УДК 004.032.26:547.022:544.351-145.83

НЕЙРОСЕТЕВОЙ ПОДХОД ДЛЯ ПРЕДСКАЗАНИЯ РАСТВОРИМОСТИ ОРГАНИЧЕСКИХ ВЕЩЕСТВ

С.А. Прохожий
Учреждение образования «Витебский государственный университет имени П.М. Машерова»

В статье исследуется применение искусственных нейронных сетей для предсказания растворимости различных органических веществ. Предлагается решение, позволяющее предсказывать растворимость, основываясь на SMILES-представлении молекул.

Цель – прогнозирование логарифмической растворимости химических соединений посредством предварительно обученных искусственных нейронных сетей, а также последующая оценка качества построенной модели.

Материал и методы. *Разработана нейросетевая модель для прогнозирования растворимости в воде различных органических соединений. Нейронные сети создавались и обучались с помощью пакета Statistica (модуль «Automated Neural Networks»). В качестве обучающей и тестовой выборки использован датасет ESOL объемом 1128 соединений. Для проверки точности прогноза применяется MAPE-метрика.*

Результаты и их обсуждение. *Для нейросетевого анализа из ESOL-датасета были выбраны два параметра – SMILES-представление и логарифмическая растворимость (logP) вещества. Нейросетевая модель обучалась на 80% от имеющихся экспериментальных данных, а оставшиеся 20% данных использованы в качестве тестовой выборки. Далее с применением модуля «Automated Neural Networks» построен ансамбль искусственных нейронных сетей для прогнозирования растворимости. Проведена оценка качества обученной модели.*

Заключение. *На основе экспериментальных данных построен ансамбль искусственных нейронных сетей для прогнозирования растворимости органических химических соединений. С помощью обученной нейронной сети сделаны прогнозы растворимости, а полученные результаты сопоставлены с экспериментальными.*

Ключевые слова: *искусственная нейронная сеть, анализ данных, растворимость*.

NEURAL NETWORK APPROACH FOR PREDICTING THE SOLUBILITY OF ORGANIC SUBSTANCES

S.A. Prokhozhiy
Education Establishment "Vitebsk State P.M. Masherov University"

The article explores the use of artificial neural networks for predicting the solubility of various organic substances. A solution is proposed that forecasts solubility based on the SMILES representation of molecules.

The purpose of the article is prediction of logarithmic solubility of chemical compounds using pre-trained artificial neural networks, as well as subsequent quality assessment of the constructed model.

Material and methods. *A neural network model has been developed to predict the solubility of various organic compounds in water. Neural networks were created and trained using Statistica package (Automated Neural Networks module). The ESOL dataset with 1128 substances is used as a training and test set. MAPE metric is applied for the forecast accuracy check.*

Findings and their discussion. *For neural network analysis, two substance parameters were selected from the ESOL dataset: the SMILES representation and the logarithmic solubility (logP). The neural network model was trained on 80% of experimental data, and the remaining 20% was used as a test set. Further, using the Automated Neural Networks module, an ensemble of artificial neural networks was built to predict solubility. The quality of the constructed model was assessed.*

Conclusion. *Based on experimental data, an ensemble of artificial neural networks is built to predict the solubility of organic chemical compounds. Using trained neural network, solubility forecast is made, and the results are compared with experimental data.*

Key words: *artificial neural network, data analysis, solubility.*

Для предсказания свойств химических соединений исследователи часто используют различные традиционные методы (например, уравнения состояния и корреляции), основанные на эмпирических моделях, которым необходим большой объем экспериментальных данных и сложных математических подходов. Однако создание и калибровка таких моделей требуют значительных усилий

и ресурсов. Кроме того, указанные модели часто ограничены в случае обобщения на новые вещества. В свете этих ограничений многие ученые и инженеры ищут альтернативные подходы для предсказания физико-химических свойств. В последнее время искусственные нейронные сети, основанные на машинном обучении, стали широко применяться в различных областях, включая промышленность и науку. Нейронные сети представляют собой альтернативный подход, который позволяет автоматически извлекать сложные зависимости из данных и строить модели, которые способны обобщаться на новые примеры. В данной статье исследуется использование искусственных нейронных сетей для предсказания растворимости различных химических веществ.

