Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Ожерельев Илья Сергеевич

Методы повышения точности моделей машинного обучения с использованием выпуклых композиций

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:

д.ф.-м.н., ведущий н.с. ВЦ РАН

О.В. Сенько

Москва, 2018

Содержание

Аннотация

Композиции алгоритмов машинного обучения часто используются для повышения обобщающей способности, позволяя входящим в композицию алгоритмам компенсировать ошибки друг друга. В работе используются два метода повышения точности моделей машинного обучения с использованием выпуклых композиций - модификация метода СВС и метод отбора аномальных наблюдений. Метод статистически взвешенных синдромов - один из методов построения выпуклых комбинаций. Метод СВС успешно применяется на малых и зашумлённых данных, способен автоматически обрабатывать пропуски в данных. В работе рассматриваются методы, позволяющие повысить качество и ускорить работу алгоритма СВС - построение синдромов по разреженной сетке и построение дополнительных разбиений. Рассматриваемый метод поиска выпадающих объектов (ВО) основан на одновременном использовании информации о величинах оценок объекта за свой и чужие классы, а также об интегральных искажениях, вносимых объектом в формируемый в результате обучения алгоритм распознавания. Возможность использования разработанного метода при высокой размерности данных была продемонстрирована на задаче прогнозирования устойчивости неорганических соединений состава A+3B+3C+2O4. Метод может быть использован для выявления ошибочных наблюдений, при поиске нетипичных объектов, при формировании обучающих выборок при решении задач распознавания или прогнозирования в различных областях.

1 Введение

Методы машинного обучения применимы для решения множества задач, не относимых к "большим данным которые, тем не менее, имеют актуальность. Распространённые причины малочисленности данных – дороговизна их извлечения, редкость исследуемых событий. В сфере "малых данных"особенно важен вопрос переобучения модели. Это создаёт спрос на простые и робастные методы, налагающие меньшие требования на структуру данных. С другой стороны, возможно использовать вычислительно сложные алгоритмы, применение которых на выборках большего объёма приведёт к большим затратам времени. Кроме того, для разметки малых данных часто приходится проводить дорогостоящие исследования. Чтобы собрать репрезентативную выборку требуется собирать данные из различных источников - статей и справочников. Данные в этих источниках представлены в различных форматах, поэтому при составлении выборки возможны ошибки.

Метод "статистически взвешенные синдромы"(СВС), был разработан в 1993-1995 гг. в ходе совместной работы группой сотрудников ВЦ РАН и Института биохимической физики им. Н. М. Эмануэля РАН. Метод основан на построении выпуклой комбинации синдромов - закономерностей, которые задаются оптимальными разбиениями признакового пространства. СВС хорошо показал себя в работе с "малыми данными"и неоднократно [1],[2] модифицировался.

В рамках данной работы поставлены следующие задачи:

- Ускорить работу метода СВС модификацией процедуры построения синдромов.
- Повысить качество работы метода СВС через включение в композицию дополнительных синдромов.
- Разработать метод для выявления ошибочных наблюдений и поиска нетипичных объектов при формировании обучающих выборок.

2 Выпуклые комбинации

Выпуклые комбинации – подход к построению ансамблей с помощью агрегирования заданных алгоритмов. Алгоритмам назначаются неотрицательные веса, с которыми они линейно входят в композицию. Проведём анализ ошибки выпуклой комбинации.

Рассмотрим задачу бинарной классификации. Задано признаковое пространство $\mathbb{X} = \mathbb{R}^n$. Существует скрытая зависимость $Y : \mathbb{X} \to \{0, 1\}$. Пусть Ω - распределение из которого получена обучающая выборка S. объекты выборки $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ имеют признаковое описание $x \in \mathbb{R}^n$ и метку класса $y \in \{0, 1\}$. Требуется восстановить зависимость Y. Пусть $f_j^i \in [0, 1]$ - прогноз кдасса объекта $s_i \in S$, вычисляемый алгоритмом A_j из композиции $A_1, \ldots, A_r$. Если $f_j^i < 0.5, A_j$ относит s_i к классу с меткой "0 если $f_j^i \geq 0.5, A_j$ относит s_i к классу с меткой "1". Тогда

$$\Delta_j = \mathbb{E}_\Omega (Y - f_j)^2$$

является математическим ожиданием квадрата ошибки прогнозирования для алгоритма A_j. Введём обозначение расстояния между прогнозами

$$\rho_{j'j''} = \mathbb{E}_\Omega \Big((f_{j'} - \mathbb{E}_\Omega f_{j'}) - (f_{j''} - \mathbb{E}_\Omega f_{j''}) \Big)^2$$

Пусть $c_1, \ldots, c_r$ - положительные коэффициенты, такие что $\sum\limits_{j=1}^{r} c_j = 1$ Обозначим через $\hat{f}$ выпуклую комбинацию прогнозов, вычисляемых алгоритмами ансамбля $A_1, \ldots, A_r$.

$$f_i = \sum_{j=1}^{r} c_j f_j$$

Для ошибки выпуклой комбинации справедливо выражение:

$$\hat{\Delta} = \mathbb{E}_\Omega (Y - \hat{f})^2$$

$$\hat{\Delta} = \mathbb{E}_\Omega (\sum_{j=1}^{r} c_j Y - \sum_{j=1}^{r} c_j f_j)^2$$

$$\hat{\Delta} = \mathbb{E}_\Omega (\sum_{j=1}^{r} c_j (Y - f_j))^2$$

$$\hat{\Delta} = \mathbb{E}_\Omega (\sum_{j=1}^{r} c_j (Y - f_j))^2 + \sum_{j=1}^{r} \sum_{i=1}^{r} c_i c_j ((\mathbb{E}_\Omega f_i - \mathbb{E}_\Omega f_j)^2 - (\mathbb{E}_\Omega f_i - \mathbb{E}_\Omega f_j)^2) =$$

$$= \sum_{j=1}^{r} \sum_{i=1}^{r} c_i c_j \mathbb{E}_\Omega ((f_i - Y)(f_j - Y)) +$$

$$+ \sum_{j=1}^{r} \sum_{i=1}^{r} c_i c_j \Big(\mathbb{E}_\Omega \big(f_i (\mathbb{E}_\Omega f_i - \mathbb{E}_\Omega f_j) \big) - \mathbb{E}_\Omega f_j (\mathbb{E}_\Omega f_i - \mathbb{E}_\Omega f_j) \big) - (\mathbb{E}_\Omega f_i - \mathbb{E}_\Omega f_j)^2 \Big) =$$

$$= \frac{1}{2} \sum_{j=1}^{r} \sum_{i=1}^{r} \mathbb{E}_\Omega \Big(2(f_i - y)(f_j - y) + (f_i - f_j)(\mathbb{E}_\Omega f_i - \mathbb{E}_\Omega f_j) + (\mathbb{E}_\Omega f_i - \mathbb{E}_\Omega f_j)^2 \Big) =$$

$$= \sum_{j=1}^{r} \sum_{i=1}^{r} c_i c_j \mathbb{E}_\Omega (f_i - Y)^2 - \frac{2}{2} \sum_{j=1}^{r} \sum_{i=1}^{r} c_i c_j \mathbb{E}_\Omega ((f_i - Y)^2 -$$

$$\frac{1}{2}\sum_{j=1}^{r}\sum_{i=1}^{r}\mathbb{E}_\Omega\Big(2(f_i-y)(f_j-y)+(f_i-f_j)(\mathbb{E}_\Omega f_i-\mathbb{E}_\Omega f_j)+(\mathbb{E}_\Omega f_i-\mathbb{E}_\Omega f_j)^2\Big)=$$

$$=\sum_{j=1}^{r}\sum_{i=1}^{r}c_ic_j\mathbb{E}_\Omega(f_i-Y)^2-\frac{1}{2}\sum_{j=1}^{r}\sum_{i=1}^{r}c_ic_j\mathbb{E}_\Omega(f_i-\mathbb{E}_\Omega f_i)-(f_j-\mathbb{E}_\Omega f_j)^2$$

$$\hat{\Delta}=\sum_{j=1}^{r}c_j\Delta_j-\frac{1}{2}\sum_{j=1}^{r}\sum_{i=1}^{r}c_ic_j\rho_{ij}$$

Так, как ρ_{ij} и коэффициенты выпуклой комбинации всегда неотрицательны,

$$\hat{\Delta}\leq\sum_{j=1}^{r}c_j\Delta_j \tag{1}$$

Вывод. Использование качественных разнородных комбинаций (с высоким расстоянием между алгоритмами, входящими в композицию) приводит к увеличению отрицательного члена в (1), уменьшая итоговую ошибку ансамбля.

