УДК 519.68

ИСПОЛЬЗОВАНИЕ РАСШИРЕННЫХ ВАЛЕНТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ (МОЛЕКУЛЯРНЫХ ПОДПИСЕЙ) В QSAR И QSPR ИССЛЕДОВАНИЯХ

Адылова Ф.Т.
д.т.н., профессор, зав. лабораторией «Мединформатика»,
Институт математики им. В.И. Романовского Академии наук РУз,
fatima_adilova@rambler.ru

Икрамов А.А.
младший научный сотрудник,
Национальный университет Узбекистана им. Мирзо Улугбека,
ikramov.alisher@list.ru

Молекулярный дескриптор позволяет кодировать химическое соединение так, чтобы задачи моделирования отношения «структура-активность» QSAR и «структура-свойство» QSPR были решены эффективно. Дескриптор молекулярной подписи является одним из немногих молекулярных дескрипторов, для которых была эффективно решена детерминированная обратная задача. Одним из этапов реализации последней является решение прямой задачи QSPR. В данной работе исследуется точность моделей SVM, построенных на дескрипторе подписи, в решении задачи QSAR. Для этого было проведено 5 вычислительных экспериментов с различными условиями в постановках задач: тип химического пространства, релевантность подписей, масштабирование выборки, взаимосвязь качества классификации на SVM с разными ядрами и коэффициентом моделеспособности выборки.

Ключевые слова: структура-активность, машинное обучение, моделеспособность, молекулярные подписи.

USE OF EXTENDED VALENCE SEQUENCES (MOLECULAR SIGNATURES) IN QSAR AND QSPR

Adilova F.T., Ikramov A.A.

A molecular descriptor allows the chemical compound to be encoded so that the tasks of modeling the QSAR structure-activity relationship and the QSPR structure-property have been effectively solved. The molecular signature descriptor is one of the few molecular descriptors for which the deterministic inverse problem solved effectively. One of the stages of implementing the inverse task is the solution of the direct QSPR problem. In this paper, we investigate the accuracy of the SVM models built on the signature descriptor in solving the QSAR problem. For this purpose, five computational experiments carried out with different conditions in the formulation of tasks: the type of chemical space, the relevance of signatures, the scaling of the sample, the relationship of the quality of classification to SVM with different kernel and the coefficient of sample modelability.

Keywords: structure-activity, machine learning, modelability, molecular signatures.

QSAR VA QSPR TADQIQODLARDA (MOLEKULYAR IMZOLAR) KENGAYTIRILGAN VALENTLI KETMA-KETLIKLARNI QO'LLASH

Adilova F.T., Ikramov A.A.

Molekulyar deskriptor kimyoviy birikmani kodlashni ta'minlaydi, shuning uchun QSAR "struktura-aktivlik" munosabatlarini modellashtirish va QSPR "struktura-xususiyat" samarali ravishda hal qilish vazifasi qo'yildi. Molekulyar imzo deskriptori teskari muammoni samarali echgan bir nechta molekulyar deskriptorlardan biridir. Ularni amalga oshirish bosqichlaridan biri QSPR muammosini echishidir. Ushbu maqolada, biz QSAR muammosini hal qilishda imzo deskriptorlariga qurilgan SVM modellarining aniqligini tekshiramiz. Buning uchun 5 ta hisoblash tajribalari turli vazifalar bilan bajarildi: kimyoviy fazoning turi, imzolarning dolzarbligi, tanlanmani mashtablash, SVM uchun turli yadrolar bilan sinflash sifati o'rtasidagi munosabatlar va tanlanmani model qurish qobiliyati koeffitsienti.

Kalit so'zlar: struktura-aktivlik, mashinali o'rganish, model qurish qobiliyati, molekulyar imzolar.

