И. С. Ожерельев, О. В. Сенько, Н. Н. Киселёва, Метод поиска выпадающих объектов с использованием параметров неустойчивости обучения, *Системы и средства информ.*, 2019, том 29, выпуск 2, 122–134

DOI: https://doi.org/10.14357/08696527190211

МЕТОД ПОИСКА ВЫПАДАЮЩИХ ОБЪЕКТОВ С ИСПОЛЬЗОВАНИЕМ ПАРАМЕТРОВ НЕУСТОЙЧИВОСТИ ОБУЧЕНИЯ*

***И. С. Ожерельев*[1], *О. В. Сенько*[2], *Н. Н. Киселёва*[3]**

Аннотация: Рассматривается метод поиска выпадающих объектов (ВО) в задачах распознавания, т. е. объектов, описания которых значительно отличаются от описаний объектов своего класса. Метод основан на одновременном использовании величин оценок принадлежности объекта к классам и интегральных искажений, вносимых объектом в формируемый в результате обучения алгоритм распознавания. Возможность использования разработанного метода при высокой размерности данных была продемонстрирована на задаче прогнозирования возможности образования неорганических соединений состава $A^{+3}B^{+3}C^{+2}\mathrm{O}_4$ при обычных условиях. Метод может быть использован с целью выявления ошибочных наблюдений для повышения качества обучающей информации при решении задач распознавания.

Ключевые слова: выпадающие объекты; базы данных; распознавание; неустойчивость обучения; неорганические соединения

DOI: 10.14357/08696527190211

1 Введение

Под выпадающими объектами обычно понимаются объекты, описание которых заметно отклоняется от основной закономерности в данных. При этом отклонение должно быть настолько значительным, что оно не могло бы быть объяснено простой случайностью и требовало бы дополнительных предположений о механизме возникновения объекта. Данное определение фактически соответствует определению, приведенному в книге [1]. Выпадающие объекты достаточно часто встречаются в базах данных. Иногда выпадающие наблюдения связаны с какими-либо неизвестными особенностями исследуемого процесса или уникальностью объектов. Идентификация таких ВО может привести к получению новых знаний об исследуемом явлении. Однако чаще всего ВО возникают в связи с различного рода ошибками, включая экспериментальные ошибки (в том числе

*Работа выполнена при частичной поддержке РФФИ (проект 17-01-00634).

[1]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, ilya365365@gmail.com

[2]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, senkoov@mail.ru

[3]Институт металлургии и материаловедения им. А. А. Байкова Российской академии наук, kis@imet.ac.ru

и ошибки в определении класса объектов или ошибки в значениях признаков), ошибки при занесении информации в базу и др. При построении прогностических моделей с помощью методов машинного обучения ВО, связанные с ошибками, могут заметно повышать неустойчивость обучения и, как следствие, снижать обобщающую способность полученной модели. Естественно, что такие ВО следует удалять из обучающей выборки после их идентификации. Следует отметить, что экспертная оценка правильности часто противоречивых экспериментальных данных разных исследователей остается наиболее сложной и плохо формализуемой задачей. Оценка правильности классификации десятков тысяч веществ в базах данных по свойствам веществ и материалов — задача крайне дорогостоящая и практически нереальная. Даже экспертная оценка данных о нескольких сотнях веществ требует многих месяцев работы, что делает необходимой автоматизацию поиска ВО, позволяющего выявить потенциально ошибочные наблюдения и после экспертной оценки сделать соответствующие исправления. Поэтому ВО продолжают привлекать внимание исследователей, в том числе при решении задач выявления связи свойств соединений со свойствами их компонентов [2, 3]. В настоящее время существует достаточно большое число подходов к идентификации ВО. В их число входят одномерные статистические тесты, оценивающие возможность интерпретации экстремальных значений переменных как ВО [4]. В многомерных данных значения каждой из переменных для ВО могут оказаться неэкстремальными, что не позволяет выявлять такие ВО с помощью одномерных тестов. Для обнаружения подобных ВО может быть использован подход, основанный на вычислении скалярной меры отклонения отдельного наблюдения от всего массива данных. В качестве меры близости может выступать, например, расстояние Махаланобиса или его робастные модификации [5].

