УДК 543.33 + 004.032.26

КЛАССИФИКАЦИЯ ХИМИКО-АНАЛИТИЧЕСКИХ ДАННЫХ НА ОСНОВЕ ОБЪЕДИНЕНИЯ НЕЙРОННОЙ СЕТИ КОХОНЕНА И ВЕРОЯТНОСТНОЙ НЕЙРОННОЙ СЕТИ

Я. Н. Пушкарева, Н. П. Титова, О. И. Юрченко, Ю. В. Холин

В статье описана и апробирована процедура классификации объектов по данным об их химико-аналитических характеристиках на основе объединения сети Кохонена и вероятностной сети. В отличие от существующих алгоритмов, предложенная процедура не требует привлекать априорную информацию ни о числе классов, ни о составе обучающей выборки. Процедура испытана при классификации образцов речных и родниковых вод г. Харькова по данным о содержании в них ионов металлов.

Ключевые слова: качественный химический анализ, классификация, сеть Кохонена, вероятностная сеть.

Введение

Современный качественный химический анализ трактуется как процедура классификации (в частности, идентификации и дискриминации) объектов по их химическим, физико-химическим и другим характеристикам [1–3]. Алгоритмы классификации составляют важный раздел хемометрии и используются для извлечения важной химической информации из многомерных массивов экспериментальных данных, характеризующих свойства и особенности веществ, материалов, продуктов питания, лекарственных препаратов и других объектов [4, 5].

Алгоритмы классификации делятся на две группы: классификация “с обучением” (дискриминантный анализ, формальное независимое моделирование аналогий классов) и “без обучения” (кластерный анализ, метод k-средних). Алгоритмы классификации “с обучением” применяют обучающий набор образцов с известной классовой принадлежностью (обучающую выборку) для выработки классификационных правил. Алгоритмы классификации “без обучения” не используют обучающую выборку, но требуют *a priori* задавать число классов [6]. Таким образом, для реализации классификационных алгоритмов как “с обучением”, так и “без обучения” необходимой информацией является число классов.

Химики-аналитики часто получают массивы экспериментальных данных, для которых число однородных групп неизвестно, а критерии отнесения образцов к тем или иным классам нечеткие или противоречивые. В этом случае обработку данных осуществляют, комбинируя различные хемометрические методы [7–9], что является трудоемким и длительным процессом.

Настоящая статья посвящена определению числа классов и нахождению устойчивой классификации с помощью процедуры на основе объединения сети Кохонена без обучения и вероятностной сети с обучением. Нейронные сети зарекомендовали себя мощным и робастным классификационным инструментом [10–12]; сведения о них, в частности, о сети Кохонена и вероятностной сети, можно найти в работах [13–16]. Предложенную процедуру верифицировали на массиве экспериментальных данных о содержании ионов 8 металлов в образцах вод из различных источников и рек г. Харькова, отобранных в разные сезоны в течение 2008–2010 годов.

Алгоритм процедуры классификации

Алгоритм классификации без априорной информации о числе классов и без наличия обучающей выборки на основе объединения нейронной сети Кохонена и вероятностной сети состоит из следующих этапов:

1) классификация данных с помощью сети Кохонена при различных значениях числа нейронов (соответственно, и числа классов);

