Anomaly Detection in Probabilistic Databases with Embeddings of Categorical Values

Daoping Wang

March 20, 2020

Technische Universität München
Mercateo Deutschland AG

Table of contents

Data as the New Oil

- Enabler for advanced analytics
- Key resource for ML-based techniques
- Requirement on data quantity and **quality**
  - Accuracy, consistency and completeness
  - Ground truth information

[1]

[1] https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Data as the New Oil

- Enabler for advanced analytics
- Key resource for ML-based techniques
- Requirement on data quantity and **quality**
  - Accuracy, consistency and completeness
  - Ground truth information

[1]

How does real-world data look like?

[1] https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Mercateo Article Data

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount	ean
102-805270-BP	1039U	9739011201	Knipex	Eckschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1528050	Knipex	Wapuzange	C62	14.82	4003773022022
102-805270-BP	108EL	158-5451	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

Mercateo Article Data

Outlier (?)

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount	
102-805270-BP	1039U	9739011201	Knipex	Eckschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1539050	Knipex	Kneifzange	C62	14.82	4003773022022
102-805270-BP	108EL	1539050	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

Missing values

Inconsistency

Conflicts

Mercateo Article Data

Outlier (?)

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount	
102-805270-BP	1039U	9739011201	Knipex	Eckschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1539059	Knipex	Kneifzange	C62	14.82	4003773022022
102-805270-BP	108EL	1539059	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

Missing values

Inconsistency

Conflicts

Goal: Automated anomaly detection!

Conventional Techniques and Limitations

- Heuristic methods — *if ... then ...*
  - Maintainability
  - Conflicting rules
  - "Knipex" vs "Primium": Who is wrong?
  - High demands on domain knowledge
- Statistical methods
  - **High-cardinality**: Most values are rarely observed
  - Occurrence information alone is not enough

Conventional Techniques and Limitations

- Heuristic methods — *if ... then ...*
  - Maintainability
  - Conflicting rules
  - "Knipex" vs "Primium": Who is wrong?
  - High demands on domain knowledge
- Statistical methods
  - **High-cardinality**: Most values are rarely observed
  - Occurrence information alone is not enough

How does a human perform anomaly detection?

Domain Knowledge as Probabilistic Relationships

Outlier (?)

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount	ean
102-805270-BP	1039U	9739011201	Knipex	Eckschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1528050	Knipex	Wapuzange	C62	14.82	4003773022022
102-805270-BP	108EL	158-5451	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

Domain Knowledge as Probabilistic Relationships

Outlier (?)

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount	ean
102-805270-BP	1039U	9739011201	Knipex	Eckschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1528050	Knipex	Wapuzange	C62	14.82	4003773022022
102-805270-BP	108EL	158-5451	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

Domain knowledge:

- $P(ek\_amount \mid keyword, manufacturer) \sim$ "Similar products have similar prices."
- $P(ek\_amount \mid set\_id) \sim$ "Articles from the same set have similar prices."
- $P(ek\_amount \mid manufacturer) \sim$ "Knipex is cheaper than Wera."
- ...

Domain Knowledge as Probabilistic Relationships

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount	ean
102-805270-BP	10390	9739011201	Knipex	Eckschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1528050	Knipex	Wapuzange	C62	14.82	4003773022022
102-805270-BP	108EL	158-5451	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

Accumulating domain knowledge:

- Modeling joint probability distribution

$$P(\mathit{set\_id}, \mathit{catalog\_id}, \mathit{manufacturer}, \mathit{keyword}, \mathit{ek\_amount})$$

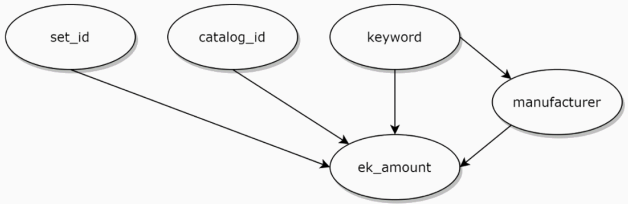
- Marginal probability = Confidence score