В рамках задачи восстановления зависимости переменной y от переменных $\{X_1, \ldots, X_p\}$ под ВО может пониматься объект $s = (\mathbf{x}, y)$, для которого отмечается существенное отклонение оценки от ожидаемого значения y в точке $\mathbf{x}$ пространства $\mathbf{R}^p$. Чтобы выявить ВО в указанном выше смысле, необходимо построить по обучающей выборке модель, связывающую ожидаемое значение y с вектором $\mathbf{x}$ из $\mathbf{R}^p$. Например, это может быть модель

$$y = \hat{y}\left(\boldsymbol{\beta}, \mathbf{x}\right) + \varepsilon\,. \tag{1}$$

Поиск вектора коэффициентов $\boldsymbol{\beta}$ может проводиться с помощью метода наименьших квадратов по обучающей выборке $\tilde{\mathbf{S}} = \{(\mathbf{x}_j, y_j)\,|j = 1, \ldots, m\}$. Однако более робастные оценки могут быть получены с помощью методов LMS (least median of squares) или LTS (least trimmed squares) [6]. Для оценки отклонения произвольного объекта $s_j = (\mathbf{x}_j, y_j)$ из обучающей выборки от зависимости (1) может быть использован, например, индекс удаленных остатков:

$$\kappa_j^r = \frac{y_j - \hat{y}\left[\boldsymbol{\beta}\left(\tilde{\mathbf{S}} \backslash s_j\right), \mathbf{x}_j\right]}{\hat{\sigma}_Y}\,,$$

где $\boldsymbol{\beta}\left(\tilde{\mathbf{S}}\backslash s_j\right)$ — вектор коэффициентов модели (1), рассчитанный по выборке $\tilde{\mathbf{S}}$ после исключения s_j; $\hat{\sigma}_Y$ — выборочное стандартное отклонение.

В качестве альтернативной меры несоответствия объекта s_j преобладающей закономерности может быть использовано расстояние Кука [7, 8], показывающее, насколько s_j искажает регрессионную модель:

$$D_i = \frac{\sum_{j=1}^{m}\left\{\hat{y}\left[\mathbf{x}_j, \beta\left(\tilde{\mathbf{S}}\right)\right] - \hat{y}\left[\mathbf{x}_j, \beta\left(\tilde{\mathbf{S}}\backslash s_i\right)\right]\right\}^2}{p\hat{\sigma}_Y}.$$

Перечисленные индексы, очевидно, могут служить количественными оценками того, что объект s_j относится к выпадающим наблюдениям. Они позволяют ранжировать объекты по степени их отклонения от закономерности, описываемой моделью (1).

2 Метод, основанный на комбинировании параметров неустойчивости и величин отступов

При решении задачи распознавания естественно считать выпадающим объект из класса K_i с описанием $\mathbf{x}$, если для него велика разность $\max_{j=1,\ldots,L}[\mathbf{P}(K_j|\mathbf{x}) - \mathbf{P}(K_i|\mathbf{x})]$ или, напротив, отрицательна и велика по модулю величина $\min_{j=1,\ldots,L}[\mathbf{P}(K_j|\mathbf{x}) - \mathbf{P}(K_i|\mathbf{x})]$. На практике количественной оценкой того, что объект с описанием $\mathbf{x}$ является выпадающим наблюдением, очевидно, служит $\max_{j=1,\ldots,L}[\hat{\mathbf{P}}(K_j|\mathbf{x}) - \hat{\mathbf{P}}(K_i|\mathbf{x})]$, где $\hat{\mathbf{P}}(K_j|\mathbf{x})$ — оценка вероятности принадлежности объекта с описанием $\mathbf{x}$ к классу K_j. Однако оценки вероятностей принадлежности к классам напрямую используют только статистические методы распознавания. В общем случае в этих целях могут быть использованы величины отступа $\Gamma(K_i, \mathbf{x}) = \max_{j=1,\ldots,L}[\gamma_i(\mathbf{x}, \tilde{\mathbf{S}}) - \gamma_j(\mathbf{x}, \tilde{\mathbf{S}})]$, где $\gamma_i(\mathbf{x}, \tilde{\mathbf{S}})$ — оценка принадлежности объекта с описанием $\mathbf{x}$ к классу K_i, которая рассчитана алгоритмом, обученным по $\tilde{\mathbf{S}}$ [9].

В зависимости от значений отступа обучающие объекты условно делятся на 5 типов в порядке убывания отступа: эталонные, неинформативные, пограничные, ошибочные, шумовые.

Эталонные объекты имеют большой положительный отступ, плотно окружены объектами своего класса и являются наиболее типичными его представителями.

Неинформативные объекты также имеют положительный отступ. Изъятие этих объектов из выборки (при условии, что эталонные объекты остаются) не влияет на качество классификации. Они не добавляют к эталонам никакой новой информации. Наличие неинформативных объектов характерно для выборок избыточного объема.

