Инженерный вестник Дона, №11 (2021)
ivdon.ru/ru/magazine/archive/n11y2021/7279

Использование критерия среднего расстояния для выявления новизны в данных

А.М. Крашенинников

Московский государственный университет пищевых производств, Москва

Аннотация: В статье рассмотрены особенности выявления новизны в данных, а также общие методы ее выявления. Поскольку отсутствие шума в обучающей информации является определяющим фактором для построения на ней качественных классификаторов в машинном обучении с учителем, то рассмотрен такой практически важный частный случай поиска новизны, когда она определяется в отдельных классах обучающих данных после того, как в этих данных устранены все выбросы. Для большей определенности при поиске новизны предложена ее геометрическая интерпретация в пространстве значений признаков объектов в виде изображающей объект класса точки, находящейся снаружи минимальной гиперокружности, описанной вокруг остальных изображающих точек объектов класса. Степень удаленности (уровень новизны) p точки относительно всех остальных выражена через радиус и центр данной гиперокружности. С целью качественной оценки уровня новизны объектов класса рассмотрены три ее градации: ближняя, средняя и дальняя. Поскольку прямое использование геометрической интерпретации при поиске новизны требует большого числа вычислений, то для практических расчетов предложен непараметрический критерий статистического характера F, названный критерием средних расстояний. Зависимость статистического критерия F от геометрической характеристики новизны p определена на специальных множествах точек - n–мерных кубах в пространстве с метрикой ”манхеттенское расстояние”. При анализе данной модели определена верхняя оценка значений F(p) для дополнительных точек, находящихся внутри куба (при $0 \le p \le 1$), и найден вид критерия для внешних по отношению к кубу точек (p > 1). По данным зависимостям найдены более общие формулы зависимости F(p) для классов произвольного вида с параметрами {n, Ncl}. Из данных формул получены значения критерия F для ближней, средней и дальней новизны. При локальном удалении новизны в классах обучающих данных рассмотрены однократный и итерационный подходы. В первом случае новизна заданного уровня определяется и удаляется из класса A только один раз. Во втором случае – итерационным образом, до получения кластера. Разработана общая структура программного обеспечения, реализующего оба метода, и алгоритмы функций.
Ключевые слова: обучающие данные, классификатор, выбросы, новизна, обнаружение новизны, геометрический подход, статистический критерий.

Введение

Для идентификации объектов в системах искусственного интеллекта их задают при помощи значений существенных свойств, которые представляют в виде вектора $\{\bar{x}\} = \{x_1,x_2,\ldots,x_n\}$. Это дает возможность взаимно-однозначно сопоставить каждому объекту с соответствующим ему вектором характеристик $\bar{a} = \{a_1,a_2,\ldots,a_n\}$ изображающую точку $\bar{a}$ в n-мерном

пространстве значений характеристик U. Для того, чтобы системы искусственного интеллекта могли применять к распознанным объектам соответствующие им методы анализа и воздействия, рассматриваемые объекты подразделяют на отдельные классы $\{A\} = \{A_1,A_2,\ldots,A_m\}$.

Задачу распознавания, заключающуюся в отнесении некоторого нового объекта, заданного вектором своих свойств $\bar{a} = \{a_1,a_2,\ldots,a_n\}$ к некоторому известному классу A_i из совокупности $\{A\}$, называют классификацией, а специальный алгоритм μ, применяемый для ее решения, называют классификатором или решающей функцией. Классификатор обеспечивает отображение объекта со свойствами $\bar{a} = \{a_1,a_2,\ldots,a_n\}$ на совокупность классов $\{A\}$: $\mu : (\bar{a}) \rightarrow \{A\}$.

В случае технологии обучения с учителем классификаторы строятся по обучающей выборке, которая состоит из примеров – объектов с заданными свойствами $\bar{a}^s$, для которых уже указаны соответствующие им классы n^s: $TE = \{ te^s \} = \{ (\bar{a}^s, n^s)\}$.

Для построения классификаторов применяются: нейронные сети, методы геометрические с использованием метрик в пространстве значений признаков, статистические подходы, методы регрессионного анализа (линейный, логистический, полиномиальный, метод опорных векторов, деревья решений, случайный лес, ридж-регрессия, лассо-регрессия), метод kближайших соседей.

