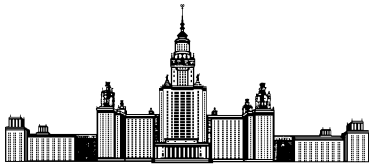
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

Колмакова Татьяна Сергеевна

«Обнаружение нестандартных неорганических веществ в моделях распознавания основанных на голосовании по системам логических закономерностей классов»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:

д.ф-м.н., профессор, академик РАЕН

Рязанов Владимир Васильевич

Москва, 2018

Содержание

Аннотация

В данной работе рассматривается возможность образования и тип кристаллической структуры оксидных соединений состава $A^{3+}B^{3+}C^{2+}O_4$, где A, B, C – элементы таблицы Меделеева при комнатной температуре и атмосферном давлении. Для этих целей используются средства машинного обучения.

Использовалась выборка данных, включающая некоторые химические характеристики элементов, входящих в состав соединения.

Предполагается, что в данных могут быть ошибки, которые ухудшают качество классификации, для их устранения используется алгоритм обнаружения аномалий, эффективность которого применительно к данной задаче показана в работе.

1 Введение

Обнаружение нестандартных объектов в данных имеет приложение во многих задачах машинного обучения. Будь то нахождение нетипичных шаблонов операций по кредитной карте, детектирование необычного трафика данных по сети интернет или обнаружение нестандартных игроков на биржевом рынке. Все эти задачи связаны с нахождением новых, еще неизвестных ранее объектов.

Нас же будет интересовать другое применение алгоритма нахождения выбросов – предварительная очистка данных. С его помощью можно повысить обучающую способность алгоритма. Этому свойству алгоритма есть простое объяснение: выбросив шумовые объекты, мы разрешаем алгоритму не подстраиваться под ненужные характеристики, тем самым упрощая модель.

В данной работе исследуется возможность улучшения качества классификации оксидных соединений состава $A^{3+}B^{3+}C^{2+}O_4$ по типу кристаллической решетки. Данные соединения исследуются для поиска новых веществ, обладающих магнитными, диэлектрическими, сверхпроводящими и другими функциональными свойствами. Некоторые соединения хорошо изучены, есть ряд работ [10, 12], в которых описаны методы их подготовки и свойства полученных соединений. Но, хотелось бы получить информацию о типе кристаллической решетки тех соединений, про которые пока ничего неизвестно. Чтобы справиться с этой задачей, нужно провести множество химических опытов, которые требуют специальных условий. Это очень дорого и, в эпоху компьютерного моделирования и анализа данных, не всегда необходимо.

Следует отметить, что, по мнению американских исследователей [7], методы машинного обучения позволяют значительно точнее предсказывать молекулярные свойства веществ, чем традиционные квантово-механические расчеты.

1.1 Используемые данные

В качестве выборки была использована база данных, содержащая информацию о 656 соединений состава $A^{3+}B^{3+}C^{2+}O_4$, полученных при комнатной температуре и атмосферном давлении. Выборка разделена на 6 несбалансированных классов по типу кристаллической решетки химического соединения.

Тип кристаллической решетки	Количество элементов	Номер класса
K_2NiF_4	266	1
$YbFe_2O_4$	84	2
Варвикит	96	3
Шпинель	91	4
$CaFe_2O_4$	55	5
Не образуют соединения	63	6

Ранее научными группами проводились исследования по определению типа решетки желаемого соединения на основе данных, включающих информацию об ионных радиусах двухвалентных и трехвалентных катионов или соотношениях этих радиусов элементов соединения. Так же предпринимались попытки найти корреляцию между типом кристаллической структуры и радиусом двухвалентного катиона. Для прогнозирования соединений, имеющих структуры типа K_2NiF_4 использовался коэффициент толерантности Гольдшмидта. Так же есть работы, использующие термодинамические характеристики соединения.

В качестве признакового описания объектов в данной работе были взяты свойства элементов, входящих в соединение перечисленные выше и также такие как: теплопроводность, энергия ионизации, квантовое число, номер группы, электроотрицательность, температура плавления, температура кипения и другие.

1.2 Постановка задачи

Цель работы заключалась в исследовании обучающей выборки на предмет аномальных объектов. В данной задаче выбросы могли появиться в следствии того, что исследователи ошиблись с описанием условий проведения эксперимента (при комнатной температуре и атмосферном давлении получается вещество другого состава). Будем решать данную задачу при помощи модели голосования по системам логических закономерностей классов.

Мерой качества нахождения аномальных объектов будем считать качество работы классификаторов на выборке, полученной из исходной выбрасыванием аномальных объектов. Классификаторы будем оценивать с помощью F-меры.

1.3 Понятие нестандартных объектов

Задача обнаружения аномалий имеет следующую специфику: в данной задаче отсутствует общее формальное определение понятий аномальность и аномалия. Обычно эти определения формализуются на этапе исследования задачи и зависят от выбранного функционала, по которому ведется оптимизация. Наиболее формально дал определение выброса Хокинс [6]:

Аномалия - это наблюдение, которое так сильно отклоняется от других наблюдений, что возникает подозрение, что оно генерируется другим механизмом.

