ХИМИЧЕСКИЙ АНАЛИЗ

УДК 543.061:004.032.26

НАДЕЖНОСТЬ ИДЕНТИФИКАЦИИ АНАЛИТОВ С ПОМОЩЬЮ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

© 2010 Я. Н. Краснянчин, А. В. Пантелеймонов, Ю. В. Холин

К решению задач идентификации объектов по данным многооткликового эксперимента привлечен аппарат искусственных нейронных сетей: алгоритмы с обучением и самоорганизующаяся сеть Кохонена. Показана удовлетворительная надежность идентификации с помощью алгоритмов искусственных нейронных сетей при разбиении на классы наборов тестовых данных со сложной структурой и при наличии в характеристиках классифицируемых объектов грубых промахов.

Ключевые слова: искусственная нейронная сеть, алгоритм, классификация, идентификация, массив данных.

ВВЕДЕНИЕ

Идентификация соединений при контроле качества пищевых продуктов, объектов окружающей среды, лекарственных препаратов, установлении структуры нового вещества и решении других задач с применением инструментальных методов анализа является актуальной областью качественного химического анализа. Среди экспериментальных методов, привлекаемых для идентификации аналитов, особое значение имеют спектроскопические и хроматографические, а также методы, основанные на использовании сенсорных систем ("электронный нос", "электронный язык") [1-3]. Под "идентификацией" понимают установление тождественности анализируемого объекта известному (эталону) или вывод о принадлежности аналита некоторому классу объектов на основе сопоставления их свойств [4]. Одна из основных проблем идентификации заключается в том, что существующие алгоритмы требуют априорной информации о характеристиках исходных данных в качестве начального приближения для расчетов (например, сведений о функции распределения экспериментальных погрешностей, числе классов и др.). Альтернативой известным параметрическим подходам служат методы, воплощающие идеи робастного оценивания и свободные от априорных предположений. Переход от обычных алгоритмов дискриминантного или кластерного анализа (например, метода k-средних) к адаптивным методам способствует повышению достоверности идентификации.

Искусственные нейронные сети нашли свое приложение в химических исследованиях. С их помощью предсказывают химические и физические свойства, проводят дизайн новых химических соединений и материалов с требуемыми характеристиками, подбирают методы синтеза, моделируют кинетику химических процессов и др. [5-8].

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ, ВАЖНЫЕ ДЛЯ КЛАССИФИКАЦИИ / ИДЕНТИФИКАЦИИ ОБЪЕКТОВ

Нейронные сети – это математические модели, состоящие из элементарных единиц обработки информации (нейронов), накапливающих экспериментальные знания и предоставляющих их для последующей обработки. К важнейшим свойствам нейронных сетей относятся параллельность обработки информации одновременно всеми нейронами, способность к обучению, абстрагированию и обобщению полученных знаний. Наличие таких характеристик обеспечивает получение ожидаемой реакции сети применительно к соответствующим данным, нечувствительность нейронной сети к малым изменениям входных сигналов, а также возможность работы с искаженными вариантами анализируемых объектов [9-11].

Представим ряд основных понятий и терминов, используемых для описания нейронносетевых алгоритмов [12, 13].

Нейрон – составная часть нейронной сети. Он состоит из элементов трех типов: синапсов, осуществляющих связь между нейронами, сумматора, который выполняет сложение поступающих сигналов, и нелинейного преобразователя результата, обеспеченного сумматором, с

применением соответствующей функции активации. Синапсы умножают входной сигнал на число, характеризующее силу связи, – вес или весовой коэффициент нейрона.

Слой – множество нейронов, имеющих общие выходные или входные сигналы. В зависимости от функций, выполняемых в сети, можно выделить входной, промежуточный (скрытый) и выходной слои.

Обучение – этап функционирования нейронной сети, в процессе которого происходит настройка параметров нейронной сети, в частности весовых коэффициентов нейронов, для получения наиболее адекватного сигнала на выходе сети.

