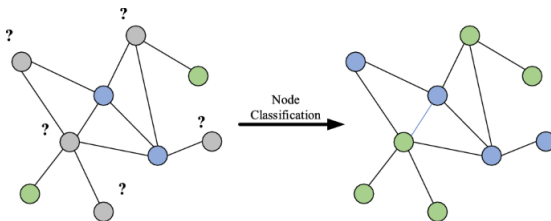
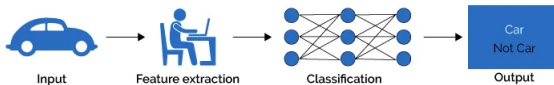


Feature engineering w problemie klasyfikacji wierzchołków dla danych grafowych

Sebastian Zając

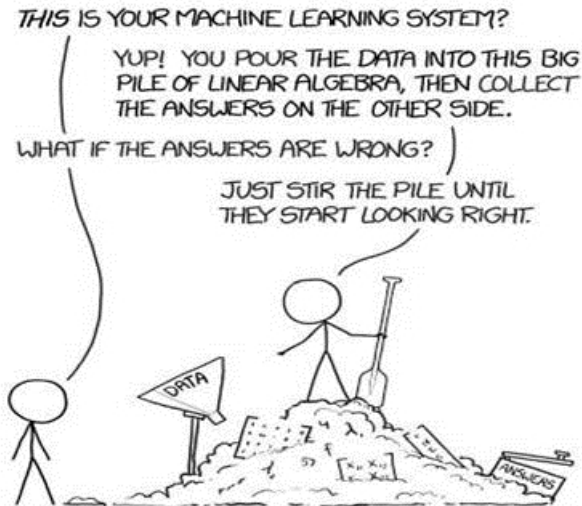
AI SPRING. SZTUCZNA INTELIGENCJA W NAUKACH EKONOMICZNYCH
08.05.2024

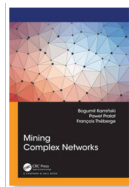
Machine Learning



Cel: przypisanie etykiety dla wierzchołków.

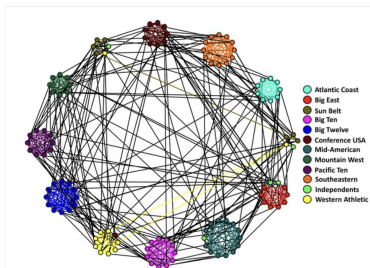
Zastosowania: systemy rekomendacyjne, wskazywanie nowych kontaktów, analizy sieci społecznościowych ...





oznaczenie	nazwa
lcc	local clustering coefficient
bc	betweenness centrality
cc	closeness centrality
dc	degree centrality
ndc	average degree centrality of neighbours
ec	eigenvector centrality
eccen	node eccentricity
core	node coreness
n2v	16-dim node2vec embedding
s2v	16-dim struc2vec embedding

Podział grafu na community



Identyfikacja communities w sieci realizowana może być w sposób nienadzorowany. Wykonywana często jako EDA.

Rozpatrujemy zmienne, które skupiają swoją uwagę na strukturze communities.

Community features literatura:

Podział $A = \{A_1, A_2, \dots, A_\ell\}$ zbioru wierzchołków V na ℓ communities.

Każda część A_i ($i \in [\ell]$) jest gęsta w porównaniu do gęstości całego grafu.

$\deg_{A_i}(v)$: liczba sąsiadów wierzchołka v w A_i

Participation coefficient wierzchołka v

$$p(v) = 1 - \sum_{i=1}^{\ell} \left(\frac{\deg_{A_i}(v)}{\deg(v)} \right)^2$$

Anomaly score CADA

$$cd(v) = \frac{\deg(v)}{d_A(v)}$$

gdzie $d_A(v)$ max liczba sąsiadów należących do tego samego community. T.J. Helling et al, A community-aware approach for identifying node anomalies in complex networks.2019

Znormalizowany anomaly score wierzchołka v

$$\overline{\text{cd}}(v) = \frac{\text{deg}_{A_i}(v)}{\text{deg}(v)}$$

Normalized within-module degree wierzchołka v

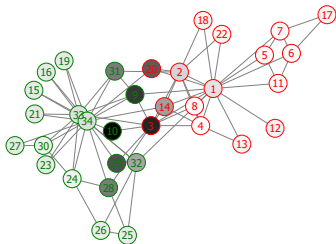
$$z(v) = \frac{\text{deg}_{A_i}(v) - \mu(v)}{\sigma(v)}$$

gdzie $\mu(v)$ i $\sigma(v)$ to , średnia i odchylenie $\text{deg}_{A_i}(u)$ ze wszystkich u wewnątrz community do którego należy v .