3 Метод статистически взвешенных синдромов

3.1 Исходный вариант метода СВС

Метод статистически взвешенных синдромов представляет собой метод классификации, основанный на процедуре статистически взвешенного голосования индикаторных функций. В качестве базовых множеств используются подобласти в пространстве признаков, внутри которых содержание объектов одного из классов значительно отличается от его содержания в смежных областях. Такие разбиения далее будут называться закономерностями или синдромами. Алгоритм можно разделить на три шага:

1. Построение синдромов.
2. Верификация синдромов.
3. Построение выпуклой комбинации синдромов.

Построение синдромов. В методе СВС используются только одномерные и двумерные синдромы. *Синдромом первого типа*, задаваемым признаком X_i, мы называем бинарный классификатор, который разбивает пространство признаков на два подпространства: $X_i < b_i'$ и $b_i' \leq X_i$. *Синдромом второго типа*, задаваемым признаком X_i мы называем бинарный классификатор, который разбивает пространство

признаков на три подпространства: $X_i < b_i'$; $b_i' \leq X_i < b_i''$ и $b_i'' \leq X_i$. *Синдромом третьего типа*, задаваемым парой признаков (X_i, X_j) - бинарный классификатор, делящий пространство признаков на четыре подпространства: $(X_i < b_i', X_j < b_j')$; $(X_i \leq b_i', X_j < b_j')$; $(X_i < b_i', X_j \leq b_j')$; $(X_i \leq b_i', X_j \leq b_j')$. В качестве оценки за класс K_l используется оценка частоты объектов класса K_l, полученная по обучающей выборке.

Границы одномерных синдромов строятся перебором всевозможных положений объектов выборки относительно границ синдрома. Для каждого признака среди всевозможных синдромов выбирается тот, который максимизирует функционал качества. Авторами метода предлагают использовать следующие функционалы:

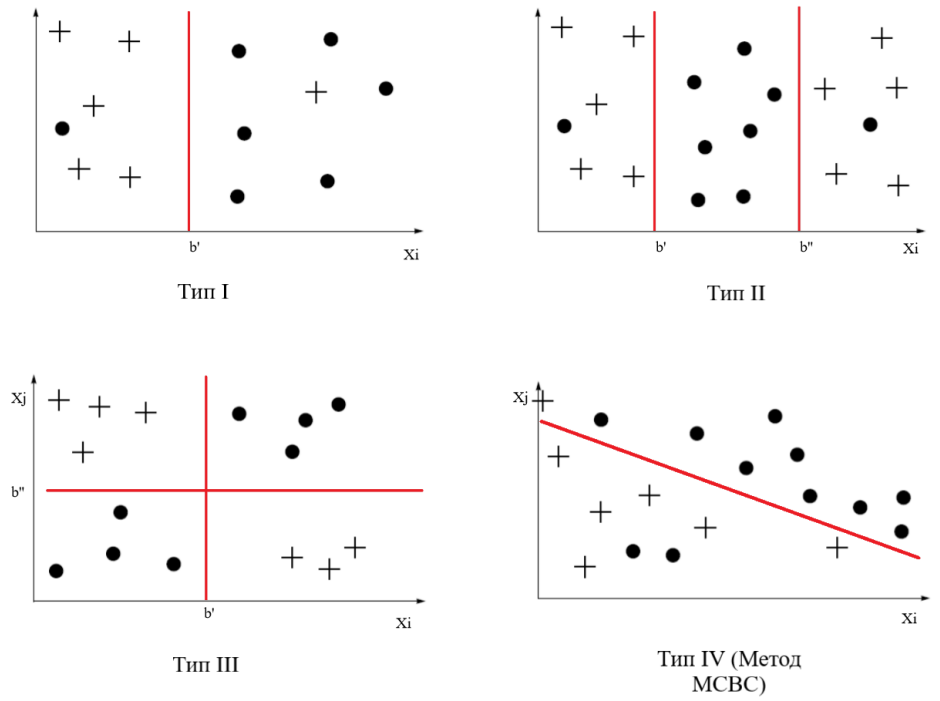

Рис. 1: Типы разбиений. Типы I и II называют одномерными разбиениями (синдромами), типы III и IV - двумерными. Метками "."и "+"обозначены объекты различных классов.

Разбиение R - покрытие признакового пространства L подобластями $q_1, \ldots, q_L$.

$$F_\chi(R, S_{train}, K_0) = \frac{1}{\nu^0(1-\nu^0)} \sum_{l=1}^{L} (\nu_l^0 - \nu^0)^2 m_l, \tag{2}$$

где m - число объектов в выборке S_{train}, m^0 - число объектов класса K_0 в выборке S_{train}, $\nu^0 = \frac{m^0}{m}$ - доля объектов класса K_0 в выборке S_{train}; m_l - число объектов S_train в подобласти q_l, m_l^0 - число объектов класса K_0 в подобласти q_l. $\nu_l^0 = \frac{m_l^0}{m_l}$ -

доля объектов класса K_0 в подобласти q_l.

$$F_{cv}(R, S_{train}, K_0) = \sum_{l=1}^{L} \sum_{i=1}^{m_l} (\alpha_{il}^0 - \frac{m_l^0 - \alpha_{il}^0}{m_l - 1}), \tag{3}$$

Где $\alpha_{il}^0 = 1$, если i-ый объект из области q_l принадлежит классу K_0, и $\alpha_{il}^0 = 0$ иначе. Двумерные синдромы строятся как всевозможные пересечения одномерных.

Верификация синдромов. В композицию включаются только те синдромы, для которых значение функционала качества выше определённого порога.

Построение композиции. Пусть $Q_0 = \{R_1, \ldots, R_k\}$ - множество синдромов построенных для класса K_0 и прошедших верификацию, пусть вектор переменных x^* объекта s^* принадлежит подобластям $q_1^*, \ldots, q_k^*$ синдромов $R_1, \ldots, R_k$, тогда оценка объекта s^* за класс K_0 вычисляется по формуле

$$\Gamma(s^*) = \frac{\sum\limits_{i=1}^{k} w_i \nu_i^0}{\sum\limits_{i=1}^{k} w_i}, \tag{4}$$

где w_i - вес подобласти $q_i^* \in R_i$.

$$w_i = \frac{m_i}{m_i + 1} \frac{1}{(1 - \nu^0)\nu^0 + \frac{1}{m_i}(1 - \nu_i^0)\nu_i^0}, \tag{5}$$

если $m_i > 0$ и 0 иначе.