Для построенного классификатора μ качество оценивается долей правильно отображенных на множествах классов $\{A\}$ новых предъявленных объектов. С этой целью обычно применяется специальный контрольный набор примеров, называемый тестовыми данными.

Общее качество классификатора при обучении с учителем в основном определяется применяемыми при его построении обучающими данными. Основным требованием к ним является следующее: примеры из одного

класса должны образовывать в пространстве признаков кластер, т.е. плотно сгруппированную совокупность точек, которая расположена отдельно от других таких же совокупностей точек, задающих объекты других классов. Данное требование называется гипотезой компактности [1]. По ней образы классов в пространстве признаков U должны составлять кластеры. Выполнение гипотезы компактности гарантирует хорошую отделимость их в пространстве U при помощи специальных гиперповерхностей.

В обучающих данных отклонения от гипотезы компактности (шум) обычно возникают из-за целого ряда ошибок, возникающих при их формировании [2]. Нарушение шумом условия компактности классов в обучающих данных ухудшает условия для построения классификатора, а также его качество - способность правильно классифицировать новые предъявляемые объекты. В частности, те нейросети, которые были построены на зашумленных данных, при последующей классификации новых объектов повторяют те же ошибки, которые содержались в их обучающих данных [3,4].

Основную долю шума в обучающих данных составляют выбросы и новизна. Выбросами называют такие примеры, для которых по тем или иным причинам неправильно заданы их классы, т.е. свойства объекта не характерны для того класса, к которому он причислен. Новизной называют примеры из обучающей выборки, у которых изображающие точки в пространстве U находятся далеко от кластеров всех классов.

Удаление выбросов и новизны при коррекции данных в общем случае может выполняться и одновременно и по отдельности. Для обучающих данных с точки сокращения объема вычислений оптимальный способ коррекции заключается в удалении всех выбросов и последующем удалении новизны.

Методы, применяемые для идентификации новизны, сходны с методами классификации и определения выбросов. Основные группы методов следующие.

1. Генерирование правил. Данные методы генерируют правила, по которым для объекта подтверждается нормальное поведение объекта либо его отклонение от нормы, при котором он является новизной [5,6,7].

2. Нейронные сети и их разновидности (многослойные персептроны; самоорганизующиеся карты (SOM); сети, основанные на привыкании; нейронные деревья; автоассоциативные сети; сети теории адаптивного резонанса (ART); сети радиальных базисных функций (RBF); сети Хопфилда; колебательные сети; байесовские сети). Сеть предварительно обучают на обычных объектах. Новизну определяют по реакции на них нейронной сети [8,9,10].

3. Статистический подход (параметрический и непараметрический), основанный на применении статистических критериев [9-13].

4. Машины опорных векторов (SVM) Подход предполагает, что обычные (нормальные) точки находятся в области с высокой их плотностью, их можно окружить сферой малого радиуса [8]. Отличающиеся данные лежат в областях с низкой плотностью [14,15].

5. Подходы на основе метода ближайшего соседа. По нему предполагается, что нормальные точки имеют близких соседей, а отличающиеся от них расположены удаленно от других точек. Точка считается новизной, если ее расстояние до k ближайших соседей превышает заданное пороговое значение, или на основании расчета фактора локального выброса (LOF) [16,17].

Исследование новизны в данных, как объектов, существенно отличающихся от уже известных, представляет собой самостоятельную

ценность в биологии, экологии, медицине, социологии, информатике и других областях науки и практики [18-20].

Исходный класс обучающей выборки с исправленными выбросами после выявления и последующего удаления в нем новизны представляет собой кластер. Данный случай является одним из возможных вариантов задачи выделения кластеров в наборах данных. Поэтому общие методы выявления новизны во многом сходны с методами кластерного анализа [21-23].

Поскольку в основе геометрических методов определения новизны в пространстве U, как и у геометрических методов классификации [24-26], лежит пространственная интерпретация объектов, то они имеют наглядный геометрический смысл, позволяющий применять для оценки новизны соответствующие аналогии.

