AGE: 2 yo JOB: Security guard



Example n.1

NAME: Paul AGE: 92-yo 20 yo (1919 --> 1991) Working Hrs/Week: 40

Example n.2



NAME: Larry AGE: 2 yo JOB: Security guard



Example n.1

NAME: Paul AGE: **92-yo** 20 yo (1919 --> 1991) Working Hrs/Week: 40

Example n.2



https://www.channel4.com/news/census-onsoffice-for-national-statistics-pets-sex-race

> Icons made by <u>Freepik</u> from <u>www.flaticon.com</u> is licensed by <u>CC 3.0 BY</u> Icons made by <u>surang</u> from <u>www.flaticon.com</u> is licensed by <u>CC 3.0 BY</u>

No Code Required questions <u>categorical space</u>

Scale Challenges

High-dimensional input data

Missing values

Skewed distributions

Challenges

The scale and nature of such data pose computational challenges to traditional OD methods.



SCALE: census is too large for a sequential execution. Most of the methods need either a distributed implementation or separate runs of the algorithm on chunks of the dataset.



MISSING VALUES: missing values tend also not to be very frequent in census and thus records containing missing values can end up flagged as outliers.

SKEWED DISTRIBUTIONS: e.g. 90% of entries



MIXED TYPE: census questions are of mixed type (numeric, categorical, ordinal, etc.) and detecting outliers in this multi-dimensional space is an open area of research.

Question 1				
Yes	No			
> 50 yes	< 50 yes			
Question 2	End Survey			

DEPENDENCIES: some variable's value depending on other variable. E.g. "No code required" if the person is not old enough.

HIGH-DIM INPUT: when the dimensionality increases, the volume of the space increases so fast that the data

Frequent Itemsets Clustering **Probability distributions Classifications** Algorithms **Decision trees** Regressors **Nearest neighbours Pattern Analysis Distributed implementations Python/Scala** Spark/PySpark

KAMILA

Iterative clustering for mixed-type data

 Integrates K-Means + Gaussian Multinomial Mixture models to balance the effect of numeric and categorical features without specifying weights



Iterative clustering method on a mixed-type dataset that equitably balances the contribution of continuous and categorical variables. KAMILA integrates two different kinds of clustering algorithms:

- · K-means algorithm: no strong parametric assumptions
- Gaussian multinomial mixture models: KAMILA uses the properties of Gaussian-multinomial mixture models to balance the effect of numeric and categorical features without specifying weights.



ODMAD

- Makes use of the concept of Frequent Itemsets for the categorical space and cosine distance to the means by category in the numerical space
- Scale well vertically (as data is traversed only once for the categorical space and twice for the numerical space) but can be a challenge to scale it horizontally when the number of columns/categories increase
- Output two scores



Based on the idea of **frequent itemsets mining**. It computes two scores for each row:

- **Categorical:** a point is likely to be an outlier if it contains single values that are infrequent or sets of values that are infrequent
- Continuous: cosine similarity between the point and the mean of the points sharing the same categorical value Can handle large dataset well, as data is traversed only once, but can fail as number of variables (and categories) increase.



SPAD

- Suitable for categorical data (numerical needs to be binned/discretised)
- Score = log-probability of the record
- Assumes independence across multiple variables
- Possible extension with random combinations of columns (tuples, triples)
- Fast and results easily interpretable



SPAD: a Simple Probabilistic Anomaly Detection. It works only with categorical data, continues data can be binned/discretised as categorical variables. A score for an example is related to the **log-probability of the record** (the lower, the higher OD score). Assumes the attributes are independent of each other, so cannot detect outliers across multiple variables. An extension addressing this drawback considers random subsets of variables.



iFOREST

- Works on numeric data, categorical columns needs to be exploded
- Can struggle with high-dimensional inputs
- Based on the idea of isolation: Samples of the dataset are recursively split on a random chosen variable. This creates sort of binary/ decision trees for each sample —> more anomalous records are more likely to be isolated earlier in the process (have shorter paths to the root)



iForest (isolation forest) works on numeric data. Categorical data can be converted in numeric format using e.g. dummy variables. Outliers are detected based on the idea of **isolation**. Samples of the dataset are recursively split on a randomly chosen variable (at a random point in the variable's range). This creates (for each sample) a sort of binary/decision tree. The dataset is then run through all such created trees. More anomalous records are more likely to be isolated earlier in the process (have shorter paths from the root).



