УДК 541.48/.486

М. А. ЗИЛЬБЕРГЛЕЙТ

ОТБОР ИСХОДНЫХ ДАННЫХ ДЛЯ ФОРМИРОВАНИЯ МАССИВОВ БАЗЫ ДАННЫХ НА ПРИМЕРЕ МНОГОКОМПОНЕНТНОЙ СИСТЕМЫ NACL–KCL–MCL_2·H_2O

Государственное научное учреждение «Институт общей и неорганической химии» Национальной академии наук Беларуси

Использование данных из различных источников при формировании БД наталкивается на естественные трудности, характеризующихся частичной противоречивостью и несопоставимостью результатов, полученных разными исследователями в разное время. В работе предложен способ поиска данных типа unusual data (данные вызывающие сомнения) при формировании баз данных из нескольких источников. В качестве критерия для поиска таких данных предложено использовать метод распознавания образов – решающее правило. При использовании двух и более решающих правил, рассматривающие различные аспекты данных, образуются пересечения, в которых присутствуют объекты, которые невозможно правильно классифицировать при помощи таких решающих правил. В качестве примера поиска данных типа unusual data показано применение этого метода при анализе данных по гетерогенному равновесию в системе NaCl–KCl–MgCl·H2O, опубликованных в 12 различных источниках за разные периоды времени. В частности предложены три критерия отбора для семи разных составов твердой фазы, трех составов, различающихся количеством компонентов в нем – т. е. одно, двух и трех компонентов, а также для составов, в которых входят и не входят кристаллогидраты. В результате получены наборы решающих правил, которые с вероятностью 92–98% правильно классифицируют отобранные объекты. Часть неправильно классифицированных объектов характерны для всех трех критериев отбора. Эти объекты предложено отмечать в БД, как данные вызывающие сомнения. Получено решающее правило, позволяющее с вероятностью 98% прогнозировать наличие или отсутствие кристаллогидрата в твердой фазе в зависимости от состава насыщенного раствора NaCl, KCl, MCl2·H2O.

Ключевые слова: *отбор исходных данных, данные вызывающие сомнения, unusual data, распознавание образов, линейные и нелинейные классификаторы, решающее правило, пересечение множеств, многокомпонентная гетерогенная химическая система.*

Введение. Предшествующий период накопления химических знаний можно очевидно охарактеризовать как период, связанный с хаотичным сбором экспериментального материала. Результатом деятельности явилось создание многочисленных химических справочников, в которых были собраны десятки, если не сотни экспериментальных данных, относящихся к одной и той же изучаемой проблеме. Эти данные собирались различными исследовательскими коллективами, различными авторами и иногда с использованием различных методов проведения эксперимента и анализа. Попытка свести эти данные воедино зачастую наталкиваются на проблемы, связанные с противоречивостью и несопоставимостью некоторых экспериментальных значений. Ручной отбор и анализ такого материала представляется крайне трудоемкой и непосильной задачей. Подход к такого рода исследованиям в плане анализа пропущенных данных развит в работах [1–4]. В тоже время работы, посвященные очистке данных, чаще всего ограничиваются исследованием на уровне непосредственного эксперимента путем применения стандартных статистических критериев для удаления выбросов.

В данном исследовании приведен подход поиска сомнительных данных из разных источников уже после того как они были собраны и опубликованы в справочных изданиях. Под термином сомнительные данные в данном случае предполагается поиск данных, которые в зарубежной статистической литературе носят название unusual data, т. е. данные вызывающие вопросы.

В качестве исходного метода исследования выдвинута концепция использования одного из методов распознавания образов – формирование решающего правила для поиска unusual data. В данном сообщении нет смысла останавливаться на основных идеях распознавания образов и получения решающего правила, так как они подробно изложены в соответствующей литературе [5–7]. Отметим лишь, что решающее правило представляет собой функцию, значение которой или ее знак позволяет отнести объект к тому или иному классу. Известно, что если решающее правило не дает полного распознавания, то существуют записи, которые не распознаются по такому правилу. Если использовать два и более решающего правила, классифицирующие совокупность объектов по различным признакам, то возможно обнаружить пересечение записей, которые не соответствуют двум и более классификациям. В последнем случае предлагается обозначать их как unusual data и помечать соответствующим образом. Иными словами существуют данные, которые не могут быть распознаны различными решающими правилами, что дает возможность предполагать о наличие в таких данных дефектов.

