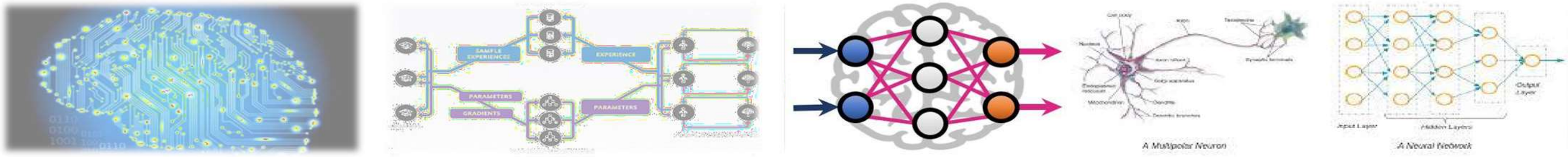


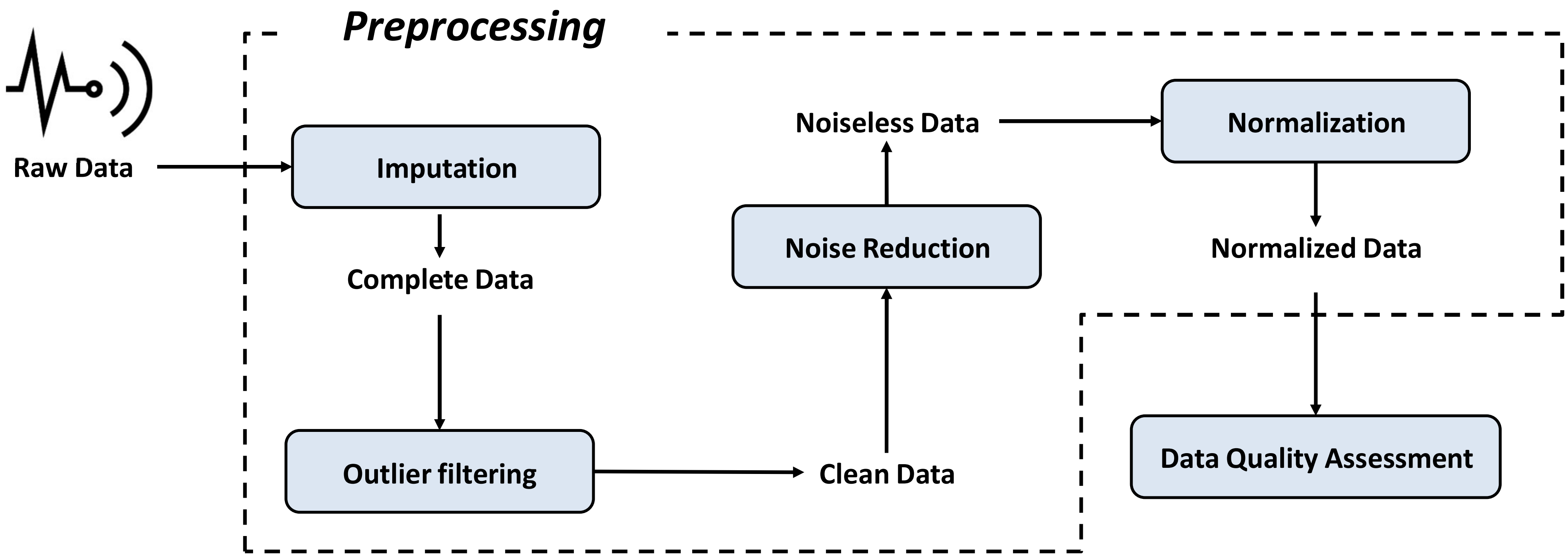
Sensor Data Preprocessing

Sensor Data 전처리

Geonil Yun,
rjsdlf45@gmail.com
2021. 07. 27.

Introduction

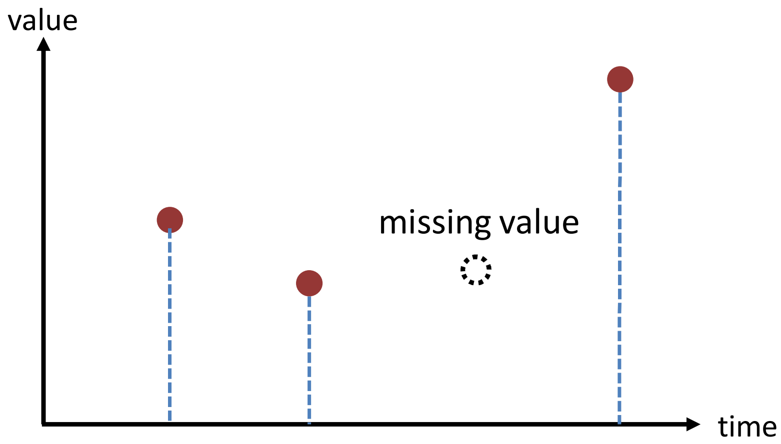
- **Sensor data preprocessing flow**
  - ➢ **4 steps preprocessing with sensor data**
  - ➢ **data quality assessment**