Bayesian Network

set_id	$P(set\_id)$
102-805270-BP	0.5
102-822855-BP	0.5

catalog_id	$P(catalog\_id)$
1039U	0.1
1082	0.1
...	...

keyword	$P(keyword)$
Wapuzange	0.08
Ratschenkabelschere	0.06
...	...



manufacturer	*keyword*	$P(manufacturer \mid keyword)$
Unbekannt	Ratschenkabelschere	0.1
Primium	Ratschenkabelschere	0.4
...	...	...

set_id	*catalog_id*	*manufacturer*	*keyword*	*ek_amount*	$P(ek\_amount \mid set\_id, catalog\_id, manufacturer, keyword)$
102-822855-BP	4445	Unbekannt	Ratschenkabelschere	152.63	0.3
102-822855-BP	986	Primium	Ratschenkabelschere	141.69	0.4
...	...	...	...	...	...

$$P(set\_id, ..., ek\_amount) = P(set\_id) \cdots P(ek\_amount \mid set\_id, ..., manufacturer)$$

- Directed acyclic graphical model
- Represents a set of **variables** and their **conditional dependencies**

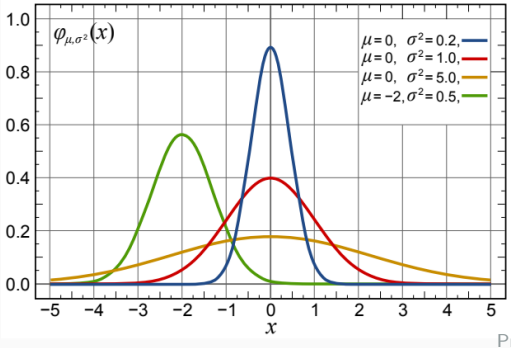
Conditional Continuous Distribution

Domain knowledge:

- "Similar products have similar prices."
- "Prices might be normally distributed."

Normal Distribution ($\mu \in \mathbb{R}$, $\sigma > 0$)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Conditional Continuous Distribution

Domain knowledge:

- "Similar products have similar prices."
- "Prices might be normally distributed."

Normal Distribution ($\mu \in \mathbb{R}$, $\sigma > 0$)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Conditional probability distribution:

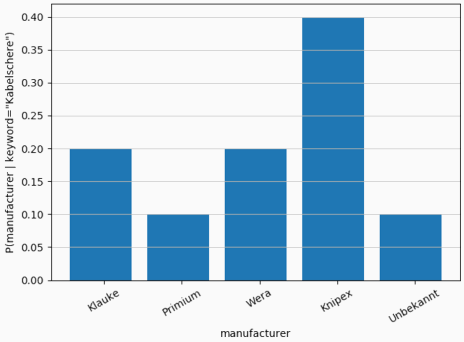
$$P(\textit{ek\_amount} \mid \textit{keyword} = \text{"Wapuzange"}, \textit{manufacturer} = \text{"Knipex"}) = \mathcal{N}(\textit{ek\_amount} \mid \mu, \sigma^2),$$

where

μ ... Mean price of " Knipex Wapuzange" articles
σ^2 ... Price variance of " Knipex Wapuzange" articles

Domain knowledge:

- "Difference manufacturers make different products."



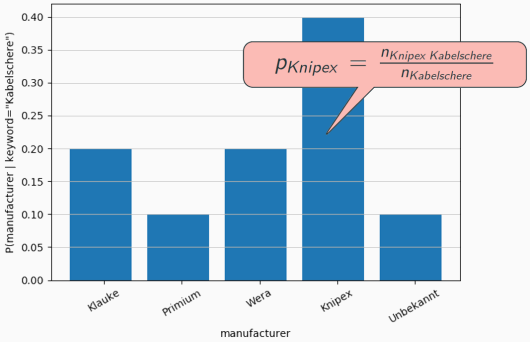
Categorical Distribution

$$P(X = x_i) = p_i,\ i \in \{1, ..., k\}$$

$k > 0$ number of events
$p_1, ..., p_k$ event probabilities
$(p_i > 0, \sum p_i = 1)$

Domain knowledge:

- "Difference manufacturers make different products."