Растворимость молекул – это важное свойство, которое необходимо знать для расчета растворов в реакциях и процессе перекристаллизации. Однако проведение экспериментов для определения растворимости довольно затратно. В данной работе предлагается решение, позволяющее предсказывать растворимость, основываясь только на SMILES-представлении молекул. Подобные модели будут полезны при создании новых материалов и лекарственных средств.

Искусственные нейронные сети выступают как система взаимосвязанных и взаимодействующих между собой простых искусственных нейронов [1]. Искусственный нейрон – это упрощенная модель естественного нейрона, базовый кирпичик нейросети. Математически искусственный нейрон представляет собой некоторую функцию от единственного аргумента – линейной комбинации всех входных сигналов. Данную нелинейную функцию называют функцией активации. Полученное значение функции посылается на единственный выход.

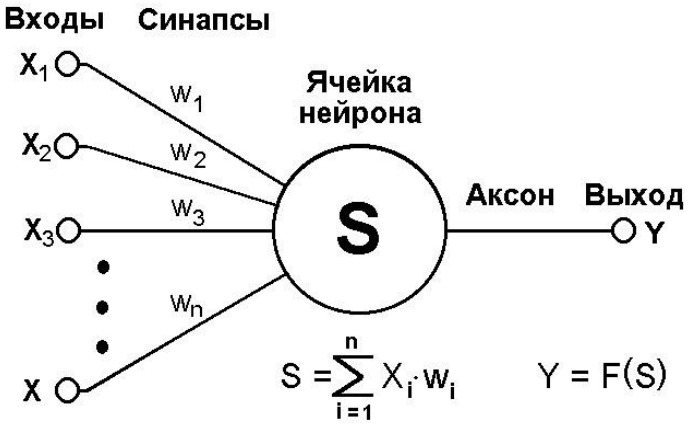


Рис. 1. **Схема искусственного нейрона**

Как видно из схемы (рис. 1), искусственный нейрон характеризуется своим текущим состоянием по аналогии с нервными клетками головного мозга, которые могут быть возбуждены или заторможены. Этот нейрон имеет группу синапсов, представляющих собой однонаправленные входные связи, соединенные с выходами других нейронов, а также имеет аксон – выходную связь данного нейрона, с которой сигнал возбуждения или торможения поступает к синапсам следующих нейронов [2].

Текущее состояние нейрона определяется как взвешенная сумма его входов:

$$S = \sum_{i=1}^{n} x_i w_i,$$

где x_i – значения входных параметров, w_i – их веса. Выход нейрона есть функция его состояния:

$$y = f(S).$$

Нейроны объединяются в так называемые слои, которые организуются определенным образом в сеть (рис. 2).

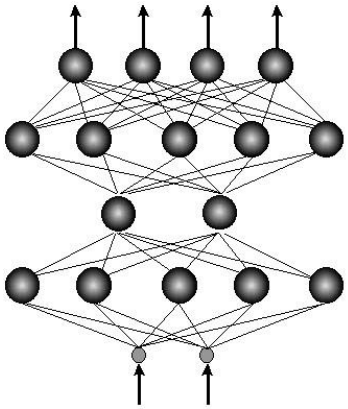


Рис. 2. **Примерная структура искусственной нейронной сети**

Обучение нейронной сети в первую очередь заключается в изменении величины синаптических связей между нейронами, т.е. в перевычислении весов w_i.

По своей природе и философии подход, основанный на использовании нейронных сетей, относится к арсеналу инструментария «Big Data». Во многих случаях точность нейросетевых прогнозов оказывается выше точности прогнозов, полученных классическими методами науки. В настоящее время можно наблюдать некоторые противоречия между нейронными сетями и традиционными научными инструментами, так как вторые проигрывают первым по точности расчетов, а первые по сравнению со вторыми не имеют под собой никакой обоснованной теории [3]. В то же время нейросетевой подход свободен от модельных ограничений, он одинаково эффективен как для линейных, так и для сложных нелинейных задач, а также задач классификации. Также следует отметить масштабируемость нейронных сетей, они способны решать как узкоспециализированные задачи, так и рассчитывать масштабные модели.