Community Association Strength

Dla każdego wierzchołka $v \in A_i$, definiujemy *the community association strength* jako:

$$\beta^*(v) = 2 \left(\frac{\deg_{A_i}(v)}{\deg(v)} - \lambda \frac{\text{vol}(A_i) - \deg(v)}{\text{vol}(V)} \right).$$



Mniejsza wartość $\beta^*(v)$, mniejsza przynależność wierzchołka v do community.

Community-aware node features w eksperymentem

abbreviation	symbol	name
CADA	$cd(v)$	anomaly score CADA
CADA*	$\overline{cd}(v)$	normalized anomaly score
WMD	$z(v)$	normalized within-module degree
CPC	$p(v)$	participation coefficient
CAS	$\beta^*(v)$	community association strength
CD_L11	$L_1^1(v)$	L^1 norm for the 1st neighbourhood
CD_L21	$L_1^2(v)$	L^2 norm for the 1st neighbourhood
CD_KL1	$kl_1(v)$	Kullback–Leibler div. 1st neigh.
CD_HD1	$h_1(v)$	Hellinger distance for the 1st neighbourhood
CD_L12	$L_2^1(v)$	L^1 norm for the 2nd neighbourhood
CD_L22	$L_2^2(v)$	L^2 norm for the 2nd neighbourhood
CD_KL2	$kl_2(v)$	Kullback–Leibler div. for the 2nd neighb.
CD_HD2	$h_2(v)$	Hellinger distance for the 2nd neighbourhood

Syntetyczne grafy ABCD + o

Artificial Benchmark for **C**ommunity **D**etection with **O**utliers

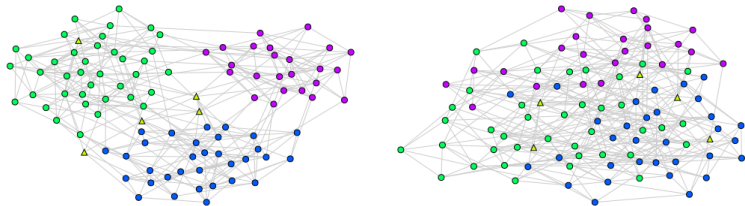
4 sieci z różnym poziomem szumu: $\chi \in \{0.3, 0.4, 0.5, 0.6\}$

N = 10,000 wierzchołków z 1,000 outlierów.

Empirical Graphs

zbiór	# nodes	śr. deg	# liczba klastrów	target
Reddit	10,980	14.30	12	3.661%
Grid	13,478	2.51	78	0.861%
LastFM	7,624	7.29	28	20.619%
Facebook	22,470	15.20	58	25.670%
Amazon	9,314	37.49	39	8.601%

The Artificial Benchmark for the Community Detection graph is a random model with community structure and power-law distribution for degrees and community sizes. It has been recently augmented to allow for the generation of outlier nodes (ABCD+o).



ABCD+o graphs with ($\xi = 0.2$, left) and ($\xi = 0.4$, right)

See also: ABCD graph generator in Julia programming language -

<https://github.com/bkamins/ABCDGraphGenerator.jl>

B. Kamiński, P. Prałat, F. Thériège: „*Mining Complex Networks*”, CRC Press (2022) or *Outliers in the ABCD Random Graph Model with Community Structure (ABCD+o)*.

Eksperyment 1

Wszystkie reprodukowalne kody: <https://github.com/sebkaz/BetaStar>

Pytanie: czy nowe zmienne mogą być modelowane klasycznymi zmiennymi?

Modele: Regresja

Linear regression, Ridge regression, Random forest, XGBoost, Lightgbm.

Miary weryfikacyjne:

Kendall correlation, spearman correlation, and R^2 score.

Wyniki 1 - ABCD +o

target	$\xi = 0.3$	$\xi = 0.4$	$\xi = 0.5$	$\xi = 0.6$
CADA	0.3305	0.2541	0.2292	0.1766
CADA*	0.3613	0.2877	0.2772	0.1713
CPC	0.3540	0.3568	0.3231	0.3106
CAS	0.4205	0.3584	0.3138	0.2167
CD_L21	0.4539	0.4043	0.3823	0.3313
CD_L22	0.6265	0.5589	0.5009	0.4492
CD_L11	0.5935	0.5571	0.5834	0.5648
CD_L12	0.6503	0.5799	0.5464	0.5188
CD_KL1	0.6991	0.6411	0.5918	0.4929
CD_HD1	0.6809	0.6334	0.6170	0.5584
CD_KL2	0.7453	0.6602	0.6090	0.5471
CD_HD2	0.7546	0.7119	0.6815	0.6352
WMD	0.7670	0.7288	0.6915	0.6387