Пограничные объекты имеют отступ, близкий к нулю. Классификация таких объектов неустойчива в том смысле, что малые изменения метрики или состава

обучающей выборки могут изменять их классификацию. Например, в химических задачах такими объектами могут быть метастабильные при определенных внешних условиях (например, при комнатной температуре и атмосферном давлении) соединения или кристаллические модификации.

Гранично-ошибочные объекты имеют небольшие отрицательные отступы и близки к пограничным. Гранично-ошибочные объекты потенциально могут быть распознаны при совершенствовании алгоритма.

Шумовые объекты — это относительно небольшое число объектов, которые плотно окружены объектами чужих классов и удалены от основной массы объектов своего класса. Многие ВО представляют собой именно шумовые объекты. Для шумовых объектов характерны большие отрицательные величины отступа, по которым они легко могут быть идентифицированы.

В условиях высокой размерности ВО могут оказывать существенное влияние на процесс обучения, существенно искажая обученный распознающий алгоритм. При этом ВО превращаются в гранично-ошибочные или пограничные объекты. Информации о величинах отступа нередко оказывается недостаточно для достоверной идентификации таких ВО.

В данной работе предлагается подход к поиску ВО при решении задач распознавания принадлежности объектов к некоторым классам $K_1, \ldots, K_L$ по признакам $X_1, \ldots, X_p$, основанный на ранжировании объектов согласно комбинированной оценке, учитывающей как величину отступа, так и величину вносимых искажений. Количественной оценкой того, что объект с описанием $\mathbf{x}$ является выпадающим, естественно считать аналог упомянутого выше расстояния Кука, используемого для описания неустойчивости линейной регрессионной модели:

$$\delta_i = \frac{\sum_{i=1}^{L} \sum_{j=1}^{m} \left[\gamma_i \left(\mathbf{x}_j, \tilde{\mathbf{S}} \right) - \gamma_i \left(\mathbf{x}_j, \tilde{\mathbf{S}} \backslash s_j \right) \right]}{Lm} .$$

Коэффициенты $\Gamma(K_i, \mathbf{x})$ и δ_i по отдельности или в комбинациях могут быть использованы для ранжирования объектов обучающей выборки по степени отклонения от существующих в данных закономерностей.

Однако одного только ранжирования объектов по мере их отклонения от аппроксимируемой зависимости недостаточно для выявления ВО. Необходимо также найти тот порог отсечения, при превышении которого объект можно было бы считать выпадающим. Естественным критерием для выбора такого порога может служить эффективность распознавания, оцениваемая по одной из стандартных метрик. Наиболее полно эффективность распознавания характеризуется с помощью AUC (area under curve) — площади под ROC (receiver operating characteristics) кривой. Будем считать, что оценка принадлежности объекта к класс — величина, изменяющаяся в диапазоне $[0, 1]$. Если это не так, оценки можно спроецировать на отрезок $[0, 1]$.

Отбор ВО происходит следующим образом.

1. Получим оценки принадлежности к классам на полной выборке:
   (а) обучим классификатор C_o на обучающей выборке $\tilde{\mathbf{S}}$;
   (б) применим C_o к $\tilde{\mathbf{S}}$ и получим оценки вероятностей принадлежности объектов к классам $[\gamma_1(\mathbf{x}_1, \tilde{\mathbf{S}}), \ldots, \gamma_L(\mathbf{x}_1, \tilde{\mathbf{S}})], \ldots, [\gamma_1(\mathbf{x}_m, \tilde{\mathbf{S}}), \ldots, \gamma_L(\mathbf{x}_m, \tilde{\mathbf{S}})]$.
2. Для каждого объекта выборки оценим, относится ли он к ВО:
   (а) построим выборку $\tilde{\mathbf{S}}_i$, исключив из $\tilde{\mathbf{S}}$ пару (x_i, y_i);
   (б) обучим классификатор C_i на выборке $\tilde{\mathbf{S}}_i$;
   (в) применим C_i к объектам из $\tilde{\mathbf{S}}$ и получим оценки $[\gamma_1(\mathbf{x}_1, \tilde{\mathbf{S}}_i), \ldots \ldots, \gamma_L(\mathbf{x}_1, \tilde{\mathbf{S}}_i)], \ldots, [\gamma_1(\mathbf{x}_m, \tilde{\mathbf{S}}_i), \ldots, \gamma_L(\mathbf{x}_m, \tilde{\mathbf{S}}_i)]$;
   (г) вычислим
   $$\delta\left(s_i\right) = \frac{\sum_{l=1}^{L} \sum_{j=1}^{m} [\gamma_l(\mathbf{x}_j, \tilde{\mathbf{S}}) - \gamma_l(\mathbf{x}_j, \tilde{\mathbf{S}}_i)]}{Lm}.$$
   Величина $\delta(s_i)$ показывает, насколько изменились оценки объектов после исключения s_i из обучающей выборки;
   (д) вычислим
   $$\omega(s_i) = \frac{\sum_{l \in \{1,\ldots,L\} \setminus \lambda(s_i)} [\gamma_l(\mathbf{x}_i, \tilde{\mathbf{S}}_i)]}{L},$$
   где $\lambda(s_i)$ — номер класса, к которому принадлежит объект s_i. Величина $\omega(s_i)$ — средняя оценка принадлежности s_i к классам, к которым он не относится, рассчитанная алгоритмом, обученным по выборке без этого объекта.
3. Отберем ВО, исходя из оценок $\delta(s_1), \ldots, \delta(s_m)$ и $\omega(s_1), \ldots, \omega(s_m)$. Для каждого объекта вычислим
   $$E(s_i, a_1, a_2, p) = [a_1 |\delta(s_i)|^p + a_2 |\omega(s_i)|^p]^{1/p}.$$
   Заметим, что $E(s_i, a_1, a_2, p)$ — модуль объекта в пространстве (δ, ω) со взвешенной метрикой Минковского.
4. Подвыборку из k объектов с наибольшими E исключаем из обучения.