На основе этого определения работает большая часть алгоритмов детектирования выбросов. Методы имеют внутреннее представление о том, что является типичным объектом выборки или же знают, из какого распределения она была получена.

Стоит знать, что природа аномалий тоже различна: это может быть как случайно сгенерированный шум, допущение ошибки при создании обучающей выборки или же умышленное вторжение в систему.

В заключении отметим, что так как алгоритмы обнаружения аномалий существенно отличаются, то и результатами их работы будут различные множества объектов.

1.4 Задача поиска логических закономерностей классов

Задачу нахождения нестандартных объектов в выборке будем решать с помощью метода поиска логических закономерностей классов. Данный метод используется для классификации данных, но обученную модель можно использовать и для построения функционала, выдающего рейтинг аномальности.

Рассмотрим «традиционную» постановку задачи поиска логических закономерностей:

Дана обучающая выборка $X = \{x_1, \ldots, x_n\}$, разбивающаяся на m непересекающихся классов $K_1, \ldots, K_m$. Каждый объект описывается l признаками. Предполагаем, что $K_i \neq \varnothing$ и выборка непротиворечива: $K_i \cap K_j = \varnothing$

Элементарным предикатом называется ограничение на признак объекта:

$$P_j^{c_j^1,c_j^2}(x) = \begin{cases} 1 & \text{если } c_j^1 \le x_j \le c_j^2 \\ 0 & \text{иначе} \end{cases}$$

где $c_j^1 \in \mathbb{R}$ и $c_j^2 \in \mathbb{R}$

Пусть $\Omega \subseteq \{1 \dots n\}$.

Определение 1.1. Логическая закономерность (ЛЗ) – предикат

$$P_\alpha^{\Omega,c^1,c^2} = \&_j P_j^{c_j^1,c_j^2}(x_j)$$

такой что:

1. $\exists x_i \in K_\alpha : P_\alpha^{\Omega,c^1,c^2}(x_i) = 1$
2. $\forall x_j \notin K_\alpha : P_\alpha^{\Omega,c^1,c^2}(x_j) = 0$
3. $F(P_\alpha^{\Omega,c^1,c^2}) = extr_{\Omega_*,c_*^1,c_*^2} F(P_\alpha^{\Omega_*,c_*^1,c_*^2})$,

   где F – оптимизируемый функционал

$$F(P_\alpha^{\Omega,c^1,c^2}) = |\{x_i \in K_\alpha : P_\alpha^{\Omega,c^1,c^2}(x_i) = 1\}|$$

Другими словами: логическая закономерность $P_\alpha^{\Omega,c^1,c^2}$– объединение некоторых предикатов, покрывающих класс α, такая что: существует элемент x_i из класса α, принадлежащий объединению данных предикатов. $P_\alpha^{\Omega,c^1,c^2}$ не покрывает объекты из «чужих» классов и данное объединение оптимизирует некий функционал. В данной постановке функционалом является количество объектов класса α, которые покрываются логической закономерностью $P_\alpha^{\Omega,c^1,c^2}$.

Таким образом, результатом обучения этого алгоритма для каждого класса являются его логические закономерности. Для этапа предсказания используется обученный функционал. Объект относится к тому классу, на пересечение предикатов которого он попал.

1.5 Критерий качества, используемый в задаче

Для измерения качества нахождения выбросов будем оценивать качество работы классификаторов на полученной выборке. Для этого существуют различные метрики.

Классическими метриками для задач классификации являются *accuracy*, *precision*, *recall* и F-мера. У каждой из рассматриваемых метрик есть свои преимущества и недостатки.

Рассмотрим каждую из перечисленных метрик в отдельности. Для этого введем вспомогательное понятие матрицы ошибок классификации. Для простоты предположим, что перед нами стоит задача классификации на два класса.

Пусть $\widetilde{y}$ – предсказанный ответ алгоритма, а y – истинная метка класса, тогда матрица ошибок выглядит следующим образом:

	$y = 0$	$y = 1$
$\widetilde{y} = 0$	True Negative (TN)	False Negative (FN)
$\widetilde{y} = 1$	False Positive (FP)	True Positive (TP)

Введем определение метрики **точность** (accuracy). Данная метрика показывает, то, сколько объектов из каждого класса алгоритм распознал верно.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Легко заметить, что метрика не подходит для случаев, когда классы несбалансированные.

Рассмотрим метрики *precision* и *recall*:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Precision показывает, какая доля элементов первого класса была определена верно из тех объектов, что классификатор определил в первый класс

Recall же показывает, какая доля объектов первого класса была распознана верно из всех объектов первого класса в обучающей выборке.

Как можно заметить, данные метрики не зависят от количества объектов в классах, а зависят только от распознавающей способности алгоритма, поэтому имеют более широкое применение, но как правило, на практике используется комбинация данных метрик, называемая F-мерой, которая в общем случае выглядит так:

$$F_\beta = (1 + \beta^2)\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

Лучшим способом оценивания качества классификации на многоклассовой выборке является матрица ошибок, но для оценивания алгоритма и подбора параметров нужно считать метрики, которые можно визуально отобразить для многих точек на графике, для этих целей и не только используются модификации данных метрик. Основная идея этого подхода: посчитать метрику для каждого класса, а потом сложить полученные значения и выдать среднее в качестве ответа.