Нейросети не предполагают никакой теории, объясняющей суть исследуемых процессов, а лишь констатируют тот факт, что набор входных переменных может быть преобразован в результирующую переменную с определенной точностью. При этом число входящих переменных может быть очень большим. Например, в языковой модели с генеративным искусственным интеллектом GPT-3 175 миллиардов параметров, а в GPT-4 – уже 100 триллионов. Для сравнения, количество нейронов в человеческом мозгу оценивается в 80–100 миллиардов, а число синапсов – в те же 100 триллионов. Для настройки или обучения сети требуется несколько тысяч эпох. Однако входные переменные, не влияющие на формирование выходной переменной, не отбрасываются, как это делается в традиционном моделировании, а нивелируются путем присвоения им минимального или нулевого весового коэффициента. Тем самым нейросетевая модель почти всегда характеризуется большой информационной избыточностью, в то же время не обладая системным описанием сущности моделируемого процесса. Вместе с тем не следует отрицать эффективность аппарата нейронных сетей ввиду отсутствия логических объяснительных схем.

Нейронные сети являются гибким и мощным инструментом, способным обрабатывать сложные зависимости между структурой вещества и его физико-химическими свойствами, а также предсказывать эти свойства. Искусственные нейросети могут быть применены для предсказания свойств на всем диапазоне значений, включая трансграничные условия, не учитывающиеся в традиционных методах. Использование искусственных нейронных сетей сокращает затраты на проведение дорогостоящих экспериментов. Вместо исследования каждого вещества (а на данный момент известны десятки миллионов органических соединений) можно обучить нейросеть на доступных данных и использовать ее для предсказания свойств новых соединений. Это позволяет сэкономить ресурсы и время, а также оптимизировать процесс разработки и тестирования новых веществ. Важно отметить, что применение нейросетевого подхода требует наличия достаточного объема качественных данных для обучения моделей. Это может включать информацию о структуре исследуемого вещества, его химических и физических свойствах, а также результаты экспериментов и испытаний. Сбор и подготовка таких данных являются ключевыми шагами для успешной реализации нейросетевого подхода [4].

Цель исследования – на базе имеющихся экспериментальных данных выполнить прогнозирование логарифмической растворимости органических соединений на основе предварительно обученных искусственных нейронных сетей и оценить качество построенной модели.

Материал и методы. Разработана нейросетевая модель для прогнозирования растворимости в воде различных химических органических соединений. Нейронные сети создавались и обучались с помощью пакета Statistica (модуль «Automated Neural Networks») – мощного инструмента анализа и прогнозирования данных, имеющего широкое применение в бизнесе, промышленности, управлении, финансах.

При обучении предназначенных для прогнозирования нейросетей использовался стандартный подход. Имеющиеся данные разбивают на две выборки: обучающую и тестовую. Обучающая выборка предназначена для подстройки синаптических коэффициентов обучаемых нейронных сетей с целью минимизации ошибки на выходе сети. Тестовая выборка, которая не применялась в процессе обучения, предназначена для контроля качества прогнозирования.

В качестве обучающей и тестовой выборки был взят датасет ESOL – сравнительно небольшой набор данных, содержащий экспериментальные сведения о растворимости в воде 1128 соединений [5]. Этот набор данных использовался для обучения моделей, оценивающих растворимость непосредственно на основе химических структур. Указанные структуры не включают 3D-координаты, поскольку растворимость – свойство молекулы, а не ее конкретных конформеров. Для каждой молекулы в наборе приводятся десять различных параметров и описание ее структуры в формате SMILES (Simplified Molecular Input Line Entry System) – системы правил однозначного описания состава и структуры молекулы химического вещества с использованием строки символов ASCII. Данная система представления была специально разработана для компьютерного применения химиками. Строка символов, составленная по правилам SMILES, может быть преобразована в двумерную или трехмерную структурную формулу молекулы. Правила кодировки этой системы можно быстро и легко изучить любому пользователю с любым уравнением начальной подготовки в области химии. Оказывается, можно извлечь много полезных молекулярных свойств, исходя только из SMILES-представления молекул.