3 Эксперименты

Вычислительный эксперимент реализован на языке программирования python версии 3.5.1 с использованием библиотеки scikit-learn версии 0.17.

3.1 Данные

В работе решалась задача выявления ВО при прогнозе возможности образования соединений состава $A^{+3}B^{+3}C^{+2}O_4$. Выборка состояла из 758 объектов двух классов. К первому классу принадлежали 695 объектов (существующие соединения), ко второму — 63 (химические системы A–B–C–O без образования соединений вышеуказанного состава). Каждый объект описывали 108 непрерывных признаков. Пропусков в данных не было. В выборке содержались объекты с неверной меткой класса.

Требовалось обнаружить в базе данных объекты, которые потенциально могли бы быть ошибочно классифицированы на этапе их внесения в базу. Принципиальная возможность решения поставленной задачи в первую очередь связана с периодичностью изменения свойств неорганических соединений в зависимости от атомных номеров элементов — компонентов химических систем.

3.2 Методы распознавания, использованные в исследовании

Для решения задачи классификации использовался градиентный бустинг (GB, gradient boosting) над решающими деревьями [10]. Градиентный бустинг был выбран после сравнения с другими популярными алгоритмами классификации, такими как решающие деревья (DT, decision trees), метод опорных векторов (SVM, support vector machine), метод ближайших соседей (KNN, k-nearest neighbors) [9, 10]. Качество оценивалось на полной выборке при помощи десятифолдовой кроссвалидации с сохранением долевого содержания классов в выборках.

Таблица 1 Результаты распознавания

Метод	AUC ROC [11]
KNN	0,82
SVM	0,85
DT	0,77
GB	0,84

Исходные (до удаления ВО) результаты распознавания для разных методов представлены в табл. 1. Градиентный бустинг показал относительно неплохие результаты по сравнению с другими алгоритмами.

3.3 Связь параметров неустойчивости и величин отступа

Дополнительным аргументом комбинирования величин неустойчивости и отступа стали результаты исследования их взаимосвязи, представленные на рис. 1.

Из рис. 1 видно, что величины δ и ω не коррелируют между собой. Низкая корреляция оценок δ и ω свидетельствует о том, что они могут служить независимыми индикаторами ВО.

3.4 Влияние числа исключенных объектов на качество классификации

В работе было изучено, как на качество классификации влияют параметры взвешенной метрики Минковского a_1, a_2, p и число исключенных объектов k.