1. Введение

Недавно был представлен новый дескриптор, названный «подпись», построенный на основе расширенной валентной последовательности. Подпись атома представляет собой каноническое представление окружения атома до заранее определенной высоты h. Подпись молекулы представляет собой вектор чисел встречаемости атомных подписей. Создание количественных моделей (QSR) отношений «структура-активность» (QSAR) и «структура-свойство» (QSPR) на новом дескрипторе показывает, что для любой молекулы диаметром D, существует молекулярная подпись высотой $h \leq D+1$, из которой может быть вычислен любой 2D дескриптор. Как следствие этого, любая модель QSAR или QSPR, включающая 2D дескрипторы, может быть построена на дескрипторах числа вхождений атомных подписей.

Представляя двумерную структуру молекулы в виде полного графа с вершинами (атомы) и ребрами (связи), можно применить к нему бесконечное число операторов для того, чтобы охарактеризовать свойства молекулы. Эти операторы, известные как дескрипторы имеют только качественный смысл [1]. Численные значения данного дескриптора на графе используются в качестве независимых переменных, которые прогнозируют различные экспери–ментальные физические свойства (QSPR) или биологическую активность (QSAR).

Хотя в трехмерной структуре молекулы содержится больше информации о пространственных отношениях атомов и связей, чем в 2D графе, топологические, или 2D дескрипторы по отношению к 3D-дескрипторам, как было показано в ряде исследований, содержат больше информации[2]. Число дескрипторов, сегодня доступных для использования, велико, появление же коммерческих пакетов позволяет выбрать подходящие для конкретной задачи [3]. Граф молекулы содержит ограниченную информацию относительно неза–висимых переменных (дескрипторов), но, как правило, многие из дескрипторов в наборе сильно коррелируют. Кроме того, часто нет взаимно-однозначного соответствия между дескриптором и молекулярным свойством [3], и потому окончательный вид QSR становится зависимым от исследователя [4].

В своей работе [4] Randic' и Basak спрашивают: "Разве нам нужны дополнительные дескрипторы, хотя сотни их ... уже доступны ...? Как мы можем определить, достаточно ли имеющихся молекулярных дескрипторов для полной характеристики молекул для QSPR и QSAR? ". Шаг к решению этих важных проблем состоит в генерации *конечного множества дескрипторов*, которые *не сильно коррелированы* и образуют *полный набор*, из которого можно вычислить все остальные дескрипторы. В работе [5] вводится понятие молекулярной подписи и показано, как вывести многие из топологических индексов, используемых в QSR, из этих подписей. Результаты с использованием подписи сравниваются с аналогичными результатами, полученными с использованием 2D дескрипторов, взятых из коммерческого пакета Molconn-Z. В качестве непосредственного применения отношений между подписью и 2D дескрипторами, приводятся результаты простого исследования, которое показывает эквивалентность между моделью QSAR, разработанной с использованием четырех индексов, и с помощью четырех атомных подписей для точки кипения 25 алифатических углеводородов.

Целью настоящей работы является оценка точности моделей SVM на дескрипторах подписи в решении задачи QSPR. Для этого было проведено 5 вычислительных экспериментов с различными условиями в постановках задач.

2. Материал и методы

Дескриптор подписи представляет собой систему кодирования над алфавитом типов атомов, описывающих расширенную валентность (т.е. соседство) атомов молекулы. Эта концепция впервые была применена в контексте структурной трактовки [6], определена для ациклических соединений и использована в анализе QSAR [7].

Молекула представлена в виде графа G= (V_G, E_G, C, C_G ()), где элементы V_G являются атомами, а E_G ребрами. Атомы молекулярного графа окрашены цветами из C, который является набором типов атомов; атомы, например, могут быть элементами периодической таблицы или любого набора типов атомов, предоставляемых силовым полем молекулы. C_G () является функцией, которая связывает атом G и тип атома. Каждый тип атома имеет валентность, которая является числом ковалентных связей, которые могут быть образованы с этим атомом. Молекулярный граф не обязательно является полным. Формально, молекулярный граф G= (V_G, E_G, C, C_G ()) представляет собой неориентированный цветной граф с функцией C_G () над элементами C, удовлетворяющий уравнению:

$$\forall x \in V_G \ \deg(x) \leq valence\big(C_G(x)\big) \qquad (1)$$

Пусть G - молекулярный граф и x - атом G. Подпись высоты h от x, $h_{\sigma_G(x)}$, является каноническим представлением подграфа G, содержащего все атомы, находящиеся на расстоянии h от x. Это каноническое представление принимает форму дерева и строится следующей пятиступенчатой процедурой.