1.6 Используемые алгоритмы классификации

Для проверки качества обнаружения выбросов, выборка полученная с помощью отбрасывания аномалий из исходной подавалась на вход классификатору. Будем считать детектирование выбросов успешным, если обучающая способность такого классификатора оказывалась больше по F-мере.

Для более качественного анализа использовались следующие классификаторы: логистическая регрессия, гребневая регрессия, метод ближайших соседей, случайный лес.

Будем рассматривать задачу классификации выборки $X = \{x_1, \ldots, x_n\}$ на два класса $y = \{-1, +1\}$. Пусть каждый объект выборки описывается m признаками, $x_i \in \mathbb{R}^m$

Опишем подробнее каждый из этих алгоритмов:

Логистическая регрессия

Решающее правило классификации задается следующим образом:

$$a(x, w) = sign(\sum_{i=1}^{m} \theta_i x_i + \theta_0) = sign\langle x, \theta \rangle$$

Задачей классификации является настройка весов θ_i. В данной задаче веса обучаются с помощью оптимизации функции эмпирического риска.

$$Q(\theta) = \sum_{1}^{n} ln(1 + exp(-y_i \cdot \langle x_i, \theta \rangle)) \longrightarrow \min_{\theta}$$

Функция эмпирического риска помогает найти гиперплосткость, разделяющую классы.

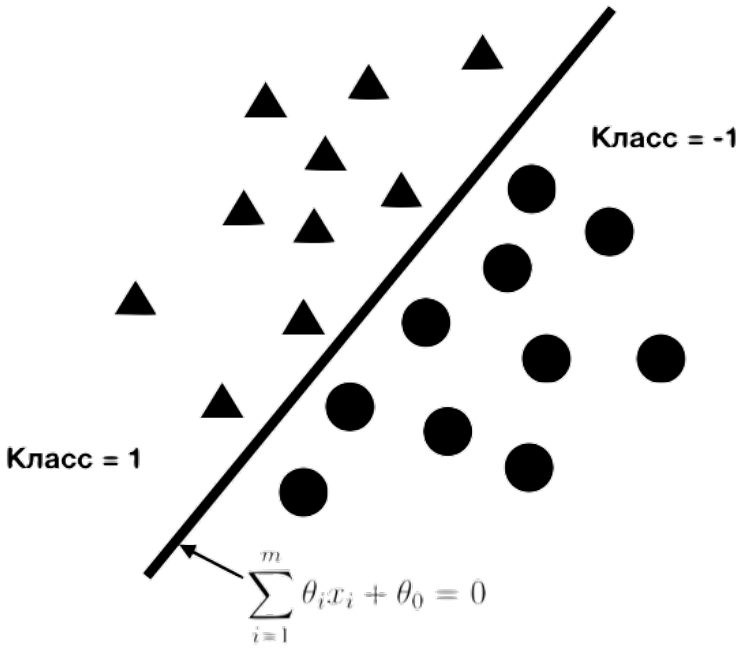


Рис. 1: Исходная выборка

Чем ближе объект находится к гиперплоскости, тем меньше вероятность принадлежности тому классу с какой стороны находится объект. Для объектов на гиперплоскости классификатор выдает вероятность 0.5, так как нельзя отнести их к одному из классов. Таким образом алгоритм может выдавать апостериорную вероятность принадлежности объекта классам.

$$\mathbb{P}(y|x) = \sigma(y\langle x, \theta\rangle) = \frac{1}{1 + e^{-y\langle x,\theta\rangle}}$$

$$\sum_{i=1}^{m} \theta_i x_i + \theta_0 = 0$$

Гребневая регрессия

Данный метод используется для того, чтобы снизить корреляцию признаков объектов. Ведь в отличии от обычной регресии в данном методе используется L_2 регуляризация, которая помогает в случае, если матрица $X^T X$ вырождена. Запишем функционал эмпирического риска гребневой регрессии и выпишем аналитическое решение для нахождения весов модели.

$$Q(\alpha) = \|X\theta - Y\|^2 + \alpha\|\theta\|^2 \longrightarrow \min_{\theta}$$

Где α коэффициент регуляризации.

$$\frac{dQ}{d\theta} = 2X^T(X\theta - Y) + 2\alpha\theta$$

$$(X^TX + \alpha I)\theta = X^TY$$

$$\widetilde{\theta} = (X^TX + \alpha I)^{-1}X^TY$$

Классификатор на основе гребневой регрессии:

$$a(x, \theta) = sign\langle x, \theta\rangle$$

Метод ближайших соседей

Данный алгоритм относится к метрическим методам классификации и исходит из предположения, что близким в метрическом пространстве объектам соответствуют похожие метки. Для нахождения метки класса метод рассматривает k ближайших в признаковом пространстве объектов, по отношению к классифицируемому и относит его к тому классу, представителей какого больше среди просмотренных соседей. Данный алгоритм предполагает задание метрики на объектах.

$$a(u) = \underset{y_i \in Y}{\operatorname{argmax}} \sum_{j=1}^{k} [x_{j,u} = y_i]$$

Обучение алгоритма состоит в запоминании обучающей выборки. На этапе прогнозирования, метод просматривает все объекты, находит k ближайших и путем голосования находит класс искомого объекта.