3.2 Метод мультимодельных статистически взвешенных синдромов

Метод мультимодельных статистически взвешенных синдромов (МСВС) является модификацией метода СВС. Синдромы третьего типа строятся не как пересечения синдромов первого типа, а перебором всевозможных координат центра креста по сетке $\{(x_{1i}+x_{2i})/2, (x_{(m-1)i}+x_{mi})/2\} \times \{(x_{1j}+x_{2j})/2, (x_{(m-1)j}+x_{mj})/2\}$, здесь x_{kl} - значение l-ого признака k-ого объекта обучающей выборки. Среди всевозможных синдромов третьего типа для пары признаков отбирается синдром с лучшим значением функционала качества.

Также в методе МСВС используются синдромы IV типа. Разбиение задаётся прямой $X_j = a'X_i + b'$. Оптимальные параметры (a', b') подбираются при оптимизации функционала (1) или (2). Рассматриваются все прямые, проходящие через пары точек $\{\big((x_{1i}+x_{2i})/2, 0\big), \big((x_{(m-1)i}+x_{mi})/2, 0\big)\} \times \{\big(0, (x_{1j}+x_{2j})/2\big), \big(0, (x_{(m-1)j}+x_{mj})/2\big)\}$. Верификация синдромов и построение композиции происходит аналогично методу ОДР.

3.3 Метод оптимальных достоверных разбиений

Суть метода оптимальных достоверных разбиений (ОДР) заключается в том, что для всех признаков и пар признаков (в зависимости от типа разбиения) выполняются следующие процедуры: **Построение синдромов.** В методе ОДР используются синдромы I, II, и III типов. по схеме метода МСВС. **Верификация синдромов.** На этом шаге с помощью перестановочного теста производится проверка достоверности найденной закономерности. Для разбиений первого типа проверяется нулевая гипотеза о независимости метки класса и расположения объекта относительно граничного значения (т.е. найденная закономерность является случайной). Под перестановочным тестом в данном случае понимается следующая процедура. Объектам присваиваются случайные метки классов (соотношение размеров классов при этом сохраняется) и для полученной выборки вычисляются новые оптимальная граница и значение функционала. Данный перерасчет повторяется некоторое количество раз (чем больше, тем достовернее оценка). Далее в качестве оценки вероятности ($p - value$) того, что нулевая гипотеза о независимости верна, используется доля испытаний, при которых новое оптимальное значение функционала оказалось не меньше, чем исходное. В случае синдромов типа II и III (модели с двумя граничными значениями) выполняется два перестановочных теста: по одному на каждое граничное значение. Суть перестановочного теста остается такой же, как и для синдромов типа I, с той лишь поправкой, что перемешивание значений целевой переменной происходит отдельно слева и справа от второй границы (той, которая не верифицируется). Допустим, мы проверяем достоверность синдрома III типа для пары признаков X_i и X_j. Обозначим b_i и b_j соответствующие оптимальные граничные значения. Тогда перестановочный тест для верификации значимости разбиения по X_i описывается следующим образом. Значения целевой переменной перемешиваются отдельно в областях $X_j < b_j$ и $X_j \geq b_j$. Далее по перемешанным данным вычисляются новое оптимальное граничное значение признака X_i и оптимальное значение функционала. Граничное значение переменной X_j при этом никак не изменяется. Аналогично верификации синдромов I типа, описанная процедура перерасчета повторяется несколько раз, и затем рассчитывается значение $p - value$ – оценка вероятности того, что целевая переменная не зависит от X_i (при фиксированной границе X_j). Повторив эту же процедуру для X_j, можно получить второе $p - value$ – оценку вероятности того, что целевая переменная не зависит от X_j (при фиксированной границе X_i). В случае синдромов

II типа процедура перестановочного теста аналогична. Таким образом, при проверке достоверности синдромов типа II и III, рассчитываются два значения $p - value$: по одному на каждое граничное значение. Такой подход позволяет верифицировать обе границы и исключить из рассмотрения фиктивные закономерности, обусловленные только лишь наличием достоверного разбиения I типа. В результате применения метода ОДР мы получаем для каждого признака (пары признаков):

- оптимальные (с точки зрения заданного функционала качества) разбиения;
- оптимальное значение функционала качества – характеризует значимость найденной закономерности; чем больше данное значение, тем более значимой считается закономерность;
- значения $p - value$ – характеризуют достоверность найденной закономерности; данную величину стоит понимать как оценку вероятности того, что целевая переменная не зависит от соответствующего признака (при фиксированной второй границе для разбиений типа II и III), т.е. чем ближе данное значение к 0, тем более достоверной является закономерность. Далее из найденных закономерностей отбираются наиболее значимые и достоверные, т.е. отбираются разбиения с большим значением функционала качества и/или низким $p - value$.

Вычисление оценки за класс. Построение композиции, вычисление весов и оценок происходит аналогично методу МСВС.

В статье [3] отмечается, что метод ОДР позволяет эффективно отбирать признаки по величине статистики F_χ.

3.4 Достоинства и недостатки методов СВС, ОДР и МСВС

Методы ОДР и МСВС легко обрабатывают данные с пропусками (при построении синдрома достаточно не учитывать объекты с пропусками в рассматриваемом признаке) и легко обобщаются на категориальные данные (достаточно перебрать всевозможные разбиения категорий на два, три или 4 множества). Методы МСВС и ОДР показали свою эффективность на зашумлённых данных[3][4]. Синдромы предоставляют интерпретируемые результаты, легко визуализирутся, качество синдрома может говорить о важности признака (пары признаков).

Тем не менее, на больших выборках построение и верификация всех двумерных разбиений является вычислительно сложной задачей. Синдромы первого, второго и

третьего типа строятся перебором. Очевидно, что часть пар признаков не проходят верификацию и время, затраченное на поиск оптимального разбиения, не приводит к улучшению модели.

4 Ускорение метода СВС

4.1 Поиск оптимального разбиения по разреженной сетке

Предположим, что решается задача классификации вещественнозначных данных. В классическом методах ОДР и МСВС при построении синдрома III типа для признаков (X_i, X_j) предлагается построить сетку, узлами которых будут точки из $\{(x_{1i}+x_{2i})/2, (x_{(m-1)i}+x_{mi})/2\}\times\{(x_{1j}+x_{2j})/2, (x_{(m-1)j}+x_{mj})/2\}$. Если предположить, что каждый объект обладает уникальным значением параметра, получаем сетку из $(m-1)^2$ точек. Полный перебор по такой сетке займёт большое число времени, а исключение объектов из обучающей выборки на "малых данных"зачастую приводит к существенным искажениям модели данных.

В работе предлагается проводить первичный поиск по разреженной сетке и полный поиск в окрестности оптимума по следующему алгоритму.

Входные параметры алгоритма: шаг сетки k, порог качества синдрома q.

1. Отсортируем значения $\{(x_{1i}+x_{2i})/2, (x_{(m-1)i}+x_{mi})/2\}$ и $\{(x_{1j}+x_{2j})/2, (x_{(m-1)j}+x_{mj})/2\}$. Получим последовательности x_i_cut и x_j_cut.

2. По сетке $\{x_i\_cut_k, x_i\_cut_{2k}, x_i\_cut_{[m/k]k}\}\times\{x_j\_cut_k, x_j\_cut_{2k}, x_j\_cut_{[m/k]k}\}$

3. По разреженной сетке найти оптимальную точку $(x_i\_cut_{preopt_i}, x_i\_cut_{preopt_j})$

4. В окрестности оптимальной точки $[x_i\_cut_{preopt_i-1}, x_i\_cut_{preopt_i+1}]\times[x_i\_cut_{preopt_j-1}, x_i\_cu$ по полной сетке найти разбиение, максимизирующее функционал качества

5. Если качество найденного на шаге 3 синдрома превосходит q, включить синдром в композицию.

