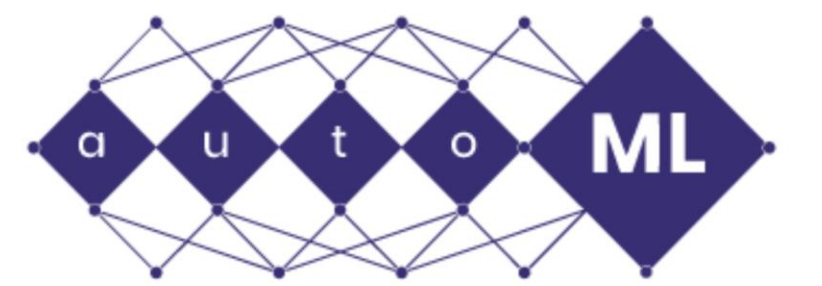


BERT-Sort: A Zero-shot MLM Semantic Encoder on Ordinal Features for AutoML

Mehdi Bahrami, Wei-Peng Chen, Lei Liu, Mukul Prasad
Fujitsu Research of America

FUJITSU



AutoML-Conf 2022

Background

Normal < mild < moderate < severe

0	1	2	3	4	5	6	7	8	9	...	61	62	63	64	65	66	67	68	69	70		
0	f	mild	f	normal	normal	?	t	?	f	f	...	f	f	normal	t	a	f	f	f	p1	cochlear_unknown	
1	f	moderate	f	normal	normal	?	t	?	f	f	...	f	f	normal	t	a	f	f	f	p2	cochlear_unknown	
2	t	mild	t	?	absent	mild	t	?	f	f	...	f	f	normal	t	a	s	f	f	f	p3	mixed_cochlear_age_fixation
3	t	mild	t	?	absent	mild	f	?	f	f	...	f	f	normal	t	b	f	f	f	p4	mixed_cochlear_age_otitis_media	
4	t	mild	f	normal	normal	mild	t	?	f	f	...	f	f	good	t	a	f	f	f	p5	cochlear_age	
...	

Ordinal Features

Target Feature

Unsupervised Approach

- Processing large context to understand the order of context

should be classified as moderate or severe. For example, relative to the study by Serlin and colleagues, a study of patients with diabetic peripheral neuropathy found a lower cutoff between mild and moderate pain (i.e., 0-3, 4-6, 7-10, for mild, moderate and severe) (11), while two replications in cancer populations (2, 15) reported a higher cutoff between moderate and severe pain (0-4, 5-7, 8-10). Based on these between-group differences, Comparison of classification systems for mild, moderate and severe pain intensity based on interference with activity

Pain type	CP 3,6	CP 3,7	CP 4,6	CP 4,7	CP 2,5	CP 3,5
Average	50.69	48.75	49.80	49.77	64.16	
Worst	46.11	46.52	46.66	49.00		

Signs and Symptoms of Mild, Moderate, and Severe Depression

Mild depression | Moderate depression | Severe (major) depression | Taking action

How depression is classified

Language Modeling

To identify empirically-derived cutoffs for mild, moderate, and severe pain often developed based upon categorical ratings of pain (e.g., mild, moderate, severe).