Если несколько классов имеют одинаковый ранг при голосовании, то есть несколько путей решения:

- Случайно выбрать из классов с одинковым рангом.
- Выбрать тот, класс, чей представитель ближе к искомому элементу.
- Выбрать тот класс, чей самый дальний представитель ближе к искомому объекту.

Случайный лес

Случайный лес – множество глубоких деревьев. Для построения дерева используется подвыборка с повторениями из исходной выборки. Более того, для построения каждой вершины используется случайное подмножество признаков. В случае

с задачей классификации решение о классе объекта принимается голосованием по большинству.

Алгоритм обучается до исчерпания обучающей подвыборки. Стоит отметить, что в данном алгоритме не используется последующее упрощение алгоритма (стрижка), так как данный алгоритм использует идею ансамблирования.

1.7 Обобщение на многоклассовую классификацию

Алгоритмы классификации можно использовать так же для многоклассовой классификации на C классов. В таком случае нужно использовать одну из двух моделей:

Один против всех

Обучается C бинарных классификаторов, каждый из которых умеет классифицировать объекты на «свои», то есть объекты принадлежащие классу i, и «чужие» – объекты всех остальных классов.

Ответ для каждого объекта находится следующим образом:

$$y(x) = \underset{c \in \{1 \dots C\}}{\operatorname{argmax}} a_c(x),$$

где a_c – классификатор отделяющий класс c от всех остальных.

Все против всех

Обучается $\frac{C(C-1)}{2}$ бинарных классификаторов для каждой пары $i, j \in \{1 \dots C\}$. Каждый из таких классификаторов умеет классифицировать объекты на два класса i и j.

На этапе обучения ответ определяется по большинству голосов:

$$y(x) = \underset{c \in \{1 \dots C\}}{\operatorname{argmax}} \sum_{i \neq j \in \{1 \dots C\}} \mathbb{I}[a_{i,j}(x) = c]$$

2 Обзор методов обнаружения аномалий

Методы обнаружения аномалий можно разделить на категории по способам их взаимодействия с объектами.

2.1 Анализ экстремальных значений

Анализ экстремальных значений очень специфичный метод для анализа аномалий. Как правило данный метод используется для нахождения выбросов по отдельным признакам. Если зафиксировать конкретный признак и построить его распределение, то аномальными будут объекты, чьи значения фиксированного признака будут находиться в хвостах полученного распределения.

Важно понимать, что аномалия может быть экстремальным значением, но не всякое экстремальное значение является аномалией. Приведем пример: $0, 1, 1, 1, 10, 21, 21, 22$. В данном примере экстремальными значениями являются 0 и 22, но значение 10 наиболее изолированное, хотя и не может являться выбросом, если исходить из данного алгоритма.

2.2 Модели основанные на глубине

Методы на основанные на глубине используют общий принцип, что выпуклая оболочка набора точек данных представляет собой оптимальные по парето экстремумы этого множества. Алгоритм, основанный на глубине, выполняется итеративно, на k-й итерации удаляются все точки в углах выпуклой оболочки набора данных. Индекс итерации k также дает оценку аномальности объекта. Меньшие значения указывают на большую тенденцию к тому, что объект будет выбросом. Эти шаги повторяются до тех пор, пока набор данных не будет пустым. Рейтинг аномальности может быть преобразован в двоичную метку, если помечать все точки данных с глубиной не более r как выбросы. Значение r может быть определено путем одномерного анализа экстремальных значений.

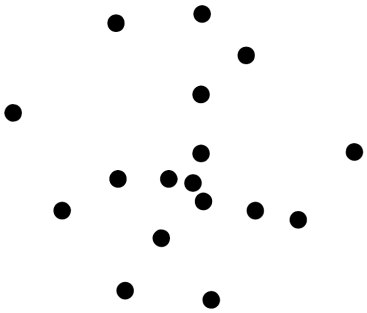


Рис. 2: Исходная выборка

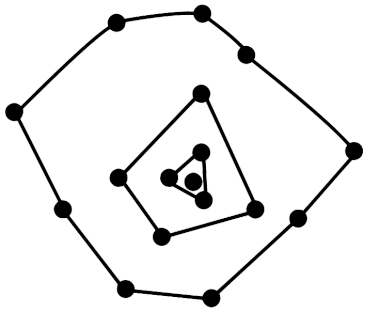


Рис. 3: Выборка с размеченными объектами

На рисунке 3 продемонстрирован результат работы алгоритма на выборке с рисунка 2.