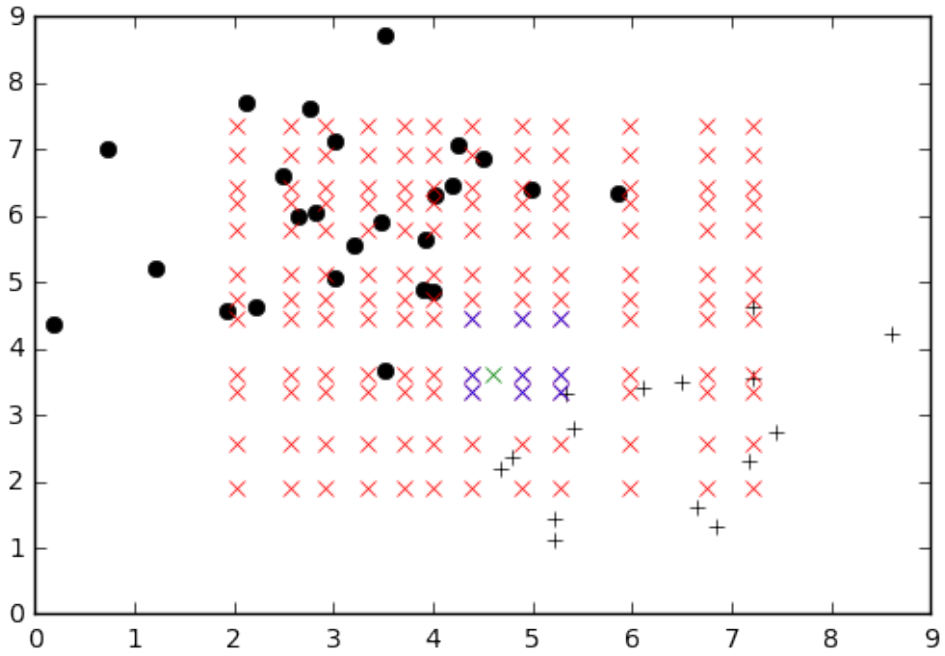

Рис. 2: Чёрные круги и крестики - объекты двух разных классов. Красные кресты - узлы разреженной сетки, синие кресты - окрестность оптимальной точки. Зелёный крест - уточнённое значение оптимума.

5 Построение дополнительных разбиений в методе СВС

5.1 Поиск оптимальных разбиений

В предшествующих модификациях метода СВС строилось по одному разбиению для каждой пары признаков. Так как для некоторых пар признаков возможно построить несколько различных, превосходящих порог качества разбиений, добавление дополнительных разбиений в выпуклую композицию может привести к повышению точности распознавания.

Предположим, что мы ищем синдром для класса K_0. Пусть R - покрытие признакового пространства L подобластями $q_1, \ldots, q_L$. Оптимальное покрытие ищется максимизацией функционала:

$$F_\chi(R, S_{train}, K_0) = \frac{1}{\nu^0(1-\nu^0)} \sum_{l=1}^{L} (\nu_l^0 - \nu^0)^2 m_l,$$

где m - число объектов в выборке S_{train}, m^0 - число объектов класса K_0 в выборке S_{train}, $\nu^0 = \frac{m^0}{m}$ - доля объектов класса K_0 в выборке S_{train}; m_l - число объектов S_train в подобласти q_l, m_l^0 - число объектов класса K_0 в подобласти q_l. $\nu_l^0 = \frac{m_l^0}{m_l}$ -

доля объектов класса K_0 в подобласти q_l.

5.2 Теорема 1

Пусть $\nu_0 \in (0,1)$, тогда $F_\chi(R, S_{train}, K_0) \in [0, m]$ при любых $S_{train}, q_1, \ldots, q_L$.

Доказательство Теоремы 1:

1) $(\nu_l^0 - \nu^0)^2 m_l \geq 0$; $\frac{1}{\nu^0(1-\nu^0)} > 0$, следовательно $F_p(R, S_{train}, K_0) \geq 0$. Ноль достигается при $\nu_1^0 = \cdots = \nu_L^0 = \nu^0$.

2) Докажем, что $F_\chi(R, S_{train}, K_0) \leq m$ Предположим, что существуют такие $q_1, \ldots, q_L$, что $F_\chi(R, S_{train}, K_0) > m$

$$\frac{1}{\nu^0(1-\nu^0)} \sum_{l=1}^{L} (\nu_l^0 - \nu^0)^2 m_l > m$$

$$\sum_{l=1}^{L} (\nu_l^0 - \nu^0)^2 m_l > m\nu^0(1-\nu^0)$$

$$(\nu^0)^2 \sum_{l=1}^{L} m_l - 2\nu^0 \sum_{l=1}^{L} \nu_l^0 m_l + \sum_{l=1}^{L} (\nu_l^0)^2 m_l > \nu^0 m - (\nu^0)^2 m$$

$$(\nu^0)^2 m - 2\nu^0 m \frac{\sum\limits_{l=1}^{L} \frac{m_l^0}{m_l} m_l}{m} + \sum_{l=1}^{L} (\nu_l^0)^2 m_l > \nu^0 m - (\nu^0)^2 m$$

$$(\nu^0)^2 m - 2\nu^0 m \frac{\sum\limits_{l=1}^{L} m_l^0}{m} + \sum_{l=1}^{L} (\nu_l^0)^2 m_l > \nu^0 m - (\nu^0)^2 m$$

$$(\nu^0)^2 m - 2\nu^0 m \frac{m^0}{m} + \sum_{l=1}^{L} (\nu_l^0)^2 m_l > \nu^0 m - (\nu^0)^2 m$$

$$(\nu^0)^2 m - 2(\nu^0)^2 m + \sum_{l=1}^{L} (\nu_l^0)^2 m_l > \nu^0 m - (\nu^0)^2 m$$

$$\sum_{l=1}^{L} (\nu_l^0)^2 m_l - (\nu^0)^2 m > \nu^0 m - (\nu^0)^2 m$$

$$\sum_{l=1}^{L} (\nu_l^0)^2 m_l > \nu^0 m$$

$$\sum_{l=1}^{L} \nu_l^0 \frac{m_l^0}{m_l} m_l > \frac{m^0}{m} m$$

$$\sum_{l=1}^{L} \nu_l^0 m_l^0 > m^0$$

$$\sum_{l=1}^{L} \nu_l^0 m_l^0 > \sum_{l=1}^{L} 1 m_l^0$$

так как $\nu_l^0 \leq 1$, пришли к противоречию. Заметим, что равенство $F_\chi(R, S_{train}, K_0) = m$ достигается тогда и только тогда, когда $\nu_l^0 \in \{0, 1\}$, то есть когда ни в одной из подобластей нет объектов S_{train} разных классов.

Вывод Вместо $F_\chi(R, S_{train}, K_0)$ лучше использовать $F_{\chi normalised}(R, S_{train}, K_0) = \frac{F_p(R, S_{train}, K_0)}{m}$. В таком случае минимальное значение функционала, использующееся для верификации разбиений не будет зависеть от размера выборки.