Геометрический подход к интерпретации уровней новизны рассмотрен в п.1. На основе идей метода опорных векторов предложена численная интерпретация понятий отсутствия новизны, а также близкой, средней и дальней новизны. Для практического определения новизны заданного уровня предложен статистический критерий, который позволяет существенно упростить методику определения новизны по сравнению с чисто геометрическими методами.

Связь введенных геометрической и статистической характеристик новизны найдена в п.2 с использованием модели в виде *n*-мерных кубов. На основе данной модели выполнен переход к общей формульной зависимости для произвольных классов, а также к пороговым значениям уровня новизны.

Для практического определения новизны в п.3 предложены два подхода – однократное удаление и итерационное. Дана структуризация необходимого программного обеспечения и алгоритм, реализующий однократное удаление.

В п.4. введено понятие кластера с фиксированным предельным уровнем близости точек. Рассмотрены кластеры первого, второго и третьего типов, соответствующие уровням близкой, средней и дальней новизны. Предложен алгоритм для определения данных кластеров в классах обучающей выборки. Даны оценки сложности алгоритмов одноразового и итеративного выделения новизны.

1. Численная оценка новизны в классах обучающей выборки

Рассмотрим обучающую выбору *TE*, состоящую из примеров, в которой на первом этапе коррекции уже удалены или скорректированы все выбросы. Поскольку в обучающих выборках плотность объектов в разных классах примерно одинакова, то в этом случае новизна будет локализована в отдельных классах обучающих примеров $\{A\} = \{A_1, A_2, \ldots, A_m\}$. Поэтому в дальнейшем будем рассматривать только данный локальный вариант определения новизны, которая содержится в фиксированном классе.

В основу численной оценки степени удаленности выделенных объектов от всех остальных объектов класса предложено принять геометрическую интерпретацию, близкую к методу опорных векторов. Представим множество изображающих точек класса A в виде $\{C\} \cup \bar{N}$, где совокупность {C} представляет собой кластер, а точка $\bar{N}$ существенно удалена от точек из {C}. Интегрально положение точек кластера {C} в пространстве U можно задать при помощи описанной вокруг них гиперсферы минимального радиуса R, а также центра гиперсферы - некоторой точки $\bar{C}_{cl}$. При этом близость точки $\bar{N}$ ко всем оставшимся точкам {C} из множества точек A можно оценить при помощи расстояния $\rho(\bar{N}, \bar{C}_{cl})$ от нее до центра кластера $\bar{C}_{cl}$.

Очевидно, что для всех точек $\bar{a}^x$ кластера {C} (внутренних) выполняется соотношение: $\rho(\bar{a}^x, \bar{C}_{cl}) \le R$. Для новизны, как внешней точки по отношению к кластеру: $\rho(\bar{N}, \bar{C}_{cl}) > R$.

В качестве относительной меры удаленной точки $\bar{N}$ ко всем оставшимся точкам {C} класса A предложено выбрать отношение, которое назовем ***уровнем близости точки $\bar{N}$ ко всем оставшимся точкам класса*** A:

$$p(\bar{a}^x) = \rho(\bar{a}^x, \bar{C}_{cl})/R. \qquad (1)$$

При этом для внутренних точек кластера: $0 \le p(\bar{a}^x) \le 1$. Для внешних точек, которые могут представлять собой новизну, $p(\bar{N}) > 1$.

Для качественной оценки наличия новизны и степени удаленности точки $\bar{a}^x$ по отношению к кластеру {C} предложены следующие градации новизны.

1. ***Отсутствие новизны*** (точка $\bar{N}$ лежит внутри кластера $\bar{C}_{cl}$ или незначительно выходит за его пределы):

$$0 < p(\bar{N}) < 1{,}5. \qquad (2а)$$

2. ***Ближняя новизна***. Назовем ближней новизной такие точки $\bar{N}$, для которых индекс $p(\bar{N})$ изменяется в пределах от 1,5 до 2:

$$1{,}5 \le p(\bar{N}) < 2. \qquad (2б)$$

3. ***Средняя новизна***. $p(\bar{N})$ изменяется в пределах:

$$2 \le p(\bar{N}) < 2{,}5. \qquad (2в)$$

4. ***Дальняя новизна***:

$$2{,}5 \le p(\bar{N}). \qquad (2г)$$

Таким образом, пороговыми значениями для качественной оценки уровня близости $p(\bar{N})$ выбраны следующие три ее значения: $p = 1{,}5$; 2 и 2,5. Выбор конкретного значения уровня близости p определяется спецификой конкретной предметной области.