(1) Подграф извлекается из G, содержащего все атомы и все связи между этими атомами, находящимися на расстоянии h от x.

(2) вершины $h_{G(x)}$ маркируются в каноническом порядке, атом x имеет метку 1.

(3) Строится дерево, охватывающее все ребра $h_{G(x)}$.

Корень дерева представляет собой сам атом x. Первый слой дерева состоит из соседей x, а второй слой состоит из соседей вершин первого слоя, за исключением атома x. Каждой вершине, добавленной к дереву, сопоставляется цвет и

каноническая метка соответствующего атома. Процесс продолжается пошагово до уровня h.

Предположим, что дерево было построено до слоя $k < h$, слой $k + 1$ строится с учетом каждой вершины y слоя k. Пусть z - сосед y в G, вершина z и ребро $[y, z]$ добавляются к слою $k + 1$, если ребра $[y, z]$ или $[z, y]$ ещё не присутствуют в дереве. Соседи y, присутствующие в слое $k + 1$, сортируются по убыванию лексикографического порядка, используя цвета и канонические метки соответствующих атомов. Как показано на рис.1, в первом слое C,3 появляется перед C,2. Следует отметить, что данная вершина z может появляться в дереве несколько раз (как C,7 на рис.1), так как это может быть сосед из нескольких вершин, присутствующих в предыдущем слое; однако, согласно процедуре построения, любое ребро появляется только один раз.

4) После того, как дерево построено до слоя h, все канонические метки, которые появляются только один раз, удаляются. Остальные метки перенумеровываются в порядке их появления во время чтения дерева на глубине первого порядка.

(5) Подпись записывается чтением дерева на глубине первого порядка и печатанием символа '(' каждый раз, когда читается ребро родитель- ребенок, и символа ')', когда ребро читается от ребенка к родителю и вершина раскрашивается в цвет в соответствии с меткой, если вершина появляется в дереве несколько раз.

В соответствии с определением, подпись атома можно рассматривать как строку символов в алфавите C типов атомов. Отметим, что для заданной высоты h, список всех возможных атомных подписей, хотя и очень большой, имеет конечный размер.

Следовательно, любая молекула может быть представлена своими координатами в векторном пространстве, где базовые векторы являются различными атомными подписями. Тогда подпись молекулы является линейной комбинацией подписей атомов. Как показано на рис.1, в первом слое, C, 3 выступает в качестве линейной комбинации ее атомных подписей

$$h_{\sigma(G)} = \sum_{x \in V_G} h_{\sigma_G(x)} = {}^{h}\alpha_G^{h} \sum \qquad (2)$$

где ${}^{h}\sum$ является основным множеством всех атомных подписей высотой h, ${}^{h}\alpha_G$ является вектором вхождения числа атомных h-подписей графа G.

Пусть $G = (V_G, E_G, C, c_G())$ является молекулярным графом, и пусть b ребро/вершина E_G. Пусть $G - b = (V_G, E_G \setminus \{b\}, C, c_G())$ является молекулярным графом, в котором ребро b было удалено. Тогда h -подпись ребра b определяется следующим образом:

$$ {}^{h}\sigma(b) = {}^{h}\sigma(G) - {}^{h}\sigma(G - b) \qquad (3)$$

При использовании атомных подписей как дескрипторов, число появлений конкретной атомной подписи используется в качестве значения дескриптора. Тем не менее, общее количество уникальных атомных подписей в наборе данных априори неизвестно, и является функцией высоты h и размера обучающего набора.