5.3 Отбор отличающихся разбиений

В главе 2 было показано, что выпуклая комбинация даёт больший прирост качества, если в неё входят алгоритмы с высоким отклонением прогнозов друг от друга. В методах СВС, МСВС, ОДР разнообразие разбиений достигается добавлением единственного синдрома от признака (пары признаков). В работе реализован метод, позволяющий строить по два синдрома каждого типа от каждого признака (пары признаков). Введём понятия расстояния между разбиениями

$$\rho_{ij}^R = \frac{1}{l} \sum_{i=1}^{l} (\nu_{il}^0 - \nu_{jl}^0) : 2 \tag{6}$$

5.4 Теорема 2

$\rho_{ij}^R \in [0, 1)$ *Доказательство Теоремы 2:*

1. Докажем, что $\rho_{ij}^R \in [0, 1]$. Разность двух чисел из отрезка $[0, 1]$ принадлежит промежутку $[-1, 1]$, следовательно квадрат разности чисел из отрезка $[0, 1]$ также принадлежит $[0, 1]$, а значит и $\rho_{ij}^R \in [0, 1]$

2. Докажем, что $\rho_{ij}^R \neq 1$. Проведём доказательство от противного. Пусть $\rho_{ij}^R = 1$. Тогда для любого l верно, что $|\nu_{il}^0 - \nu_{jl}^0| = 1$. Такое возможно в двух случаях: $\nu_{il}^0 = 1, \nu_{jl}^0 = 0$ и $\nu_{il}^0 = 0, \nu_{jl}^0 = 1$. Для одномерных достаточно рассмотреть объекты с минимальным значением признака. Очевидно, что этот объект всегда будет находиться в области $X_i < b_i'$, следовательно для этой области $|\nu_{il}^0 - \nu_{jl}^0| < 1$. Для синдромов третьего типа достаточно провести аналогичные рассуждения для левого нижнего объекта выборки. Пришли к противоречию.

Таким образом, можно ввести обобщённый функционал качества модели:

$$F_{model}(R_i, R_j, S_{train}, K_0) = \alpha F_{\chi normalised}(R_j, S_{train}, K_0) + (1 - \alpha)\rho_{ij}^R, \quad (7)$$

Где R_i - отобранное и верифицированное при помощи функционала $F_{\chi normalised}$ разбиение. R_j - разбиение того же типа, что и R_j для тех же признаков. Даный функционал позволяет получать дополнительную информацию из некоторых признаков (пар признаков).

5.5 Быстрый СВС с дополнительными разбиениями

В рамках работы был реализован метод "быстрый СВС с дополнительными разбиениями". В методе используются синдромы I, II, и III типа. Синдромы I и II типа строятся с использованием функционала $F_{\chi normalised}$

Построение разбиений В методе используются синдромы I, II, и III типа. Синдромы I и II типа строятся с использованием функционала $F_{\chi normalised}$ по полной сетке. Для каждой пары признаков строится по два синдрома III типа. Синдромы строятся по разреженной сетке.

Верификация разбиений Синдромы первого и второго типа отсеиваются по порогу $F_{\chi normalised}$, первый синдром третьего типа максимизирует $F_{\chi normalised}$ и отсеивается по порогу $F_{\chi normalised}$, второй синдром третьего типа максимизирует функционал F_{model} и также строится по разреженной сетке. Второй синдром третьего типа также отсеивается по порогу $F_{\chi normalised}$.

Построение выпуклой комбинации Построение выпуклой комбинации и процедура голосование не отличаются от стандартного СВС.

6 Поиск выпадающих наблюдений

Под выпадающим объектом (ВО) обычно понимается объект, описание которого заметно отклоняется от основной закономерности в данных. При этом отклонение должно быть настолько значительным, чтобы оно не могло быть объяснено простой случайностью, и требовало бы дополнительных предположений о механизме возникновения объекта. Данное определение фактически соответствует определению, приведённому в книге [11]. В статье [10] ВО определяют как объект, сильно отличающийся от остальной выборки по некоторой метрике. ВО 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 и ошибки в определении класса объектов или ошибки в значениях признаков), ошибки при занесении информации в базу и др. Такие ВО оказывают существенное влияние на точность прогнозирования. Так при построении прогностических моделей с помощью методов машинного обучения ВО, связанные с ошибками, могут заметно повысить неустойчивость обучения и заметно снизить обобщающую способность полученной модели. Естественно что такие ВО следует удалить из обучающей выборки после их идентификации. Следует отметить, что экспертная оценка правильности часто противоречивых экспериментальных данных разных исследователей остается наиболее сложной, длительной и плохо формализуемой задачей, поэтому необходима автоматизация поиска ВО, которая позволяет выявить потенциально ошибочные наблюдения и после экспертной оценки сделать соответствующие исправления. В связи с этим проблема ВО продолжает привлекать внимание исследователей, в том числе при решении задач выявления связи свойств соединений со свойствами их компонентов [12]. В настоящее время существует достаточно большое число подходов для идентификации ВО. В их число входят одномерные статистические тесты, оценивающие возможность интерпретации экстремальных значений переменных как ВО. В качестве примера можно привести критерии 3σ, Диксона, Граббса [14]. В многомерных данных значения каждой из переменных для ВО могут оказаться не экстремальными, что не позволяет выявлять такое ВО с помощью одномерных тестов. Для обнаружения подобных ВО может быть использован подход, основанный на вычислении скалярной меры отклонения отдельного наблюдения x_i от всего массива данных $X = \{x_1, ..., x_N\}$. Для каждого объекта $x_i \in X$ вычисляется метрика отклонения $\rho(x_i, X)$, далее проводится ранжирование объектов по величине ρ, и объекты с максимальным отклонением рассматриваются как предполагаемые ВО. В качестве меры близости может выступать, например, расстояние Махаланобиса или его робастные модификации [16]. Другим популярным подходом является поиск объектов с малым числом соседей в некоторой его окрестности. Если используемая метрика основана на суммировании различий по всем признакам, а число признаков очень велико, то все объекты выборки могут оказаться практически одинаково далеки друг от друга. Поэтому упомянутые методы малоприменимы для данных с признаковыми описаниями большой размерности. В

рамках задачи восстановления зависимости переменной y от переменных $x_1, \dots x_p$ под выпадающим объектом может пониматься объект $s = (x, y)$, для которого существует существенное отклонение оценки от ожидаемого значения y в точке x. Для того, чтобы выявить ВО в указанном выше смысле, необходимо построить по обучающей выборке модель, связывающую ожидаемое значение y с вектором $x = x_1, \dots, x_p$. Например, это может быть модель:

$$y = \hat{y}(y, \beta) + \epsilon \tag{8}$$

Поиск вектора коэффициентов β может производиться с помощью метода наименьших квадратов по обучающей выборке $S_train = \{(x_j, y_j)|j = 1, \dots, m\}$. Однако, более робастные оценки могут быть получены с помощью методов LMS или LTS [15]. Для оценки отклонения произвольного объекта $s_j = (x_j, y_j)$ из обучающей выборки от зависимости (1) может быть использован, например, индекс удалённых остатков:

$$K_j^r = \frac{y_j - \hat{y}[\beta(S \backslash s_j), x_j]}{\hat{\sigma_y}}, \tag{9}$$

где $S \backslash s_j$- вектор коэффициентов модели (1), рассчитанный по выборке 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 [13], показывающее насколько s_j искажает регрессионную модель.

$$D_i = \frac{\sum_{j=1}^{m} \Big(\hat{y}\big(\beta(S, x_j)\big) - \hat{y}\big(\beta(S \backslash s_j, x_j)\big) \Big)^2}{p\hat{\sigma_y}} \tag{10}$$

Перечисленные индексы, очевидно, являются количественными оценками того, что объект s_j является выпадающим наблюдением. Они позволяют ранжировать объекты по степени их отклонения от закономерности, описываемой моделью (1). В рамках задачи распознавания поиск ВО основан на идее измерения расстояния от объекта до его соседей из своего . Предполагается, что в окрестностях типичных объектов, то есть не являющихся ВО, преобладают объекты из того же класса, которому принадлежит сам объект. ВО же, напротив, находятся вдалеке от объектов своего класса. Поэтому объект считается ВО, если сумма расстояний до K ближайших соседей больше некоторого значения или если в некоторой окрестности объекта количество соседей меньше определённого значения. На этой идее основано большое число алгоритмов. [18] Если используемая метрика основана на суммировании различий по всем признакам, а число признаков очень велико, то все точки выборки