Ряд алгоритмов искусственных нейронных сетей реализован в пакете MATLAB 6.5 [14]. Среди них алгоритмы обучаемых нейронных сетей и самоорганизующаяся сеть Кохонена. Обучаемые нейронные сети являются двухслойными сетями прямого распространения сигнала (все сигналы в нейронной сети движутся только в сторону от входов к выходам). В табл. 1 представлены типы реализованных в MATLAB нейронных сетей и соответствующие функции активации.

Таблица 1. Нейронные сети (НС) и соответствующие функции активации

<table>
<tr><th rowspan="2">Название</th><th colspan="2">Функции активации</th></tr>
<tr><th>Промежуточный слой</th><th>Выходной слой</th></tr>
<tr><td>НС радиального основания (Radial Basis Function Networks, RBN)</td><td rowspan="3">Функция Гаусса
$y_{Gauss} = \exp\left[-\left(\frac{d_i}{\sigma_i}\right)^2\right]$
$d_i = \sqrt{\sum_{j=1}^{N}(x_j - w_{ij})^2}$
$i = 1,...,L$</td><td rowspan="2">Линейная функция
$y_{lin} = k \cdot NET$,
$NET = \sum_{j=1}^{N} x_j w_{ij}$</td></tr>
<tr><td>НС радиального основания с нулевой ошибкой (Radial Basis Function Networks with Zero Error, RBEN)</td></tr>
<tr><td>Вероятностная НС (Probabilistic Neural Network, PNN)</td><td>Конкурирующий слой: подсчитывает вероятность принадлежности входного вектора к тому или иному классу</td></tr>
<tr><td>НС прямой передачи (Feed Forward Networks, FFN)</td><td rowspan="3">Гиперболический тангенс
$y_{\tanh} = \frac{e^{NET} - e^{-NET}}{e^{NET} + e^{-NET}}$</td><td rowspan="3">Линейная функция
$y_{lin} = k \cdot NET$,
$NET = \sum_{j=1}^{N} x_j w_{ij}$</td></tr>
<tr><td>Каскадная НС прямой передачи (Cascade Feed Forward Networks, CFFN)</td></tr>
<tr><td>НС прямой передачи с временной задержкой (Time Delay Feed Forward Neural Networks, FFTDN)</td></tr>
<tr><td>Сеть Кохонена</td><td>Не содержит скрытого слоя</td><td>Конкурирующий слой: подсчитывает вероятность принадлежности входного вектора к тому или иному классу</td></tr>
</table>

$X = (x_1,...,x_N)$ – массив входных характеристик; $W_i = (w_{i1},...,w_{iN})$ – весовой массив i-го нейрона скрытого слоя; σ_i, k – параметры активационных функций; L – число эталонов; NET – уровень активации нейрона.

Сеть Кохонена принципиально отличается от других представленных нейронных сетей, поскольку использует неконтролируемое обучение. Для реализации алгоритма необходимо определить меру соседства нейронов (окрестность "нейрона-победителя"). "Нейрон-победитель" характеризуется минимальным расстоянием до входного сигнала, т.е. его вектор весов ближе всех к входному вектору. На текущей итерации (t) корректируются веса только "нейрона-победителя" и нейронов из его зоны соседства по формуле:

$$w_{ij}(t+1) = w_{ij}(t) + \eta(x_j(t) - w_{ij}(t)), \quad (1)$$

где x_j – j-й элемент входного сигнала; w_{ij} – весовой коэффициент, характеризующий связь между j-м элементом входного сигнала и нейроном i выходного слоя; η – шаг обучения.

В настоящей работе аппарат искусственных нейронных сетей привлечен к решению задач идентификации объектов различной структуры по данным многооткликового эксперимента.

