Метод автоматизованої класифікації текстових даних на основі гібридних моделей

Виконав: Бегерський М.В.
Науковий керівник: к.т.н, доцент Заболотня Т.М.

Мета

Створити новий алгоритм побудови прогностичної моделі, що буде демонструвати точність передбачення не меншу, ніж аналогічні моделі для схожого роду вхідних даних, та мати просту реалізацію.

Об’єкт та предмет дослідження

Об’єкт дослідження:

Процес побудови алгоритму універсальної прогностичної моделі

Предмет дослідження:

Методи побудови прогностичних моделей та алгоритми класифікації даних

Термінологія

Прогностична модель - набір математичних методів, що використовуються статистику для передбачення майбутніх значень досліджуваної величини.

Лінійна регресія - підхід в математичній статистиці для побудови зв’язків між скалярною залежною величиною та однією чи більше додаткових незалежних величин

Критерії оцінки

- швидкість виконання алгоритму (виконання CPU-інструкцій на рядок вхідних даних таблиці)
- часова складність алгоритму
- простота реалізації (Cyclomatic complexity)
- точність передбачення

Існуючі підходи

Використання алгоритмів класифікації

- SVM
- Elastic Net classifier
- XGBoost
- Regularized Logistic Regression

Підбір оптимального алгоритму методом “trial and error” **або з урахуванням доменної галузі**

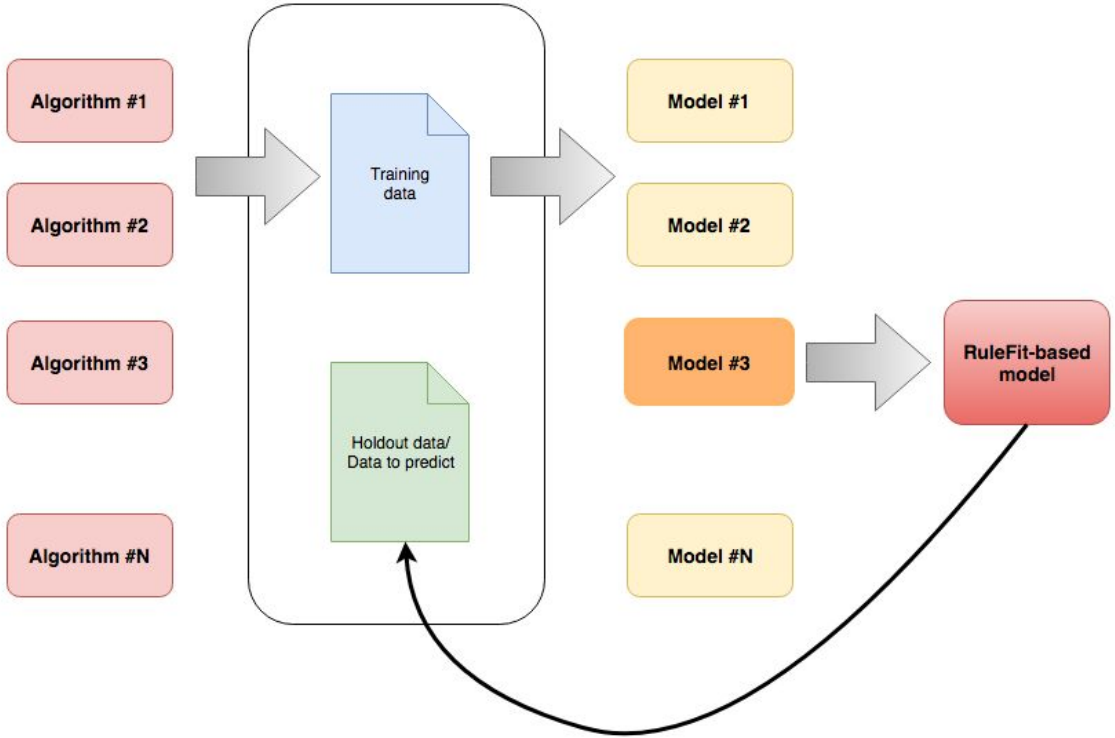
Недоліки існуючих реалізаціи

- високий рівень “ізольованості” внутрішньої реалізації
- висока алгоритмічна складність під час обчислення
- високий рівень використання обчислювальних ресурсів під час запуску на великих об’ємах даних
- ручне втручання для оптимізації роботи

Запропоновании підхід

1. Використати алгоритм, що має найкращі показники для заданого набору вхідних даних
2. Здійснити тренування `RuleFit` моделі на основі вихідних даних цього алгоритму
3. Створити модель, використовуючи результати попереднього кроку
4. Надалі для всіх вхідних даних для передбачень використовувати створену модель

Візуалізація алгоритму



Приклад вхідних даних (imdb movie dataset)