Таблица 1. Результаты классификации образцов речных вод

Река, год отбора и анализа пробы	Концентрации ионов металлов, мг/л								Класс
	Zn	Cu	Mn	Fe	Cd	Pb	Co	Ni	
Немышля, 2008 г.	0.0150	0.0055	0.0230	0.1000	0.0021	0.0350	0.0084	0.0200	1
	0.0130	0.0045	0.0410	0.1200	0.0015	0.0440	0.0088	0.0180	1
	0.0080	0.0073	0.0170	0.0280	0.0024	0.0350	0.0094	0.0140	1
	0.0120	0.0073	0.0130	0.0380	0.0022	0.0470	0.0063	0.0130	1
	0.0120	0.0078	**0.4330**	0.0190	0.0022	0.0460	0.0140	0.0200	2
Харьков, 2009 г.	0.0050	0.0034	0.0020	0.0220	0.0017	0.0410	0.0105	0.0160	3
	0.1040	0.0103	0.0070	0.0190	0.0019	0.0410	0.0112	0.0120	3
	0.0150	0.0043	0.0030	0.0170	0.0017	0.0340	0.0090	0.0080	3
	0.0070	0.0052	0.0030	0.0160	0.0039	0.0310	0.0085	0.0100	3
	0.0100	0.0052	0.0030	0.0170	0.0030	0.0230	0.0097	0.0070	3
Лопань, 2009 г.	0.0470	0.0043	0.0050	0.0110	0.0032	0.0310	0.0070	0.0120	3
	0.1090	0.0069	0.0030	0.0170	0.0052	0.0430	0.0068	0.0140	3
	0.0650	0.0069	0.1330	0.0220	0.0060	0.0380	0.0078	0.0130	3
	0.0100	0.0078	0.0060	0.0170	0.0030	0.0280	0.0075	0.0190	3
	0.0110	0.0087	0.0400	0.0230	0.0023	0.0380	0.0107	0.0160	3
Уды, 2010 г.	0.0120	0.0056	0.2790	0.0690	н/о*	н/о	0.0125	0.0110	4
	0.0170	0.0056	0.0030	0.1460	н/о	н/о	н/о	0.0040	4
	0.0050	0.0032	н/о	0.0110	н/о	н/о	0.0071	0.0040	4
	0.0300	0.0286	0.0150	0.0970	0.0090	**0.2000**	**0.0554**	0.0300	2
	0.0090	0.0063	0.0230	0.0460	н/о	0.0380	0.0089	**0.0390**	2
	0.0080	0.0110	0.0130	0.0430	0.0021	0.0100	0.0047	0.0070	4
	0.0110	0.0080	0.0280	0.0460	0.0015	0.0250	0.0078	0.0090	4

* Здесь и в табл. 2 – не обнаружено

Таблица 2. Результаты классификации образцов родниковых вод

Источник, год отбора и анализа пробы	Концентрации ионов металлов, мг/л								Класс
	Zn	Cu	Mn	Fe	Cd	Pb	Co	Ni	
Саржин Яр “Харьковская-1”, 2010 г.	0.0090	0.0100	0.0300	0.0480	0.0024	0.0250	0.0084	0.0080	1
	0.0130	0.0080	0.0240	0.0720	0.0024	0.0280	0.0063	0.0090	1
	0.0120	0.0080	0.0260	0.0500	0.0018	0.0250	0.0078	0.0090	1
	0.0120	0.0050	0.0200	0.1020	0.0019	0.0230	0.0084	0.0110	1
“Харьковская -2”, 2010 г.	0.0070	0.0050	0.0180	0.0570	0.0012	0.0230	0.0069	0.0120	1
	0.0050	0.0040	0.0130	0.0590	0.0016	0.0190	0.0069	0.0110	1
	0.0070	0.0040	0.0080	0.0330	0.0018	0.0160	0.0103	0.0120	1
	0.0080	0.0050	0.0190	0.0570	0.0021	0.0230	0.0063	0.0120	1
	0.0190	0.0180	0.0080	0.0640	0.0029	0.0350	0.0078	0.0200	1
	0.0080	0.0750	0.0610	0.0760	0.0024	0.0290	0.0056	0.0200	1
	0.0080	0.0230	0.0700	0.1200	0.0022	0.0280	0.0063	0.0200	1
Пантелеймоновская церковь, 2010 г.	0.0080	0.0190	0.0220	0.0550	0.0019	0.0350	0.0094	0.0200	1
	0.0100	0.0930	0.0180	0.0370	0.0025	0.0250	0.0078	0.0200	1
	0.0050	0.0070	0.0410	0.0280	0.0019	**н/о**	0.0078	0.0090	2
	0.0060	0.0080	0.0210	0.0280	0.0026	**н/о**	0.0063	0.0100	2
Завод пищевых кислот, 2010 г.	0.0100	0.0400	0.0140	0.0260	0.0012	0.0160	0.0078	0.0100	3
	0.0140	0.0150	0.0240	0.0780	0.0026	0.0160	0.0078	0.0140	3
	0.0230	0.0220	0.0130	0.0430	0.0024	0.0130	0.0100	0.0140	3
ул. Уборевича, 2009 г.	0.0230	0.0050	0.0110	0.0080	0.0083	0.0110	0.0133	0.0120	4
	0.0140	0.0040	0.0020	0.0150	0.0053	0.0150	0.0125	0.0140	4
	0.0180	0.0050	0.0030	0.0150	0.0049	0.0270	0.0063	0.0140	4
	0.0140	**0.0020**	0.0040	0.0100	0.0052	0.0170	**н/о**	0.0110	2
Парк “Юность”, 2009 г.	0.0340	0.0040	0.0150	0.0520	0.0123	0.0210	0.0125	0.0380	4
	0.0150	0.0030	0.0030	0.0290	0.0086	0.0070	0.0125	0.0060	4

2) определение групп образцов, которые независимо от числа задаваемых нейронов отнесены сетью Кохонена к одному и тому же классу; использование этих образцов в качестве первой обучающей выборки для обучения вероятностной сети;

3) случайное формирование небольших выборок из образцов, не вошедших в первую обучающую выборку, и последовательное их предъявление на вход вероятностной сети в качестве тестовых выборок;

4) включение образцов каждой тестовой выборки в обучающую выборку после их классификации вероятностной сетью (обучающая выборка увеличивается, что обеспечивает адекватную классификацию последующих тестовых выборок);

5) проведение кросс-валидации ("leave-one-out" cross validation [17]) для проверки и уточнения полученной классификации.