Экспериментальный материал для анализа

В качестве исходного материала нами выбраны данные по гетерогенному рав-новесию в системе NaCl–KCl–MgCl–H_2O. Выбор такой системы обусловлен тем, что на технологические процессы галургического разделения солей данного состава в зависимости их растворимости и соотношения компонентов, а также температуры до сих пор являются предметом изучения. Источник информации – «Справочник экспериментальных данных по растворимости многокомпонентных водно-солевых систем», т. 2, Четырехкомпонентные и более сложные системы, составители А. Б. Здановский, Е. И. Ляховская, Р. Э. Шлеймович, Государственное издательство научно-технической литературы, Ленинград, 1964 г. В данном издании использовались данные E. Kayser, Kali, № 17, 7–8, 38, 1923; Н. С. Курнаков, Н. А. Осокорева, Калий, № 2, 27, 1932, Труды ГИПХ, вып. 16, 42–45, 1932, Соликамские карналиты, 61–67, 1935; Е. А. Ахумов, М. П. Головков, ЖОХ, т. 5, вып. 4, 507, 1935; Н. И. Хайдуков, Э. Г. Линецкая, Калий, № 8, 33, 1935; Н. С. Курнаков, А. И. Заславский, Е. И. Лукьянова, ГИПХ, 1936. F. Frowein, E. von Muehlendahtl, Z. anorg. Chem., v. 39, 1492–1493б 1926; G. Leimbach, Kali, № 1, 12–13, 1926; J. D Ans, A. Bertsch, A. № 9, 153–154, 1915; H/ Keitel, Mitt, KFA, v. 32, 112–117, 1922; W. Feit, K. Przibylla, Kali, № 18, 394, 1909; № 14, 300, 1910; Michels Przibylla, Die Kalirohsalze, Ihre Gewinnung und Verarbeitung, Leipzig, 106, 1916; H. Precht, B/Wittjen, Ber., v. 14, 1673, 1881; Serowy, Kali, № 17, 347–348, 1923, а также единичные дополнительные данные ряда авторов за 1898–1913 годы, приведенные на 743 стр. данного справочника. Количество отобранных записей составило 231. *Запись базы данных* – это строка таблицы, содержащая набор значений свойств.

Авторы справочника в предисловии к изданию упоминают, что они на основании полученных данных создают сводную вероятностную таблицу, однако способ сведения данных так и не указан.

Фрагмент таблицы из упомянутого выше справочника приведен на рис. 1.

Температурные интервалы выбранных записей находились в пределах 0–105 ^oC. Качественный состав твердой фазы характеризовался следующим образом: NaCl+KCl, NaCl+KCl+$KCl{\cdot}MgCl_2{\cdot}6H_2O$, NaCl+$KCl{\cdot}MgCl_2 6H_2O$+$MgCl_2{\cdot}6H_2O$, NaCl+$KCl{\cdot}MgCl_2{\cdot}6H_2O$, KCl+$KCl{\cdot}MgCl_2{\cdot}6H_2O$, NaCl, KCl.

Для разделения использовались линейные и нелинейные дискриминаторы, которые давали наилучшее в данных условиях разделение.

Обсуждение полученных результатов

Таким образом, для классификации использовались три вида решающего правила: для 7 разных составов твердой фазы, трех составов, различающихся количеством компонентов в нем – т. е. 1, 2 и 3 компонента, а также для составов, в которых входят и не входят кристаллогидраты.

В табл. 1–3 приведены результаты классификации для предложенных выше решающих правил.

Очевидно, что даже самый трудный способ классификации (7 различных составов твердой фазы) дал вполне удовлетворительные результаты – суммарный итог разделения составил около 91%, при этом наихудшие результа-