Для вышеупомянутых алгоритмов контролируемых (обучаемых) нейронных сетей применили алгоритм обучения методом обратного распространения ошибки (back propagation). Указанный алгоритм минимизирует полусумму квадратов разностей между желаемой величиной выхода d_k и реально полученными на выходах сети значениями y_k для каждого объекта k:

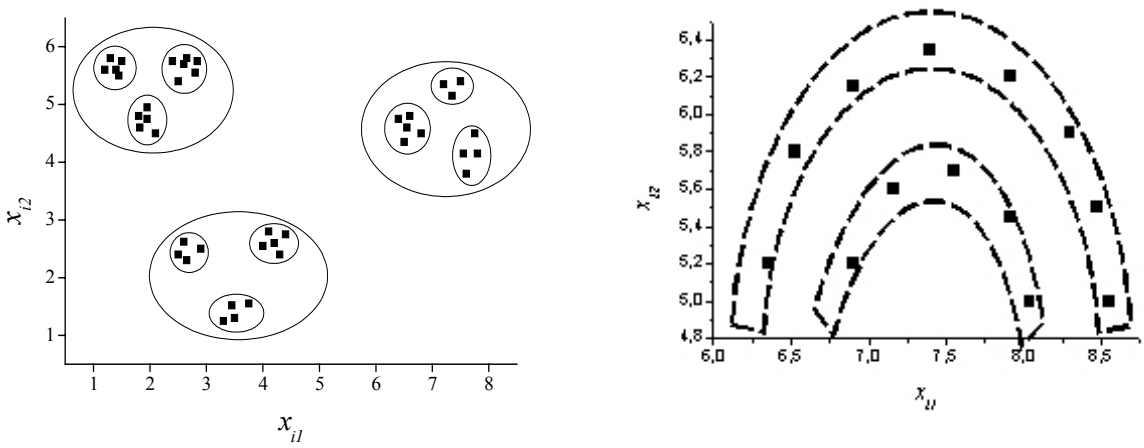
$$E = \frac{1}{2}\sum_{k=1}^{Q}\sum_{i=1}^{Y}(d_k^i - y_k^i)^2 \ , \qquad (2)$$

где Q – число объектов в обучающем множестве; индекс Y соответствует числу выходов многослойной сети.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Изучая возможность применения алгоритмов нейронных сетей для идентификации аналитов по данным многооткликового эксперимента, мы испытали указанные выше алгоритмы на модельных данных.

Использовали данные со сложными структурами (двуиерархической и дугообразной) (рис. 1) [15].



а) б)

Рис. 1. Наборы данных с двуиерархической (а) и дугообразной (б) структурами.

О надежности алгоритмов судили по значениям долей неправильно классифицированных объектов

$$P = \frac{n}{N} \times 100, \% \, , \qquad (3)$$

где n – число неправильно классифицированных объектов, N – их общее число. Также испытали работоспособность традиционных методов классификации методом k-средних и с помощью дискриминантного анализа (табл. 2). Полученные результаты свидетельствуют о применимости алгоритмов искусственных нейронных сетей для классификации данных со сложной структурой в ситуации, когда стандартные методы классификации недостаточно эффективны.

Важной характеристикой алгоритмов классификации является их устойчивость к наличию в массивах численных характеристик классифицируемых объектов "грубых промахов". В качестве тестовых использовали два набора данных. Первый содержит характеристики цветков ириса (известный тестовый набор для испытания алгоритмов классификации и кластерного анализа) [16], второй – характеристики образцов итальянских вин (также популярный набор данных для тестирования алгоритмов классификации) [17]. Массив данных о цветках ириса содержал сведения о 150 образцах, каждый из которых характеризуется четырьмя числовыми признаками (длиной и шириной чашелистика, длиной и шириной лепестка). Образцы ирисов разделяются на три класса, в каждом из которых содержится по 50 образцов. В массиве данных о винах со-

держались результаты определения 13 признаков 178 образцов вин, принадлежащих к трем классам. В исходные данные характеристик образцов ириса и образцов вин вносили погрешности (ε), рассчитываемые по формуле

$$\varepsilon = [(100 - q) \cdot \varepsilon_{Gauss}(0, \sigma) + q \cdot \varepsilon_{Logist}(0, \sigma)] / 100, \quad (4)$$

