TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Fraunhofer ITWM

Unsupervised Anomaly Detection for Auditing Data and Impact of Categorical Encodings

Ajay Chawda, Stefanie Grimm, Marius Kloft.

Motivation – Synthetic Data

Dataset	*Rows*	*Numerical attributes*	*Categorical attributes*	*Anomaly ratio*
Credit Card Fraud	284207	30	0	0.001
Arrhythmia	452	279	0	0.26
KDD	494021	34	7	0.2
Car Insurance	1000	16	17	**0.25**
Vehicle Insurance	15420	6	26	**0.06**
Vehicle Claims	268255	5	13	**0.21**

- Table 1 : Auditing datasets are in the form of mixed tabular dataset. *More categorical attributes than numerical attributes.*

- Auditing datasets are different from available Tabular benchmark anomaly detection datasets.
- Unavailable Public benchmark dataset. Therefore, we generate a synthetic dataset.

Motivation – Unsupervised Methods

- Labeling of data is inefficient.
- Requires a domain expert to audit the journals and find discrepancies.
- Time consuming and Expensive.
- Unsupervised methods are suitable for auditing data.

Vehicle Claims - Source

- DVM-Car dataset consists of 1.4 million car images.
- The metadata consists of tabular data containing information about the cars in the dataset.
- *Basic table* contains attributes 1,011 generic models from 101 automakers used as an identifier.
- *Price table* contains price information of models.
- *Trim table* includes information such as selling price, fuel type and engine size.

Vehicle Claims – Feature Engineering

With previous domain knowledge and feature engineering we create anomalous points in the data.

```
final_data['Color'].fillna(value='Gelb', inplace=True)
final_data['Reg_year'].fillna(value= 3010, inplace=True)
final_data['Bodytype'].fillna(value='Wood', inplace=True)
final_data['Engin_size'].fillna(value= '999.0L', inplace=True)
final_data['Gearbox'].fillna(value= 'Hybrid', inplace=True)
final_data['Fuel_type'].fillna(value='Hydrogen', inplace=True)
final_data['Door_num'].fillna(value= 0, inplace=True)
```

Figure 3: NA values are replaced with anomalous values.

1) Categorical features *issue*, *issue_id* are added. *issue_id* is a subcategory of the issue column.
2) The *repair_complexity* column is added based on the maker of the vehicle. *The common brands like Volkswagen have complexity 1 whereas Ferrari has complexity 4*.
3) *repair_hours* and *repair_cost* are calculated based on the *issue*, *issue_id* and *repair_complexity*. Every tenth row in *repair_cost* and $20th$ row in *repair_hours* is an anomaly.
4) Anomalous values are also inserted while replacing NA values.
5) Labels are added for evaluation.

Vehicle Claims – Challenges

Maker	Color	Runned_Miles	Engin_size	Price	Seat_num	Door_num	issue	issue_id	Adv_day	breakdown_date	repair_complexity	repair_cost	repair_hours	repair_date	Label
Bentley	Silver	60000	6.8L	21500.0	5.0	4.0	Starter Motor Issue	0	19	2018-04-19	3	395.0	9.0	2018-04-21	?
Bentley	Grey	44000	6.8L	28750.0	5.0	4.0	Radiator Leaking	0	15	2018-06-15	3	695.0	6.0	2018-06-16	?
Audi	Blue	55000	6.8L	29999.0	5.0	4.0	Steering Wheel Shaking	0	10	2017-11-10	3	89999.00	3.0	2017-11-10	?
BMW	Green	14000	6.8L	34948.0	5.0	4.0	Electrical Issue	4	14	2018-04-14	3	224844.00	6.0	2018-04-15	?
Volkswagen	Grey	61652	6.8L	26555.0	5.0	4.0	Windscreen Crack	0	6	2017-11-06	3	75.93299999999999	3.0	2017-11-06	?
Audi	Blue	55000	6.8L	24950.0	5.0	4.0	Tyre Alignment	0	6	2017-12-06	3	32495.00	1.5	2017-12-06	?
Bentley	Green	67000	6.8L	29995.0	5.0	4.0	Sensor Malfunction	0	2	2018-08-02	3	1679.75	9.0	2018-08-04	?
BMW	Black	99200	6.8L	22450.0	5.0	4.0	Sensor Malfunction	0	3	2018-04-03	3	1302.5	9.0	2018-04-05	?
Bentley	Silver	27541	6.8L	26990.0	4.0	4.0	Windscreen Crack	0	18	2018-02-18	3	76194.00	3.0	2018-02-18	?
Ferrari	Silver	38000	6.8L	29995.0	5.0	4.0	Flat Tyres	0	3	2017-11-03	3	244.98721913926553	3.0	2017-11-03	?

Figure 1: A small subset of Synthetic dataset created for evaluation of unsupervised models.