736

$NaCl—KCl—MgCl_2—H_2O$

Н. С. Курнаков, Н. А. Осокорева, Калий, № 2, 27 (1932) [25 и 100°]; Труды ГИПХ, вып. 16, 42—45 (1932); Соликамские карналлиты, стр. 61—64, 1935

t, °C	Жидкая фаза: вес. % NaCl	вес. % KCl	вес. % $MgCl_2$	г/100 г солей NaCl	г/100 г солей KCl	г/100 г солей $MgCl_2$	M/1000 M H_2O 2NaCl	M/1000 M H_2O 2KCl	M/1000 M H_2O $MgCl_2$	индексы 2NaCl	индексы 2KCl	индексы H_2O	d	Твердая фаза
10	12,57	6,59	9,50	43,86	22,99	33,14	27,15	11,16	25,19	42,76	17,57	1575	1,235	$NaCl + KCl$
	5,42	4,43	18,59	19,06	15,57	65,37	11,67	7,48	49,10	17,10	10,96	1465	1,245	" "
	1,92	2,67	25,43	6,41	8,92	84,67	4,22	4,60	68,41	5,46	5,96	1295	1,270	$NaCl + KCl + KCl \cdot MgCl_2 \cdot 6H_2O$
	1,09	0,57	29,66	3,48	1,82	94,70	2,45	1,05	81,69	2,88	1,23	1174	1,289	$NaCl + KCl \cdot MgCl_2 \cdot 6H_2O$
	0,68	2,63	26,08	2,31	8,95	88,74	1,48	4,50	69,86	1,95	5,93	1319	1,265	$KCl + KCl \cdot MgCl_2 \cdot 6H_2O$
	0,27	0,07	34,78	0,80	0,20	99,00	0,67	0,13	101,41	0,66	0,13	978	1,334	$NaCl + KCl \cdot MgCl_2 \cdot 6H_2O + MgCl_2 \cdot 6H_2O$
20	13,85	8,33	7,60	46,51	27,97	25,52	30,39	14,33	20,47	46,62	21,98	1534	1,235	$NaCl + KCl$
	1,88	3,23	25,44	6,15	10,57	83,28	4,17	5,62	69,29	5,27	7,11	1265	1,275	$NaCl + KCl + KCl \cdot MgCl_2 \cdot 6H_2O$
	1,42	1,29	28,30	4,58	4,16	91,26	3,17	2,26	77,59	3,82	2,72	1205	1,283	$NaCl + KCl \cdot MgCl_2 \cdot 6H_2O$
	0,35	0,08	35,22	0,98	0,24	98,80	0,84	0,16	103,54	0,80	0,15	957	1,337	$NaCl + KCl \cdot MgCl_2 \cdot 6H_2O + MgCl_2 \cdot 6H_2O$
25	6,87	6,39	15,87	23,58	21,94	54,48	14,97	10,89	42,76	21,81	15,87	1457	1,244	$NaCl + KCl$

Рис. 1. Снимок таблицы экспериментальных данных из [8]

Таблица 1. **Результаты классификации солей для 7 разных составов твердой фазы**

Исходные группы	Размер групп	Полученные группы						
		1	2	3	4	5	6	7
1	116	108 (93,10%)	8 (6,90%)	0 (0,00%)	0 (0,00%)	0 (0,00%)	0 (,00%)	0 (0,00%)
2	44	0 (0,00%)	40 (90,91%)	0 (0,00%)	1 (2,27%)	3 (6,82%)	0 (0,00%)	0 (0,00%)
3	24	0 (0,00%)	1 (4,17%)	23 (95,83%)	0 (0,00%)	0 (0,00%)	0 (0,00%)	0 (0,00%)
4	6	0 (0,00%)	0 (0,00%)	1 (16,67%)	4 (66,67%)	1 (16,67%)	0 (0,00%)	0 (0,00%)
5	5	0 (0,00%)	1 (20,00%)	0 (0,00%)	0 (0,00%)	4 (80,00%)	0 (0,00%)	0 (0,00%)
6	17	2 (11,76%)	0 (0,00%)	0 (0,00%)	0 (0,00%)	0 (0,00%)	15 (88,24%)	0 (0,00%)
7	19	2 (10,53%)	0 (0,00%)	0 (0,00%)	0 (0,00%)	1 (5,26%)	0 (0,00%)	16 (84,21%)

Таблица 2. **Результаты классификации солей для разных составов твердой фазы, различающихся количеством компонентов в нем**

Исходные группы	Размер групп	Полученные группы		
		1	2	3
1	36	30 (83,33%)	6 (16,67%)	0 (0,00%)
2	127	0 (0,00%)	115 (90,55%)	12 (9,45%)
3	68	0 (0,00%)	1 (1,47%)	67 (98,53%)

Таблица 3. **Результаты классификации солей для разных составов твердой фазы, различающихся наличием или отсутствием кристаллогидратов**

Исходные группы	Размер групп	Полученные группы	
		1	2
1	79	79 (100,00%)	0 (0,00%)
2	152	4 (2,63%)	148 (97,37%)

ты были получены для групп четыре и пять. Это и неудивительно, так как их количество среди всей выборки составило примерно 2,3–3%. Можно было бы выбросить эти результаты, т. к. они были найдены только у одного автора. Однако учитывая, что их опубликовал классик науки о физико-химическом анализе водно-солевых систем – академик Н. С. Курнаков, они были оставлены.