2.3 Вероятностные модели

Вероятностные модели предполагают, что выборка была сгенерирована из смеси k распределений $G_1, \ldots, G_k$ с помощью следующего процесса:

1. У каждой компоненты смеси есть априорная вероятность того, что эта компонента будет выбрана. Предполагаем, что на этом шаге выбрали m распределение с априорной вероятностью α_m
2. Генерируем объект из распределения G_m

После оценки параметров модели $\alpha_i, i \in \{1 \ldots k\}$ и параметров распределений $G_i, i \in \{1 \ldots k\}$ выбросы определяются как те точки, что имеют малую вероятность генерации описанным путем.

2.4 Кластеризация

Алгоритмы кластеризации находят совокупности объектов, находящихся близко друг к другу и объединяют их в один кластер, выбросы же наоборот являются уединенными обектами. Таким образом, можно сказать, что любой объект выборки либо входит в какой-то кластер, либо является выбросом.

Аномалии сами могут образовывать небольшие кластеры, такое случается если процесс в следствии которого появляются аномалии повторяется несколько раз.

Поэтому кластеры с количеством объектов ниже порогового значения можно считать выбросами. Слабыми аномалиями могут являться объекты, находящиеся на границе кластеров.

В качестве рейтинга аномальности можно использовать расстояние до ближайшего центроида, посчитанное на кластеризованных данных.

Алгоритмы кластеризации не могут отличать выбросы типа шум и настоящие аномалии.

2.5 Модели основанные на растоянии

Наиболее распространенным является метод подсчета рейтинга аномальности как расстояние до k ближайшего соседа. Иногда используются вариации, например, подсчет среднего арифметического расстояний до k ближайших соседей. Параметр k является настраиваемым. Если выбрать $k > 1$, то таким образом можно найти изолированную группу объектов. Важно, что параметр k не включает в себя искомый элемент, так как иначе для метода одного соседа для всех объектов рейтинг бы равнялся нулю.

Данная модель может отличать шум от аномалий, дело в том, что у шумовых объектов дистанция до k-го соседа будет меньше, чем у изолированных точек. В алгоритмах кластеризации такое разделение выбросов не достигается из-за того, что расстояние до ближайшего центроида не дает информации о ближайшей окрестности точки.

2.6 Модели основанные на плотности

Основная идея данного типа моделей – выделить регионы, плотно заполненные объектами и далее с помощью данных блоков сконструировать кластеры.

Одним из примеров является сеточный алгоритм, идеей которого является разбиение пространства на n-мерные гиперкубы и, если в гиперкуб попало больше чем k точек, то данную вершину объявляем вершиной графа. Между вершинами прово-

дится ребро, если они смежны по l измерениям из d. Ищем компоненты связности в полученном графе. Каждая компонента связности является кластером.

Компоненты связности с малым количеством вершин объявляются аномалиями.

3 Метод решения

Пусть имеется выборка $X = \{x_1 \dots x_n\}$. На ней запустим метод поиска логических закономерностей классов. Будем использовать полученные предикаты $P_1^{c_1^2,c_1^2} \dots P_m^{c_m^2,c_m^2}$ для подсчета рейтинга аномальности.

Весом v_i предиката $P_i^{c_i^1,c_i^2}$ будем называть долю объектов, которые покрывает данный предикат. Вес описывает «важность» предиката, которая интерпретируется следующим образом: чем больше объектов покрывает предикат, тем лучше он описывает обучающую выборку.

$$\widetilde{C}_i = |\{x_j : x_j \in K_k, P_i^{c_i^1,c_i^2} \in K_k\}|$$

$$v_i = \frac{1}{\widetilde{C}_i} \sum_1^n \mathbb{I}[P_i^{c_i^1,c_i^2}(x_n) = 1]$$

Введем функцию рейтинга аномальности:

$$f(x) = -\sum_1^m v_i \cdot \mathbb{I}[P_i^{c_i^1,c_i^2}(x) = 1]$$

Чем выше рейтинг аномальности, тем больше вероятность того, что объект является выбросом. Таким образом, мы получили интуитивно понятный функционал: чем больше важных предикатов покрывают объект, тем меньше шансов на то, что он является выбросом. Такой объект находится в скоплении других объектов. Тем самым, рассматриваемая функция ранжирует объекты в порядке их наибольшей удаленности от остальных.

Рассмотрим следующий функционал:

$$\mathcal{F}(x) = P_{base} \cdot \big(1 - \frac{|outliers|}{|base|}\big) + P_{outliers} \cdot \frac{|outliers|}{|base|}$$

Где P_{base} – вероятность правильной классификации выборки без выбросов. $P_{outliers}$ – вероятность правильной классификации выбросов.

Будем проводить одномерную оптимизацию. На каждом шаге будем удалять из выборки элемент с самым высоким рейтингом аномальности. Для полученной выборки посчитаем функционал $\mathcal{F}(x)$.

Вероятность правильной классификации будем считать при помощи следующих моделей: логистической регресии, гребневой регрессии, классификатора, основанного на случайном лесе и метода ближайших соседей.

