Algorithms

Alex,Jung,Muhamed,Pallavi

Anomaly detection by Classification

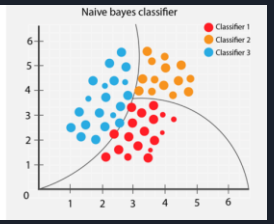
Finding anomalous category of data using data classification.

- Rule Based - if/then -> Rule tree - conditions(T/F)
- Bayesian - calculates statistical probability

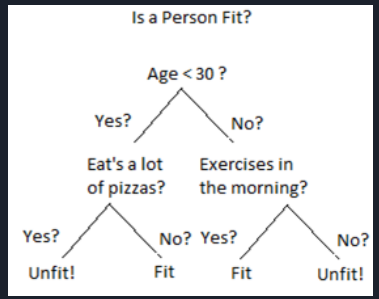
$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

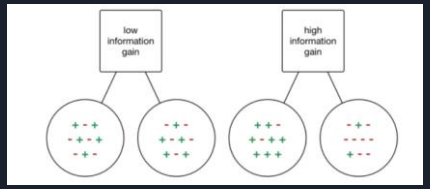


- Decision Tree



Decision Tree

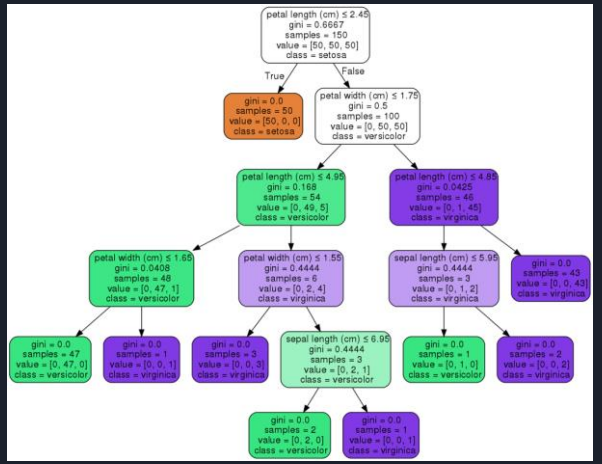
- Induction
  - Creating the tree based on the data.
    - Using greedy algorithm, select best feature to split the tree.



- Pruning
  - Removing the unnecessary branches.
    - Use greedy algorithm to evaluate each rules by removing one, and select the most effective ones.

- Pros
  - Easy to understand.
  - Requires very little data preparation.
  - Cost of creating a tree is log(N) which is little.
- Cons
  - Overfitting is common
  - If training dataset is biased, the result will be biased too. Needs class balancing to mitigate the effect.
  - Suitable to machine created logs. Low performance for Natural Language.

- Iris dataset

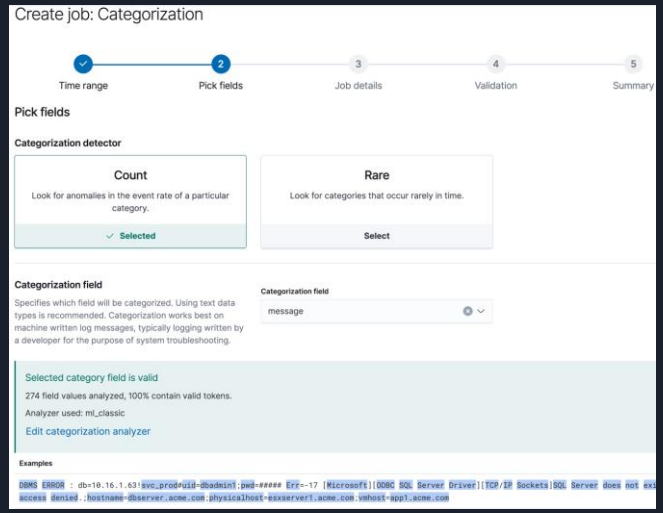


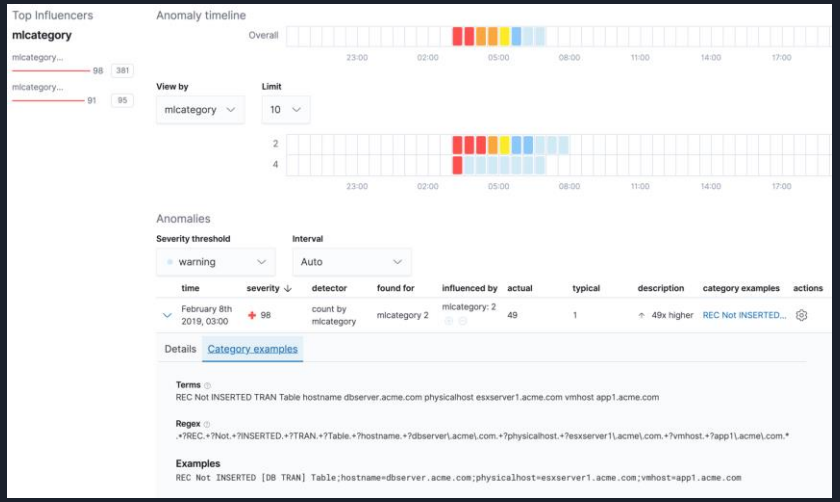
Anomaly detection by Categorization

- In Kibana, there is a categorization wizard to help you create this type of anomaly detection job.