Ниже представлены все номера всех записей, качество классификации которых оказалось неудовлетворительным для варианта раз-

деления семи солей в твердой фазе: 10, 17, 64, 65, 82, 83, 84, 101, 102, 120, 121, 127, 129, 137, 137, 144, 171, 173, 175, 185, 187, 188, 201, 208, 209, 223, 225.

Суммарный результат классификации составил 92%, среди которых оказались неудовлетворительными следующие записи: 3, 84, 121, 126, 127, 129, 177, 185, 189, 193, 194, 198, 208, 209, 218. Очевидно, что уже имеется пересечение с предыдущей таблицей, состоящее из записей: 84, 121, 127, 129, 185. Следовательно, эти записи уже два раза не смогли войти в результаты удачной классификации.

В этом случае (табл. 3) результат классификации равен 98%, а в качестве неудовлетворительных были выделены следующие записи – 64, 65, 83, 84, 102, 120, 201. Как и в предыдущем случае существуют результаты, которые дважды или трижды повторяются как неудачные, например, 64, 65, 83, 84, 126, 129, 185, 193, 194, 198, 102, 120, 201. Следовательно, наборы этих записей следует пометить в базе данных как вызывающие вопросы.

Кроме того предлагаемый нами способ оценки данных в данном конкретном сообщении позволит ответить также и на ряд практических вопросов, связанных с определением качественного состава твердой фазы, находящимся в равновесии с насыщенным раствором при разных температурах. Например, для последней, наиболее удачной классификации в качестве примера приведено решающее правило, которое имеет вид:

$$F_1 = -135{,}761 - 0{,}565196 \cdot Col\_1 + 8{,}70281 \cdot Col\_2 + 7{,}66225 \cdot Col\_3 + 8{,}69064 \cdot Col\_4,$$

$$F_2 = -105{,}677 - 0{,}490061 \cdot Col\_1 + 7{,}9137 \cdot Col\_2 + 7{,}01466 \cdot Col\_3 + 7{,}53163 \cdot Col\_4,$$

где *Col_1* – температура раствора, °С; *Col_2* – NaCl, %мас; *Col_3* – KCl, %мас; *Col_4* – $MgCl_2$, %мас.

Подставив в уравнение соответствующие значения концентраций равновесного состава, с вероятностью 98% можно прогнозировать наличие кристаллогидрата в составе твердой фазы. Прочие решающие правила позволяют с меньшей вероятностью (91–92%) дать прогноз о качественном составе твердой фазы.

Заключение

В представленной работе рассмотрена возможность поиска сомнительных данных в справочниках при объединении их БД. В качестве метода исследования предлагается использовать один из алгоритмов распознавания образов – разработка решающего правила. На наш взгляд такой подход оказался полезен для выявления данных типа unusual data в многокомпонентной гетерогенной системе NaCl–KCl–$MCl_2 \cdot H_2O$, собранной из 12 различных источников.

Получено решающее правило, позволяющее с вероятностью 98% прогнозировать наличие или отсутствие кристаллогидрата в твердой фазе в зависимости от состава насыщенного раствора NaCl, KCl, $MCl_2 \cdot H_2O$.

Литература

1. **Россиев, А. А.** Моделирование данных при помощи кривых для восстановления пробелов в таблицах / А. А. Рос–сиев // Методы нейроинформатики: сборник научных трудов / Красноярский гос. тех. унив.; под ред. А. Н. Горбаня. – Красноярск, 1998. – С. 6–22.

2. **Зильберглейт, М. А.** Восстановление пропущенных данных при изучении свойств совмещенных эластомеров и пластиков / М. А. Зильберглейт, Р. М. Долинская // Материалы. Технологии. Инструменты. – 2011. –Т. 16. – № 1. – С. 111–114.