Следует заметить, что ESOL – достаточно маленький датасет, как и большинство бесплатных наборов для машинного обучения. Для доступа к специализированным датасетам большого размера требуется платная подписка. Описываемая нейросетевая модель обучалась на 80% от этих данных (т.е. на 902 записях), а оставшиеся 20% (т.е. 226 записей) данных были использованы в качестве тестовой выборки.

Результаты и их обсуждение. Для нейросетевого анализа из ESOL-датасета были выбраны два параметра – SMILES-представление и логарифмическая растворимость (logP) химического соединения. Величина logP напрямую связана с растворимостью вещества в воде и определяется как обычная единица растворимости, соответствующая десятичному логарифму растворимости молекулы, измеренной в молях на литр.

Далее полученные данные после переформатирования экспортируются в пакет Statistica. На основе обучающей выборки, содержащей 902 записи, при помощи модуля «Automated Neural Networks» происходит тренировка нейронных сетей. Мы обучаем 25 нейросетей (включая как MLP-, так и RBF-сети), из которых выбираются 3 наилучшие по критерию Test perfection (рис. 3).

В процессе обучения нейросети находят скрытые зависимости между SMILES-представлением молекулы и ее растворимостью. Нейросеть сама по себе не «знает» химию и теорию растворов, и ее обучение – это процесс перевычисления весовых синаптических коэффициентов w_i, в ходе которого компьютерная модель учится выполнять определенные задачи посредством предоставленных ей данных, в том числе делать прогнозы или принимать решения на основе новых данных. Например, в данном обучающем датасете нейросеть «замечает» закономерность: чем выше молекулярная масса спирта (т.е. вещества, имеющего в SMILES-представлении гидроксильный анион [OH-]), тем меньше его растворимость. Эта закономерность уже присутствовала в обучающем наборе, и сложная математическая модель «уловила» набор «ассоциаций» между входными данными и необходимой реакцией на выходе (рис. 4).