Categorical Distribution

$$P(X = x_i) = p_i,\ i \in \{1, ..., k\}$$

$k > 0$ number of events
$p_1, ..., p_k$ event probabilities
$(p_i > 0, \sum p_i = 1)$

Conditional Discrete Distribution

Domain knowledge:

- "Difference manufacturers make different products."

Conditional probability distribution:

$$P(\textit{manufacturer} \mid \textit{keyword} = \text{"Kabelschere"}) = \textit{categorical}(\mathbf{p}),$$

where

$\mathbf{p} = \{p_1, ..., p_k\}$ ... Probabilities of manufacturers given *keyword* = "Kabelschere"
$\Rightarrow$ Calculated from occurrence statistics

Probabilistic Programming:

- Paradigm for specifying probabilistic models

Our implementation:

- A sampler script to resemble the Bayesian network
- Simulating the data generation process

Model Implementation with Probabilistic Programming

Outlier (?)

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount	ean
102-805270-BP	1039U	9739011201	Knipex	Eckschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1528050	Knipex	Wapuzange	C62	14.82	4003773022022
102-805270-BP	108EL	158-5451	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

```
@gen function mercateo_data_model(observation)
    # No prior knowledge on root nodes => Assume uniform distribution
    set_id = @trace(choose_uniformly(observation.set_id), :set_id)
    catalog_id = @trace(choose_uniformly(observation.catalog_id), :catalog_id)
    keyword = @trace(choose_uniformly(observation.keyword), :keyword)

    # Given keyword, sample manufacturer based on occurrence statistics
    manufacturer = @trace(categorical(observation.manufacturer[keyword]), :manufacturer)

    # Given sampled values above, sample ek_amount based on observed mean and std
    ek_mu = mean(observation.ek_amount[set_id, catalog_id, keyword, manufacturer])
    ek_sigma = std(observation.ek_amount[set_id, catalog_id, keyword, manufacturer])
    ek_amount = @trace(normal(ek_mu, ek_sigma), :ek_amount)

    return set_id, catalog_id, keyword, manufacturer, ek_amount
end
```

article_id	$\log(P(set\_id, ..., ek\_amount))$
9739011201	-32.7
1528050	-19.7
158-5451	-19.2
201096-YY	-30.2
K-95 31 250	-23.1
0184472	-27.2

Model Implementation with Probabilistic Programming

Outlier (?)

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount	ean
102-805270-BP	1039U	9739011201	Knipex	Eckschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1528050	Knipex	Wapuzange	C62	14.82	4003773022022
102-805270-BP	108EL	158-5451	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

```
@gen function mercateo_data_model(observation)
    # No prior knowledge on root nodes => Assume uniform distribution
    set_id = @trace(choose_uniformly(observation.set_id), :set_id)
    catalog_id = @trace(choose_uniformly(observation.catalog_id), :catalog_id)
    keyword = @trace(choose_uniformly(observation.keyword), :keyword)

    # Given keyword, sample manufacturer based on occurrence statistics
    manufacturer = @trace(categorical(observation.manufacturer[keyword]), :manufacturer)

    # Given sampled values above, sample ek_amount based on observed mean and std
    ek_mu = mean(observation.ek_amount[set_id, catalog_id, keyword, manufacturer])
    ek_sigma = std(observation.ek_amount[set_id, catalog_id, keyword, manufacturer])
    ek_amount = @trace(normal(ek_mu, ek_sigma), :ek_amount)

    return set_id, catalog_id, keyword, manufacturer, ek_amount
end
```

article_id	$\log(P(set\_id, ..., ek\_amount))$
9739011201	-32.7
1528050	-19.7
158-5451	-19.2
201096-YY	-30.2
K-95 31 250	-23.1
0184472	-27.2