Расчеты выполняли в пакете MATLAB 6.5.

Результаты и обсуждение

Анализируемый массив данных включает 22 образца речных вод и 24 образца родниковых вод, отобранных в г. Харькове. В образцах были измерены концентрации ионов меди, цинка, свинца, кадмия, марганца, железа, кобальта и никеля [18] (табл. 1, 2). Относительное стандартное отклонение определяемых концентраций не превышало 0.03.

Поскольку концентрации ионов металлов в образцах вод варьировались в диапазоне от тысячных до десятых мг/л, перед применением процедуры классификации провели автомасштабное преобразование исходных данных [19]:

$$x_i^{norm} = \frac{x_i - \bar{x}}{std(x)}, i = 1,2,...,N,$$

где x^{norm} – преобразованная безразмерная концентрация данного элемента в *i*-м образце вод (величины x^{norm} имеют нулевое среднее и единичную дисперсию), x_i – концентрация данного элемента в i-м образце, $\bar{x}$ – среднее значение концентрации данного элемента в образцах, $std(x)$ – стандартное отклонение концентрации данного элемента в образцах, N – число образцов вод.

Классификацию массивов данных провели согласно процедуре, описанной выше.

На первом этапе варьировали число нейронов от 3 до 7. В результате применения сети Кохонена для образцов речных вод выявили 4 группы (10 образцов), которые независимо от числа нейронов отнесены к одному и тому же классу, для образцов родниковых вод – 5 групп (14 образцов). Из оставшихся образцов сформировали 3 тестовые выборки для образцов речных вод и 2 тестовые выборки для образцов родниковых вод. Окончательная классификация образцов вод в результате выполнения этапов 3–5 представлена в табл. 1, 2. Следует отметить, что две группы образцов родниковых вод были объединены, т.к. одна из них включала только два образца, что не позволило подтвердить их выделение в отдельный класс.

Полученная классификация образцов речных и родниковых вод соответствует их происхождению. Образцы, отобранные из различных рек или источников, не перемешаны между собой; наблюдается только объединение некоторых образцов, отобранных из различных рек и источников (реки Харьков и Лопань, источники "Харьковская-1", "Харьковская-2" и в районе Пантелеймоновской церкви) в силу близости их характеристик.

В случае образцов речных вод класс № 2 включает наиболее загрязненные образцы, характеризующиеся наибольшими содержаниями марганца, свинца, кобальта, никеля, существенно превышающими содержание этих металлов в других образцах.

Для родниковых вод класс № 2 включает наименее загрязненные образцы вод (анализ родниковых вод был направлен на обнаружение источника, вода из которого наиболее пригодна для употребления) с наименьшими концентрациями цинка, меди, свинца и кобальта.

Заключение

Показана эффективность процедуры классификации на основе объединения сети Кохонена и вероятностной сети для определения однородных групп образцов на примере обработки масси-

вов многомерных результатов химического анализа. Алгоритм можно рекомендовать для эксплораторного анализа (предварительной обработки) химико-аналитических данных и для решения задач дискриминации и идентификации в качественном химическом анализе.

Работа выполнена при финансовой поддержке Фонда фундаментальных, прикладных и поисковых научно-исследовательских работ ХНУ имени В. Н. Каразина (номер государственной регистрации 0112U003024).