Imputation

- **Reasons of missing data**
  - **Sensor malfunction**
  - **Unstable internet connection**
  - **Other technical difficulties**

missing some points in time



ignore missing data → low efficiency and unreliable results

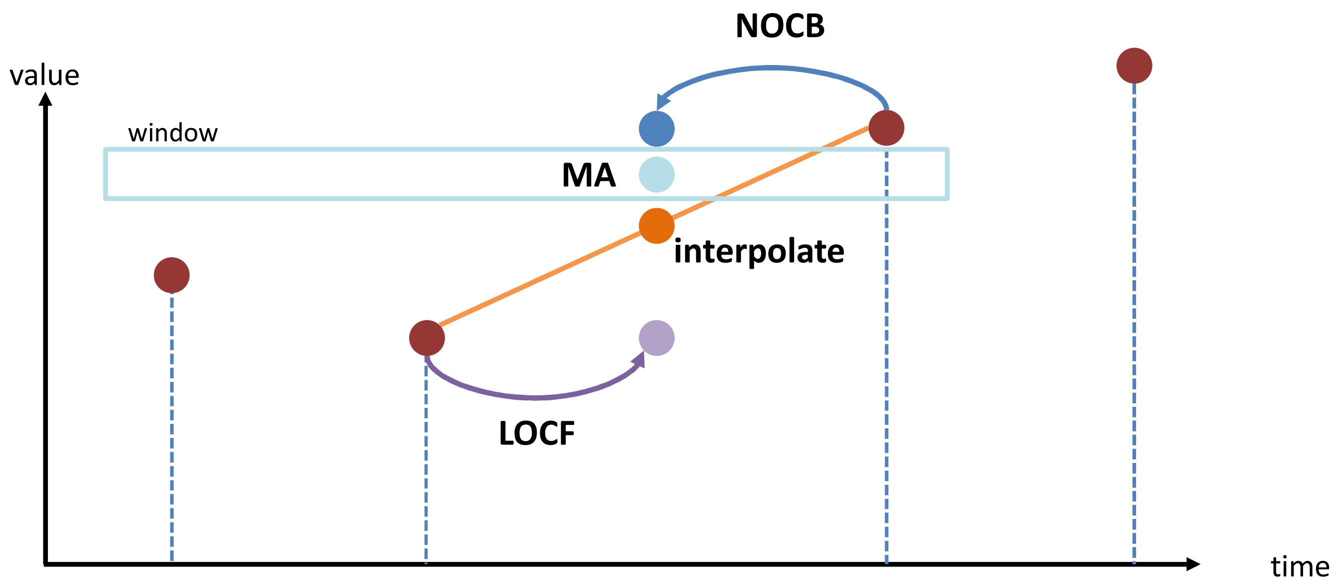
Imputation

- **Simple methods**
  - **Median, Mode, Mean, Random imputation**
  - **Not just for time series data**
  - **Very fast and simple, but low accuracy**

raw data	Median	Mode	Mean	Random
Score	Score	Score	Score	Score
100	100	100	100	100
90	90	90	90	90
90	90	90	90	90
	90	**90**	**80**	**20**
80	80	80	80	80
40	40	40	40	40