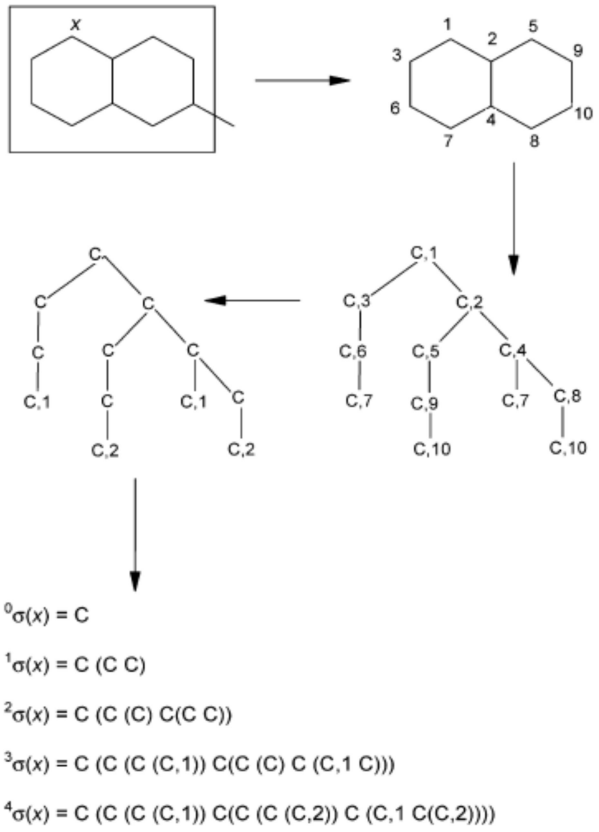

*Рис. 1. Показаны пять этапов процедуры вычисления подписи атома **x** в methylnaphthene*

В данном исследовании высота подписи была зафиксирована $h = 3$.

Материалами для исследований послужили данные из баз данных ChEMBL[8] и VAMMPIRE[9].

3. Результаты и обсуждение

Вычислительный эксперимент 1

Цель – сравнение точности моделей SVM в двух разных химических пространствах. Вначале была использована выборка, состоящая из групп MMP (matched molecular pairs) из 54 соединений, взятая из VAMMPIRE Database. Было создано два файла – в одном дескрипторы ChEMBL, в другом – подписи. Дополнительно был использован набор из 442 соединений из базы ChEMBL. На них также были обучены модели SVM. Сравнительные результаты тестирования созданных моделей на самих же обучающих выборках представлены в таблице 1.

Таблица 1

Сравнительные результаты точности моделей SVM, обученных на дескрипторах из ChEMBL и подписях и проверенных на обучающих выборках

	ChEMBL	Подписи
54 соединения	85%	81%
442 соединения	62%	48%

Полученные данные не дают убедительного доказательства преимущества одного из подходов.

Поэтому было выбрано 1794 соединений из базы ChEMBL, данные (InChi) по которым передали на обработку программному обеспечению (https://sourceforge.net/projects/molsig/), разработанному J.-L. Faulon для построения подписей. После постобработки была создана матрица соединение-подпись.

С помощью случайного перемешивания исходная выборка была разбита на подвыборки по 200 соединений в каждой. После этого каждая подвыборка была записана в два файла – в одном с физическими дескрипторами, в другом с подписями. На каждой из подвыборок была обучена модель SVM и проверена её точность на других подвыборках в качестве тестовых. Результаты представлены в таблице 2.

Таблица 2

Точность моделей SVM (в %), обученных на дескрипторах из ChEMBL и подписях на выборках по 200 соединений в каждой

№ выборки	Дескрипторы ChEMBL			Подписи		
	Минимальное значение точности	Максимальное значение точности	Среднее значение точности	Минимальное значение точности	Максимальное значение точности	Среднее значение точности
1	36.5	50.0	43.2	45.0	58.0	51.9
2	34.5	49.0	43.2	45.5	60.0	52.5
3	43.0	52.0	46.6	52.0	60.5	55.3
4	38.5	51.5	43.4	45.0	61.0	53.2
5	33.5	50.5	41.9	46.5	59.5	53.7
6	42.0	51.5	45.3	46.0	61.0	53.2
7	39.0	51.0	43.7	43.0	58.0	51.5
8	39.0	50.0	42.5	47.5	60.0	54.7