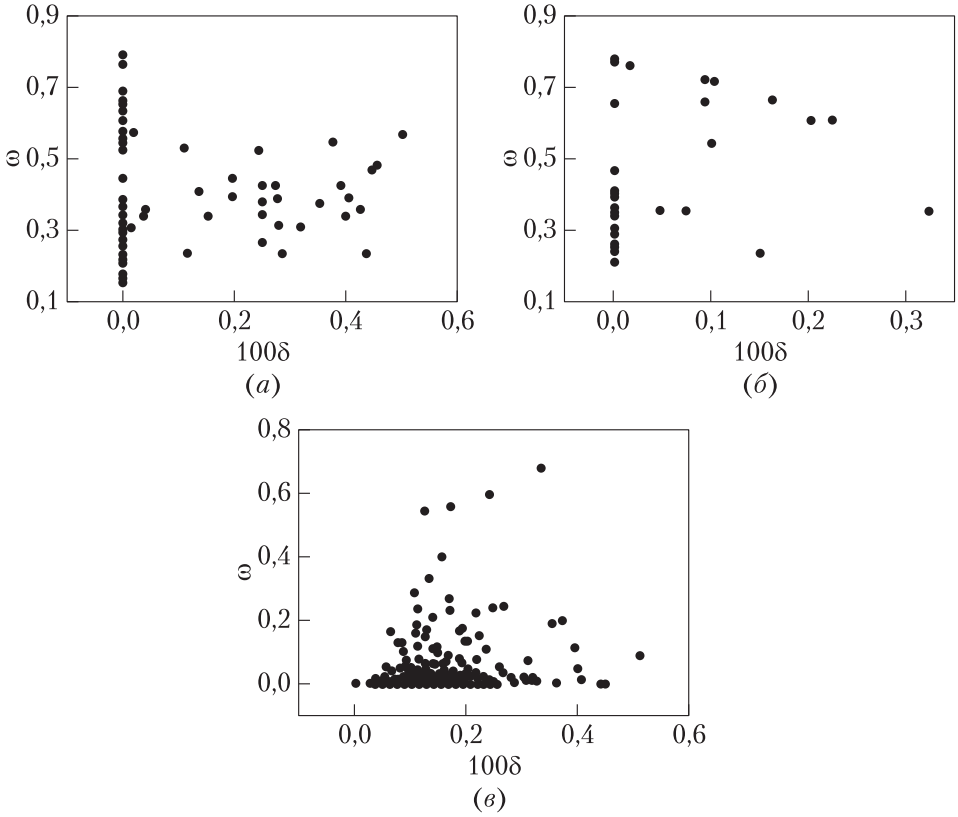

Рис. 1 Зависимость между величинами отступа ω и неустойчивости δ

Для оценки качества оптимизации процедуры отбора ВО и оценки влияния отбора на точность классификации использовались внешняя и внутренняя процедуры скользящего контроля. На каждом из шагов внешней процедуры формировались обучающая и контрольная выборки. Идентификация ВО внутри обучающих выборок производилась по показателям (δ, ω), рассчитанным с использованием процедур внутреннего скользящего контроля. Внутренний скользящий контроль использовался для подбора параметров градиентного бустинга — оптимальной скорости и числа деревьев. Параметры отбора, включая число ВО k, степенной показатель метрики Минковского p и соотношение весов a_1/a_2, также подбирались в ходе внутреннего скользящего контроля, исходя из требования максимизации точности распознавания, оцениваемой с помощью ROC AUC. Отметим, что поиск параметров отбора ВО осуществлялся при заранее найденных фиксированных оптимальных параметрах градиентного бустинга. После удаления из обучающей выборки выявленных ВО алгоритм распознавания обучался за-

Таблица 2 Сценарий экспериментов

№ эксперимента	Число блоков внутреннего скользящего контроля (N_{in})	Число блоков внешнего скользящего контроля (N_{out})
1	10	10
2	10	20
3	30	10

ново. Результаты распознавания оценивались на соответствующих контрольных выборках. Эксперименты проводились с различным числом блоков скользящего контроля (табл. 2).

При выборе p достаточно перебрать числа от одного до пяти, соотношение a_1/a_2 выбрать сложнее. Эмпирически установлено, что качество распознавания максимально, когда отношение $\text{med}_{\text{x}\in\text{X}}\delta(\text{x})a_1/\text{med}_{\text{x}\in\text{X}}\omega(x)a_1 \in [1,2]$, где med — 0,5-квантиль. Таким образом, число выполняемых циклов внутреннего скользящего контроля составило $N_{\text{in}}N_{\text{out}}N_pN_{a_1/a_2}k_{\max}$, где N_p — число перебираемых степенных показателей; N_{a_1/a_2} — число перебираемых соотношений a_1/a_2; $k_{\max}$ — предполагаемое максимальное число ВО. Заметим, что подбор параметров градиентного бустинга занял гораздо больше времени, чем вычисление значения метрики Минковского в двухмерном пространстве для каждого объекта выборки.

Из рис. 2 видно, что изменение a_1 значительно влияет на качество. При этом виден отчетливый экстремум при $a_1 = 300$.