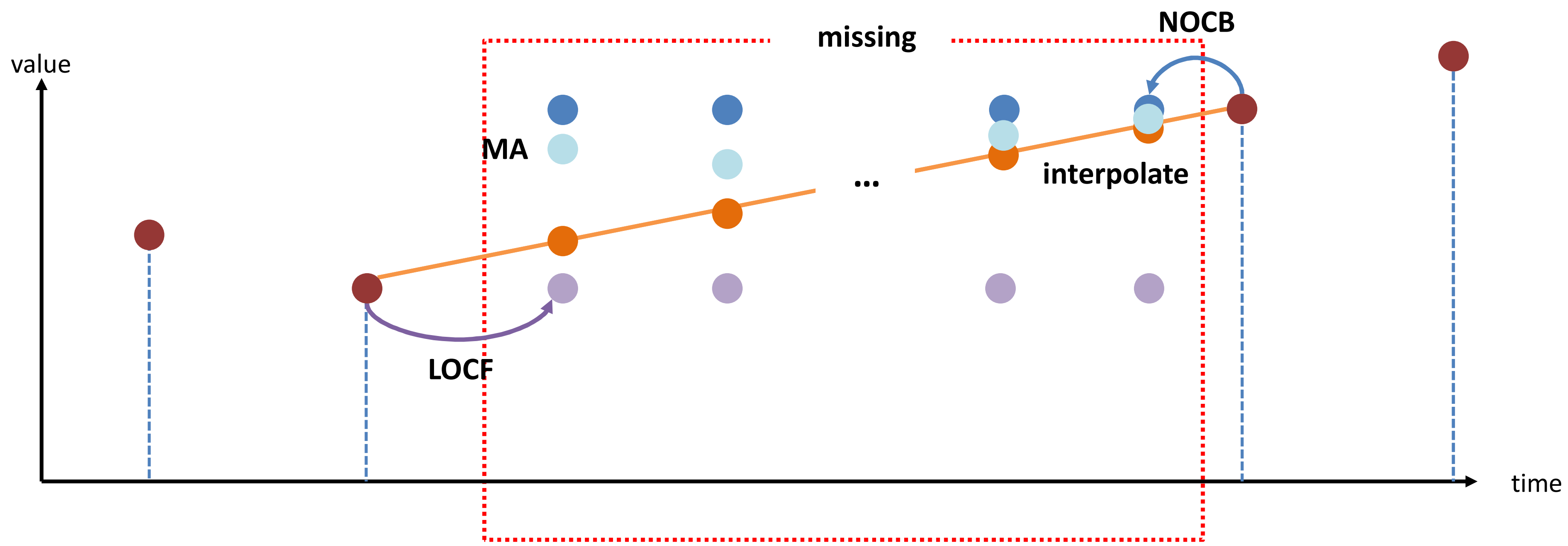
Imputation

- **Simple methods for time series**
  - **LOCF (Last Observation Carried Forward)**
  - **NOCB (Next Observation Carried Backward)**
  - **Interpolate**
  - **Moving average (simple, weighted, exponential)**



Imputation

- **Simple methods for time series**
  - **Large gap sub-sequences**



Imputation

- **Metrics**
  - **RMSE (Root Mean Square Error)**

$$RMSE = \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{n}}$$

  - **MAPE (Mean Absolute Percentage Error)**

$$MAPE = \frac{100\%}{n} \times \sum_i \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