Wyniki 1 - Rzeczywiste grafy

target	Amazon	Facebook	Grid	LastFM	Reddit
CADA	0.5830	0.5666	0.2156	0.4815	0.6826
CADA*	0.6058	0.5828	0.2174	0.5058	0.6867
CPC	0.6338	0.5992	0.2193	0.5175	0.7193
CAS	0.6538	0.6257	0.2999	0.5594	0.7306
CD_L21	0.7052	0.6464	0.3496	0.5698	0.7574
CD_L22	0.7554	0.7355	0.3557	0.6295	0.7941
CD_L11	0.7251	0.7041	0.6978	0.6220	0.7735
CD_L12	0.7794	0.7785	0.6447	0.6884	0.7810
CD_KL1	0.7176	0.7516	0.7394	0.6289	0.7755
CD_HD1	0.7383	0.7482	0.7168	0.6459	0.7853
CD_KL2	0.7706	0.7826	0.7292	0.6853	0.8097
CD_HD2	0.8212	0.8173	0.6930	0.7369	0.8221
WMD	0.8447	0.8456	0.8488	0.8531	0.7638

Eksperyment 2: jakość przewidywania zmiennej celu.

Weryfikacja użyteczności nowych zmiennych w procesie klasyfikacji z określonym targetem.

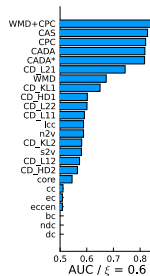
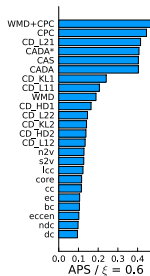
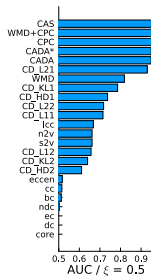
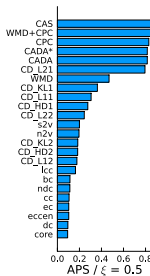
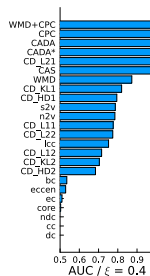
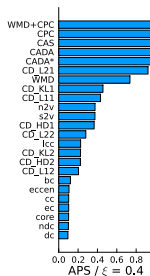
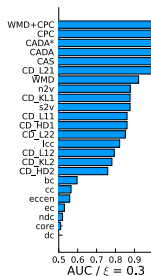
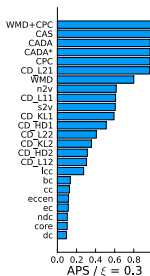
Models: Klasyfikacja binarnej zmiennej

Logistic regression, Random forest, XGBoost, Lightgbm.

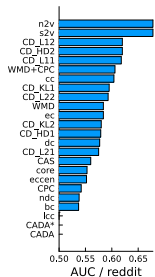
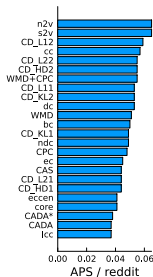
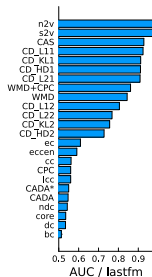
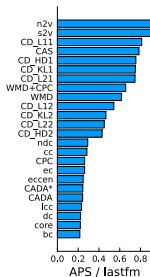
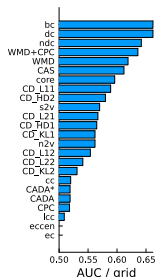
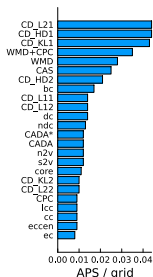
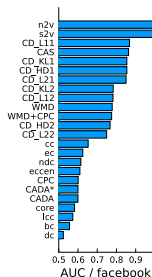
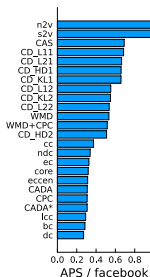
Miary jakości modelu:

ROC AUC score, Average Precision Score.

ABCD + o Wyniki



Wyniki dane empiryczne



Najważniejsze wyniki:

Zweryfikowano:

- 1 nowe zmienne bazujące na informacji o community niosą informacje, których nie da się odtworzyć na podstawie standardowych zmiennych lub wykorzystując zanurzenia.
- 2 Utworzenie nowych zmiennych bazujących na community, dla dużych grafów ($> 100K$) jest mniej czasochłonne niż korzystanie z metod embeddingów.
- 3 Istnieją procesy klasyfikacyjne wierzchołków dla których zastosowanie nowych zmiennych opartych o community poprawia jakość predykcyjną modeli.

Dziękuję za uwagę!
sebastian.zajac@sgh.waw.pl