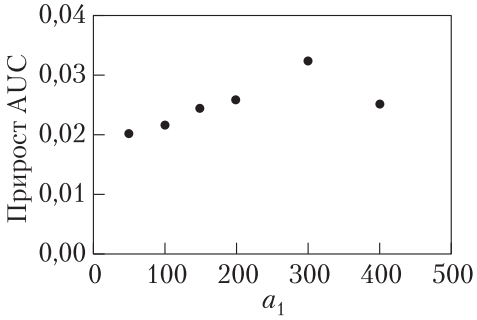

Рис. 2 График зависимости прироста AUC от a_1 ($a_2 = 1$; $p = 1$)

Метод поиска ВО с использованием параметров неустойчивости обучения сравнивался с другими популярными методами поиска ВО. Первый метод основан на исключении объектов с максимальным по модулю отрицательным отступом. Предполагается, что такие наблюдения лежат в гуще объектов противоположного класса. Второй метод основан на исключении объектов с малым по модулю отступом. Предполагается, что объекты с малым отступом лежат на границе двух классов, не являются эталонами класса и снижают обобщающую способность алгоритма. Результаты использования этих алгоритмов продемонстрированы на рис. 3.

Метод поиска ВО с использованием параметров неустойчивости обучения показывал более высокие результаты при различных наборах гиперпараметров (рис. 4).

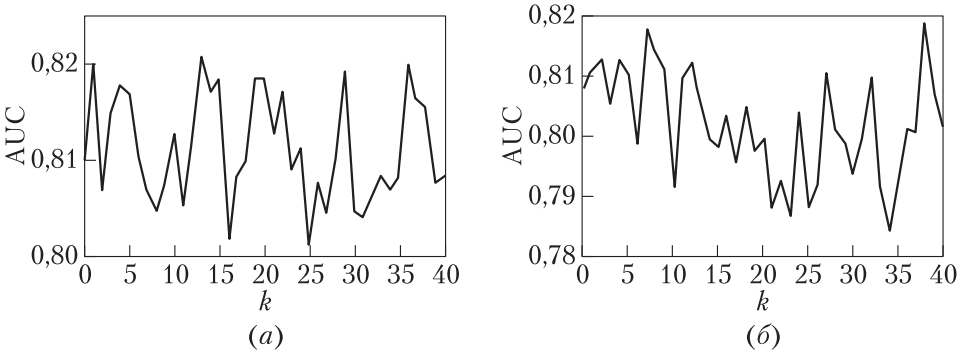

Рис. 3 AUC ROC после удаления k объектов с наибольшим (*а*) и с наименьшим (*б*) по модулю отступом. Вычисление AUC проводилось в режиме скользящего контроля с учетом исключенных объектов (эксперимент № 3). Видно, что зависимость неустойчива и нет выраженной тенденции изменения AUC: (*а*) прирост AUC составил 0,012; (*б*) прирост AUC составил 0,011

Описанный в работе алгоритм сильнее повышает AUC ROC, однако имеет большую вычислительную сложность. Из рис. 4 видно, что оптимальные результаты достигнуты при $p = 2$ и $a_1/a_2 = 0{,}0025$ и при $p = 3$ и $a_1/a_2 = 0{,}0025$. При $p = 4$ и $a_1/a_2 = 0{,}0025$ прирост AUC ROC ниже, чем в других случаях. При $p = 1$ и $a_1/a_2 = 0{,}005$ нет выраженной тенденции повышения точности при исключении объектов по порядку ранжирования. Кроме k необходимо подобрать p и верное соотношение a_1/a_2. Основную сложность представляет подбор соотношения a_1/a_2 и числа исключенных объектов k. Поскольку нельзя сделать никаких предположений о зависимости параметров и качества обнаружения ВО, оптимальные значения параметров метода приходится подбирать при помощи процедуры скользящего контроля.

Из рис. 4 видно, что максимальное качество классификации (AUC = = 0,836) достигается при исключении из полной обучающей выборки подвыборки из 21 объекта, которая была проанализирована экспертом. Оказалось, что восемь объектов имели ошибочную метку первого класса, один объект имел ошибочную метку второго класса, о принадлежности одного объекта не было найдено достоверных данных. Таким образом, продемонстрирована способность разработанного метода выявлять именно ошибочные наблюдения.