где $0 \le q \le 100$, % – интенсивность "грубых промахов"; ε_{Gauss} – случайная величина, распределенная по закону Гаусса с нулевым средним и стандартным отклонением σ; ε_{Logist} – случайная величина, подчиняющаяся логистическому распределению с нулевым средним и стандартным отклонением σ:

$$p(\varepsilon) = \frac{\pi \cdot \exp\left\{-\frac{\pi \varepsilon}{\sigma\sqrt{3}}\right\}}{\sigma\sqrt{3} \cdot \left(1 + \exp\left\{-\frac{\pi \varepsilon}{\sigma\sqrt{3}}\right\}\right)^2}, \quad x \in (-\infty, \infty), \quad (5)$$

где $p(\varepsilon)$ – плотность распределения. Поскольку коэффициент эксцесса логистического распределения ($\gamma_2 = 1.2$) больше, чем нормального ($\gamma_2 = 0$), с ростом q возрастает вероятность появления среди ε "грубых промахов" – погрешностей, более чем в два-три раза превышающих стандартные отклонения σ [18].

Таблица 2. Результаты классификации объектов при дугообразной и двуиерархической структурах данных

Алгоритм классификации	*P*, %	
	Данные с двуиерархической структурой	Данные с дугообразной структурой
RBN, RBEN, PNN	0	0
FFN	19.5	0
CFFN	4.9	0
FFTDN	0	7.7
сеть Кохонена	0	38.5
метод k-средних	24.4	38.5
дискриминантный анализ	0	38.5

Рис. 2. Длины чашелистиков цветов ириса (x_i).

Проиллюстрируем внесение в данные "грубых промахов" на примере одной из характеристик цветков ириса – длин чашелистиков (рис. 2). В эти длины вносили погрешности, вычисляемые согласно (4), для q = 0; 10; 30; 50; 80 и 100 % (внесенные погрешности показаны на рис. 3 для q = 10% и q = 80 %).

На рис. 4 и рис. 5 приведены значения долей неправильно идентифицированных образцов ириса и образцов вин с применением алгоритмов нейронных сетей, метода k-средних и дискриминантного анализа в зависимости от доли грубых промахов внесенных в исходные данные погрешностей.

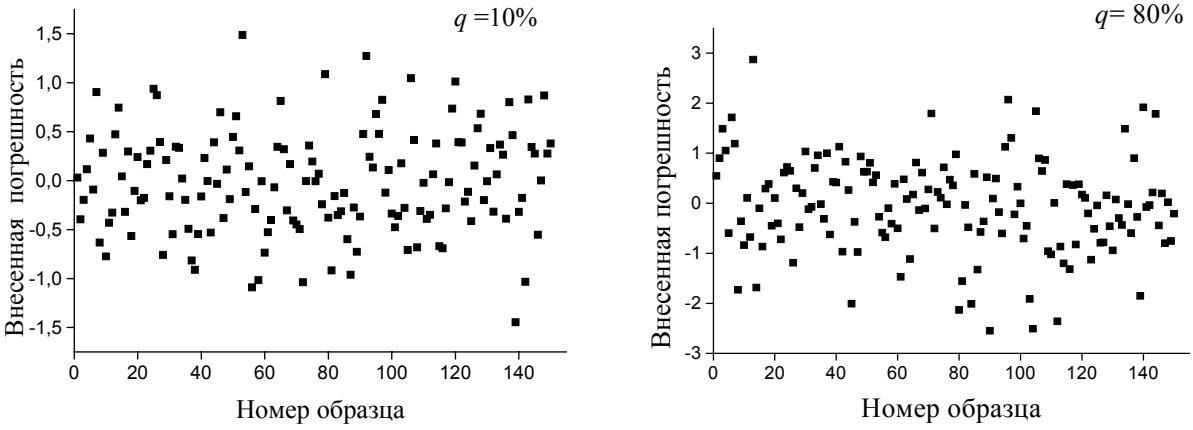


Рис. 3. Погрешности, внесенные в длины чашелистиков цветов ириса при $\sigma = 0.1 \cdot \bar{x}_i$.