Using the composite decision tree algorithm called **Boosted tree regression** model

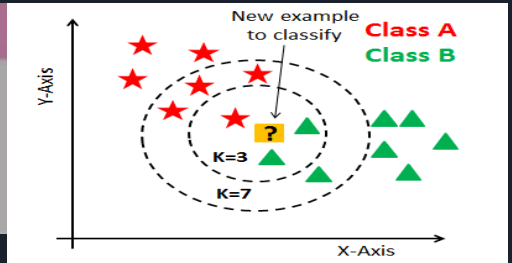
- Analyzes the content in the message field to find out the machine generated log's category.
- Can customize the categorization_analyzer to have better results.

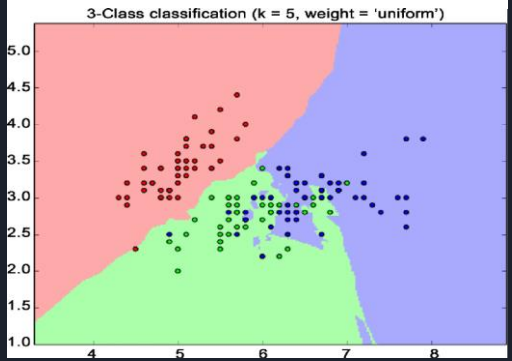




Distance of K-nearest neighbors (KNN)

Algorithm	Pros	Cons
K-NEAREST NEIGHBOR: K-NN	• VERY EASY TO UNDERSTAND • GOOD FOR CREATING MODELS THAT INCLUDE NONSTANDARD DATA TYPES SUCH AS TEXT	• LARGE STORAGE REQUIREMENTS • COMPUTATIONALLY-EXPENSIVE • SENSITIVE TO THE CHOICE OF THE SIMILARITY FUNCTION FOR COMPARING INSTANCES





- Density-based anomaly detection methods
- Continuous data uses Euclidean distance
- Discrete data uses Hamming distance
- Creates models for non-standard data
- It's a supervised algorithm
- Great for fraud detection
- Business and Finance
- Little to no training phase
- Benchmark for other algorithms(Support Vehicle Machine or Artificial Neural Network)
- Good only for prediction and Quick calculation time

Local Outlier Factor

- Density based method
- Based on this approach, a metric is computed called local outlier factor (lof) for each data point.
  - ^ LOF = data point is more outlying
- Computed by taking the average ratio of the local reachability density of a point and its k-nearest neighbors.
- Contamination : proportion of the most isolated points (points that have the highest LOF scores) to be predicted as anomalies.

Pros:

- Well-known and good algorithm for local anomaly detection

Cons:

- Only relies on its direct neighborhood
- Performs poorly on datasets with global anomalies

Osti.gov: Research by IEEE members: an event detection LOF is proposed to detect abnormal events in power systems.

Local Outlier Factor

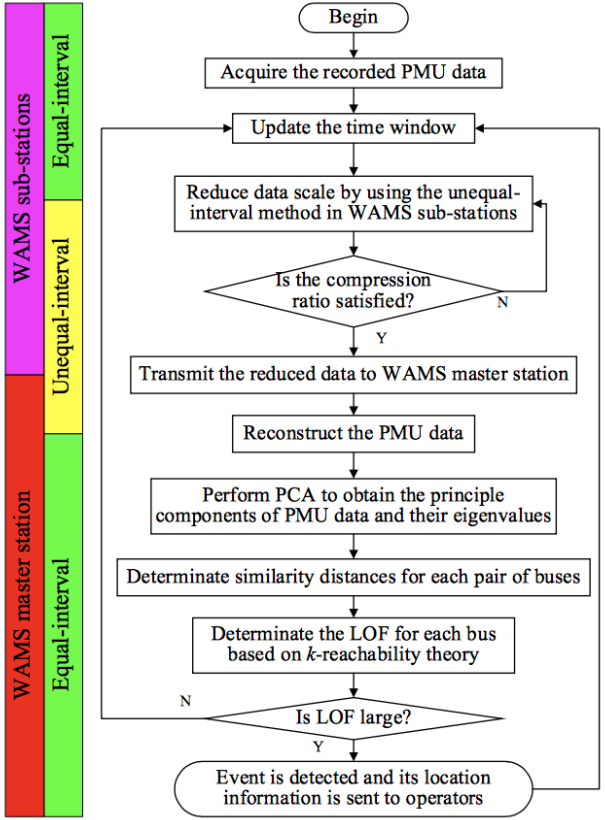


Fig. 5. Flow chart of the proposed SS-LOF algorithm.