Вычислительный эксперимент показал значительную разницу (17.9%) между точностью моделей на 19 дескрипторах ChEMBL и на 794 подписях в пользу последних. Однако размерность пространства, в котором строились модели SVM на подписях, значительно превышала количество соединений в выборке, поэтому пришлось уменьшить размерность пространства. Для этого из 1794 соединений последовательным поиском были удалены все соединения, у которых имелись подписи, встречавшиеся в выборке не более 5 раз. Этот процесс повторялся до тех пор, пока не осталось 1358 соединений и 342 подписи, встречавшиеся, по крайней мере, у 6 соединений каждая. Полученная выборка была поделена на 3 подвыборки по 440, 2 подвыборки по 650 и 2 подвыборки по 678 соединений каждая.

Внутри каждого семейства были обучены модели SVM на дескрипторах ChEMBL и подписях и проверена на других подвыборках из того же семейства (с тем же количеством соединений в файле). Результаты представлены в таблице 3.

Таким образом, уменьшение выборки за счёт удаления соединений с редко встречающимися подписями увеличило точность не только модели SVM, обучаемой на подписях, но и модели, обучаемой на дескрипторах ChEMBL. Это свидетельствует о существенных отличиях удаленных соединений от оставшихся. Это означает, что можно использовать подход с подписями для выделения «шума» из выборки с разным порогом (в нашем исследовании этот порог был равен 5). С увеличением числа соединений в выборке возрастает разница в точности моделей SVM (21,98%), обученных на подписях и на дескрипторах ChEMBL.

Таблица 3

Точность моделей SVM на выборках разного размера и разных дескрипторах после уменьшения размерности пространства подписей

Семейство	№ обучающей/ № тестовой	ChEMBL	Подписи
440 соединений	1/2	47.7%	61.8%
	1/3	48.9%	63.6%
	2/1	50.5%	60.2%
	2/3	49.8%	65.5%
	3/1	51.1%	63.2%
	3/2	51.1%	63.2%
650 соединений	1/2	48.8%	66.8%
	2/1	53.7%	64.9%
678 соединений	1/2	47.2%	65.6%
	2/1	50.3%	66.1%

Следовательно, подход обучения моделей на подписях в качестве дескрипторов даёт лучшие результаты.

Вычислительный эксперимент 2

Цель ВЭ_3,-выделение релевантных подписей. Ранее мы использовали модели, обученные на дескрипторах, предоставляемых ChEMBL. В данном эксперименте оценивается возможность повышения точности решения задачи QSPR за счёт просеивания и удаления из выборки тех соединений, которые относятся к слабо представленным классам.

Было выбрано 1794 соединений из базы ChEMBL. Их данные (InChi) были переданы в программу https://sourceforge.net/projects/molsig/ для построения дескрипторов подписей. Были установлены три различных порога – 5, 9 и 13. Это минимальное

число соединений, у которых должна встретиться подпись высотой 3, чтобы считаться релевантной. Все нерелевантные подписи удалялись из рассмотрения, а вместе с ними все соединения, их содержащие. Таким образом, среди оставшихся соединений каждая подпись была релевантной.

Из 1794 соединений последовательным поиском были удалены все соединения, у которых имелись подписи, встречавшиеся в выборке не более 5 раз. Этот процесс повторялся до тех пор, пока не осталось 1358 соединений и 342 подписи, встречавшиеся по крайней мере у 6 соединений каждая (выборка 1). Процесс был повторён с границей 9,- получено 1004 соединений и 246 подписей (выборка 2), и с границей 13, получено 643 соединения и 151 подпись (выборка 3).

Каждая из выборок была случайным образом перемешена и разделена на парные файлы по 200 соединений в каждом (в паре – одни и те же соединения, только в одном дескрипторами являются данные ChEMBL, в другом – подписи). На каждом файле обучена модель SVM и протестирована на других файлах (с такими же дескрипторами, как и в обучающем). Результаты средних значений точности представлены в таблице 4.