POD

- Uses a subset of variables (Xs) to predict another variable (Y)
- Predictions are then used to calculate an Outlier score (how different is the predicted score from the original value)
- The process is repeated for different combinations of Xs and Y: finally all scores are used to calculate a final one (can be intensive)
- Any algorithm can be plugged
- Offer lots of insights —> you can drill down to find why the record is an outlier



Pattern based Outlier Detection (POD) is a technique where group of variables (**Xs**) are used to predict another variable (**Y**). The prediction is then compared with the actual value and an OD score is computed for the given variable. The process is repeated for different combination of Xs and Y and single OD scores are combined into a final one. Any algorithm can be used to predict the Y and scalability, how mixed type and high-dim data are handled depend on that choice (mainly SVM, logistic regression or random forests).



SNLA_SPR

- Based on the idea of K-nearest neighbours
- Euclidean distance and hamming distance (but any can be used)
- The complexity is greatly reduced via a pruning technique
- Return only a pre-fixed number of outliers



SNLA_SPR is based on the idea of **K-nearest neighbours** for continuous variables. Categorical data can be converted in numeric format using e.g. dummy variables. It uses two nested loops, however, the complexity is greatly reduced via a **pruning technique**: it keeps track of closest neighbours found so far for an example and, once an exclusion condition gets satisfied, it stops computing additional neighbours and remove this example from future computations as it cannot be an outlier. Return only a prefixed number of outliers.



Summary

	KAMILA	ODMAD	SPAD	iFOREST	POD	SNLA_SPR
SCALABILITY	0000	00000	00000	00000	0000	00000
MIXED TYPE	ୢ୶ୖୢ୶ୢୄ୶ୖୢୢ୶ଡ଼ୢୢ୕ୢ୵ଡ଼ୢୢୖ୵ଡ଼ୢୖ	ୢ୶ୖୢ୶ୢ୵ୖ _୵ ଡ଼ୢୄ୵ଡ଼ୢୖୄ୵ଡ଼	ୢୢ୶ୖୢ୵୶ୢ୵ଡ଼ୢୖ	ୢ୶ୢୢୖ୶ୢୢ୶ୖ	_x ଟ୍ଟ' _x ଟ୍ଟ' _x ଟ୍ଟ	_୳ ଟ୕ୢୢ୳ଡ଼୕ୢ ^୲ ୵ଟ
HIGH - DIM INPUT	***	•••	••••		6666	****
OD SCORE	~	**	Å	*	*	~

Wrap up

Wrap up

- Ongoing project
- Potential not just for Census!
- Currently testing methods on raw Census 2011 data
- Plan to test the methods in the Census rehearsal —> end of this year
- Plan to publish algorithms on <u>GitHub</u>
- Promising results (found the obvious) but no found, yet!

References

- **iFOREST:** L.F. Tony, T.K. Ming and Z. Zhi-Hua, Isolation-based anomaly detection, ACM Transactions on Knowledge Discovery from Data (TKDD) 6.1 (2012), 3.
- **SNLA_SPR:** S.D. Bay and M.A. Schwabacher, Mining distance-based outliers in near linear time with randomization and a simple pruning rule, ACM SIGKDD international conference on Knowledge discovery and data mining (2003).
- **ODMAD:** A. Koufakou and M. Georgiopoulos, A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes, Data Mining and Knowledge Discovery (2010), 20: p. 259–289.
- SPAD: S. Aryal, K.M. Ting and G. Haffari, Revisiting Attribute Independence Assumption in Probabilistic Unsupervised Anomaly Detection, PAISI (2016).
- POD: M.S. Mausam, R. Bart, S. Soderland and O. Etzioni, Open language learning for information extraction, Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2012), p. 523-534.
- KAMILA: A. Foss, M. Markatou, B. Ray and A. Heching, A semiparametric method for clustering mixed data, Machine Learning (2016).

Thank you!