B	C	D	E	F	H	I	J	K	L	M	N	Q	S	T	U	V	W	X	Y	Z
director_nan	num_critic_f	duration	director_face	actor_3_face	actor_1_face	gross	genres	actor_1_nam	movie_title	num_voted_	cast_total_fa	plot_keywor	num_user_for	language	country	content_ratii	budget	title_year	actor_2_face	imdb_score
James Camei	723	178	0	855	1000	760505847	Action|Adve	CCH Pounde	AvatarÂ	886204	4834	avatar|futur	3054	English	USA	PG-13	237000000	2009	936	7.9
Gore Verbins	302	169	563	1000	40000	309404152	Action|Adve	Johnny Depp	Pirates of the	471220	48350	goddess|mai	1238	English	USA	PG-13	300000000	2007	5000	7.1
Sam Mendes	602	148	0	161	11000	200074175	Action|Adve	Christoph W	SpectreÂ	275868	11700	bomb|espio	994	English	UK	PG-13	245000000	2015	393	6.8
Christopher I	813	164	22000	23000	27000	448130642	Action|Thrill	Tom Hardy	The Dark Kni	1144337	106759	deception|in	2701	English	USA	PG-13	250000000	2012	23000	8.5
Doug Walker			131		131		Documentar	Doug Walker	Star Wars: Ep	8	143								12	7.1
Andrew Stan	462	132	475	530	640	73058679	Action|Adve	Daryl Sabara	John CarterÂ	212204	1873	alien|americ	738	English	USA	PG-13	263700000	2012	632	6.6
Sam Raimi	392	156	0	4000	24000	336530303	Action|Adve	J.K. Simmons	Spider-Man 3	383056	46055	sandman|sp	1902	English	USA	PG-13	258000000	2007	11000	6.2
Nathan Gren	324	100	15	284	799	200807262	Adventure|A	Brad Garrett	TangledÂ	294810	2036	17th century	387	English	USA	PG	260000000	2010	553	7.8
Joss Whedor	635	141	0	19000	26000	458991599	Action|Adve	Chris Hemsw	Avengers: Ag	462669	92000	artificial inte	1117	English	USA	PG-13	250000000	2015	21000	7.5
David Yates	375	153	282	10000	25000	301956980	Adventure|F	Alan Rickmar	Harry Potter	321795	58753	blood|book|	973	English	UK	PG	250000000	2009	11000	7.5
Zack Snyder	673	183	0	2000	15000	330249062	Action|Adve	Henry Cavill	Batman v Su	371639	24450	based on cor	3018	English	USA	PG-13	250000000	2016	4000	6.9
Bryan Singer	434	169	0	903	18000	200069408	Action|Adve	Kevin Spacey	Superman Re	240396	29991	crystal|epic|	2367	English	USA	PG-13	209000000	2006	10000	6.1
Marc Forster	403	106	395	393	451	168368427	Action|Adve	Giancarlo Gia	Quantum of	330784	2023	action hero|	1243	English	UK	PG-13	200000000	2008	412	6.7
Gore Verbins	313	151	563	1000	40000	423032628	Action|Adve	Johnny Depp	Pirates of the	522040	48486	box office hit	1832	English	USA	PG-13	225000000	2006	5000	7.3
Gore Verbins	450	150	563	1000	40000	89289910	Action|Adve	Johnny Depp	The Lone Rar	181792	45757	horse|outlav	711	English	USA	PG-13	215000000	2013	2000	6.5
Zack Snyder	733	143	0	748	15000	291021565	Action|Adve	Henry Cavill	Man of Steel	548573	20495	based on cor	2536	English	USA	PG-13	225000000	2013	3000	7.2
Andrew Adar	258	150	80	201	22000	141614023	Action|Adve	Peter Dinklag	The Chronicl	149922	22697	brother brot	438	English	USA	PG	225000000	2008	216	6.6
Joss Whedor	703	173	0	19000	26000	623279547	Action|Adve	Chris Hemsw	The Avenger	995415	87697	alien invasio	1722	English	USA	PG-13	220000000	2012	21000	8.1
Rob Marshal	448	136	252	1000	40000	241063875	Action|Adve	Johnny Depp	Pirates of the	370704	54083	blackbeard|c	484	English	USA	PG-13	250000000	2011	11000	6.7
Barry Sonner	451	106	188	718	10000	179020854	Action|Adve	Will Smith	Men in Black	268154	12572	alien|crimina	341	English	USA	PG-13	225000000	2012	816	6.8
Peter Jackso	422	164	0	773	5000	255108370	Adventure|F	Aidan Turner	The Hobbit: T	354228	9152	army|elf|ho	802	English	New Zealand	PG-13	250000000	2014	972	7.5
Marc Webb	599	153	464	963	15000	262030663	Action|Adve	Emma Stone	The Amazing	451803	28489	lizard|outcas	1225	English	USA	PG-13	230000000	2012	10000	7
