ACM2

АКТУАЛНИ РЕГУЛАТИВНИ ИЗИСКВАНИЯ КЪМ ХИМИЧНИЯ АНАЛИЗ И МОДЕРНИ ИНСТРУМЕНТАЛНИ ТЕХНОЛОГИИ ЗА ТЯХНОТО ПОКРИВАНЕ
12 Юни 2019, ПЛОВДИВ

Химически Факултет

МИКРОБНА МУТАГЕННОСТ. ПРОУЧВАТЕЛЕН АНАЛИЗ НА КОМПЮТЪРНИ ПОДХОДИ ЗА КЛАСИФИКАЦИЯ.

J. Statieva, V. Paskaleva

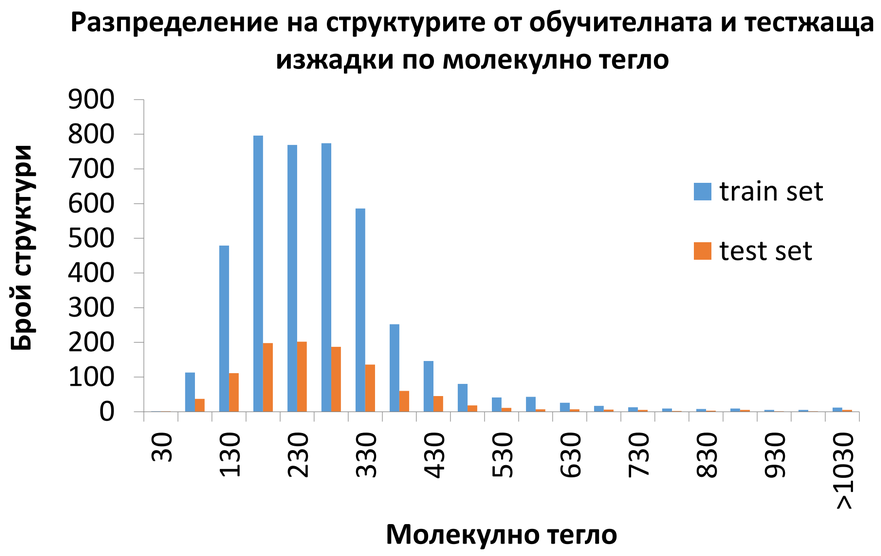
University of Plovdiv, Department of Analytical Chemistry and Computer Chemistry, 4000, Plovdiv, Bulgaria.

Микробната мутагенност е важна харектеристика, която носи токсилогична информация и е необходима при регистрирането на нови съединеня по регламента на REACH. Също така ICH изисква информация за канцерогенните и мутагенните свойства за „Оценка на опасността“ (Hazard Assessment) на фармацевтичните продукти, а при липса на експериментални данни, препоръчват използването на (Q)SAR методи за предсказване на микробната мутагенност. В настоящата работа представяме извършения проучвателен анализ на компютърните подходи за класифициране на химичните съединения спрямо микробната им мутагенност. Извършеният анализ включва подготовка и описание на работните извадки, построяване, валидиране и тестване на множество модели, анализ на резултатите и избор на модели за последваща оптимизация.

Структурни извадки

- Benchmark data set
- Брой структури – 5232 (Активни – 2530; Неактивни – 2702)

Структурната извадка е разделена на обучителна и тестваща извадки в съотношение 80:20.



Изчисляване на молекулни дескриптори и фингърпринти

Фингърпринти:

- PaDEL-Descriptor v.2.21
- ФП1 – Fingerprinter, PubchemFingerprinter, SubstructureFingerprinter, KlekotaRothFingerprinter
- ФП2 - Fingerprinter, ExtendedFingerprinter, EStateFingerprinter, GraphOnlyFingerprinter, MACCSFingerprinter, PubchemFingerprinter, SubstructureFingerprinter, KlekotaRothFingerprinter

Дескриптори:

- Dragon v.7
- Брой изчислени молекулни дескриптора – 2498 (МД1, МД2)

