Propositionalization for Unsupervised Outlier Detection in Multi-Relational Data

Fatemeh Riahi, Oliver Schulte

Flairs, May 2016

The Short Version

Our Goal

Most of outlier detection method work with a single data table or attribute value format.

One of the main data models for structured data is the relational data model.

We develop a preprocessing method that leverages single-table methods.

Player

Player ID
112
123

Match

Match ID
1
2

Team

Team ID
1
20

AppearsPlayerInMatch

Player ID	Match ID	ShotEff(T,M)
112	1	High
112	2	High
123	1	Low
123	2	Med
151	1	low

AppearsTeamInMatch

Team ID	Match ID	ShotEff(T,M)
20	1	Med.
20	2	Med.
1	1	Low

Approach

Fix a target class of individuals (e.g. players).

Learn informative features from the relational data.

Combine the learned features in a single table, one column for each learned feature.

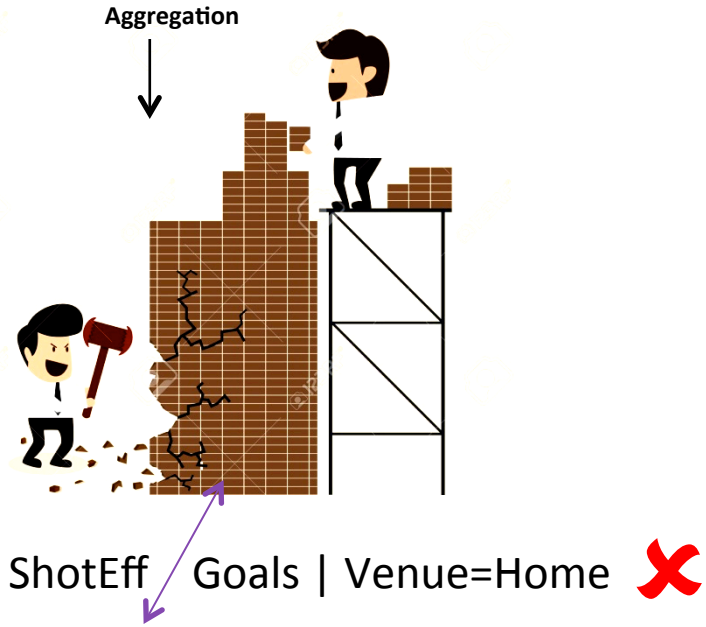
- This is called “propositionalization”.

Apply standard single-table outlier detection methods to the learned feature table.

Previous Work

Manually construct new features by aggregating single attributes (Breunig et al.).

Information loss: misses interactions among features.



Avg(ShotEff)=0.45 Avg(Goals)=0.5

PlayerID	MatchID	Venue	Shot_ Eff	Goals
1123	M1	Home	0.9	1
1123	M2	Away	0	0
1123	M3	Home	0.9	1
1123	M4	Away	0	0

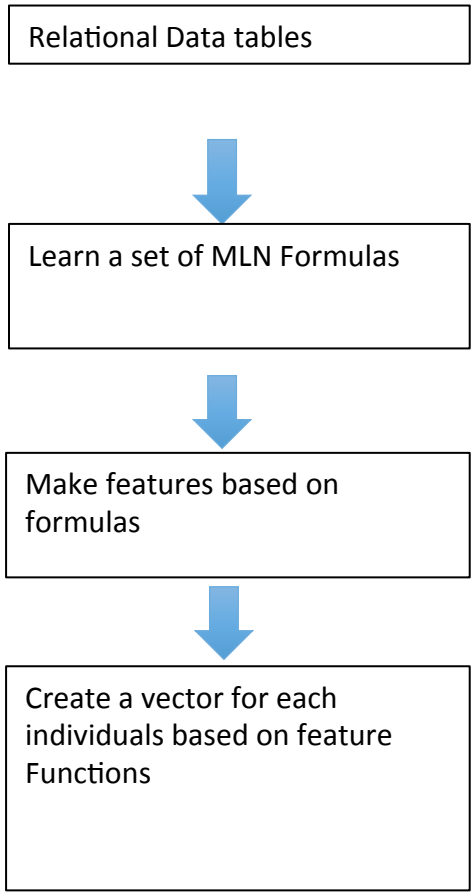
Our Approach

Learn conjunctions of associated attributes.