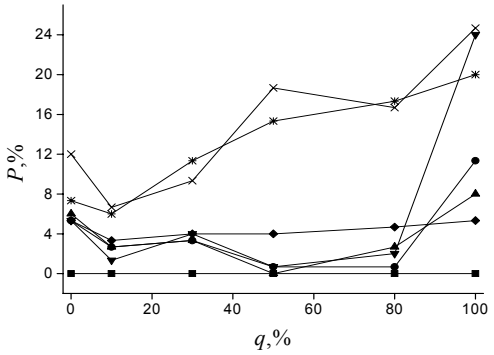


Рис. 4. Зависимость значений долей неправильно идентифицированных образцов ириса от доли грубых промахов, внесенных в исходные данные погрешностей
(—■— RBN, RBEN, PNN; —●— FFN; —▲— CFFN; —▼— FFTDN; —*— сеть Кохонена; —×— метод k-средних; —◆— дискриминантный анализ).

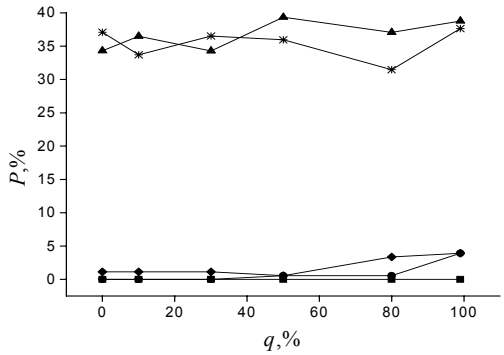


Рис. 5. Зависимость значений долей неправильно идентифицированных образцов вин от доли грубых промахов, внесенных в исходные данные погрешностей
(—■— RBN, RBEN, PNN, FFN, FFTDN; —●— CFFN; —▲— сеть Кохонена; —*— метод k-средних; —◆— дискриминантный анализ).

Сравнение долей неправильно идентифицированных образцов ириса и вин, полученных при использовании алгоритмов обучаемых нейронных сетей и в результате применения стандартных методов классификации, свидетельствует о большей устойчивости, и, как следствие, о предпочтительности использования алгоритмов искусственных нейронных сетей. Наиболее адаптивными алгоритмами (из проверенных авторами) являются нейронные сети радиального основания, которые безошибочно классифицируют/идентифицируют наборы данных различной структуры. Нейронную сеть Кохонена не рекомендуется использовать для работы с данными, характеризуемыми большим числом признаков (например, при идентификации образцов вина эффективность сети низка).

Алгоритмы, показавшие максимальную адаптивность, применили для классификации ряда растворителей, характеризующихся тремя сольватохромными параметрами (эмпирический параметр кислотности растворителя как донора водородных связей, эмпирический параметр основности растворителя как акцептора водородной связи, эмпирический параметр полярности и поляризуемости). Для обучения сети была использована выборка, включающая 45 веществ, принадлежащих к 6 различным классам [19, 20]. Обученную сеть тестировали на выборке из 10 веществ [20]. Состав обучающей выборки и результаты отнесения к различным классам веществ тестируемой выборки приведены в табл. 3.

Значения долей неправильно идентифицированных растворителей (табл. 4) указывают на то, что безошибочная идентификация достигается с применением каскадной сети прямой передачи и дискриминантного анализа.