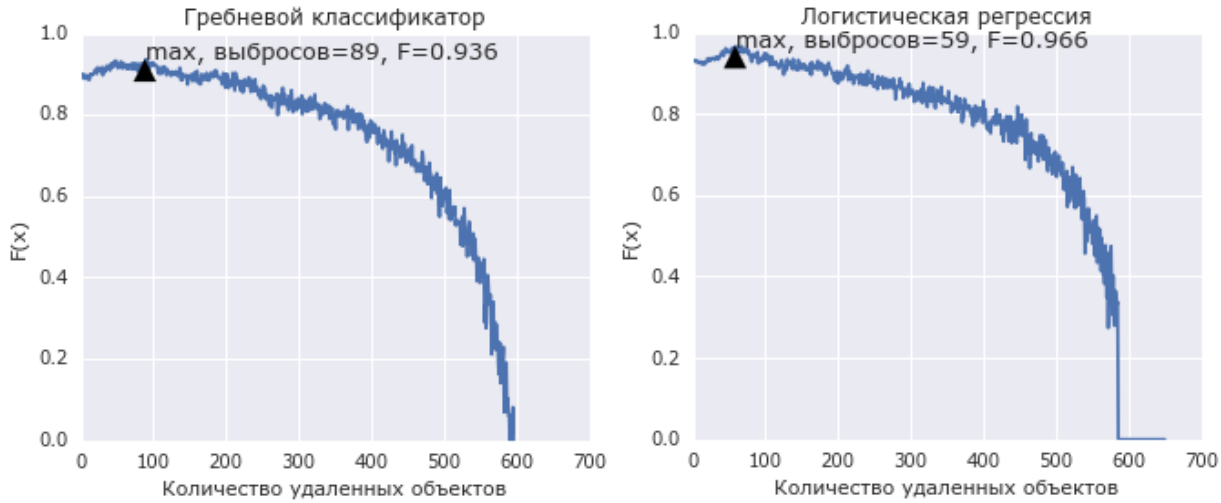
Начнем с подобра параметров моделей. Для этого будем использовать исходную выборку, чтобы понять какие параметры классификаторов лучше подходят для данной задачи. Найдем их с помощью скользящего контроля.

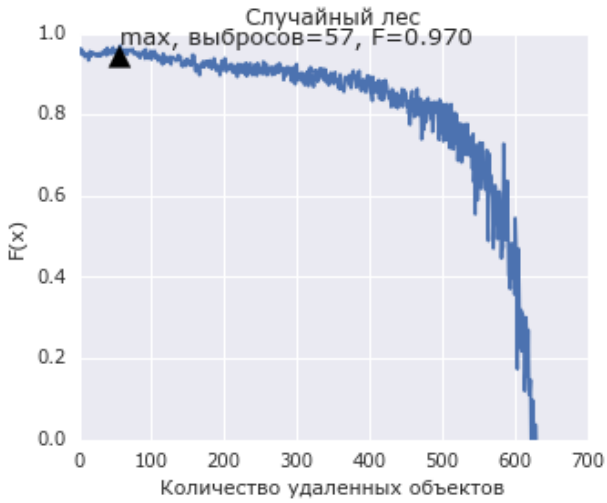
Рассмотрим шаг s и любую из моделей с подобранными параметрами. На данном шаге у нас имеется $|outliers| = s$ объектов выбросов и $|base| = n - s$ объектов исходной выборки, назовем их базовыми. Вероятность P_{base} будем считать по базовым элементам, а $P_{outliers}$ по выбросам. Для оценивания вероятности будем использовать скользящий контроль и оценивать качество алгоритма с помощью F-меры. Отдельно стоит уточнить, что для выбросов мы обучаем классификатор отдельно.

4 Вычислительные эксперименты

4.1 Анализ функции аномальности

Посмотрим на поведение функционала для различных алгоритмов.





По графикам заметим что, на всех моделях функционал определяет меньше 100 выбросов, но тем не менее, количество аномальных объектов разнится. Для того, чтобы найти оптимальное количество выбросов, посмотрим на поведение функции рейтинга аномальности.

Как видно из графика, первый скачок уменьшения аномальности объектов функция совершает на 100 элементе. Вероятно, это самые аномальные объекты. Удалим их из обучающей выборки и посмотрим на качество классификации по F-мере моделей на полученной подвыборке.

Но качество по F-мере измеряется только на тех объектах, что не являются выбросами, а $\mathcal{F}(x)$ – функционал дает нам взвешенную оценку качества на базовых

объектах и выбросах. В общем случае выбросы могут быть и в выборке, метки которой нам нужно предсказать.