Литература

1. Vlasov Yu., Legin A., Rudnitskaya A., Di Natale C., D'Amico A. Nonspecific sensor arrays ("electronic tongue") for chemical analysis of liquids // Pure Appl. Chem. 2005. 77(11). P. 1965.
2. Hardcastle W. A. Qualitative analysis: a guide to best practice. Cambridge: Royal Society of Chemistry, 1998, 24 p. ISBN 0-85404-462-0.
3. Milman B. L. Identification of chemical compounds // Trends Anal. Chem. 2005. 24(6). P. 493.
4. Родионова О. Е., Померанцев А. Л. Хемометрика в аналитической химии, 2006, 61 с. http://www.chemometrics.ru/materials/articles/chemometrics_review.pdf
5. Adams M. J. Chemometrics in analytical spectroscopy (2nd ed.). Cambridge: Royal Society of Chemistry, 2004, 238 p. ISBN 0-85404-555-4.
6. Mutihac L., Mutihac R. Mining in chemometrics // Anal. Chim. Acta. 2008. 612. P. 1.
7. de Juan A., Fonrodona G., Casassas E. Solvent classification based on solvatochromic parameters: a comparison with the Snyder approach // Trends Anal. Chem. – 1997. 16(1). P. 52.
8. Simeonov V., Simeonova P., Tsakovskii S., Lovchinov V. Lake water monitoring data assessment by multivariate statistics // J. Water Resource Protect. 2010. 2. P. 353.
9. Skorek R., Jablonska M., Polowniak M., Kita A., Janoska P., Buhl F. Application of ICP-MS and various computational methods for drinking water quality assessment from the Silesian District (Southern Poland) // Centr. Eur. J. Chem. 2010. 10(1). P. 71.
10. Balabin R. M., Safieva R. Z., Lomakina E. I. Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques // Anal. Chim. Acta. 2010. 671. P. 27.
11. Galao O. F., Borsato D., Pinto J. P., Visentainer J. V., Carrao-Panizzi M. C. Artificial neural networks in the classification and identification of soybean cultivars by planting region // J. Braz. Chem. Soc. 2011. 22(1). P. 142.
12. Pushkarova Ya., Kholin Yu. The classification of solvents based on solvatochromic characteristics: the choice of optimal parameters for artificial neural networks // Centr. Eur. J. Chem. 2012. 10(4). P. 1318.
13. Краснянчин Я. Н., Пантелеймонов А. В., Холин Ю. В. Надежность идентификации аналитов с помощью искусственных нейронных сетей // Вісник Харківського національного ун-ту. 2010. № 895. Хімія. Вип. 18(41). С. 39.
14. Краснянчин Я. Н., Пантелеймонов А. В., Холин Ю. В. Некоторые аспекты параметризации искусственных нейронных сетей в задачах качественного химического анализа // Вісник Харківського національного ун-ту. 2010. № 932. Хімія. Вип. 19(42). С. 170.
15. Круглов В. В., Борисов В. В. Искусственные нейронные сети. Теория и практика. М.: Горячая линия-Телеком, 2002, 382 с. ISBN 5-93517-031-0.
16. Осовский С. Нейронные сети для обработки информации / Пер. с польского. М.: Финансы и статистика, 2002, 344 с. ISBN 5-279-02567-4.
17. Dong M., Wang N. Adaptive network-based fuzzy inference system with leave-one-out cross-validation approach for prediction of surface roughness // Appl. Math. Model. 2011. 35(3). P. 1024.
18. Пушкарева Я. Н., Следзевская А. Б., Пантелеймонов А. В., Титова Н. П., Юрченко О. И., Иванов В. В., Холин Ю. В. Идентификация образцов воды источников и рек г. Харьков: сравнение методов многомерного анализа данных // Вестн. Моск. ун-та. Серия 2. Химия. 2012. 53(6). С. 405.
19. Шараф М. А., Иллмэн Д. Л., Ковальски Б. Р. Хемометрика / Пер. с англ. Ленинград: Химия, 1989, 272 с. ISBN 5-7245-0361-1.