This can be done by applying Markov Logic Network Structure Learning.

Example:

- Formula: $ShotEff = high\ and\ PassEff = low$ is a new feature indicating when both conditions are true.



Contributions

First propositionalization method for supporting relational outlier detection (as opposed to classification).

Find informative conjunctive features for relational outlier detection: A novel application of Markov Logic Network structure learning.

Propositionalization With Markov Logic Networks

Logical Concepts

Population is a set of individuals.

- Example: all players in the database.

Relationship shows which objects are linked.

- Example: shotEfficiency(Player, Match) links Player *p* and Match *m.*

Formula is conjunction of assignments. $f(\sigma_1,...\sigma_n) = v$

- Example: shotEfficiency(Player, Match)=low/high/medium
- Grounding of a formula or term means assigning constant values to its logical variables.
  - Example: shotEfficiency(Player='wayne rooney', Match=1)

A summary of the model

Markov Logic Network is a set $\{(\phi_1, \omega_1), \ldots, (\phi_n, \omega_n)\}$ where ϕ_i is a formula and ω_i is the weight of the formula.

Markov Logic Learning:

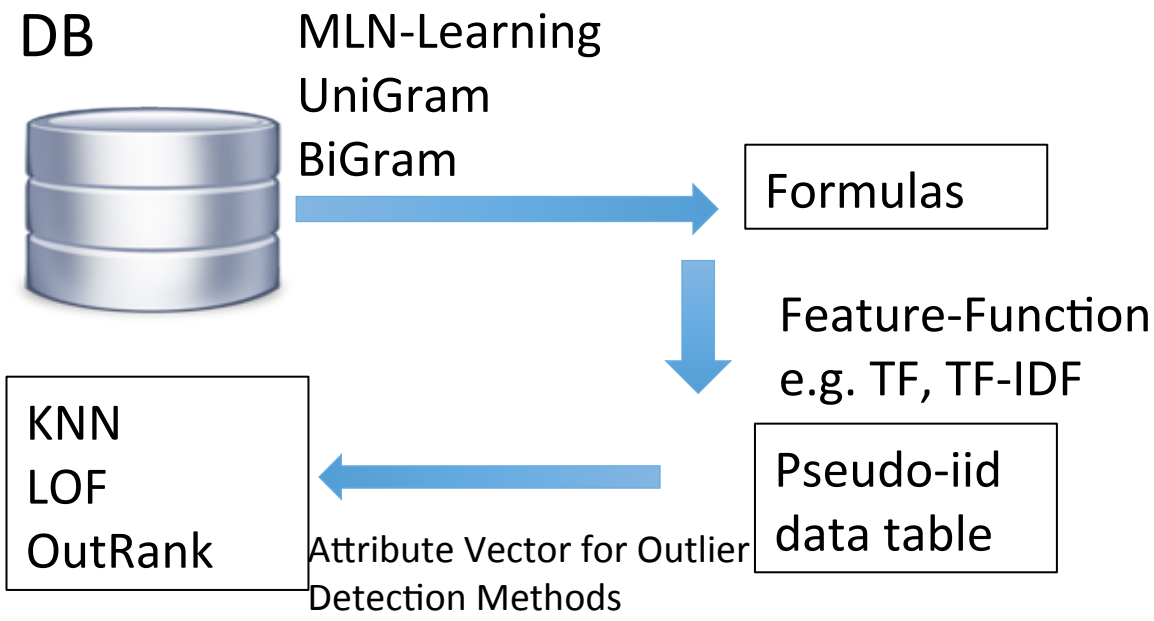
- Input: A relational database.
- Output: A set of conjunctive formulas that describe statistical patterns in the relational data.
- We use the moralization method (Khosravi et al. AAAI 2010, Schulte and Khosravi MLJ 2012).