Problem of High-cardinality

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount	ean
102-805270-BP	1039U	9739011201	Knipex	Eckschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1528050	Knipex	Wapuzange	C62	14.82	4003773022022
102-805270-BP	108EL	158-5451	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

Conditional distribution:

$$P(ek\_amount \mid keyword = \text{“Ratschen-Kabelschere”}) = \mathcal{N}(ek\_amount|\mu, \sigma^2)$$

Problem of High-cardinality

Too many distinct values!

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount	ean
102-805270-BP	1039U	9739011201	Knipex	Eckschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1528050	Knipex	Wapuzange	C62	14.82	4003773022022
102-805270-BP	108EL	158-5451	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

Conditional distribution:

$$P(ek\_amount \mid keyword = \text{“Ratschen-Kabelschere”}) = \mathcal{N}(ek\_amount|\mu, \sigma^2)$$

$\Rightarrow$ Not enough evidence to justify μ and σ!

Problem of High-cardinality

Evaluate μ and σ from similar keywords

set_id	catalog_id				unit	ek_amount	ean
102-805270-BP	1039U	973901		ksenschwedenzange	C62	55.7	4003773022022
102-805270-BP	1082	1528050	Knipex	Wapuzange	C62	14.82	4003773022022
102-805270-BP	108EL	158-5451	Knipex	Eck-Rohrzange	C62	18.33	
102-822855-BP	4445	201096-YY	Unbekannt	Ratschenkabelschere	C62	152.63	4003773043935
102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0	
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69	4003773043935

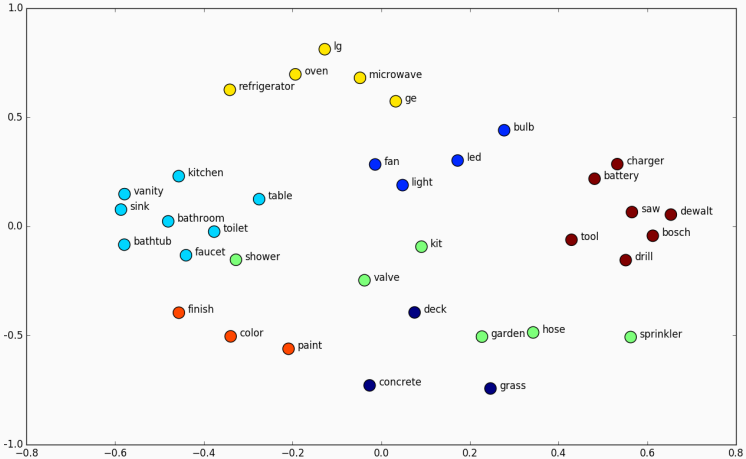
Conditional distribution:

$$P(ek\_amount \mid keyword = \text{“Ratschen-Kabelschere”}) = \mathcal{N}(ek\_amount|\mu, \sigma^2)$$

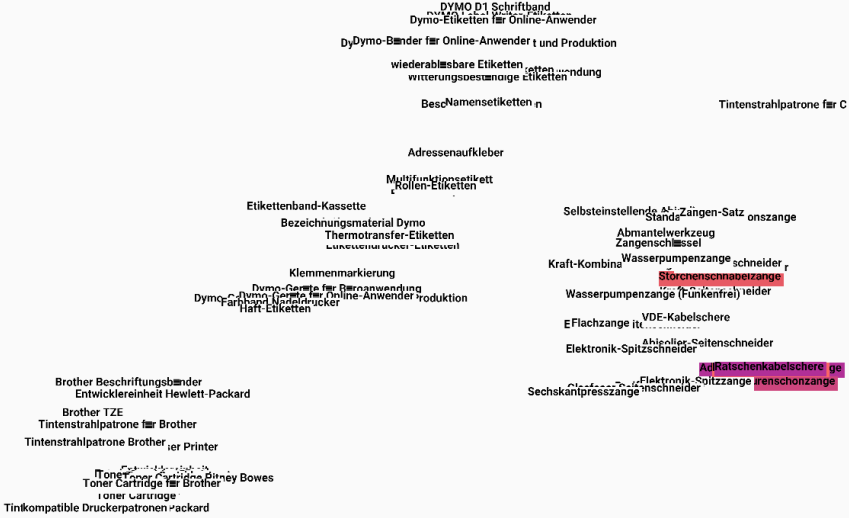
$\Rightarrow$ Not enough evidence to justify μ and σ!