4 Заключение

В ходе работы был разработан алгоритм отбора ВО, основанный на исключении из выборки объектов, наиболее сильно искажающих разделяющую поверхность. Данный метод позволяет добиться большего улучшения качества, чем его аналоги, однако он требует довольно тщательного подбора гиперпара-

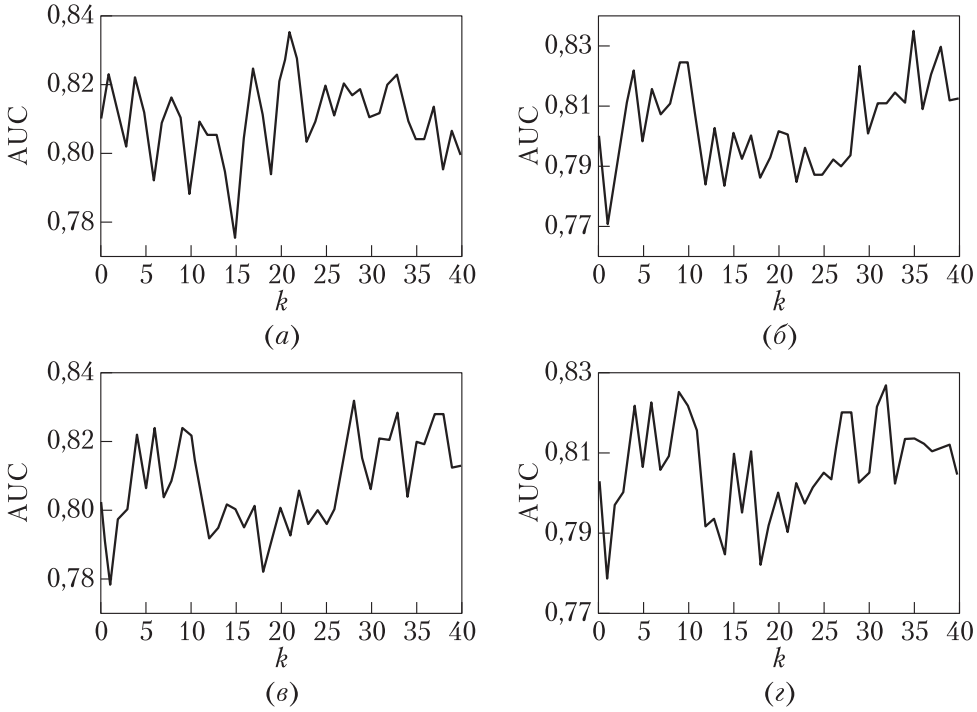

Рис. 4 Зависимость величины AUC от числа исключенных по порядку ранжирования объектов обучающей выборки по комбинированным оценкам, включающим отступ и неустойчивость ($a_2 = 1$): (*а*) $a_1 = 200$, $p = 1$ — прирост AUC на 0,026; (*б*) $a_1 = 400$, $p = 2$ — прирост AUC на 0,028; (*в*) $a_1 = 400$, $p = 3$ — прирост AUC на 0,024; (*г*) $a_1 = 400$, $p = 4$ — прирост AUC на 0,019. Видно, что, несмотря на сохранение неустойчивости, появляется выраженная тенденция повышения точности при исключении объектов по порядку ранжирования

метров, что создает сложности при использовании этого алгоритма для больших данных. С другой стороны, предложенный алгоритм может быть легко выполним параллельно. Применение разработанного алгоритма при фильтрации ошибок в базах данных по свойствам неорганических соединений позволило значительно сократить время и трудозатраты на выявление ошибок в определении статуса химических объектов и повысить точность прогнозирования при конструировании новых неорганических соединений. Следует отметить, что анализ выявленных ВО дает стимул к дополнительному изучению соответствующих соединений.

Литература

1. *Aggarwal C. C.* Outlier analysis. — New York, NY, USA: Springer-Verlag, 2013. 446 p.

2. *Киселёва Н. Н.* Компьютерное конструирование неорганических соединений. — М.: Наука, 2005. 289 с.
3. *Киселева Н. Н., Столяренко А. В., Рязанов В. В., Сенько О. В., Докукин А. А.* Прогнозирование новых соединений состава $A^{3+}B^{3+}C^{2+}O_4$ // Ж. неорганической химии, 2017. Т. 62. № 8. С. 1068–1077.
4. *Grubbs F. E.* Procedures for detecting outlying observations in samples // Technometrics, 1969. Vol. 11. No. 1. P. 1–21.
5. *Rousseeuw P. J., Van Driessen K.* Computing LTS regression for large data sets // Data Min. Knowl. Discovery, 2006. Vol. 12. P. 29–45.
6. *Rousseeuw P. J.* Least median of squares regression // J. Acoust. Soc. Am., 1984. Vol. 79. P. 871–880.
7. *Cook R. D.* Influential observations in linear regression // J. Acoust. Soc. Am., 1979. Vol. 74. P. 169–174.
8. *Cao D. S., Liang Y. Z., Xu Q. S, Li H. D., Chen X.* A new strategy of outlier detection for QSAR/QSPR // J. Comput. Chem., 2010. Vol. 31. P. 592–602.
9. *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. 176 с.
10. *Hastie T., Tibshirani R., Friedman J. H.* The elements of statistical learning: Data mining, interference, and prediction. — 2nd ed. — New York, NY, USA: Springer, 2009. 767 p.
11. *Zweig M. H.* Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine // Clin. Chem., 1993. Vol. 39. P. 561–577.