Очевидно, что для каждой проверяемой точки $\bar{N}$ построение описанной гиперсферы вокруг оставшихся точек класса с определением ее радиуса R и

центра $\bar{C}_{cl}$ требует значительного числа вычислений. Для существенного их сокращения т унификации для каждого объекта $\bar{a}^x$ класса A_f относительный уровень близости $\rho(\bar{a}^x(p), A_f \setminus \bar{a}^x)/R$ предложено косвенно оценивать при помощи отношения следующих статистических характеристик рассматриваемой точки $\bar{a}^x$ и всего класса A_f:

1) $\bar{\rho}(\bar{a}^x(p)) = \Sigma\rho(\bar{a}^x(p), \bar{a}^u)/N_{cl}$ - частное среднее выборочное значение расстояния от выделенного объекта $\bar{a}^x(p)$ до всех N_{cl} объектов $\bar{a}^u$ класса A_f (в том числе – до самого объекта $\bar{a}^x$),

2) $\bar{\rho} = \Sigma\rho(\bar{a}^v, \bar{a}^u)/N_{cl}^2$ – общее среднее выборочное расстояний между всеми парами объектов $\{\bar{a}^v, \bar{a}^u\}$ класса A_f.

Для упрощения расчетов в общем среднем выборочном значении $\bar{\rho}$ учитываются, как пары значений $\{\rho(\bar{a}^v, \bar{a}^u); \rho(\bar{a}^u, \bar{a}^v)\}$, так и значения $\rho(\bar{a}^u, \bar{a}^u) = 0$.

С использованием данной замены вместо геометрической характеристики близости $p(\bar{a}^x)$ точки класса $\bar{a}^x$ ко всем оставшимся предложено рассмотреть статистический критерий $F(\bar{a}^x(p), A_f)$ оценки положения объекта $\bar{a}^x(p)$ относительно всего анализируемого класса A_f:

$$F(\bar{a}^x(p), A_f) = \bar{\rho}(\bar{a}^x(p)) / \bar{\rho}. \qquad (3)$$

Величина $F(\bar{a}^x(p), A_f)$ названа ***критерием средних расстояний***. Данный критерий по принципу действия сходен с критерием метода k ближайших соседей [27,28] с той разницей, что в нем суммируются расстояния не до части (k), а до всех объектов рассматриваемого класса.

Таким образом, использование геометрической характеристики (уровня) новизны позволяет дать четкую ее пространственную интерпретацию и, в частности, обоснованно выбирать пороговые значения при ее поиске. В то же время, применение статистического критерия F позволяет существенно упростить поиск новизны в вычислительном плане.

Очевидно, существует монотонная зависимость критерия F($\bar{a}^x(p)$, A_f) от уровня близости p($\bar{a}^x$). Для обоснованного применения критерия требуется определить вид этой зависимости F($\bar{a}^x(p)$, A_f) от p($\bar{a}^x$), т.е. связать между собой рассмотренные геометрическую и статическую оценку новизны. При этом, в частности, возникает вопрос, какие значения критерия F соответствуют пороговым значениям p = 1,5; 2,0; 2,5 для произвольных классов, имеющих параметры {n, N_{cl} }? Для определения зависимости F от p используем специальную модель.

2. Определение связи между геометрическим уровнем новизны *p* и статистическим критерием F средних расстояний

В качестве модели гиперсферы в n–мерном пространстве, равномерно заполненной точками, принят n–мерный куб. В пространстве U введена метрика ”манхеттенское расстояние”. Связь значений критерия средних расстояний **F** и геометрического уровня близости точек p рассмотрена на этой геометрической модели.

2.1. Куб C^n_q. Суммы расстояний между его точками

Рассмотрим в n - мерном пространстве значений характеристик объектов U ($Ox_1x_2\ldots x_n$) n–мерный куб C^n_q, имеющий следующую структуру.