Таблица 1: Качество классификации на исходной выборке и на выборке с количеством аномалий в точке максимума функционала

Алгоритм классификации	F-мера полной выборки	F-мера на подвыборке
Логистическая регрессия	0.929035	0.984699
Метод ближайших соседей	0.943460	0.969372
Гребневой классификатор	0.902585	0.944731
Случайный лес	0.944703	0.993272

Таблица 2: Изменение качества алгоритма по F-мере и по введеному функционалу после отбрасывания 100 выбросов

Алгоритм классификации	F-мера на подвыборке	$\mathcal{F}(x)$ на подвыборке
Логистическая регрессия	0.996546	0.940259
Метод ближайших соседей	0.987488	0.950848
Гребневой классификатор	0.980902	0.920423
Случайный лес	0.996551	0.948142

Тем самым, можно наблюдать, что качество классификации заметно повысилось. Мы рассмотрели первый скачок функции, который меньше единицы. Но на графике видно еще два скачка с более значительным изменением функции. Они находятся на 137 и 227 объектах соответственно. Рассмотрим оценку качества алгоритма с помощью скользящего контроля при удалении данного количества объектов.

Таблица 3: Изменение качества алгоритма по F-мере при отбрасывании 137 аномалий

Алгоритм классификации	F-мера	$\mathcal{F}(x)$
Логистическая регрессия	0.996767	0.928263
Метод ближайших соседей	1.000000	0.930447
Гребневой классификатор	0.992561	0.923450
Случайный лес	0.998129	0.933427

Таблица 4: Изменение качества алгоритма по F-мере при отбрасывании 227 аномалий

Алгоритм классификации	F-мера	$\mathcal{F}(x)$
Логистическая регрессия	1.0	0.869904
Метод ближайших соседей	1.0	0.907333
Гребневой классификатор	1.0	0.864520
Случайный лес	1.0	0.892921

Заметим, что модели при удалении 227 объектов начали идеально распознавать оставшуюся выборку, но точность распознования выбросов снизилась. Но посмотрим на соотношение классов в выборке.