могут оказаться практически одинаково далеки друг от друга, что делает Поэтому методы, основанные исключительно на вычислении расстояний, малоприменимы для данных с признаковыми описаниями большой размерности. В пространствах большой размерности углы между векторами вычисляются стабильнее, чем расстояние между точками. Метод Angle-Based Outlier Detection (ABOD) [17] строится на простом предположении - если объект находится в центре класса, он окружён соседями. ВО находится в стороне от своих соседей. Для каждого объекта вычисляется величина:

$$ABOD(x) = \frac{1}{|ne(x)|(ne(x)-1)-1} \sum_{x',x''\in ne(x),x'\neq x''} \frac{< x'-x, x''-x >}{||x'-x||^2||x''-x||^2} -$$

$$-\frac{1}{|ne(x)|(ne(x)-1)} \sum_{x',x''\in ne(x),x'\neq x''} \frac{< x'-x, x''-x >}{||x'-x||^2||x''-x||^2},$$

Здесь $ne(x)$ - множество соседей объекта x. $ABOD$ - оценка дисперсии углов между соседями объекта x. Если оценка дисперсии мала, объект лежит в стороне от своих соседей и предположительно является ВО.

7 Метод, основанный на комбинировании параметров неустойчивости и величин отступов

При решении задачи распознавания естественно считать выпадающим объект из класса K_i с описанием x, если для него велика разность $max_{j=1,\dots,L}P(K_j|x) - P(K_i|x)$. На практике количественной оценкой того, что объект с описанием является выпадающим наблюдением, очевидно является $\lambda(K_i,x) = max_{j=1,\dots,L}P(K_j|x,S) - P(K_i|x,S)$, где $P(K_j|x,S)$ - оценка вероятности принадлежности объекта с описанием x классу K_j. Однако, оценки вероятностей принадлежности классам напрямую используют только статистические методы распознавания. В общем случае в этих целях могут быть использованы величины $\Gamma(K_i,x) = max_{j=1,\dots,L}\gamma_j(x,S) - \gamma_i(x,S)$, где $\gamma_j(x,S)$ - оценка принадлежности объекта с описанием x классу K_j которая рассчитана алгоритмом, обученным по обучающей выборке. Некоторые методы отбора ВО основаны на ранжировании объектов по отступу. Отступом объекта $(x_i,y_i) \in S$ относительно алгоритма классификации, имеющего вид $a(x) = argmax_{y\in Y}\gamma^y(x)$, называется величина

$$M(s) = \gamma^{y_i(x_i)} - max_{y\in Y\setminus y_i}\gamma^y(x_i), \tag{11}$$

где $\gamma^y(x_i)$ - оценка принадлежности объекта x_i к классу y. Отступ показывает степень типичности объекта. Отступ отрицателен тогда и только тогда, когда алгоритм допускает ошибку на данном объекте. В зависимости от значений отступа обучающие объекты условно делятся на пять 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. Фактически, они не добавляют к эталонам никакой новой информации. Наличие неинформативных объектов характерно для выборок избыточно большого объёма. 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. Ошибочные объекты имеют небольшие отрицательные отступы и близки к пограничным. 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 являются именно шумовыми объектами. Для шумовых объектов характерна 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 ошибочные или пограничные объекты. Информации о величинах отступа нередко оказывается недостаточно для достоверной идентификации таких ВО. Нами предлагается подход для поиска ВО, при решении задач распознавания принадлежности объектов некоторых классов $K_1, \ldots, K_L$ по признакам $x_1, \ldots, x_p$, основанный на ранжировании объектов согласно комбинированной оценке, учитывающий как величину отступа, так и величину вносимых искажений. Количественной оценкой того, что объект с описанием x является выпадающим, естественно считать аналог упомянутого выше расстояния Кука, используемого для описания неустойчивости линейной

регрессионной модели:

$$\delta_i = \frac{\sum\limits_{j=1}^{m}\sum\limits_{l=1}^{L}\big(\gamma_l(x_j, S) - \gamma_l(x_j, S\backslash s_i)\big)^2}{Lm} \tag{12}$$

Коэффициенты $\Gamma(K_i, x)$ и γ_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 является эффективность распознавания, оцениваемая одной из стандартных метрик. Наиболее полно эффективность распознавания характеризуется с помощью AUC - площади под ROC кривой. Будем считать, что оценка объекта за 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 за класс на полной выборке
   (a) Обучим классификатор C_0 на обучающей выборке S
   (b) Применим C_0 к X. Получим оценки вероятностей принадлежности объектов к классам $\gamma_1(x_1, S), \ldots, \gamma_L(x_1, S), \ldots, \gamma_1(x_m, S), \ldots, \gamma_L(x_m, S)$
2. Для каждого объекта выборки оценим, является ли он ВО. Для этого:
   (a) Построим выборку S_i, исключив из S пару (x_i, y_i)
   (b) Обучим классификатор C_i на выборке S_i
   (c) Применим C_i к X. Получим оценки $\gamma_1(x_1, S\backslash s_i), \ldots, \gamma_L(x_1, S\backslash s_i), \ldots, \gamma_1(x_m, S\backslash s_i), \ldots$
   (d) Вычислим $\delta_i = \frac{\sum\limits_{j=1}^{m}\sum\limits_{l=1}^{L}\left(\gamma_l(x_j,S)-\gamma_l(x_j,S\backslash s_i)\right)^2}{Lm}$ Величина δ_i показывает насколько изменились оценки объектов после исключения x_i из обучающей выборки.
   (e) Вычислим $\omega(x_i) = \frac{\sum\limits_{l\in\{1,\ldots,L\}\backslash y_i}\gamma_l(x_i,S\backslash s_i}{Lm}$ Величина $\omega(x_i)$ - оценка x_i за противоположный класс.
3. Отберём ВО исходя из оценок $\overline{\delta}$ и $\overline{\omega}$. Для каждого объекта вычислим $E(x_i, a_1, a_2, p) = (a_1|\delta_i|^p + a_2|\omega(x_i)|^p)^{\frac{1}{p}}$ Предположим, что чем больше $E(x_i, a_1, a_2, p)$, тем скорее

объект x_i является ВО. Заметим, что $E(x_i, a_1, a_2, p)$- является модулем объекта в пространстве (δ, ω) со взвешенной метрикой Минковского $\rho^*(obj, obj') = \left(\sum_i (a_i|obj_i - obj'_i|)^p\right)^{\frac{1}{p}}$

4. Отсортируем объекты по убыванию E.

5. Пусть (X^*, y^*) - подвыборка из k объектов с наибольшими E. Исключим эти объекты из обучения.

8 Вычислительные эксперименты с методом СВС

В ходе вычислительных экспериментов сравнивались следующие методы

- SVM с RBF ядром
- Random Forest
- СВС с синдромами I, II, III, максимизирующими $F_{\chi normalised}$, на полной сетке
- СВС с синдромами I, II, III, максимизирующими $F_{\chi normalised}$, на разреженной сетке
- Быстрый свс с дополнительными разбиениями

Качество измерялось площадью под ROC кривой.

8.1 КардиоКВАРК

Задача - выявить пациентов с повышенным давлением. Обучающая выборка - 832 объекта, 114 вещественнозначных признаков. Соотношение классов 594 к 238. Тестовая выборка - 495 объектов. Соотношение классов 405 к 90.

Обучение - 10-fold CV на тренировочной выборке. Результат замерялся на тестовой выборке.