What are Embeddings?

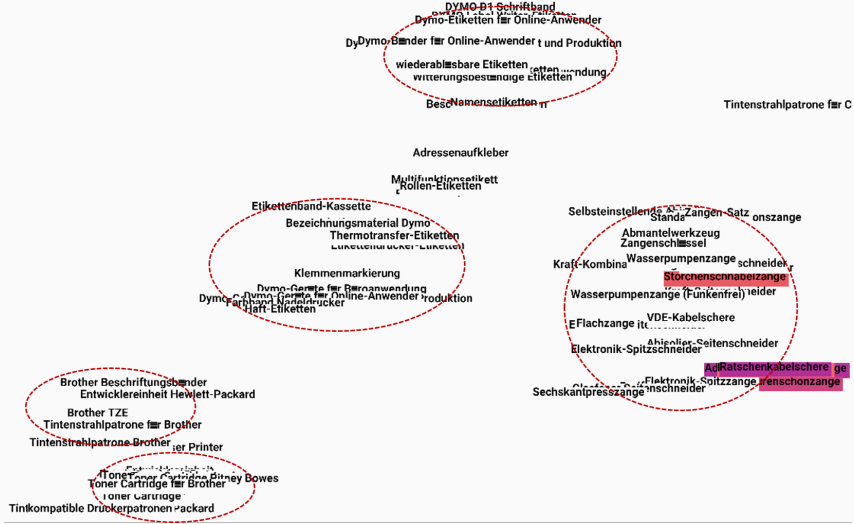
- In NLP: High-dimensional vector representations for words
- Words that share common contexts are close to each other in the vector space



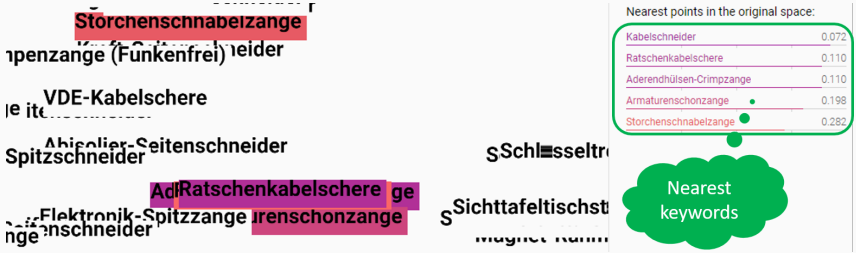
Neighbors in High-dimensional Space



Neighbors in High-dimensional Space



Neighbors in High-dimensional Space

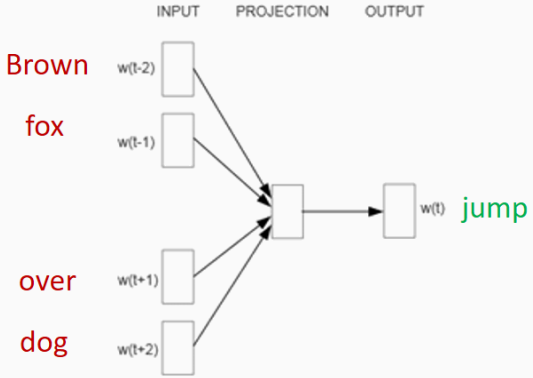


Conditional distribution:

$$P(ek\_amount \mid keyword = \text{“Ratschen-Kabelschere”})$$
$$= P(ek\_amount \mid keyword \in \{\text{“Ratschen-Kabelschere”, “Kabelschneider”, ...}\})$$

Embedding Training Model

- Derived from *Word2vec*[2] models
- Given "context", predict "target"



[2]Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space."