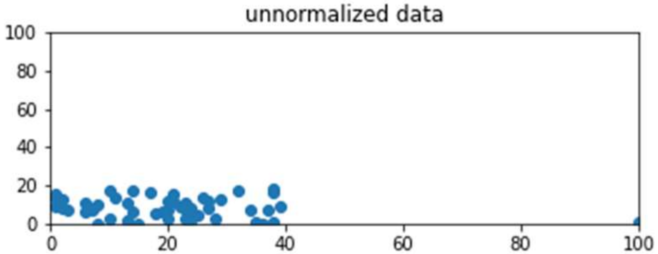
Noise Reduction

- **Methods**

  - **Frequency domain approaches**
  - **Time domain approaches**

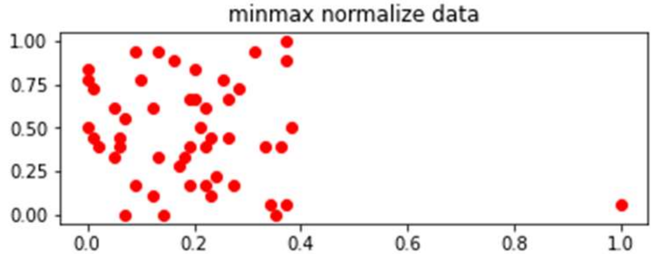
Data Normalization

- **Data Normalization methods**



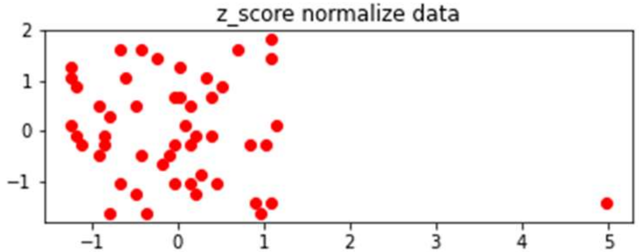
  - **min-max normalization**

$$\hat{y}_{minmax} = \frac{y - \min(Y)}{\max(Y) - \min(Y)}$$



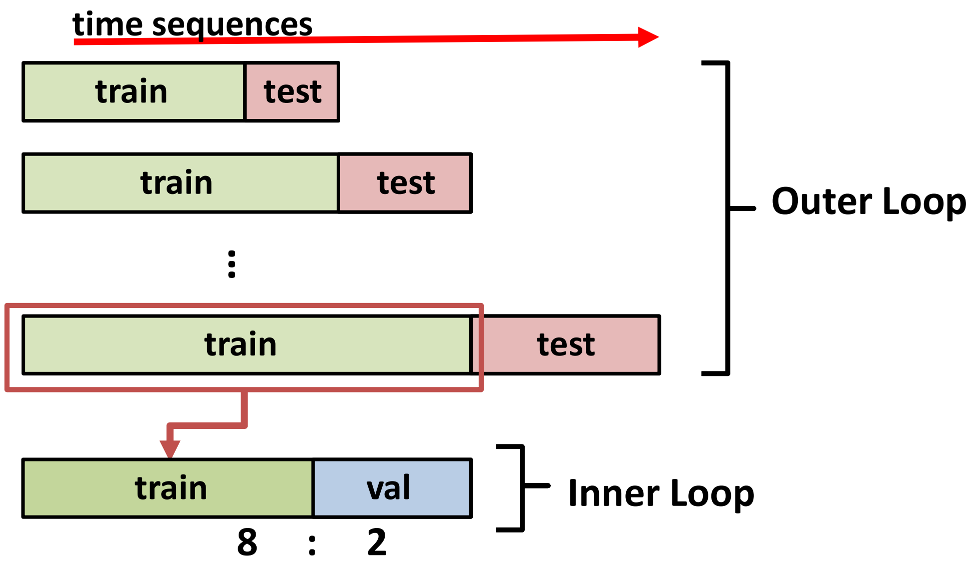
  - **z-score normalization**

$$\hat{y}_{z-score} = \frac{y - \mathrm{mean}(Y)}{std(Y)}$$



Experimental setting

- **Nested Cross Validation for time series**



➢ **train past data & predict future data**
➢ **handling time series data in Nested CV**

- hyperparameter

Model	CNN (ShuffleNet)	RNN(LSTM, GRU)
fold num	10	
hidden size	64	
input length	24	
output length	1	
epoch	30	
batch size	32 or 64	
learning rate	**0.01, 0.001**	

➢ batch size 32 for **ISE, Chickenpox** dataset
➢ batch size 64 for others

Experimental setting

- **Statistics**

Dataset		Mean	Standard Deviation	Skewness	Kurtosis	Jarque Bera	ADF
Air Quality		1099.833	217.080	0.755	0.334	897.811	-9.823
Appliances Energy		97.694	102.524	3.386	13.664	191240.781	-21.616
Beijing Air Pollution		98.613	92.050	1.802	4.768	62162.743	-20.606
Hungarian Chickenpox		101.245	76.354	0.948	1.400	121.010	-6.933
Electric Consumption		1.483	1.194	1.141	1.732	3421.560	-5.360
Eye State		0.448	0.497	0.205	-1.957	2497.788	-4.803
Istanbul Stock Exchange	ISE	0.0016	0.016	-0.094	1.391	44.022	-22.746
	SP	0.0006	0.014	-0.090	3.068	211.013	-11.031
	DAX	0.0007	0.014	-0.106	2.104	99.894	-23.056
Metro Volume		3259.818	1986.860	-0.089	-1.309	3506.110	-28.016
Occupancy Detection		0.212	0.408	1.406	-0.020	2686.285	-3.430
SML 2010		20.493	3.312	-0.169	-0.105	21.792	-4.314

*ISE : Istanbul Stock Exchange, *SP : Standard & Poor's, *DAX : Germany

Comparison results

- Training time [s]

Data		CNN (ShuffleNet)	LSTM	GRU
Air Quality		85.945	111.179	204.589
Appliances Energy		193.282	253.574	466.305
Beijing Air Pollution		99.431	127.850	237.616
Hungary Chickenpox		8.842	11.399	20.961
Electric Consumption		100.391	130.031	241.473
Eye State		154.765	198.695	361.457
Istanbul Stock Exchange	ISE	9.133	12.116	21.743
	SP	9.142	11.667	22.085
	DAX	9.271	11.812	22.005
Metro Interstate Traffic Volume		492.337	615.670	1135.469
Occupancy Detection		81.002	105.606	192.808
SML 2010		41.084	52.834	98.692

Comparison results

- RMSE (Root Mean Squared Error)