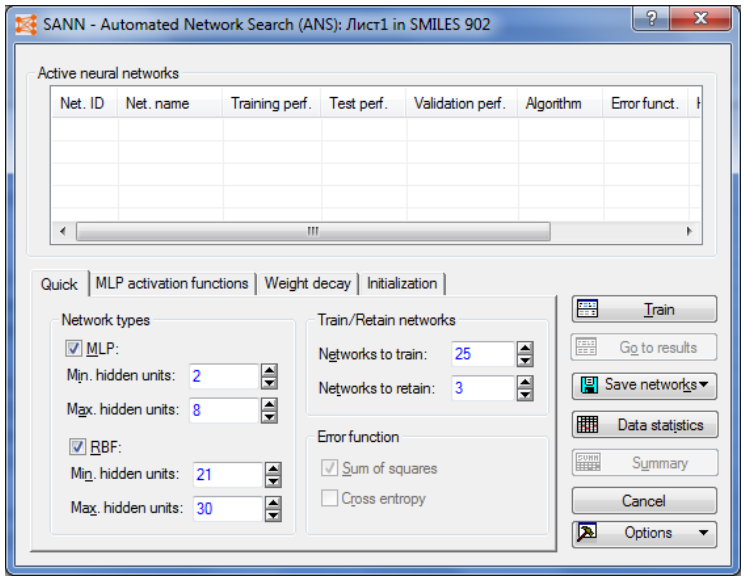


Рис. 3. **Параметры обучения**

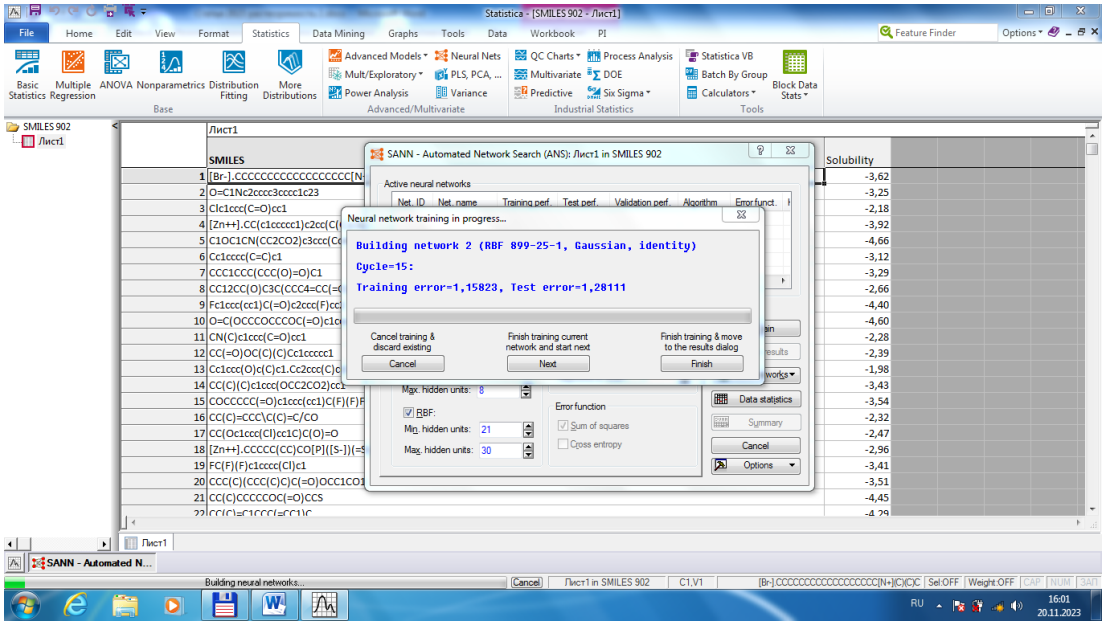
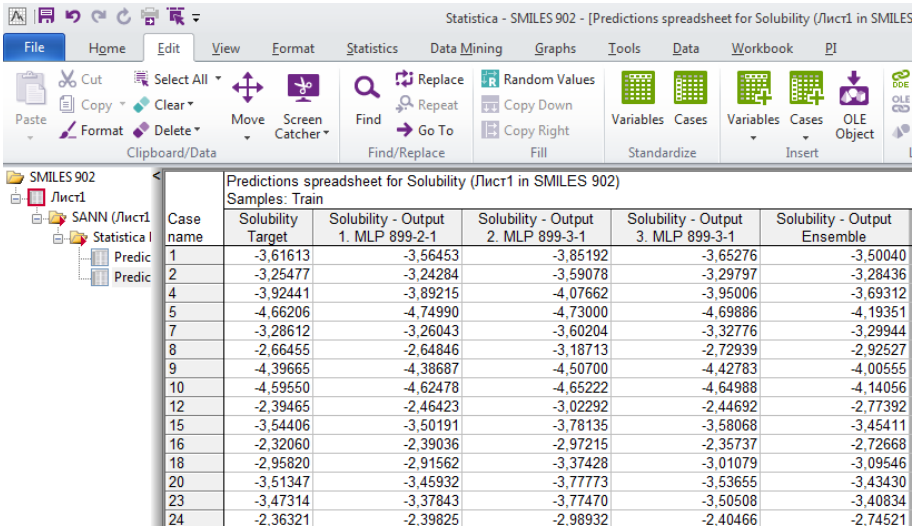


Рис. 4. **Процесс обучения нейросетей**

После обучения проводятся тестирование и проверка качества обученных нейросетей. На их входы подаются SMILES-представления веществ из тестовой выборки. Каждая из тренированных сетей вычисляет прогнозное значение логарифмической растворимости этого химического соединения. Консолидированное мнение всего ансамбля искусственных нейросетей (а именно, их среднее арифметическое) мы и считаем искомым прогнозом (рис. 5).