По каждой оси куба, начиная от начальной О, равномерно с шагом 1 расположено q точек. Начальная точка куба имеет координаты (0,0,…,0). В метрике ρ «Манхэттенское расстояние» куб C^n_q является аналогом гипершара - кластера, равномерно заполненного точками. Общее число V точек в кубе равно q^n. Начальная точка куба имеет координаты (0,0,…,0). Радиус куба R равен $n(q-1)/2$. Его центр $\bar{C}_{cl}$ имеет координаты ((q -1)/2, (q -1)/2, …,(q -1)/2).

С учетом величины радиуса куба R = $n(q-1)/2$ в качестве реперной точки $\bar{a}_{nov}(p)$, соответствующей выбранному значению p($\bar{a}^v$) некоторой точки $\bar{a}^v$ принята точка с координатами: $\bar{a}_{nov}(p)$ =(- $pn(q-1)/2$,0,…,0). Величина Δ_p = - $pn(q-1)/2$ задает сдвиг точек $\bar{a}_{nov}(p)$ относительно

начальной точки $\bar{O}$ по оси x_1. Пороговым значениям p = 1,5; 2 и 2,5 соответствуют точки с координатами:

$\bar{a}_{nov}(1.5) = (-\Delta_1, 0, \ldots, 0) = (-n(q-1)/4, 0, \ldots, 0)$,

$\bar{a}_{nov}(2.0) = (-\Delta_2, 0, \ldots, 0) = (-n(q-1)/2, 0, \ldots, 0)$,

$\bar{a}_{nov}(2.5) = (-\Delta_3, 0, \ldots, 0) = (-3n(q-1)/4, 0, \ldots, 0)$.

Для примера на рис.1 в системе координат Ox_1x_2 показан куб C^2_3 для случая n = 2, q = 3. В нем число точек: $V = 3^2 = 9$. Радиус $R = n(q-1)/2 = 2$. Центр $\bar{C}_{cl}$ имеет координаты (1, 1). Координаты реперных точек $\bar{a}_{nov1} = \bar{a}_{nov}(1.5)$, $\bar{a}_{nov2} = \bar{a}_{nov}(2.0)$, $\bar{a}_{nov3} = \bar{a}_{nov}(2.5)$ будут следующие:

$\bar{a}_{nov1} = (-n(q-1)/4; 0) = (-1; 0)$, $\bar{a}_{nov2} = (-n(q-1)/2; 0) = (-2; 0)$, $\bar{a}_{nov3}(-3n(q-1)/4; 0) = (-3; 0)$.

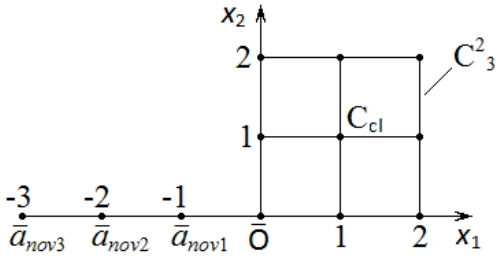

Рис. 1. Куб C^2_3 (n = 2, q = 3) с реперными точками $\bar{a}_{nov1}$, $\bar{a}_{nov2}$, $\bar{a}_{nov3}$

Для выбранных реперных точек $\bar{a}_{nov1}$, $\bar{a}_{nov2}$, $\bar{a}_{nov3}$ в рассмотренной модели в точности выполняются геометрические требуемые соотношения: $\rho(\bar{a}_{nov1}, \bar{C}_{cl})/R = 1,5$; $\rho(\bar{a}_{nov2}, \bar{C}_{cl})/R = 2$; $\rho(\bar{a}_{nov3}, \bar{C}_{cl})/R = 2,5$.

Определим связь между геометрической характеристикой (уровнем новизны p) и статистическим критерием средних расстояний F.

В множестве точек $A_f = \{C^n_q \cup \bar{a}_{nov}(p)\}$, моделирующем класс, содержащий кластер (C^n_q) и возможную новизну ($\bar{a}_{nov}(p)$), общее число элементов: $N_{cl} = V+1 = q^n+1$.

Сумма всех расстояний между точками в кубе C^n_q равна:

$\bar{\rho}(C^n_q) N_{cl}^2 = \Sigma(C^n_q)\rho(\bar{a}^v, \bar{a}^u) = nq^{2n-1}(q^2-1)/3$.