Таблица 4

Точности моделей SVM на выборках по 200 и разных дескрипторах после уменьшения размерности пространства подписей

№ выборки (порог для подписи)	ChEMBL	Подписи
1 (5)	44.7%	55.0%
2 (9)	50.5%	61.2%
3 (13)	59.1%	65.0%

Таким образом, уменьшение выборки за счёт удаления соединений с редко встречаемыми подписями можно рассматривать как подход для выделения «шума» из выборки с разным порогом.

Показана эффективность применения разного порога (с увеличением вырастает и точность), причём на точность SVM по дескрипторам ChEMBL это оказывает даже больший эффект, чем на SVM по дескрипторам подписей.

Следовательно, оба типа дескрипторов можно использовать при построении моделей для последующей классификации. Однако, подписи показывают больший процент правильно определенных значений. Это связано в первую очередь с переобучением – большая размерность пространства признаков (подписей) и не столь большое количество соединений.

Вычислительный эксперимент 3

Целью эксперимента являлось более масштабное исследование зависимости наличия «редких» подписей в базе и точности моделей SVM, прогнозирующих активность соединений. Было взято 23448 соединения из базы ChEMBL (выборка 1), тогда как в ВЭ_1и ВЭ_2 использовали выборку из 1794 соединений. На 23448 соединениях было получено в общей сложности 2157 различных подписей.

Три различные границы были установлены – 10, 30 и 60. Это минимальное число соединений, у которых должна встретиться подпись, чтобы считаться релевантной. Из общего числа соединений последовательным поиском были удалены все соединения, у которых имелись подписи, встречавшиеся в выборке не более 10 раз. Этот процесс повторялся до тех пор, пока не осталось 21153 соединений и 916 подписей, встречавшиеся по крайней мере у 11 соединений каждая (выборка 2). Процесс был повторён с границей 30,-получено 16509 соединений и 546 подписей (выборка 3), и с границей 60,-получено 9349 соединений и 299 подписей (выборка 4).

Каждая из выборок была случайным образом перемешена и разделена на парные файлы по 2000 соединений в каждом (в паре – одни и те же соединения, только в одном дескрипторами являются данные ChEMBL, в другом – подписи). На каждом файле обучена модель SVM и протестирована на других файлах (с такими же дескрипторами, как и в обучающем). Результаты средних значений точности представлены в таблице 5.

Таблица 5

Точности моделей SVM на выборках по 2000 и разных дескрипторах после уменьшения размерности пространства подписей

№ выборки (порог для подписи)	ChEMBL	Подписи
1 (0)	23.4	34.2
2 (10)	22.9	20.6
3 (30)	22.5	20.3
4 (60)	20.4	20.5

Подписи показывают больший процент правильно определенных значений на выборке без отсеивания редко встречающихся подписей. Как видно, определение оптимального значения порога может быть произведено путём перебора различных значений и определения точности моделей SVM.

Вычислительный эксперимент 4

Целью ВЭ было определить взаимосвязь между точностью моделей SVM, построенных на выборках из 2000 соединений, полученных из совокупной выборки путём случайного перемешивания, и коэффициента MODI, вычисленного на совокупной выборке. Отметим, что индекс "моделе-способности" (MODelabIlity, MODI) является количественной мерой быстрой оценки того, можно ли получить модель (и) прогноза для данного набора химических данных [10]. В таблице 6 приведены оценки точности моделей SVM и значения MODI на разных дескрипторах после уменьшения размерности пространства подписей.

Был вычислен квадрат коэффициента корреляции Пирсона для обоих видов дескрипторов. Для дескрипторов ChEMBL результат 0,93, для подписей, – 0,99995.