Ridley Scott	343	156	0	738	891	105219735	Action|Adve	Mark Addy	Robin HoodÂ	211765	3244	1190s|arche	546	English	USA	PG-13	200000000	2010	882	6.7
Peter Jackso	509	186	0	773	5000	258355354	Adventure|F	Aidan Turner	The Hobbit: T	483540	9152	dwarf|elf|lal	951	English	USA	PG-13	225000000	2013	972	7.9
Chris Weitz	251	113	129	1000	16000	70083519	Adventure|F	Christopher I	The Golden C	149019	24106	children|epi	666	English	USA	PG-13	180000000	2007	6000	6.1
Peter Jackso	446	201	0	84	6000	218051260	Action|Adve	Naomi Watts	King KongÂ	316018	7123	animal name	2618	English	New Zealand	PG-13	207000000	2005	919	7.2
James Camei	315	194	0	794	29000	658672302	Drama|Rom	Leonardo Di	TitanicÂ	793059	45223	artist|love|s	2528	English	USA	PG-13	200000000	1997	14000	7.7
Anthony Rus	516	147	94	11000	21000	407197282	Action|Adve	Robert Dowr	Captain Ame	272670	64798	based on cor	1022	English	USA	PG-13	250000000	2016	19000	8.2
Peter Berg	377	131	532	627	14000	65173160	Action|Adve	Liam Neeson	BattleshipÂ	202382	26679	box office flo	751	English	USA	PG-13	209000000	2012	10000	5.9
Colin Trevorr	644	124	365	1000	3000	652177271	Action|Adve	Bryce Dallas	Jurassic Wor	418214	8458	dinosaur|dis	1290	English	USA	PG-13	150000000	2015	2000	7
Sam Mendes	750	143	0	393	883	304360277	Action|Adve	Albert Finney	SkyfallÂ	522030	2039	brawl|childh	1498	English	UK	PG-13	200000000	2012	563	7.8
Sam Raimi	300	135	0	4000	24000	373377893	Action|Adve	J.K. Simmons	Spider-Man 3	411164	43388	death|docto	1303	English	USA	PG-13	200000000	2004	11000	7.3
Shane Black	608	195	1000	3000	21000	408992272	Action|Adve	Robert Dowr	Iron Man 3Â	557489	30426	armor|explo	1187	English	USA	PG-13	200000000	2013	4000	7.2
Tim Burton	451	108	13000	11000	40000	334185206	Adventure|F	Johnny Depp	Alice in Won	306320	79957	alice in wonc	736	English	USA	PG	200000000	2010	25000	6.5
Brett Ratner	334	104	420	560	20000	234360014	Action|Adve	Hugh Jackma	X-Men: The L	383427	21714	battle|mutai	1912	English	Canada	PG-13	210000000	2006	808	6.8
Dan Scanlon	376	104	37	760	12000	268488329	Adventure|A	Steve Buscer	Monsters Un	235025	14863	cheating|frat	265	English	USA	G	200000000	2013	779	7.3
Michael Bay	366	150	0	464	894	402076689	Action|Adve	Glenn Morsh	Transformers	323207	3218	autobot|dec	1439	English	USA	PG-13	200000000	2009	581	6
Michael Bay	378	165	0	808	974	245428137	Action|Adve	Bingbing Li	Transformers	242420	3988	blockbuster|	918	English	USA	PG-13	210000000	2014	956	5.7
Sam Raimi	525	130	0	11000	44000	234903076	Adventure|F	Tim Holmes	Oz the Great	175409	73441	circus|magic	511	English	USA	PG	215000000	2013	15000	6.4
Marc Webb	495	142	464	825	15000	202853933	Action|Adve	Emma Stone	The Amazing	321227	28631	costumed he	1067	English	USA	PG-13	200000000	2014	10000	6.7

Наукова новизна

- Універсальність підходу
- Використання будь-якого алгоритму (моделі) чи їх композиції
- Підвищення швидкодії за рахунок однократного запуску базового алгоритму
- Використання гомогенних інструкцій CPU

Результати

- Часові показники використання CPU (менше 1 мс на рядок вхідних даних)
- Складність O(M*N) - M (кількість рядків вхідних даних), N (кількість порівнянь)
- Відхилення від еталонної моделі ~0.001-0.1%

Model Name	Validation	Cross Validation	Holdout
Nystroem Kernel SVM Classifier	0.6075	0.6109	0.6120
RuleFit-based model	0.6065	0.6106	0.6120

Використане ПЗ

- Python3
- NumPy, Pandas
- scikit-learn
- mccabe/radon/flake8
- Kaggle datasets
- time/perf

Висновки

1. Проаналізовано існуючі алгоритми класифікації даних та методи побудови моделей на основі них
2. Запропоновано алгоритм побудови універсальної прогностичної моделі
3. Здійснено порівняння та підтвердження ефективності запропонованого методу

Дякую за увагу