... The quick brown fox jumps over the lazy dog. ...

102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0

Similarities:

- Sequences with contextual information
- Limited knowledge source for high-dimensional mapping

... The quick brown fox jumps over the lazy dog. ...

102-822855-BP	381	K-95 31 250	Knipex	Ratschen-Kabelschere	C62	118.0

Similarities:

- Sequences with contextual information
- Limited knowledge source for high-dimensional mapping

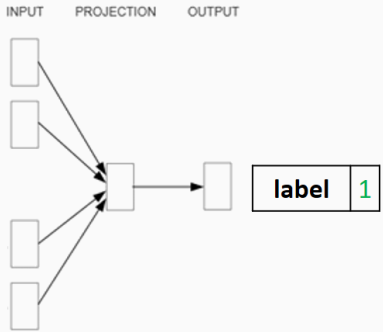
Differences:

- Variable vs. fixed sequence length
- Single vs. multiple vocabularies
- Numerical values in tabular data

Embedding Training Model

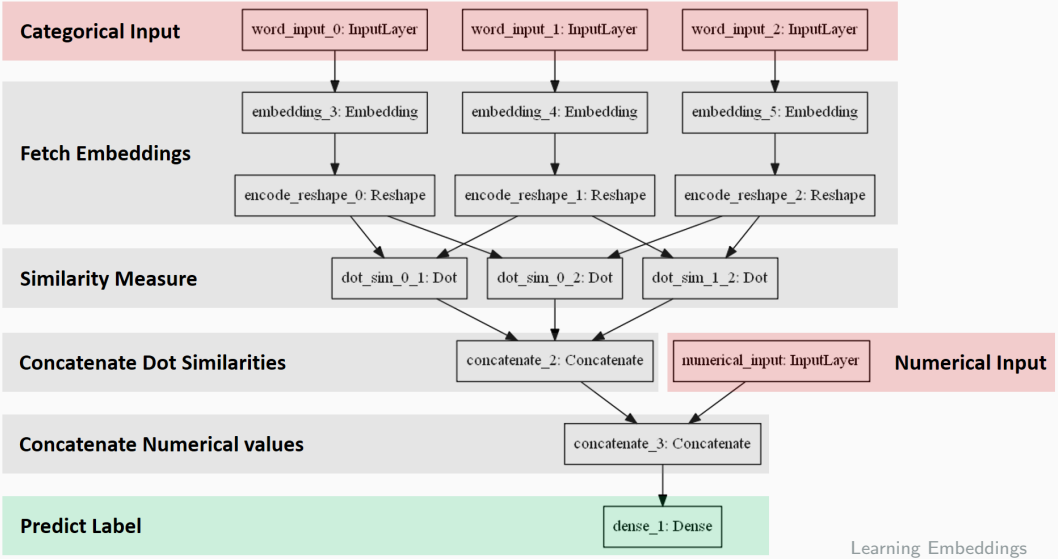
- Derived from *Word2vec*[3] models
- Given "context", predict "target"

set_id	102-822855-BP
catalog_id	381
...	...
keyword	Ratschen-Kabelschere
ek_amount	118.0



[2]Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space."