Таблица 6

Точность моделей SVM на выборках по 2000 и разных дескрипторах после уменьшения размерности пространства подписей

№ выборки (порог для подписей)	ChEMBL		Подписи	
	SVM	MODI	SVM	MODI
1 (0)	23.4	49.5466	34.2	57.8054
2 (10)	22.9	50.0494	20.6	20.2572
3 (30)	22.5	50.4559	20.3	19.7109
4 (60)	20.4	51.2277	20.5	20.1002

Дополнительно были изучены модели линейной регрессии на тех же выборках. Для этого был использован MATLAB. Каждая выборка (№№ 1, 2, 3 и 4) была случайным образом перемешана. Первые 60 % соединений выступили в качестве обучающей выборки. В результате получены модели, которые затем были проверены на тестовых выборках.

Ошибка считалась по формуле:

$$Error = \sum_i (pred_i - act_i)^2$$

где $pred_i$ – значение, вычисленное линейной моделью на основе дескрипторов, act_i – настоящее значение активности данного соединения.

В таблице 7 приведены ошибки и их нормированные значения.

Таблица 7

Результаты тестирования линейной регрессии на 4 совокупных выборках соединений (две модели – линейная и логарифмическая)

№ выборки	Нормированная ошибка на простой модели	Ошибка на логарифмированной модели	Нормированная ошибка на логарифмированной модели
1 (0)	$3{,}6 \cdot 10^{16}$	29198	1,2452
2 (10)	$3{,}97 \cdot 10^{16}$	25727	1,2162
3 (30)	$4{,}89 \cdot 10^{16}$	20555	1,2451
4 (60)	$1{,}09 \cdot 10^{15}$	7385	0,79

Так как ошибки на обычной линейной модели были очень большими, значения активности прологарифмировали и построили модели на основании полученных данных (столбцы 3 и 4 таблицы 7).

Таким образом, получен способ применения MODI для определения точности моделей, которые можно построить на данной выборке. Способ применим лишь к тестированию модели, а не к апробации на соединениях с неизвестным значением активности, так как здесь MODI не сможет проанализировать совокупную выборку.

Вычислительный эксперимент 5

Цель ВЭ_5 была та же, что и в ВЭ_4, – определить взаимосвязь между точностью моделей SVM, построенных на выборках из 2000 соединений, полученных из совокупной выборки путём случайного перемешивания, и коэффициента MODI, вычисленного на совокупной выборке. Но в качестве «границ по подписям» были взяты пороги: 10, 20, 30, 40, 50, 60, 70. В итоге было создано 8 совокупных выборок, отличающихся набором дескрипторов и количеством соединений. Каждая совокупная выборка с помощью случайного перемешивания разделялась на выборки по 2000 соединений каждая. На каждой выборке были обучены модели SVM и протестированы на других выборках, сгенерированных из той же совокупной выборки. Это было проделано отдельно для дескрипторов ChEMBL и подписей. Дополнительно на каждой совокупной выборке был рассчитан индекс MODI.

Таблица 8

Средняя точность моделей SVM на выборках по 2000 и разных дескрипторах после уменьшения размерности пространства подписей

№ выборки (пороги для подписей)	ChEMBL		Подписи	
	SVM	MODI	SVM	MODI
1 (0)	26.73	39.44	34.21	58.11
2 (10)	26.69	39.68	20.54	19.72
3 (20)	26.74	39.98	20.41	20.01
4 (30)	26.96	39.78	20.56	20.07
5 (40)	26.94	39.98	20.39	19.91
6 (50)	26.79	40.59	20.62	20.34
7 (60)	27.61	42.56	20.23	19.28
8 (70)	28.63	41.44	20.82	20.02

Между значениями MODI и средней точностью SVM-моделей был высчитан коэффициент Пирсона, равный 0,51. Между точностью на подписях и MODI коэффициент Пирсона равен 0,999.

Таким образом, на линейном ядре SVM созданные модели на подписях предсказуемы по точности с помощью коэффициента MODI. В то же время при увеличении числа замеров (ранее было всего 3 границы по подписям) коэффициент Пирсона между MODI и моделями на линейных ядрах SVM, обученных на выборках с дескрипторами ChEMBL, значительно уменьшился.