References

1. Vlasov Yu., Legin A., Rudnitskaya A., Di Natale C., D'Amico A. Nonspecific sensor arrays ("electronic tongue") for chemical analysis of liquids // Pure Appl. Chem. 2005. 77(11). P. 1965.
2. Hardcastle W. A. Qualitative analysis: a guide to best practice. Cambridge: Royal Society of Chemistry, 1998, 24 p. ISBN 0-85404-462-0.
3. Milman B. L. Identification of chemical compounds // Trends Anal. Chem. 2005. 24(6). P. 493.
4. Rodionova O. Ye., Pomerantsev A. L. Hemometrika v analiticheskoy himii, 2006, 61 p. http://www.chemometrics.ru/materials/articles/chemometrics_review.pdf [in Russian]
5. Adams M. J. Chemometrics in analytical spectroscopy (2nd ed.). Cambridge: Royal Society of Chemistry, 2004, 238 p. ISBN 0-85404-555-4.
6. Mutihac L., Mutihac R. Mining in chemometrics // Anal. Chim. Acta. 2008. 612. P. 1.
7. de Juan A., Fonrodona G., Casassas E. Solvent classification based on solvatochromic parameters: a comparison with the Snyder approach // Trends Anal. Chem. – 1997. 16(1). P. 52.
8. Simeonov V., Simeonova P., Tsakovskii S., Lovchinov V. Lake water monitoring data assessment by multivariate statistics // J. Water Resource Protect. 2010. 2. P. 353.
9. Skorek R., Jablonska M., Polowniak M., Kita A., Janoska P., Buhl F. Application of ICP-MS and various computational methods for drinking water quality assessment from the Silesian District (Southern Poland) // Centr. Eur. J. Chem. 2010. 10(1). P. 71.
10. Balabin R. M., Safieva R. Z., Lomakina E. I. Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques // Anal. Chim. Acta. 2010. 671. P. 27.
11. Galao O. F., Borsato D., Pinto J. P., Visentainer J. V., Carrao-Panizzi M. C. Artificial neural networks in the classification and identification of soybean cultivars by planting region // J. Braz. Chem. Soc. 2011. 22(1). P. 142.
12. Pushkarova Ya., Kholin Yu. The classification of solvents based on solvatochromic characteristics: the choice of optimal parameters for artificial neural networks // Centr. Eur. J. Chem. 2012. 10(4). P. 1318.
13. Krasnianchyn Ya. N., Panteleimonov A. V., Kholin Yu. V. // Visn. Hark. nac. univ., 2010, № 895, Ser. Him., issue. 18(41), P. 39. [ISSN 2220-637X (print), ISSN 2220-6396 (online), http://chembull.univer.kharkov.ua/archiv/2010/05.pdf] [in Russian].
14. Krasnianchyn Ya. N., Panteleimonov A. V., Kholin Yu. V. // Visn. Hark. nac. univ., 2010, № 932, Ser. Him., issue. 19(42), P. 170. [ISSN 2220-637X (print), ISSN 2220-6396 (online), http://chembull.univer.kharkov.ua/archiv/2010_2/21.pdf] [in Russian].
15. Kruglov V. V., Borisov V. V. Iskusstvenny'e neyronny'e seti. Teoriya i praktika. M.: Goryachaya liniya-Telekom, 2002, 382 p. ISBN 5-93517-031-0. [in Russian]
16. Osowski S. Sieci neuronowe do przetwarzania informacji. Warszawa: Oficyna wydawnicza politechniki Warszawskiej, 2000. ISBN 83-7207-187-X.
17. Dong M., Wang N. Adaptive network-based fuzzy inference system with leave-one-out cross-validation approach for prediction of surface roughness // Appl. Math. Model. 2011. 35(3). P. 1024.
18. Pushkarova Ya. N., Sledzevska A. B., Panteleimonov A. V., Titova N. P., Yurchenko O. I., Ivanov V. V., Kholin Yu. V. // Moscow Univ. Chem. Bull. 2012. 67(6). P. 287.
19. Sharaf M. A., Illman D. L., Kowalski B. R. Chemometrics. New York, Chichester, Brasbane, Totonto, Singapore: John Wiley & Sons, 1986, 332 p. ISBN 0471831069.

Поступила в редакцию 17 июня 2012 г.

Я. М. Пушкарьова, Н. П. Тітова, О. І. Юрченко, Ю. В. Холін. Класифікація хіміко-аналітичних даних на основі поєднання нейронної мережі Кохонена та імовірнісної нейронної мережі.

В статті описано й апробовано процедуру класифікації об'єктів за даними про їх хіміко-аналітичні характеристики на основі об'єднання мережі Кохонена та ймовірнісної мережі. На відміну від існуючих алгоритмів, запропонована процедура не вимагає залучати апріорну інформацію ані про число класів, ані про склад навчальної вибірки. Процедуру випробувано при класифікації зразків річкових і джерельних вод м. Харкова за даними про вміст у них іонів металів.

Ключові слова: якісний хімічний аналіз, класифікація, мережа Кохонена, імовірнісна мережа.

Ya. N. Pushkarova, N. P. Titova, O. I. Yurchenko, Yu. V. Kholin. Classification of chemical analytical data with the use of a combination of the Kohonen and the probabilistic neural networks.

The paper presents a novel procedure capable to classify objects proceeding from their chemical characteristics. The approach is based on a combinantion of the Kohonen and the probabilistic neural networks. In contrast to existing analogs, the procedure does not require any a priori information about the number of classes and the patterns in the training set. To verify the procedure, the problem of the classification of water samples from different Kharkiv springs and rivers has been considered. The initial experimental data set consisted of concentrations of metal ions in water samples.

Key words: qualitative chemical analysis, classification, Kohonen neural network, probabilistic neural network.

Kharkov University Bulletin. 2012. № 1026. Chemical Series. Issue 21 (44).