Проучвателен анализ

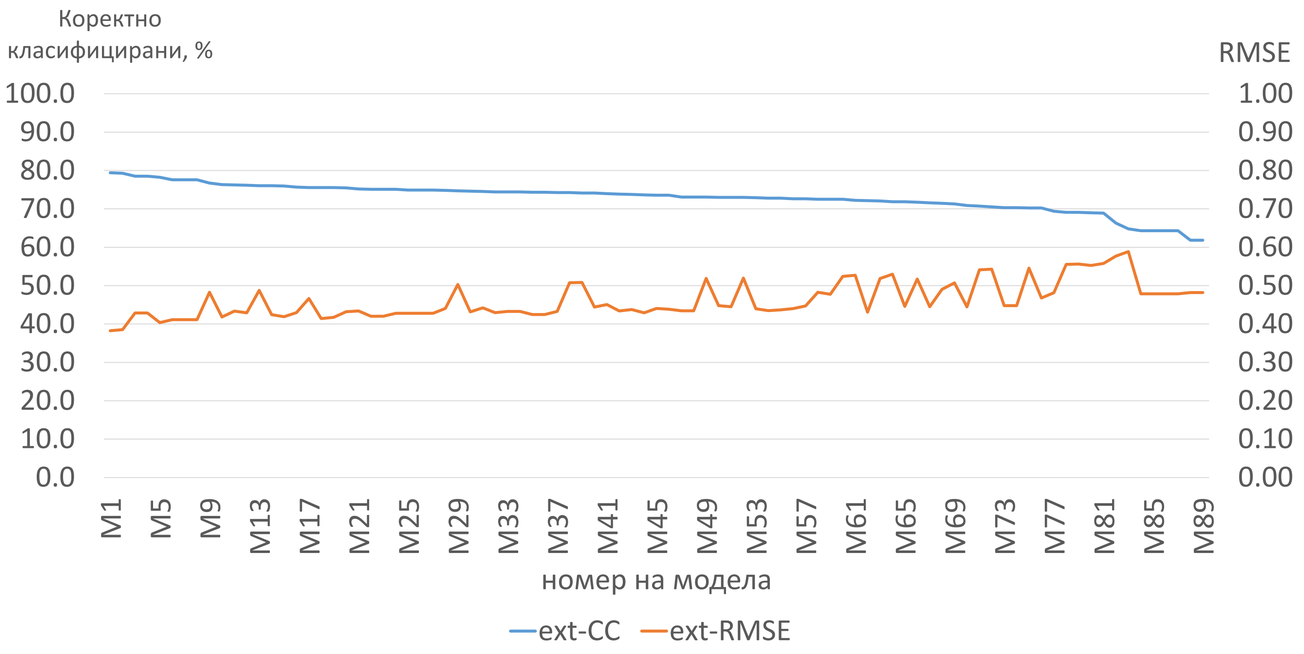
За построяване на моделите е използван софтуерът със свободен код Weka v.3.8.2 В Проучвателния анализ са използвани 14 метода за класифициране. Структурната информация е кодирана чрез използване на фингърпринти и молекулни дескриптори: две комбинации фингърпринти (ФП1, ФП2), молекулни дескрптори (МД1, МД2). В таблицата е представена схема за проведения проучвателен анализ.

	Кодиране на СИ	ФП-1		ФП-2		МД-1	МД-2
	Метод за избор на дескриптори	По подразбиране	CAE+Ranker	По подразбиране	CAE+Ranker	По подразбиране	По подразбиране
Класификационни алгоритми	BFTree	+		+		+	+
	CDT	+	+	+	+	+	+
	DecisionStump	+	+	+	+	+	+
	Ibk	+	+	+	+	+	+
	J48	+	+	+	+	+	+
	KernelLogisticRegression	+	+	+	+	+	+
	LibSVM	+	+	+	+	+	+
	Logistic	+	+			+	+
	MLPClassifier	+	+	+		+	+
	NaiveBayes	+	+	+	+	+	+
	RandomForest	+	+	+	+	+	+
	RandomTree	+	+	+	+	+	+
	REPTree	+	+	+	+	+	+
	SMO	+	+	+	+	+	+