Embedding Training Model



Evaluation Goal

To investigate:

- Effectiveness of probabilistic model
- Benefits from embeddings regarding anomaly detection performance

Evaluation Goal

To investigate:

- Effectiveness of probabilistic model
- Benefits from embeddings regarding anomaly detection performance

No ground truth information available

Evaluation Goal

No ground truth information available

To investigate:

- Effectiveness of probabilistic model
- Benefits from embeddings regarding anomaly detection performance

Procedure:

1. Consider observed data as "normal"
2. Manually generate "anomalous" tuples
3. Build probabilistic model and train embeddings using combined data
4. Evaluate confidence scores
5. Check if manual anomalies have low confidence scores

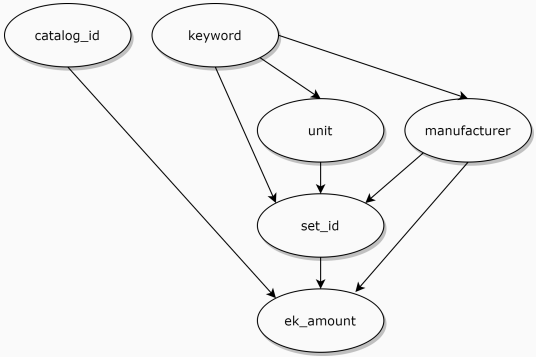
Dataset and External Knowledge

Mercateo article database

- Top 200 sets of articles
- 9049 data tuples as observation
- 452 manually generated anomalies (5% of observation)

Attribute	Attribute type	Cardinality
catalog_id	Categorical	210
unit	Categorical	10
keyword	Categorical	167
manufacturer	Categorical	35
set_id	Categorical	200
ek_amount	Numerical	n.a.

Bayesian network:



Evaluation Measures

Precision

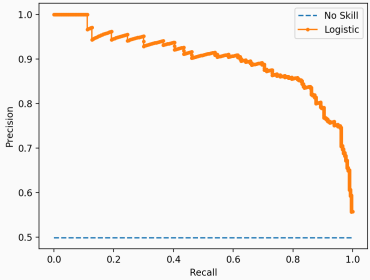
$$Precision = \frac{TP}{TP + FP}$$

Recall

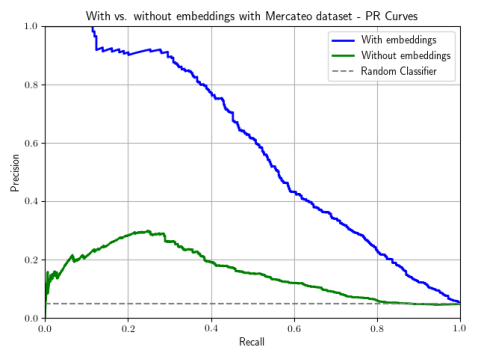
$$Recall = \frac{TP}{TP + FN}$$

Precision-recall curve:

- Describes how good a model predicts positive class (anomalies)
- Goal: Keep precision high while recall increases