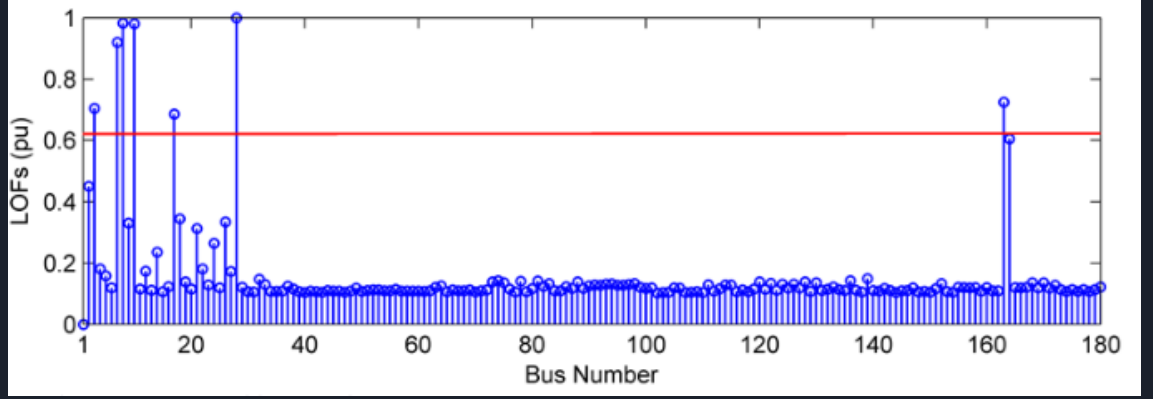
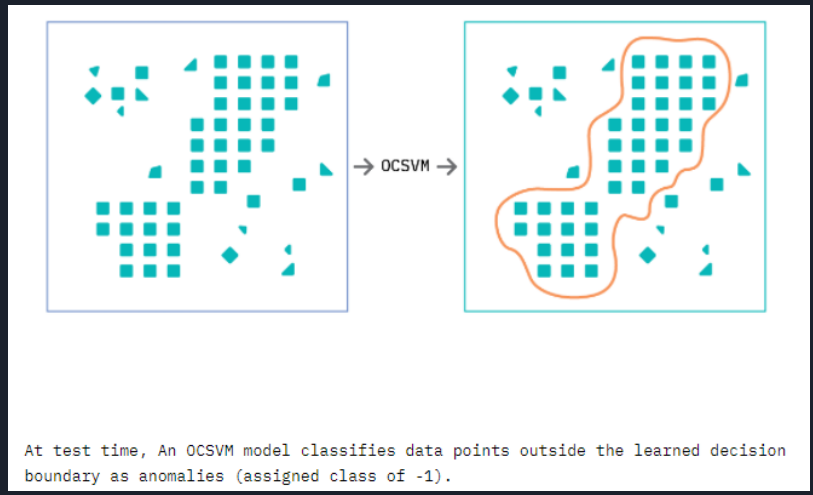


TABLE VIII
COMPARISONS OF THE PROPOSED SS-LOF ALGORITHM WITH THE OTHERS

Functions	VMD [14]	MW-PCA [16]	RQA [21]	MVEE [23]	ST-LOP [35]	Proposed SS-LOF
Event detection	✓	✓	✓	✓	✓	✓
Multiple events		✓	✓	✓		✓
Event location			✓		✓	✓

Support Vector Machine

- Classification via hyperplane
- Finding the optimal hyperplane
- Supervised or unsupervised
- One-class SVM



At test time, An OCSVM model classifies data points outside the learned decision boundary as anomalies (assigned class of -1).

Table 4: Comparing the AUC of SVM based algorithms against other anomaly detection algorithms

Dataset	One-class	Robust one-class	Eta one-class	k-NN	LOF	COF	INFLO	LoOP	Histogram	CBLOF	u-CBLOF	LDCOF
ionosphere	0.9878	0.9956	**0.9972**	0.9933	0.9178	0.9406	0.9406	0.9211	0.7489	0.3183	0.9822	0.9306
shuttle	0.9936	0.9597	**0.9941**	0.9208	0.6072	0.5612	0.5303	0.5655	0.9889	0.8700	0.8739	0.5312
breast-cancer	0.9843	0.9734	0.9833	0.9826	0.9916	0.9888	**0.9922**	0.9882	0.9829	0.8389	0.9743	0.9804
satellite	0.8602	0.8861	0.8544	**0.9003**	0.8964	0.8708	0.8592	0.8664	0.8862	0.4105	0.9002	0.8657

Autoencoder

	False Positive Rate(%)	Detection Accuracy(%)
k-NN	38.02	88.91
SVM	6.91	92.98
TANN	3.83	96.91
AE	4.09	95.85
CAE	3.44	96.87

- Deep Learning
- Dimensionality reduction
- Reconstruction error as AS
- Noise removal
- Importance of training