Одной из наиболее распространенных метрик, используемых для измерения точности прогнозирования модели, является MAPE (Mean Absolute Percentage Error), что означает среднюю абсолютную ошибку в процентах. Данная метрика рассчитывается по формуле $\varepsilon_t = \frac{1}{n}\sum\left|\frac{y_i - \overline{y_i}}{y_i}\right|$, где y_i – фактическое значение показателя, $\overline{y_i}$ – прогнозное значение, n – объем тестовой выборки. Прогнозные значения ансамбля искусственных нейросетей и эмпирические данные из ESOL-набора экспортируются в табличный редактор Excel, где при помощи стандартных функций и вычисляется MAPE-метрика (рис. 6).



Statistica - SMILES 902 - [Predictions spreadsheet for Solubility (Лист1 in SMILES

File | Home | Edit | View | Format | Statistics | Data Mining | Graphs | Tools | Data | Workbook | PI

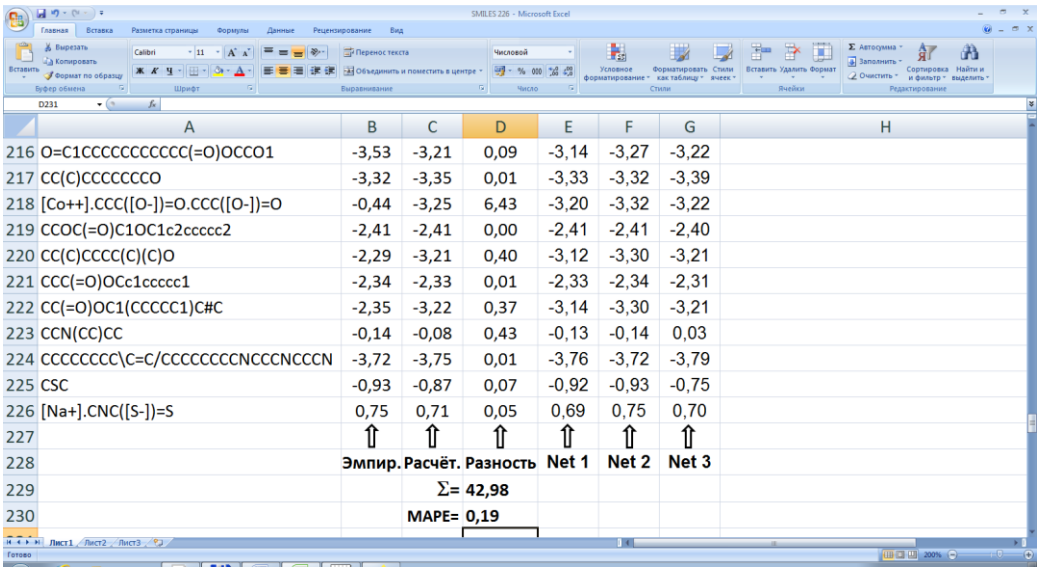
Cut, Copy, Format, Paste, Select All, Clear, Delete, Move, Screen Catcher — Clipboard/Data
Find, Replace, Repeat, Go To — Find/Replace
Random Values, Copy Down, Copy Right — Fill
Variables, Cases — Standardize
Variables, Cases, OLE Object — Insert