Data		CNN (ShuffleNet)	Standard Deviation	LSTM	Standard Deviation	GRU	Standard Deviation	LSTM Enc-Dec	Standard Deviation
Air Quality		147.489	33.383	103.805	25.297	295.904	256.525		
Appliances Energy		72.338	8.558	72.563	22.345	184.684	244.671		
Beijing Air Pollution		48.369	11.659	26.796	8.089	99.672	61.693		
Hungary Chickenpox		63.992	17.450	53.865	15.022	80.651	30.074		
Electric Consumption		0.848	0.357	0.294	0.238	1.228	1.397		
Eye State		0.059	0.074	3.012	4.770	3.531	3.895		
Istanbul Stock Exchange	ISE	0.028	0.007	6.235	6.343	1.259	0.585		
	SP	0.022	0.009	5.557	5.713	2.803	3.791		
	DAX	0.024	0.009	4.771	5.510	2.582	3.354		
Metro Interstate Traffic Volume		864.935	252.857	527.695	125.450	2261.804	1003.633		
Occupancy Detection		0.388	0.151	1.737	2.729	0.955	0.847		
SML 2010		2.512	1.128	0.310	0.219	5.659	7.295		

Comparison results

- MAE (Mean Absolute Error)

Data		CNN (ShuffleNet)	Standard Deviation	LSTM	Standard Deviation	GRU	Standard Deviation	LSTM Enc-Dec	Standard Deviation
Air Quality		112.998	25.412	79.423	24.508	258.354	264.234		
Appliances Energy		34.821	6.233	39.905	20.294	150.641	253.860		
Beijing Air Pollution		31.581	5.880	16.690	4.281	81.923	63.357		
Hungary Chickenpox		47.789	11.065	36.731	10.603	65.710	25.898		
Electric Consumption		0.623	0.294	0.121	0.086	1.035	1.296		
Eye State		0.469	0.068	2.690	4.789	3.499	3.909		
Istanbul Stock Exchange	ISE	0.022	0.006	6.233	6.344	1.254	0.587		
	SP	0.017	0.008	5.555	5.713	2.796	3.787		
	DAX	0.019	0.007	4.768	5.511	2.576	3.349		
Metro Interstate Traffic Volume		576.552	166.768	343.280	70.116	1925.708	918.449		
Occupancy Detection		0.222	0.097	1.492	2.202	0.836	0.851		
SML 2010		2.321	1.146	0.256	0.188	5.322	7.368		

References

[1] : Latyshev, E. (2018, October). Sensor Data Preprocessing, Feature Engineering and Equipment Remaining Lifetime Forecasting for Predictive Maintenance. In *DAMDID/RCDL* (pp. 226-231).

[2] :

Thank you

Appendix : Shuffle Net Architecture

Layer	Output size	KSize	Stride	Repeat	Output channels (g groups)				
					$g=1$	$g=2$	$g=3$	$g=4$	$g=8$
Image	224×224				3	3	3	3	3
Conv1	112×112	3×3	2	1	24	24	24	24	24
MaxPool	56×56	3×3	2						
Stage2	28×28		2	1	144	200	240	272	384
	28×28		1	3	144	200	240	272	384
Stage3	14×14		2	1	288	400	480	544	768
	14×14		1	7	288	400	480	544	768
Stage4	7×7		2	1	576	800	960	1088	1536
	7×7		1	3	576	800	960	1088	1536
GlobalPool	1×1	7×7							
FC					1000	1000	1000	1000	1000
Complexity					143M	140M	137M	133M	137M

Appendix : Dataset

Dataset	references
Air Quality	https://archive.ics.uci.edu/ml/machine-learning-databases/00360/
Appliances Energy	https://archive.ics.uci.edu/ml/machine-learning-databases/00374/
Beijing Air Pollution	https://archive.ics.uci.edu/ml/machine-learning-databases/00381/
Hungarian Chickenpox	https://archive.ics.uci.edu/ml/machine-learning-databases/00580/
Electric Consumption	https://archive.ics.uci.edu/ml/machine-learning-databases/00235/
Eye State	https://archive.ics.uci.edu/ml/machine-learning-databases/00264/
Istanbul Stock Exchange	https://archive.ics.uci.edu/ml/machine-learning-databases/00247/
Metro Volume	https://archive.ics.uci.edu/ml/machine-learning-databases/00492/
Occupancy Detection	https://archive.ics.uci.edu/ml/machine-learning-databases/00357/
SML 2010	https://archive.ics.uci.edu/ml/machine-learning-databases/00274/