Чтобы определить взаимосвязь между точностью моделей SVM, построенных на выборках из 2000 соединений, полученных из совокупной выборки путём случайного перемешивания, в сравнении между разными наборами дескрипторов и разными ядрами, выборки были случайным образом перемешаны и разделены снова (таблица 9).

Таблица 9

Точность моделей SVM с различными ядрами на выборках по 2000

	Сигмоидное ядро		Радиальное ядро	
№	Chembl	Подписи	Chembl	Подписи
1	24,7	24,7	26,2	25,6
2	24,5	24,9	26,3	24,4
3	24,2	25,1	26,4	23,7
4	24,7	25,6	27,2	23,8
5	23,5	25,2	26,9	22,8
6	23,6	25,9	27,2	23,2
7	23,4	24,6	28,6	22
8	24,1	23,7	30,4	21,9

Как видно из значений, точность моделей повышается с увеличением порога до определенного момента, а затем снижается. Коэффициент корреляции Пирсона между точностями на дескрипторах ChEMBL и подписях 0,003.

В этом случае точность на дескрипторах подписи практически постоянно уменьшается с ростом порога, в то же время точность моделей, построенных на радиальном ядре с дескрипторами ChEMBL практически постоянно возрастает с ростом порога. Коэффициент корреляции Пирсона 0,64.

4. Заключение

Проведена оценка эффективности дескриптора подписи в решении прямой задачи QSPR. Для этого было проведено 5 вычислительных экспериментов с различными условиями в постановках задач. Суммируя выводы по результатам вычислительных экспериментов, можно утверждать, что нет стабильного преимущества дескриптора подписи перед стандартными дескрипторами ChEMBL. Эффективней всего среди ядер SVM созданные модели лучше всего себя показали с дескрипторами ChEMBL на радиальном ядре. В то же время лучшее значение на дескрипторах подписи достигается на сигмоидном ядре.

Также в ВЭ 4 была показана экспоненциальная зависимость значения активности соединений от исследованных дескрипторов (при переходе к логарифму значения активности линейные модели продемонстрировали резкое снижение значения ошибки).

Литература

[1] *Kier L.B.*, *Hall L.H.* Intermolecular accessibility : The meaning of molecular connectivity // J. Chem. Inf. Comput. Sci. – 2000. – №40. – Pp. 792-795.

[2] *Brown R.D.*, *Martin Y.C.* The information content in 2D and 3D structural descriptors relevant to ligand-receptor binding // J. Chem. Inf. Comput. Sci. – 1997. – №37. – Pp. 1-9.

[3] *Randic M.*, *Zupan J.* On interpretation of well-known topological indices // J. Chem. Inf. Comput. Sci. – 2001. – №41. – Pp. 550-560.

[4] *Randic M.*, *Basak S.C.* A new descriptor for structure-property and structure-activity correlations // J. Chem. Inf. Comput. Sci. – 2001. – №41. – Pp. 650-656.

[5] *Faulon J.L.* The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies // J. Chem. Inf. Comput. Sci. – 2003. – №43. – Pp. 707-720.

[6] *Faulon J.L.* Stochastic generator of chemical structure: 1. Application to the structure elucidation of large molecules // J. Chem. Inf. Comput. Sci. – 1994. – №34. – Pp. 1204-1218.

[7] *Faulon J.L.*, *Visco J.*, *Pophale R.S.* Developing a Methodology for an Inverse Quantitative Structure-Activity Relationship Using the Signature Molecular Descriptor // J. Molecular Graphics Modeling. – 2002. – №20. – Pp. 429-438.

[8] ChEMBL. – https://www.ebi.ac.uk/chembl/compound.

[9] VAMMPIRE. – https://www.ncbi.nlm.nih.gov/pubmed/23734609.

[10] *Golbraikh A.*, *Muratov E.*, *Fourches D.*, *Tropsha A.* Data Set Modelability by QSAR // J. Chem. Inf. Model. – 2014. – №4. – Pp. 1-4.