SMILES 902
Лист1
SANN (Лист1
Statistica I
Predic
Predic

Predictions spreadsheet for Solubility (Лист1 in SMILES 902)
Samples: Train

Case name	Solubility Target	Solubility - Output 1. MLP 899-2-1	Solubility - Output 2. MLP 899-3-1	Solubility - Output 3. MLP 899-3-1	Solubility - Output Ensemble
1	-3,61613	-3,56453	-3,85192	-3,65276	-3,50040
2	-3,25477	-3,24284	-3,59078	-3,29797	-3,28436
4	-3,92441	-3,89215	-4,07662	-3,95006	-3,69312
5	-4,66206	-4,74990	-4,73000	-4,69886	-4,19351
7	-3,28612	-3,26043	-3,60204	-3,32776	-3,29944
8	-2,66455	-2,64846	-3,18713	-2,72939	-2,92527
9	-4,39665	-4,38687	-4,50700	-4,42783	-4,00555
10	-4,59550	-4,62478	-4,65222	-4,64988	-4,14056
12	-2,39465	-2,46423	-3,02292	-2,44692	-2,77392
15	-3,54406	-3,50191	-3,78135	-3,58068	-3,45411
16	-2,32060	-2,39036	-2,97215	-2,35737	-2,72668
18	-2,95820	-2,91562	-3,37428	-3,01079	-3,09546
20	-3,51347	-3,45932	-3,77773	-3,53655	-3,43430
23	-3,47314	-3,37843	-3,77470	-3,50508	-3,40834
24	-2,36321	-2,39825	-2,98932	-2,40466	-2,74521

Рис. 5. **Прогнозы нейросетей**



	A	B	C	D	E	F	G	H
216	O=C1CCCCCCCCCC(=O)OCCO1	-3,53	-3,21	0,09	-3,14	-3,27	-3,22	
217	CC(C)CCCCCCO	-3,32	-3,35	0,01	-3,33	-3,32	-3,39	
218	[Co++].CCC([O-])=O.CCC([O-])=O	-0,44	-3,25	6,43	-3,20	-3,32	-3,22	
219	CCOC(=O)C1OC1c2ccccc2	-2,41	-2,41	0,00	-2,41	-2,41	-2,40	
220	CC(C)CCCC(C)(C)O	-2,29	-3,21	0,40	-3,12	-3,30	-3,21	
221	CCC(=O)OCc1ccccc1	-2,34	-2,33	0,01	-2,33	-2,34	-2,31	
222	CC(=O)OC1(CCCCC1)C#C	-2,35	-3,22	0,37	-3,14	-3,30	-3,21	
223	CCN(CC)CC	-0,14	-0,08	0,43	-0,13	-0,14	0,03	
224	CCCCCCCC\C=C/CCCCCCCCNCCCNCCCN	-3,72	-3,75	0,01	-3,76	-3,72	-3,79	
225	CSC	-0,93	-0,87	0,07	-0,92	-0,93	-0,75	
226	[Na+].CNC([S-])=S	0,75	0,71	0,05	0,69	0,75	0,70	
227		⇧	⇧	⇧	⇧	⇧	⇧	
228		**Эмпир.**	**Расчёт.**	**Разность**	**Net 1**	**Net 2**	**Net 3**	
229			**Σ= 42,98**					
230			**MAPE= 0,19**					

Рис. 6. **Вычисление MAPE-метрики**

Для нашей выборки получено значение $\varepsilon_t = 0{,}19$. Это означает, что средняя разница между прогнозом и фактической растворимостью составляет приблизительно 19%. Полученное значение в теории прогнозирования чаще всего интерпретируется как «среднее качество прогнозирования». С одной стороны, этого явно недостаточно для научных или промышленных проектов, требующих высокой точности, например, для производства лекарственных препаратов. С другой стороны, для многих проектов (например, прогнозирования объема продаж [6], предсказания уровня инфляции [3; 7] или прогнозирования скорости восстановления функционального состояния спортсменов после физической нагрузки [8]) с учетом относительной простоты данного метода полученная точность вполне приемлема. Кроме того, текущая модель – базовая, и есть много возможностей для ее улучшения. Данный метод можно усовершенствовать, использовав не только свойства атомов, но и свойства связей в молекуле, а также большее количество параметров из обучающей выборки. Еще одним существенным улучшением станет обучение модели на датасете большего размера.

Заключение. Молекулярное машинное обучение быстро развивается в последнее время. Усовершенствованные методы и наличие достаточных наборов данных позволяют алгоритмам машинного обучения делать все более точные прогнозы о свойствах молекул.