Алгоритм	Максимальный AUC ROC
SVM с RBF ядром:	0.731
Random Forest:	0.778
свс на полной сетке:	0.770
свс на разреженной сетке (шаг сетки 8):	0.773
свс с дополнительными разбиениями (шаг сетки 8):	0.785

8.2 Смертность пациентов с тяжёлыми формами туберкулёза

Задача - по анализам пациента спрогнозировать ремиссию. Выборка 246 объектов, 34 признака.

Обучение и контроль качества - контроль по отдельным объектам(LOO)

Алгоритм	Максимальный AUC ROC
SVM с RBF ядром:	0.94
Random Forest:	0.96
свс на полной сетке:	0.96
свс на разреженной сетке (шаг сетки 8):	0.95
свс с дополнительными разбиениями (шаг сетки 8):	0.97

8.3 Анализ результатов метода свс

Построение разбиений на разреженной сетке приводит к значительному ускорению работы метода, но требует подбора шага сетки и может привести к нестабильной работе метода. Наиболее качественные композиции строятся при более высоком пороге функционала.

Не имеет смысла тщательным образом подбирать шаг сетки и порог качества. Метод СВС достаточно робастный.

Использование дополнительных разбиений приводит к повышению качества ансамбля. Построение второго разбиение требует лишь затрат памяти на параметров синдромов для одной пары признаков и дополнительного вычисления ρ^R для качественных синдромов, что не представляет особой трудности.

9 Эксперимент по отбору ВО

Вычислительный эксперимент реализован на языке программирования python версии 3.5 [8] с использованием библиотеки scikit-learn[9]. Данные. В работе решается задача выявления ВО при прогнозе возможности образования соединений состава A+3B+3C+2O4. Выборка состоит из 758 объектов двух классов. К первому классу принадлежали 695 объектов (существующие соединения), ко второму - 63 (химические системы A-B-C-O без образования соединения вышеуказанного состава). Каждый объект описывают 108 непрерывных признаков. Пропусков в данных нет. В выборке содержатся объекты с неверной меткой класса. Требуется:

1. Обнаружить объекты с ошибочными метками

2. Верно классифицировать объекты

Методы распознавания, используемые в исследовании. Для решения задачи классификации использовался градиентный бустинг над решающими деревьями [7]. Градиентный бустинг был выбран после сравнения с другими популярными алгоритмами классификации: решающие деревья (DT), машины опорных векторов (SVM), метод ближайших соседей (KNN) [5]. Качество измерялось на полной выборке при помощи десятифолдовой кроссвалидации с сохранением долевого содержания классов в выборках. Исходные (до удаления ВО) для разных методов результаты распознавания представлены в таблице 1.

Таблица 1

	GB	KNN	SVM	DT
ROC AUC[6]	0.82	0.85	0.77	0.84

Градиентный бустинг показывает относительно неплохие результаты по сравнению с другими алгоритмами и имеет большую обобщающую способность. 3.3 Связь параметров неустойчивости и величин отступов Дополнительным аргументом комбинирования величин неустойчивости и отступа явились результаты исследования их взаимосвязи, представленные на рисунке 1. Из рисунка видно, что величины δ и ω не

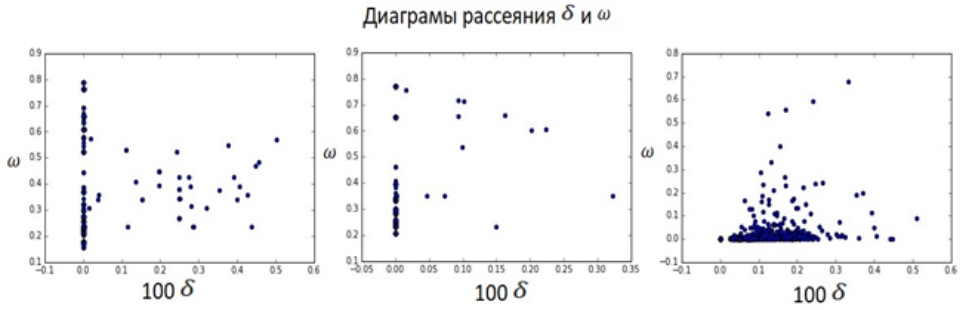

Рис. 3: Зависимость между величинами отступа ω и неустойчивости δ .

коррелируют между собой. Низкая корреляция оценок δ и ω свидетельствует о том, что они могут являться независимыми индикаторами ВО. Влияние числа исключённых объектов на качество классификации

В работе изучено, как на качество классификации влияют параметры взвешенной метрики Минковского a_1, a_2, p и количество исключённых объектов k. Для оценки качества оптимизации процедуры отбора ВО и оценки влияния отбора на точность классификации использовался 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 $\frac{\alpha_1}{\alpha_2}$ 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 заново. Результаты распознавания оценивались на соответствующих контрольных выборках. Эксперименты проводились с числом различным блоков скользящего контроля (см. таблицу 2).

Таблица 2

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

При выборе p достаточно перебрать числа от одного до пяти, соотношение $\frac{\alpha_1}{\alpha_2}$ выбирать сложнее. Эмпирически установлено, что качество распознавания максимально, когда отношение $\frac{med_{x \in X} \delta(x) a_1}{med_{x \in X} \omega(x) a_2} \in [1, 2]$, где med - 0.5-квантиль. Таким образом, число выполняемых циклов внутреннего скользящего контроля составило $N_in * N_out * N_p * N_{(\frac{\alpha_1}{\alpha_2})} * k_max$, где N_p - число перебираемых степенных показателей, $N_{(\frac{\alpha_1}{\alpha_2})}$ - число перебираемых соотношений $\frac{\alpha_1}{\alpha_2}$, k_max – предполагаемое максимальное число ВО. Заметим, что подбор параметров градиентного бустинга занимает гораздо больше времени, чем вычисление значения метрики Минковского в двухмерном пространстве для каждого объекта выборки. Двадцать один исключённый из выборки объект, давший наибольший прирост качества классификации, был проанализирован экспертами. Из них восемь объектов имели ошибочную метку первого класса, один объект имел ошибочную метку второго класса, о принадлежности одного объекта не было найдено достоверных данных. Метод поиска ВО с использованием параметров неустойчивости обучения сравнивался с другими популярными методами поиска ВО. Первый

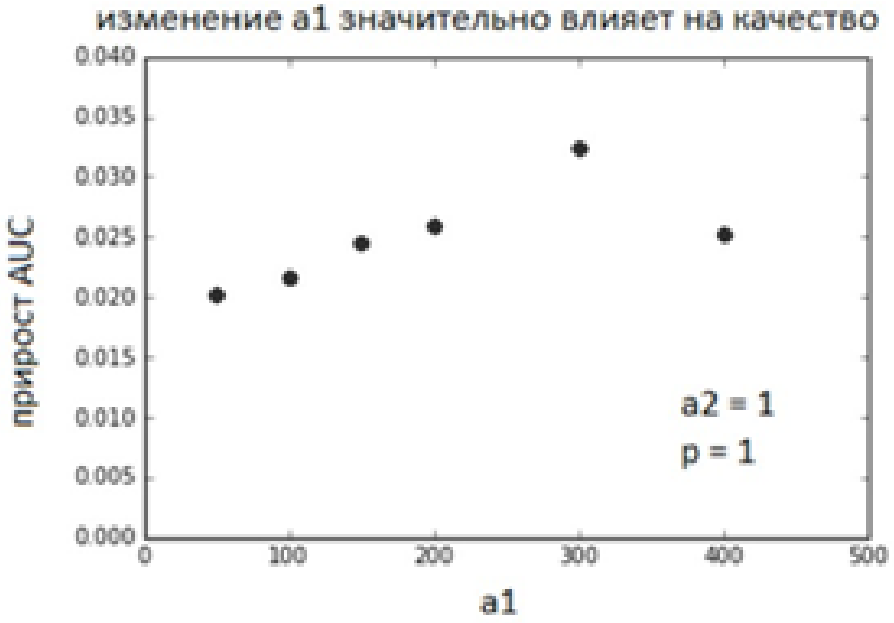

Рис. 4: График зависимости прироста AUC от a_1. Виден отчётливый экстремум при $a_1 = 300$.