Таблица 4. Результаты идентификации растворителей

Алгоритм	RBEN, RBN,	FFN, FFTDN	PNN	Дискриминантный анализ, CFFN
P, % (в обучаемой / тести-руемой выборке)	0/20	0/10	0/50	0/0

ЗАКЛЮЧЕНИЕ

Приведенные результаты характеризуют нейронные сети как перспективную альтернативу традиционным методам классификации. Эффективность работы алгоритмов нейронных сетей определяется свойствами образующих ее нейронов и индивидуальной архитектурой. Показана возможность применения алгоритмов искусственных нейронных сетей к данным с различной структурой, к модельным данным, содержащим "грубые промахи", а также для решения задач идентификации объектов по данным многооткликового эксперимента.

Авторы выражают благодарность репозитарию Center for Machine Learning and Intelligent Systems (School of Information and Computer Science, University of California, Irvine, CA) за предоставление доступа к массивам данных о характеристиках цветов ириса и образцов вин.

ЛИТЕРАТУРА

1. Cárdenas S., Valcárcel M. Analytical features in qualitative analysis // Trends Anal. Chem. – 2005. – V. 24, № 6. ─ P. 477–487.
2. Вершинин В.И. Дерендяев Б.Г., Лебедев К.С. Компьютерная идентификация органических соединений. М.: Академкнига, 2002. – 197 с.
3. Vlasov Y., Legin A. Non-selective chemical sensors in analytical chemistry: from "electronic nose" to "electronic tongue" // Fresenius J. Anal. Chem. – 1998. – V. 361. – P. 255-260.
4. Milman B.L. Identification of chemical compounds // Trends Anal. Chem. – 2005. – V. 24, № 6. – P. 493-508.
5. Taskinen J., Yliruusi J. Prediction of physicochemical properties based on neural network modelling // Advanced Drug Delivery Reviews. – 2003. – V. 55. – P. 1163-1183.
6. Kiss I.Z., Mandi G., Mihdly T. Beck. Artifical neural network approach to predict the solubility of C_{60} in various solvents // J. Phys. Chem. A. – 2000. – V. 104. – P. 8081-8088.
7. Sivaraman N., Srinivasan T. G., Vasudeva Rao P.R. QSPR modeling for solubility of fullerene (C_{60}) in organic solvents // J. Chem. Inf. Comput. Sci. – 2001. – V. 41 – P. 1067-1074.
8. Дюк В.А., Самойленко А.П. Data mining: учебный курс. – СПб.: Питер, 2001. –367 с.
9. Agatonovic-Kustrin S., Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research // J. of Pharm. and Biomed. Analysis. – 2000. – V. 22. – P. 717-727.
10. Хайкин С.. Нейронные сети: полный курс, 2-е изд., испр.: Пер. с англ. – М.: ООО "И.Д. Вильямс", 2006. – 1104 с.
11. Заенцев И.В. Нейронные сети: основные модели. Учебное пособие к курсу "Нейронные сети". – Воронеж: Воронежский Гос. ун-т, 1999. – 76 с.
12. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика, 2-е изд., стереотип. – М.: Горячая линия – Телеком, 2002. – 382 с.
13. Бэстенс Д.-Э., ван дер Берг В.-М., Вуд Д. Нейронные сети и финансовые рынки: принятие решений в торговых операциях. Пер. с англ. – М.: ТВП, 1997. – 236 с.
14. Медведев В.С., Потемкин В.Г. Нейронные сети. MATLAB 6 / Под общ. ред. к.т.н. В.Г. Потемкина. – М.: ДИАЛОГ-МИФИ, 2002. – 496 с.

15. Коняев Д.С. Методы анализа данных и химической информатики в исследовании комплексообразования в растворах и на поверхности химически модифицированных кремнеземов. Дисс. … канд. хим. наук. – Х.: – 1999. – с. 139-141.
16. Iris Data Set (1988). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Iris].
17. Wine Data Set (1991). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/wine].
18. Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. Пер. с болг. – М.: Финансы и статистика, 1987. – 239 с.
19. de Juan A., Fonrodona G., Casassas E. Solvent classification based on solvatochromic parameters: a comparison with the Snyder approach // Trends in Anal. Chem. – 1997. – V. 16, № 1. – P. 52-62.
20. Barwick V. J. Strategies for solvent selection – a literature review // Trends in Anal. Chem. – 1997. – V. 16, № 6. – P. 293-309.