3. **Загоруйко, Н. Г.** Алгоритмы обнаружения эмпирических закономерностей/ Н. Г. Загоруйко, В. Н. Ёлкина. – Новосибирск.: Наука, 1985. – 110 с.

4. **Горелик, А. Л.** Методы распознавания / А. Л. *Горелик, В*. А. С*крипкин.* – М.: Высшая школа, 1984. – 262 с.

5. **Вапник, В. Н.** Теория распознавания образов / В. Н. *Вапник*, *А*. Я. *Червоненкис*. – М.: Наука, 1974. – 416 с.

6. **Фомин, Я. А.** Распознавание образов: теория и применения / Я. А. Фомин. – М.: ФАЗИС, 2012. – 429 с.

7. **Фомин, Я. А.** Статистическая теория распознавания образов / Я. А. *Фомин, Г. Р. Тарловски*. – М.: Радио и связь, 1986. – 624 с.

8. **Здановский, А. Б.** Справочник экспериментальных данных по растворимости многокомпонентных водно-солевых систем /А. Б. Здановский, Е. Е. Ляховская, Р. Э. Шлеймович. – М.: Государственное научно-техническое издательство химической литературы, 1954. – 1270 с.

References

1. **Rossiev, A. A.** Simulation data using recovery curves for gaps in the tables / A. A. Rossiev // The neuroinformatics methods: collection of scientific papers / Krasnoyarsk State. Technical. Univer.; – Krasnojarsk, 1998. – P. 6–22.

2. **Zil'berglejt, M. A.** Recovery of missing data in the study of the properties of elastomers and plastics, combined / M. A. Zil'berglejt, R. M. Dolinskaja // Materials. Technologies. Instruments. – 2011. – V.. 16. – № 1. – P. 111–114.

3. **Zagorujko, N. G.** Detection algorithms empirical regularities / N. G. Zagorujko, V. N. Jolkina. – Novosibirsk.: Nauka, 1985. – 110 p.

4. **Gorelik, A. L.** Detection Methods / A. L. Gorelik, V. A. Skripkin. – M.: Vysshaja shkola, 1984. – 262 p.

5. **Vapnik, V. N.** The theory of pattern recognition / V. N. Vapnik, A. Ja. Chervonenkis. – M.: Nauka, 1974. – 416 p.

6. **Fomin, Ja. A.** Pattern Recognition: Theory and application / Ja. A. Fomin. – M.: FAZIS, 2012. – 429 p.

7. **Fomin, Ja. A.** Statistical theory of pattern recognition / Ja. A. Fomin, G. R. Tarlovski. – M.: Radio i svjaz', 1986. – 624 p.

8. **Zdanovsky, A. B.** /Reference book of experimental data on solubility of multicomponent water-salt systems – M.: State scientific and technical publishing house of chemical literature, 1954. – 1270 p.

Поступила	*После доработки*	*Принята к печати*
25.03.2016	*15.04.2016*	*10.05.2016*

M. A. Zilbergleit

SELECTION OF BASIC INFORMATION FOR FORMATION OF THE DB BY WAY OF EXAMPLE OF MULTICOMPONENT SYSTEM NACL–KCL–MGCL2·H2O

In article is offered the way of a search of the facts like unusual data (these raising doubts) by forming databases from several sources. As criterion for search of such data it is offered to use a method of pattern recognition – decision rule. By using two and more decision rules there are crossings at which objects which can't be classified correctly by means of such decisive rules are formed. As an example of a search of the data like unusual data application of this method is shown in the analysis of the data of heterogeneous balance in NaCl–KCl–MgCl H_2O system published in 12 various sources for the different periods of time. Accuracy of classification has made 92–98%.

Keywords: *selection of the source data, the data is questionable, unusual data, pattern recognition, linear and non-linear classifiers, decision rule, set intersection of multicomponent heterogeneous chemical system.*

Зильберглейт М. А. – доктор химических наук, заведующий лабораторией Института общей и неорганической химии НАНБ. Закончил Белорусский государственный технологический университет по специальности химическая технология переработки нефти и газа. Работал в ИФОХ НАНБ, Главгазе БССР, БГТУ. Подготовил 9 кандидатов наук в области химии, полиграфии и информационных технологий. Более 100 публикаций, патентов, учебных пособий. Специализация – оптимизация и управление химико-технологическими процессами. E-mail: mazi@list. ru.