Baselines

Wordification (Perovsek 2013): Using the concept of n-grams from NLP:

- Unigram: All single literals.
- Bigram: All conjunctions of two literals.
- We can use either term frequency (TF) or term frequency/inverse document frequency (TF-IDF).
- We call TF and TF-IDF that map multiple instances of a formula to real values **Feature Functions**

tlier Detection using
ppositionalization



Example

CP Table

ShotEff(P,M)	PassEff(P,M)	CP	Prior
Low	High	0.5	0.5
Low	Low	1	0.5
High	High	0.95	0.5

Smooth CP Table

ShotEff(P,M)	PassEff(P,m)	CP	Prior
Low	High	0.5	0.5
Low	Low	1	0.5
High	High	0.95	0.5
High	Low	0	0

Extract Formulas

$$f1: ShotEff(P,M) = low \text{^} PassEff(P,M) = high$$
$$f2: ShotEff(P,M) = low \text{^} PassEff(P,M) = low$$
$$f3: ShotEff(P,M) = high \text{^} PassEff(P,M) = high$$
$$f4: ShotEff(P,M) = high \text{^} PassEff(P,M) = low$$

Choose weighting function

Prior Vector

f1	f2	f3	f4
0.5	0.5	0.5	0

CP Vector

f1	f2	f3	f4
0.5	1	0.95	0

Apply feature function
e.g. TF,TFIDF

KNN
LOF
OutRank

Evaluation Methodology

Synthetic Datasets

- Synthetic Datasets: Should be easy! Two Features per player per match.

tio	ShotEff	Match Result
	1	1
	1	1
	0	0
	0	0
	1	1
	1	0
	0	0
	0	1

low correlation	ShotEff	Match Result
Normal	1	1
	1	0
	0	0
	0	1
Outlier	1	1
	1	1
	0	0
	0	0

Single Feature	ShotEff	Match Result
Normal	0	0
	0	0
	0	0
	1	1
Outlier	1	1
	1	1
	1	1
	1	1

Real-World Datasets

Real Datasets:

- Soccer Data
  - Strikers vs. Goalies
  - Midfielders vs. Strikers
- IMDb data
  - Drama vs. Comedy

Evaluation

Dimensionality: The number of attributes in the final attribute table

Attribute Complexity: The length of conjunctions that define attributes

Outlier Analysis Run Time

Attribute Construction Time

Apply state-of-the-art outlier detection methods to the attribute table in order to compare the performance of different feature generation methods.

- Performance accuracy score: AUC

Evaluation Results

Evaluation Results

A propositionalization method is scored 1 point if it produces the bes accuracy on a dataset, and 0.5 points if it ties. The table shows the total number of wins and average of AUC over all datasets.

ropositonalization → / utlier Detection Method ↓	MLN-TF		Bigram-IDF		Unigram-TF		Treeliker	
	Wins	AVG(AUC)	Wins	AVG(AUC)	Wins	AVG(AUC)	Wins	AVG(AU
utRank	**2.50**	**0.79**	**2.50**	0.70	1.00	0.64	0	NA
NN	**3.50**	**0.78**	1.50	0.67	1.50	0.67	0	0.64
DF	**4.00**	**0.63**	1.00	0.55	1.00	0.61	1	0.61

Evaluation Results

Comparison of complexity, dimensionality and construction time for the attributes produced by different propositionalization methods.

	MLN			Bigram			Unigram		
ataset	Formula Length	Dimensionality	Construction Time	Formula Length	Dimensionality	Construction Time	Formula Length	Dimensionality	Construct
rikers vs Goalies	3.55	331	5.24	2	1825	1.2	1	63	0.
idfielders vs Strikers	3.27	198	4.92	2	1762	0.85	1	62	0.
rama vs Comedy	4.20	930	10.80	2	1991	2.87	1	47	0.

Conclusions

Summary and Future Work

Impedance mismatch: Standard outlier methods assume single-table data, but relational databases maintain multiple interrelated tables.

AI-based solution: Discover informative features in the relational database use them to construct a single-data table.

Efficient conjunctive feature discovery = Markov Logic Network structure learning.

Works better than baselines with isolated attributes (unigrams) or enumerating all binary conjunctions of attributes (bigrams).

More results on comparing with supervised propositionalization (not in paper).

Other ways of generating unsupervised formula: WARMR