Handling of categorical attributes

Categorical Encodings

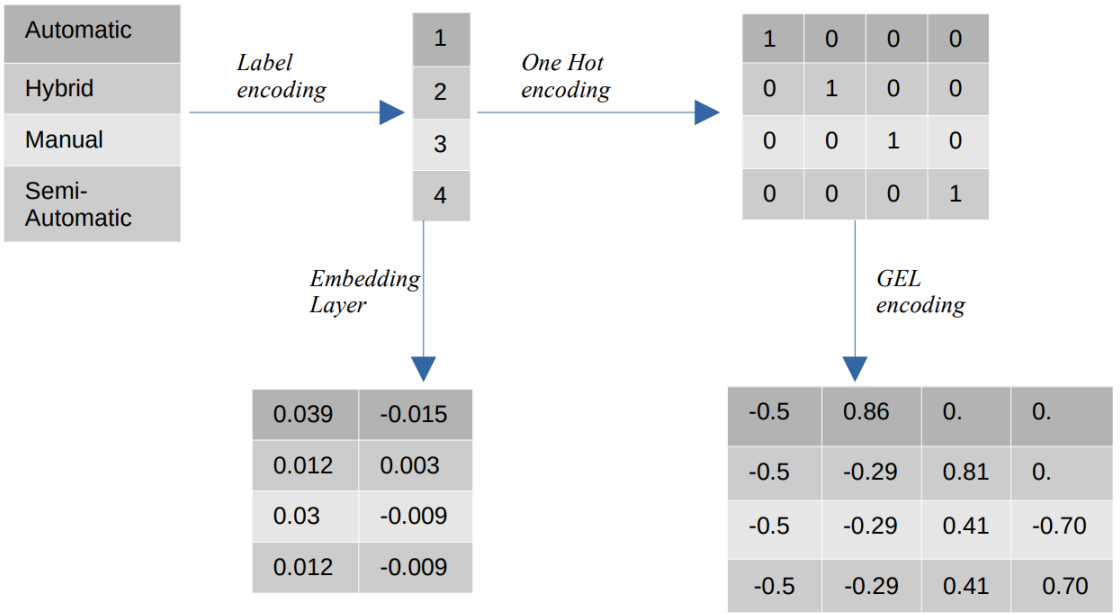

Figure 2: Representation of encodings and embeddings for an example containing four unique values.

- **Embedding** Layer – Random weights are
- initialized for the embeddings. Trained on the
- reconstruction loss.

- **Label** encoding – Suitable for Ordinal values.
- **One Hot** encoding – Each categorical value is
- transformed into a distinct features. Increases
- dimensionality for large datasets.
- **GEL** encoding – Low rank representation of One Hot encoding.

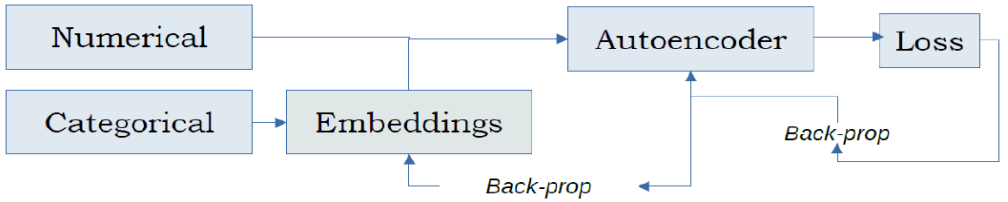

Figure 3: Training embeddings with auotencoder .

Unsupervised Methods - Models

- Competitive Learning - SOM - Self Organizing Maps.

- Density Estimation - 1) DAGMM – Deep Autoencoding Gaussian Mixture Model.
  2) SOM-DAGMM – Self Organizing Map - Deep Autoencoding Gaussian Mixture
  3) Model.

- Reconstruction Error - RSRAE – Robust Surface Recovery Layer for Unsupervised Anomaly Detection.

- Contrastive Learning - 1) NeuTral-AD – Neural Transformation Learning for Anomaly Detection.
  2) LOE – Latent Outlier Exposure for Anomaly Detection with Contaminated Data.

Results

Method	**Model**	*Label Encoding*	*One Hot Encoding*	*GEL Encoding*	*Embedding Layer*
Supervised	Random Forest	**98.65**	92.35	66.87	-
Supervised	Gradient Boosting	93.26	**95.88**	85.43	-
Unsupervised	Isolation Forest	**59.42**	49.56	52.45	-
Unsupervised	Local Outlier factor	51.78	53.05	52.86	-
Unsupervised	One Class - SVM	51.68	50.42	51.12	-
Unsupervised	SOM	56.64	**57.67**	58.32	**65.43**
Unsupervised	RSRAE	53.44	52.51	55.38	52.68
Unsupervised	DAGMM	50.86	48.79	48.86	51.22
Unsupervised	SOM-DAGMM	49.53	50.22	49.97	53.82
Unsupervised	NeuTraL-AD	54.82	53.71	57.03	-
Unsupervised	LOE	57.03	55.32	**58.59**	-

Table 2 : AUC Scores for Vehicle Claims dataset. The performance of the model depends on the encoding of categorical variables. We observer poor performance for One Hot encoding which is preferred approach for categorical attributes.

Conclusion

- Our dataset has comparable performance to other auditing datasets.
- Helps in identifying categorical encoding issue.
- For large dataset, One Hot encoding is not suitable.
- Code to generate data is available on Github (https://github.com/ajaychawda58/UADAD).
- Can be used to create dataset suitable to the learning task.

Thank You.