Поступила в редакцию 29.03.18

METHOD FOR SEARCHING OUTLIER OBJECTS USING PARAMETERS OF LEARNING INSTABILITY

***I. S. Ozhereliev*[1], *O. V. Senko*[2], *and N. N. Kiseleva*[3]**

[1]Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

[2]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

[3]A. A. Baikov Institute of Metallurgy and Materials Science of the Russian Academy of Sciences, 49 Leninskiy Prosp., Moscow 119991, Russian Federation

Abstract: The paper describes a new method of outliers detection in pattern recognition tasks. The authors define an outlier as an object which deviates

significantly from the other objects of the same class. The method is based on simultaneous use of evaluated object estimates for classes and integral distortion of recognition algorithm that is caused by evaluated object. Usefulness of the developed technique was shown for the task of predicting if an inorganic compound of composition $A^{+3}B^{+3}C^{+2}O_4$ is formed under ordinary conditions. The method may be used for erroneous observations detection that is aimed to improve training information in different recognition tasks.

Keywords: outliers; data bases; recognition; instability of training; nonorganic compounds

DOI: 10.14357/08696527190211

Acknowledgments

The work was partly supported by the Russian Foundation for Basic Research (project 17-01-00634).

References

1. Aggarwal, C. C. 2013. *Outlier analysis.* New York, NY: Springer-Verlag. 446 p.
2. Kiseleva, N. N. 2005. *Komp'yuternoe konstruirovanie neorganicheskikh soedineniy* [Computer design of nonorganic compounds]. Moscow: Nauka. 289 p.
3. Kiseleva, N. N., A. V. Stolyarenko, V. V. Ryazanov, O. V. Sen'ko, and A. A. Dokukin. 2017. Prediction of new $A^{3+}B^{3+}C^{2+}O_4$ compounds. *Russ. J. Inorg. Chem.* 62:1058–1066.
4. Grubbs, F. E. 1969. Procedures for detecting outlying observations in samples. *Technometrics* 11(1):1–21.
5. Rousseeuw, P. J., and K. Van Driessen. 2006. Computing LTS regression for large data sets. *Data Min. Knowl. Discovery* 12:29–45.
6. Rousseeuw, P. J. 1984. Least median of squares regression. *J. Acoust. Soc. Am.* 79:871–880.
7. Cook, R. D. 1979. Influential observations in linear regression. *J. Acoust. Soc. Am.* 74:169–174.
8. Cao, D. S., Y. Z. Liang, Q. S. Xu, H. D. Li, and X. Chen. 2010. A new strategy of outlier detection for QSAR/QSPR. *J. Comput. Chem.* 31:592–602.
9. Zhuravlev, Yu. I., V. V. Ryazanov, and O. V. Sen'ko. 2006. *"Raspoznavanie." Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya* ["Recognition." Mathematical methods. Program system. Applications]. Moscow: Fazis. 159 p.
10. Hastie, T., R. Tibshirani, and J. H. Friedman. 2009. *The elements of statistical learning: Data mining, interference, and prediction*. 2nd ed. New York, NY: Springer. 767 p.
11. Zweig, M. H. 1993. Receiver-operating Characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin. Chem.* 39:561–577.

Received March 29, 2018

Contributors

Ozhereliev Ilya S. (b. 1994) — master student, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; ilya365365@gmail.com

Senko Oleg V. (b. 1957) — Doctor of Science in physics and mathematics, leading scientist, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; senkoov@mail.ru

Kiseleva Nadezhda N. (b. 1949) — Doctor of Science in chemistry, head of laboratory, A. A. Baikov Institute of Metallurgy and Materials Science of the Russian Academy of Sciences, 49 Leninskiy Prosp., Moscow 119991, Russian Federation; kis@imet.ac.ru