Поступила в редакцию 10 марта 2010 г.

Я. М. Краснянчин, А. В. Пантелеймонов, Ю. В. Холін. Надійність ідентифікації аналітів за допомогою штучних нейронних мереж.

Апарат штучних нейронних мереж (алгоритми з навчанням і самоорганізована мережа Кохонена) застосували для вирішення задач ідентифікації об'єктів за даними багатовідгукового експерименту. Показано задовільну надійність ідентифікації за допомогою алгоритмів штучних нейронних мереж при розбитті на класи наборів тестових даних із складною структурою і за наявності в характеристиках об'єктів, що класифікуються, грубих промахів.

Ключові слова: штучна нейронна мережа, алгоритм, класифікація, ідентифікація, масив даних.

Ya. N. Krasnianchyn, A. V. Panteleimonov, Yu. V. Kholin. Reliability of identification of analytes in terms of artificial neural networks.

Artificial neural networks, namely various learned algorithms and self-organizing Kohonen's network, have been used to solve identification problems of objects on the basis of experimental multi-parameter data. The adequacy of algorithms under study has been demonstrated for classification of test samples and identification of iris samples, wines samples and solvents series. Stability of algorithms was also tested for the presence of gross errors in data sets.

Keywords: artificial neural networks, algorithm, classification, identification, data array.

Таблица 3. Состав обучающей выборки и результаты отнесения растворителей тестируемой выборки (неправильное отнесение отмечено знаком *)

Номер класса	Классы веществ по классификации Снайдера [20]	Обучающая выборка	Тестируемая выборка			
			RBN, RBEN	FFN, FFTDN	CFFN, дискриминантный анализ	PNN
1	алифатические эфиры и замещенные алифатические амины	диизопропиловый эфир, ди-н-бутиловый эфир, диэтиловый эфир, триэтиламин				циклогексанон*, γ-бутиролактон*, нитрометан*, ацетофенон*, фторбензол*,
2	алифатические циклические эфиры, вещества, содержащие карбонильную функциональную группу (сложные эфиры, кетоны), нитрилы	диоксан, тетрагидрофуран, дибензиловый эфир, 2-бутанон, ацетон, этилацетат, этилбензоат, пропиленкарбонат, нитробензол, бензонитрил, ацетонитрил	п-ксилол*, циклогексанон, γ-бутиролактон, нитрометан, ацетофенон, фторбензол*	циклогексанон, γ-бутиролактон, ацетофенон	циклогексанон, γ-бутиролактон, нитрометан, ацетофенон	
3	пиридины, амиды, сульфоксиды, мочевины, фосфортриамиды	диметилацетамид, диметилформамид, N-метилпирролидон, тетраметилмочевина, диметилсульфоксид, гексаметилфосфортриамид, пиридин, 2,6-диметилпиридин, хинолин				
4	ароматические соединения (эфиры, углеводороды и их галогенпроизводные), алифатические галогензамещенные углеводороды	анизол, дифениловый эфир, этилфениловый эфир, толуол, бензол, хлорбензол, бромбензол, карбонтетрахлорид, 1,2-дихлорэтан, метиленхлорид, хлороформ	хлорэтан,	нитрометан*, п-ксилол, хлорэтан, фторбензол	п-ксилол, хлорэтан, фторбензол	п-ксилол, хлорэтан,
5	спирты, вода	трет-бутанол, изопропанол, н-бутанол, этанол, метанол, этиленгликоль, вода				
6	алифатические углеводороды	пентан, гексан, гептан	декан, гексадекан, изооктан	декан, гексадекан, изооктан	декан, гексадекан, изооктан	декан, гексадекан, изооктан