Легенда:
СИ – структурната информация
ФП-1 – Fingerprinter, PubchemFingerprinter, SubstructureFingerprinter, KlekotaRothFingerprinter
ФП-2 – Fingerprinter, ExtendedFingerprinter, EStateFingerprinter, GraphOnlyFingerprinter, MACCSFingerprinter, PubchemFingerprinter, SubstructureFingerprinter, KlekotaRothFingerprinter
МД-1 – NormalizedDescr_Binned
МД-2 – numeric_Binned

Извадка от най-добрите получени модели при провеждане на проучвателния анализ.

N	D/F	Метод за избор на дескриптори	Класификатори	train-CC	train-MAE	train-RMSE	cv5-CC	cv5-MAE	cv5-RMSE	ext-CC	ext-MAE	ext-RMSE
M1	ФП-2	ClassifierAttributeEval+Ranker	RandomForest	99.904	0.113	0.143	79.230	0.316	0.388	79.389	0.311	0.383
M2	**ФП-1**	**ClassifierAttributeEval+Ranker**	**RandomForest**	**99.904**	**0.116**	**0.146**	**78.800**	**0.324**	**0.392**	**79.294**	**0.317**	**0.385**
M3	ФП-1	ClassifierAttributeEval+Ranker	MLPClassifier	96.893	0.057	0.169	74.689	0.266	0.456	78.531	0.240	0.429
M4	ФП-1	ClassifierAttributeEval+Ranker	J48	91.659	0.136	0.261	72.275	0.299	0.488	78.531	0.240	0.429
M5	ФП-2	По подразбиране	RandomForest	91.205	0.175	0.262	76.052	0.309	0.411	78.244	0.300	0.403
M9	ФП-2	ClassifierAttributeEval+Ranker	Ibk	99.904	0.001	0.022	76.052	0.239	0.488	76.718	0.235	0.483
M11	ФП-2	По подразбиране	kNN	91.205	0.116	0.241	75.645	0.283	0.441	76.241	0.271	0.434



Оптимизиране на модел базиран на метода на най-близки съседи

Брой съседи: 65, 33, избрани чрез крос-валидиране
Метрики за разстояние: Евклидово, Манхатан и др.
Теглова схема по разстоянието: без тегло, 1/разстоянието, 1- разстоянието

Модел 1 – метод за избор на дескриптори По подразбиране; брой съседи 3 (избрани чрез крос-валидиране); Евклидово разстояние; без теглова схема по разстоянието.

Модел 2 – метод за избор на дескриптори ClassifierAttributeEval Ranker; ; брой съседи 5 (избрани чрез крос-валидиране); Манхатан; 1/разстояние.

Модел 3 – метод за избор на дескриптори ClassifierAttributeEval Ranker; ; брой съседи 5 (избрани чрез крос-валидиране); Евклидово разстояние; 1/разстояние.

Модел 4 – метод за избор на дескриптори По подразбиране; брой съседи 3 (избрани чрез крос-валидиране); Манхатан; 1/разстояние.

Модел 5 – метод за избор на дескриптори По подразбиране; брой съседи 3 (избрани чрез крос-валидиране); Евклидово разстояние; 1/разстояние.

	Обучение			**Валидиране**			**Тестване**		
Модел	CC	MAE	RMSE	CC	MAE	RMSE	CC	MAE	RMSE
Модел 1	87.3805	0.1683	0.29	74.7132	0.2967	0.4474	74.5229	0.2971	0.4423
Модел 2	99.9044	0.0032	0.0226	77.2467	0.2896	0.4071	76.5267	0.2836	0.4052
Модел 3	99.9044	0.0128	0.0289	76.0516	0.3025	0.4122	76.3359	0.2902	0.4067
Модел 4	87.1893	0.1796	0.2936	75.3824	0.3056	0.4282	75.0954	0.3048	0.4273
Модел 5	87.1893	0.1801	0.2936	75.1434	0.3057	0.4296	75	0.3042	0.4276

Acknowledgements: project MU19-HF-003