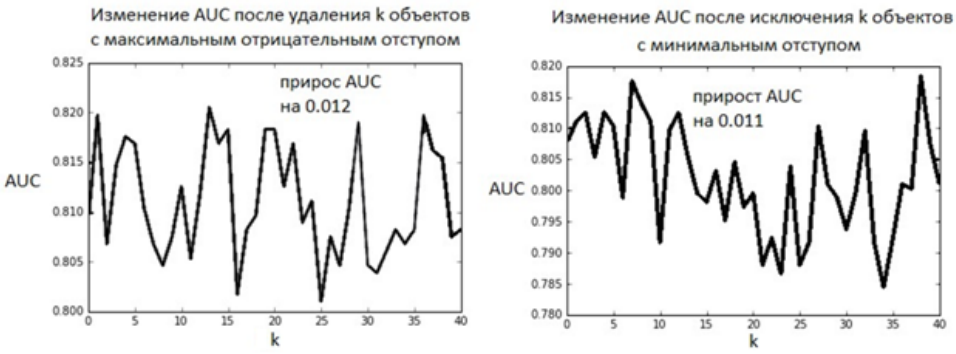

Рис. 5: AUC ROC после удаления k объектов с наименьшим по модулю отступом и с наибольшим по модулю отрицательным отступом. Вычисление AUC производилось в режиме скользящего контроля учётом исключённых объектов. Эксперимент №3. Из рисунка видно, что зависимость является неустойчивой и выраженная тенденция изменения AUC отсутствует.

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 рисунке

2.

Метод поиска ВО с использованием параметров неустойчивости обучения показывает более высокие результаты при различных гипаерпараметрах. Описанный в

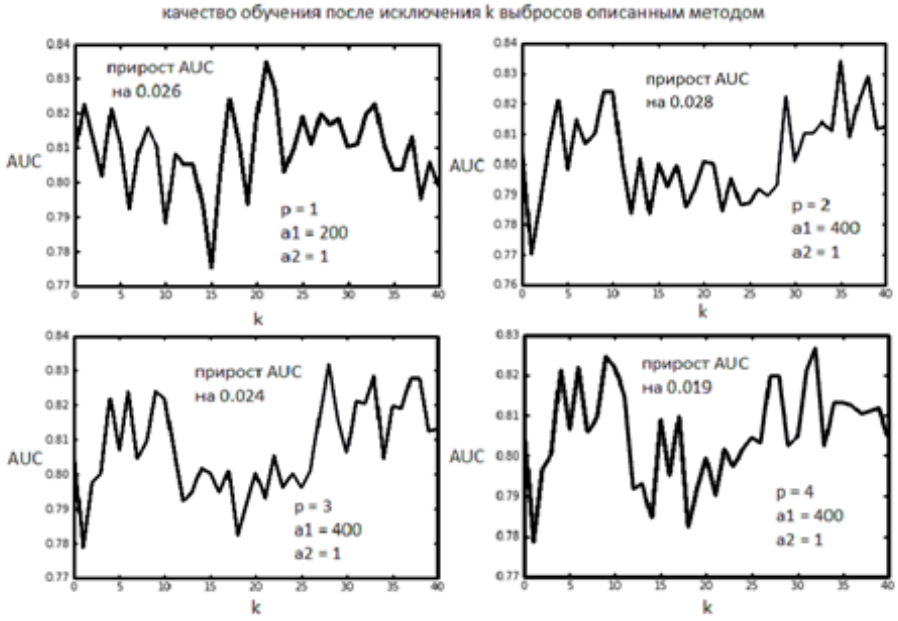

Рис. 6: На рисунке 3 изображена зависимость величины AUC от числа исключённых по порядку ранжирования объектов обучающей выборки по комбинированным оценкам, включающим отступ и неустойчивость. Из рисунка 3 видно, что несмотря на сохранение неустойчивости, появляется выраженная тенденция повышения точности при исключении объектов по порядку ражирования.

работе алгоритм лучше повышает AUC ROC, однако имеет большую вычислительную сложность. Из рисунка 3 видно, что оптимальные результаты достигнуты при $p = 2, \frac{\alpha_1}{\alpha_2} = 0.0025$ и при и $p = 3, \frac{\alpha_1}{\alpha_2} = 0.0025$ При $p = 4, \frac{\alpha_1}{\alpha_2} = 0.0025$ прирост AUC ROC ниже, чем в других случаях. При $p = 1, \frac{\alpha_1}{\alpha_2} = 0.005$ нет выраженной тенденции повышения точности при исключении объектов по порядку ранжирования. Кроме k необходимо подобрать p и верное соотношение $\frac{\alpha_1}{\alpha_2}$. Основную сложность представляет подбор соотношения $\frac{\alpha_1}{\alpha_2}$, и числа исключённых объектов k. Поскольку мы не можем делать никаких предположений о зависимости параметров и качества обнаружения ВО, оптимальные значения параметров метода приходится подбирать при помощи процедуры скользящего контроля.

10 Анализ результата отбора ВО

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, этот метод требует довольно тщательного подбора гиперпараметров, что создаёт сложности для применения 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.

11 Заключение

В ходе работы:

- Была разработана модификация метода свс, работающая быстрее оригинального метода и не уступающая по качеству как оригинальному методу, так и другим известным методам машинного обучения.
- Была доказана Теорема 1, дающая обоснование функционала F_{model}
- Был разработан алгоритм отбора ВО
- Алгоритм отбора ВО позболил обнаружить значительное число ошибок в данных.

Список литературы

[1] *Senko O., Kuznetsova A.* A recognition method based on collective decision making using systems of regularities of various types. Pattern Recognition and Image Analysis. 2010. V. 20. № 2. P. 152–162.

[2] *Senko O.V., Kuznetsova A.V.* The Optimal Valid Partitioning Procedures, «InterStat», Statistics in Internet. URL: http://interstat.statjournals.net/YEAR/2006/articles/0604002.pdf (дата обращения: 18.02.2013). Математические модели в экономике и биологии. Материалы научного семинара. Планерное. Московская обл. МАКС Пресс, 2003, с. 57–67.

[3] *Leo Breiman* Bagging predictors. Machine Learning, 24(2):123–140, 1996.

[4] *А. В. Кузнецова, И. В. Костомарова, Н. Н. Водолагина, Н. А. Малыгина, О. В. Сенько* Изучение влияния клинико-генетических факторов на течение дисциркуляторной энцефалопатии с использованием методов распознавания, Матем. биология и биоинформ., 2011, том 6, выпуск 1, 115–146

[5] *К. В. Воронцов* Математические методы обучения по прецедентам (теория обучения машин). Москва, 2011.

[6] (1993) «Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine». Clinical Chemistry 39 (8): 561–577. PMID 8472349

[7] *Friedman J. Greedy* Function Approximation: A Gradient Boosting Machine. — IMS 1999 Reitz Lecture.

[8] https://www.python.org/downloads/release/python-350/

[9] https://pypi.python.org/pypi/scikit-learn/

[10] *Aggarwal C C and Yu P S* Outlier detection for high dimensional data In Proc ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX.

[11] *C. C. Aggarwal* Outlier analysis. Springer-Verlag New York, 2013