reported a higher cutoff between moderate and severe pain

$$P(w_t | context) \forall t \in V$$

$$P(w_t | w_{t-k}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k})$$

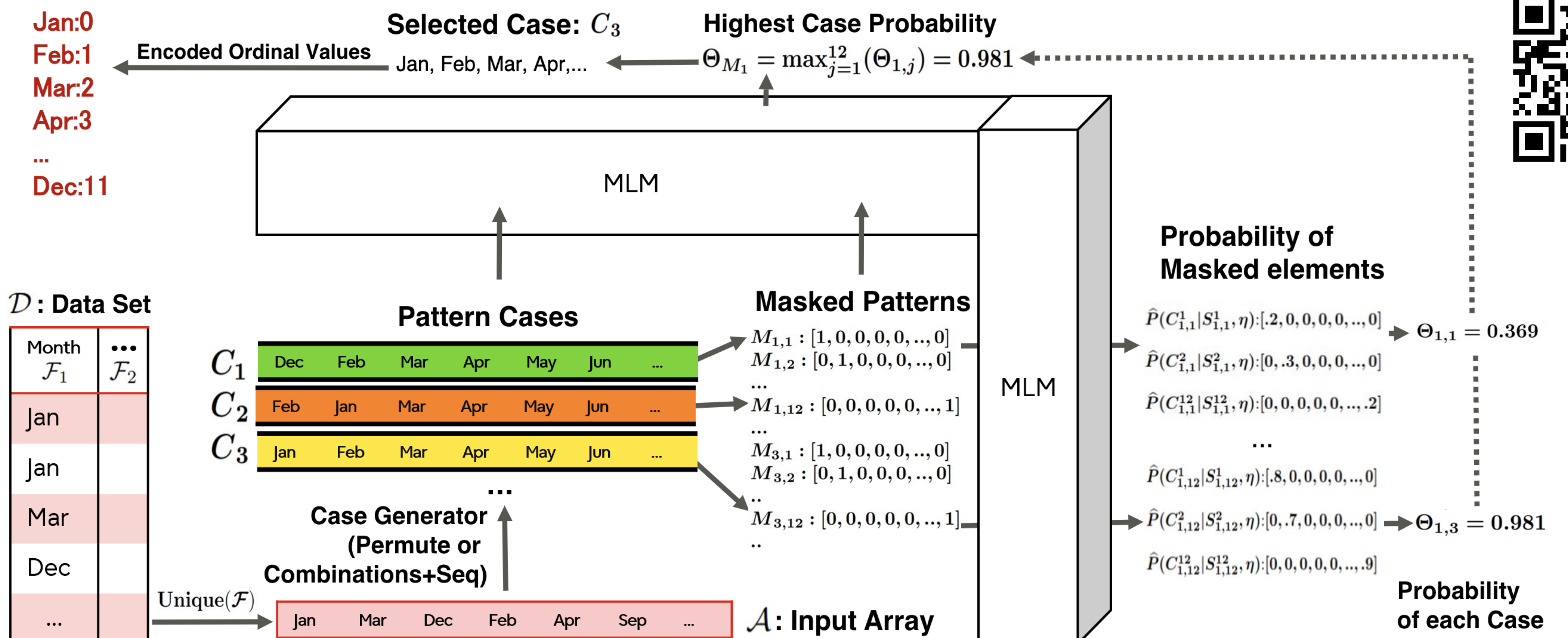
Evaluation

$$\Theta_{i,j} = \begin{cases} \frac{\sum_{k=1}^{|A_i|} \hat{P}(C_{i,j}^k | S_{i,j}^k, \eta)}{|A_i|}, & \text{if } C_{i,j}^k \in \mathcal{W}_\eta \\ 0, & \text{otherwise} \end{cases}$$

$$Ord_{Acc} = \sum_{i=1}^{|A|} \frac{\|A\| - |\mathcal{L}_i - \hat{\mathcal{L}}_i|}{\|A\|^2}$$

#	Ranked Values	Acc	Ord _{Acc}
1	[Feb < Jan < Mar < Apr]	0.5	0.87
2	[Mar < Feb < Jan < Apr]	0.5	0.75

BERT-Sort



Semantic Evaluation

A Comparisons of semantic ordinal value evaluation of BERT-Sort with initiation on 4 different MLMs of DistilBERT, RoBERTa, XLM, BERT-base-uncased, and OrdinalEncoder

Evaluation	Ord_{Acc}				Acc			
Approach	BERT-Sort				OrdinalEncoder			
Model	M_1	M_2	M_3	M_4	M_1	M_2	M_3	M_4
Feature								
#champions	30	35	31	32	3	23	31	27
	w.r.t. OrdinalEncoder				w.r.t. M_2			
Improvement	0.20	0.27	0.25	0.20	baseline	0.31	0.55	0.49

Multilingual Multi-Domain BERT Sort

#	Input	Model	BERT-Sort (Top 1)	OrdinalEncoder
2	[Lava Hot, Hot, Boiling Hot]	RoBERTa-large	[Hot<Boiling Hot <Lava Hot]	[Boiling Hot<Hot<Lava Hot]
3	[Eight, Four, Two, Six, Twelve]	RoBERTa-large	[Two < Four< Six< Eight < Twelve]	[Eight<Four<Six< Twelve<Two]
4	[Low, Medium, High]	RoBERTa-large	[Low < Medium < High]	[High < Low < Medium]
6	[Leukemia, Cancer, Melanoma]	RoBERTa-large	N/A	[Cancer < Leukemia < Melanoma]
7	[Leukemia, Cancer, Melanoma]	BioClinical BERT	[Melanoma < Leukemia < Cancer]	[Cancer < Leukemia < Melanoma]
8	[優れた, 貧しい, 良い]	Japanese BERT-MLM	[貧しい < 良い < 優れた]	[優れた < 良い < 貧しい]
9	[Muy Buena, Normal, Buena]	Spanish BERT-MLM	[Normal < Buena < Muy Buena]	[Buena < Muy Buena < Normal]
10	[差, 好, 优秀]	Chinese BERT-WWM	[优秀 < 好 < 差]	[优秀 < 好 < 差]

AutoML Evaluation

Overall average F1 score and average Accuracy score performance of 8 original data sets, and its 4 other methods of ordinal value encoders on 4 AutoML platforms of AutoGluon, FLAML, H2O, and MLJAR with 4 different randomization experiences

Method	F1 Score	Accuracy Score
Encoded BERT	0.520	0.728
OrdinalEncoder	0.615	0.764
Original	0.625	0.769
BERT-Sort	0.636	0.784
Human Annotation	0.637	0.785

Performance of 11 ML algorithms on encoded the original UCI Car Evaluation data set through: i) BERT-Sort Encoder and ii) OrdinalEncoder