Таким образом, применение модуля «Automated Neural Networks» системы статистического анализа Statistica позволило на основе экспериментальных данных построить ансамбль искусственных нейронных сетей для прогнозирования растворимости органических химических соединений. Обученная нейросетевая модель была протестирована на ранее неизвестных ей данных. С помощью обученной нейронной сети сделаны прогнозы растворимости, а полученные результаты сопоставлены с экспериментальными.

Автор благодарит доцента кафедры химии и естественнонаучного образования ВГУ имени П.М. Машерова, кандидата педагогических наук, доцента Алексея Александровича Белохвостова за идею исследования, а также ценные замечания при подготовке данной статьи.

ЛИТЕРАТУРА

1. Хайкин, С. Нейронные сети: полный курс / С. Хайкин. – 2-е изд. – М.: Издательский дом «Вильямс», 2008. – 1104 с.
2. Рассел, С. Искусственный интеллект: современный подход / С. Рассел, П. Норвиг. – 2-е изд. – М.: Издательский дом «Вильямс», 2006. – 1424 с.
3. Балацкий, Е.В. Использование нейронных сетей для прогнозирования инфляции: новые возможности / Е.В. Балацкий, М.А. Юревич // Вестник УрФУ. Сер.: Экономика и управление. – 2018. – Т. 17, № 5. – С. 823–838.
4. Дударов, С.П. Теоретические основы и практическое применение искусственных нейронных сетей: учеб. пособие / С.П. Дударов, П.Л. Папаев. – М.: РХТУ им. Д.И. Менделеева, 2014. – 104 с.
5. jxie/esol – Datasets at Hugging Face [Electronic resource] // huggingface.co: The AI community building the future. – Mode of access: https://huggingface.co/datasets/jxie/esol/viewer/default/train_0?p=8. – Date of access: 05.12.2023.
6. Ливенцева, А.В. Использование нейронной сети при прогнозировании объема продаж торговой фирмы / А.В. Ливенцева // Вестник науки и образования. – 2017. – № 2(26). – С. 24–28.
7. Choudhary, M.A. Neural network models for inflation forecasting: an appraisal / M.A. Choudhary, A. Haider // Applied Economics. – 2012. – Vol. 44, iss. 20. – P. 2631–2635.
8. Прохожий, С.А. Прогнозирование восстановления функционального состояния организма после истощающей физической нагрузки / С.А. Прохожий, Э.С. Питкевич // Весн. Віцеб. дзярж. ун-та. – 2020. – № 1. – С. 16–20.

REFERENCES

1. Haikin S. *Neironnie seti: polnii kurs* [Neural networks], Moscow, Williams, 2008, 1104 p.
2. Rassel S. *Iskusstvennii intellect: sovremennii podhod* [Artificial intelligence: modern approach], Moscow, Williams, 2006, 1424 p.
3. Balatskii E.V. *Vestnik UrFU* [Ural Federal University Journal], 2018, 17(5), p. 823–838.
4. Dudarov S.P., Papaev P.L. *Teoreticheskie osnovi i prakticheskoe primenenie iskusstvennyh neironnyh setei* [Theoretical foundations and practical application of artificial neural networks], Moscow, Mendeleev University of Chemical Technology, 2014, 104 p.
5. jxie/esol – Datasets at Hugging Face [Electronic resource] // huggingface.co: The AI community building the future. – Mode of access: https://huggingface.co/datasets/jxie/esol/viewer/default/train_0?p=8. – Date of access: 05.12.2023.
6. Liventseva A.V. *Vestnik nauki I obrazovaniya* [Science and Education Journal], 2017, 2(26), p. 24–28.
7. Choudhary, M.A. Neural network models for inflation forecasting: an appraisal / M.A. Choudhary, A. Haider // Applied Economics. – 2012. – Vol. 44, iss. 20. – P. 2631–2635.
8. Prokhozhiy S.A., Pitkevich E.S. *Vesnik VDU* [Bulletin of Vitebsk State University], 2020, 1, p. 16–20.

Поступила в редакцию 08.12.2023
Адрес для корреспонденции: e-mail: ProkhozhiySA@vsu.by – Прохожий С.А.