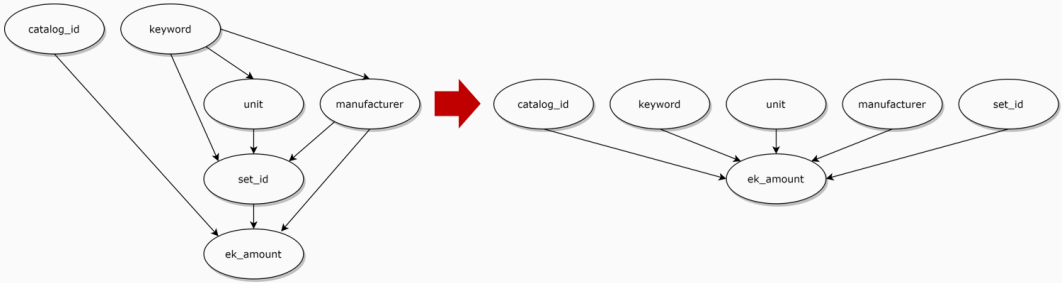


Results - With vs. Without Embeddings

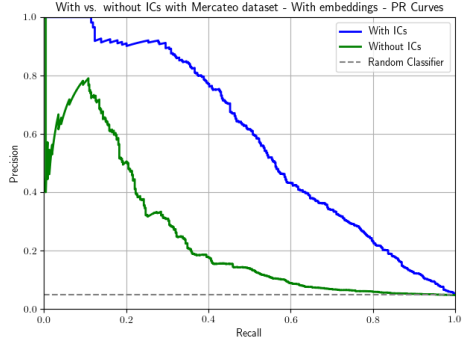


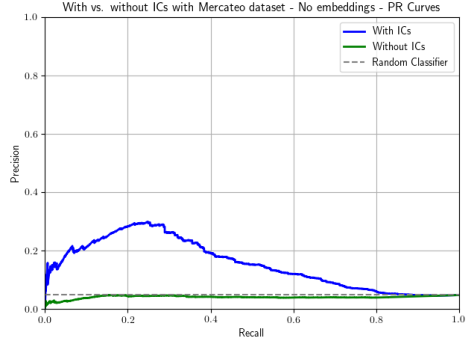
	With embeddings	Without embeddings
Recall@$P = 452$	0.54	0.29

Results - Minimizing External Knowledge



Results - Minimizing External Knowledge





	With embeddings	Without embeddings
Recall@$P = 452$	0.54	0.29
Recall@$P = 452$, minimal EK	0.30	0.02

Summary

Embeddings are helpful considering

- Improvement in anomaly detection under high-cardinality
- Compensation for limited external knowledge

Anomaly detection is feasible with probabilistic models!

Outlook

Future steps:

- Further examination of the neural network model
- Experiments with larger datasets
- Automated repair suggestion

Thank you! (Questions?)

Model Implementation with Probabilistic Programming

```
@gen function mercateo_data_model(observation)
    # No prior knowledge on root nodes => Assume uniform distribution
    set_id = @trace(choose_uniformly(observation.set_id), :set_id)
    catalog_id = @trace(choose_uniformly(observation.catalog_id), :catalog_id)
    keyword = @trace(choose_uniformly(observation.keyword), :keyword)

    # Given keyword, sample manufacturer based on occurrence statistics
    manufacturer = @trace(categorical(observation.manufacturer[keyword]), :manufacturer)

    # Given sampled values above, sample ek_amount based on observed mean and std
    ek_mu = mean(observation.ek_amount[set_id, catalog_id, keyword, manufacturer])
    ek_sigma = std(observation.ek_amount[set_id, catalog_id, keyword, manufacturer])
    ek_amount = @trace(normal(ek_mu, ek_sigma), :ek_amount)

    return set_id, catalog_id, keyword, manufacturer, ek_amount
end
```

Using Embeddings

```
# Original keyword
keyword = @trace(choose_uniformly(observation.keyword), :keyword)

# Find neighboring keywords
keyword_neighborhood = find_neighbors(keyword, keyword_embeddings)

# P(manufacturer | keyword ∈ keyword_neighborhood)
manufacturer = @trace(categorical(observation.manufacturer[keyword_neighborhood]),
                      :manufacturer)

# Find neighboring manufacturers
manufacturer_neighborhood = find_neighbors(manufacturer, manufacturer_embeddings)

# P(ek_amount | keyword ∈ keyword_neighborhood, manufacturer ∈ manufacturer_neighborhood)
ek_mu = mean(observation.ek_amount[keyword_neighborhood, manufacturer_neighborhood])
ek_sigma = std(observation.ek_amount[keyword_neighborhood, manufacturer_neighborhood])
ek_amount = @trace(normal(ek_mu, ek_sigma), :ek_amount)
```

Examples from “False Positives”

set_id	catalog_id	article_id	manufacturer	keyword	unit	ek_amount
102-805270-BP	1039U	9739011201	Knipex	Eckschwedenzange	C62	55.7
102-822855-BP	986	0184472	Primium	Ratschenkabelschere	C62	141.69
1029-1017005	1039U	9859711444	Brother	Etiketten-Papier	PK	20.45
102-823785-BP	1039U	9739011068	Knipex	Stillson-Rohrzange	C62	42.58
102-801909-BP	1039U	9739011049	Knipex	Abmantelwerkzeug	C62	80.7