Рис. 4: Соотношения классов

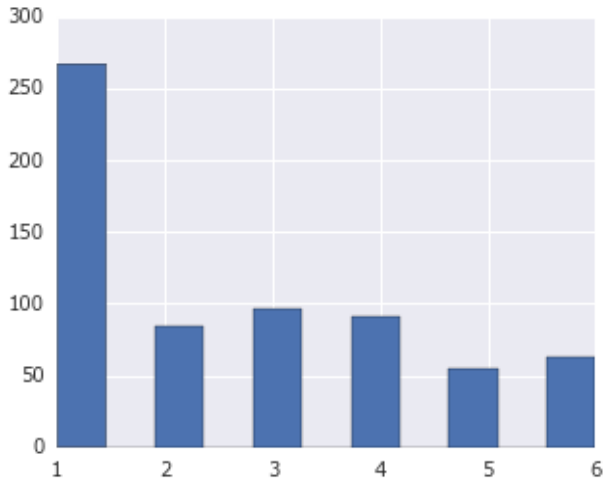


Рис. 5: В исходной выборке

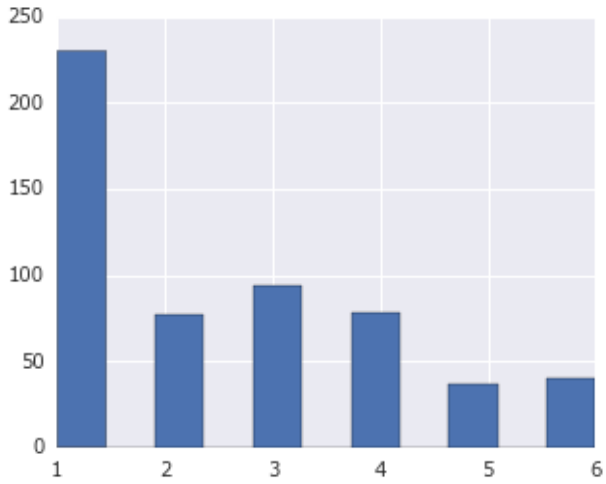


Рис. 6: При удалении 100 выбросов

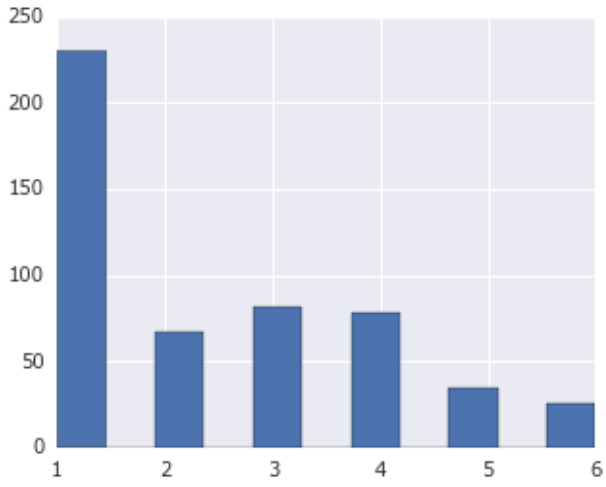


Рис. 7: При удалении 137 выбросов

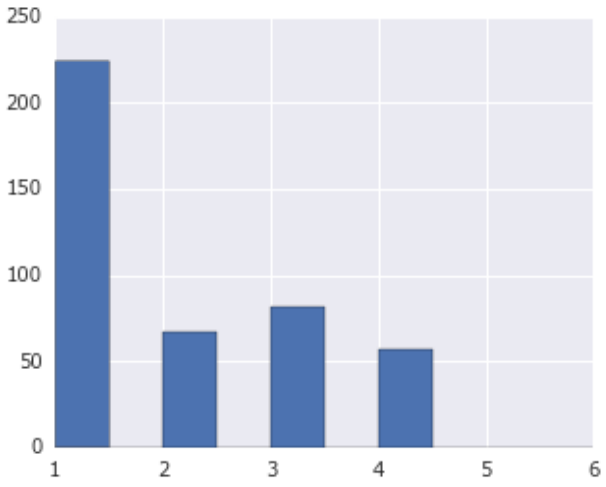


Рис. 8: При удалении 227 выбросов

На гистограммах видно, что соотношение классов остается примерно таким же как в исходной выборке при удалении 100 и 137 объектов, при этом обобщающая способность лучше у алгоритмов, обученных на выборке без 137 аномалий. При удалении 227 выбросов исчезают два класса, но точность распознования оставшихся максимальна.

Таким образом, для того чтобы найти оптимальное количество выбросов, нужно рассмотреть все значимые скачки функции аномальности и исследовать соотношения классов при удалении выбранного количества объектов. Или же искать скачки функции наиболее близкие к максимуму функционала $\mathcal{F}(x)$

4.2 Выводы

Алгоритм нахождения выбросов, основанный на логических закономерностях классов показал свою состоятельность применительно к задаче классификации химических соединений по типу кристаллической решетки. С его помощью удалось повысить качество классификации данных без выбросов минимум на 0.04 пункта.

Был придуман функционал применительно к задаче детектирования аномалий, взвешенно оценивающий качество классификации на выборке без выбросов и выборке состоящей только из выбросов, тем самым помогая подобрать оптимальное количество выбросов.

Было показано, что так же можно выбирать количество выбросов по функции рейтинга аномальности, данный метод помогает найти более оптимальное значение F-меры.

Тем самым, была решена важная прикладная задача по выяснению заранее неизвестных свойств химических соединений, опираясь лишь на данные о характеристиках каждого элемента. Данный метод хорош для выборки по которой нужно сделать прогноз: если мы хотим хорошо классифицировать хотя бы небольшую ее часть, то нужно найти количество выбросов, опираясь на функцию аномальности.

5 Заключение

Алгоритмы обнаружения выбросов хорошо себя показывают применительно к задачам классификации, убирая выбросы из данных они помогают получить модель с лучшими обобщающими характеристиками, которая не подстраивается под шумовые объекты и не пытается вписать лишние данные.

В данной работе реализован алгоритм детектирования аномалий, основанный на логических закономерностях классов. Алгоритм использует построенные предикаты и на их основе строит рейтинг аномальности для каждого объекта.

Были проведены эксперименты, показывающие возможности данного алгоритма по улучшению классификации путем удаления выбросов из данных.

Так же была решена важная прикладная задача по определению типа кристаллической решетки элементов $A^{3+}B^{3+}C^{2+}O_4$, что очень важно для поиска новых веществ с различными функциональными свойствами.

Список литературы

[1] Журавлев Ю. И., Рязанов В. В., Сенько О. В. *«РАСПОЗНАВАНИЕ». Математические методы. Программная система. Практические применения.* М.: ФАЗИС. 2006. 176 с

[2] Charu C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015.

[3] Александр Дьяконов. *Анализ малых данных. КвазиНаучный блог* `https://alexanderdyakonov.wordpress.com`

[4] Н.В.Ковшов, В.Л.Моисеев, В.В.Рязанов. *Алгоритмы поиска логических закономерностей в задачах распознавания*

[5] Журавлев Ю.И., *Об алгебраическом подходе к решению задач распознавания или классификации. Проблемы кибернетики.* М.: Наука, 1978. Вып.33.

[6] Douglas M. Hawkins, *Identification of Outliers* Chapman and Hall, 1980

[7] Cundari T.R., Moody E.W. *A Comparison of Neural Networks versus Quantum Mechanics for Inorganic Systems* J. Chem. Inf. Comput. Sci. 1997

[8] Сообщество Open Data Science, Цикл статей по машинному обучению `https://habr.com/company/ods/blog`

[9] R. Patel, C. Simon, and M. T. Weller, J. *Solid State Chem.* (2007)

[10] D. Ganguly, J. *Solid State Chem.* (1979)

[11] N.Kimizuka and T.Mohri, J. *Solid State Chem.* (1989)

[12] R. Roy, *The Major Ternary Structural Families* Springer, Berlin/Heidelberg/New York, 1974.

[13] В.В.Рязанов.*Логические закономерности в задачах распознавания (параметрический подход)* Журнал вычислительной математики и математической физики, 2007, Т. 47.

[14] Журавлев Ю.И., Никифоров В.В. *Алгоритмы распознавания, основанные на вычислении оценок*, Кибернетика. 1971. №3. С. 1-11.

[15] Christopher M. Bishop *Pattern Recognition and Machine Learning*, Springer, 2006