Также важной характеристикой модели является сумма всех расстояний от начальной точки $\bar{O} = (0,0,\ldots,0)$ до всех точек кластера C^n_q:

$$\bar{\rho}(\bar{O})N_{cl} = \Sigma\rho(\bar{O}, \bar{a}^u) = nq^n(q\text{-}1)/2.$$

В множестве точек $A_f = \{C^n_q \cup \bar{a}_{nov}(p)\}$ с точки зрения анализа удаленности дополнительной точки $\bar{a}_{nov}(p)$ ко всему кубу C^n_q возможны две принципиально отличные возможности.

1. $0 \le p \le 1$. Новая точка находится внутри кубу, т.е. является внутренней точкой кластера.

2. $1 < p$. Новая точка лежит за пределами куба, является внешней точкой кластера.

2.2. Анализ внутренних точек кластера

Рассмотрим в качестве дополнительной точки $\bar{a}_{nov}(p)$ внутреннюю точку кластера, для которой $0 \le p \le 1$. Сумма расстояний $\Sigma\rho(\bar{a}_{nov}(p), \bar{a}^u)$ до всех точек куба у нее не может превосходить $\Sigma\rho(\bar{O}, \bar{a}^u)$, поскольку угловая точка $\bar{O}$ является наиболее суммарно удалённой от всех точек куба.

Соответственно, для значения критерия у внутренней точки выполняется следующее неравенство:

$$F(\bar{a}_{nov}(p), A_f) = \Sigma\rho(\bar{a}_{nov}(p), \bar{a}^u) /[2\Sigma\rho(\bar{a}_{nov}(p), \bar{a}^u) + \Sigma(C^n_q)\rho(\bar{a}^v, \bar{a}^u)] \le \Sigma\rho(\bar{O}, \bar{a}^u) /[2\Sigma\rho(\bar{O}, \bar{a}^u) + \Sigma(C^n_q)\rho(\bar{a}^v, \bar{a}^u)] = F(\bar{O}, A_f).$$

Величина критерия для дополнительной точки, совпадающей с точкой $\bar{O}$, на множестве точек $A_f = \{C^n_q \cup \bar{O}\}$ равна:

$$F(\bar{O}, A_f) = [nq^n(q\text{-}1)/2]/[2nq^n(q\text{-}1)/2 + nq^{2n-1}(q^2\text{-}1)/3] = 1/[2 + (2/3)\, q^{n-1}(q+1)].$$

Учитывая, что общее число элементов в множестве $\{C^n_q \cup \bar{a}_{nov}(p)\}$: $N_{cl} = V+1 = q^n + 1$, формула для предельного значения критерия для внутренних дополнительных точек у произвольных классов с параметрами $\{n, N_{cl}\}$ принимает вид: $F(\bar{O}, A\{n, N_{cl}\}) = 1{,}5/[N_{cl} + (N_{cl} - 1)^{(n-1)/n} + 2]$.

Таким образом, в качестве критерия попадания дополнительной точки во внутреннюю часть кластера для произвольного класса с параметрами $\{n, N_{cl}\}$ можно принять значение:

$$F_{in}(n, N_{cl}) = F(\bar{O}, A\{n, N_{cl}\}) = 1{,}5/[N_{cl} + (N_{cl} - 1)^{(n-1)/n} + 2]. \quad (4)$$

При $F(\bar{a}_{nov}(p), A_f) \le F_{in}(n, N_{cl})$ анализируемая точка $\bar{a}_{nov}(p)$ является внутренней, для ее уровня близости точек p выполняется: $0 \le p \le 1$. Такая точка не может быть новизной в классе с параметрами $\{n, N_{cl}\}$.

При $F(\bar{a}_{nov}(p), A_f) > F_{in}(n, N_{cl})$ точка $\bar{a}_{nov}(p)$ является внешней ($p > 1$). Такая точка может быть новизной в классе с параметрами $\{n, N_{cl}\}$.

2.3. Анализ дополнительных точек, внешних по отношению к кластеру.

У внешних точек $\bar{a}_{nov}(p)$ при $F(\bar{a}_{nov}(p), A_f) > F_{in}(n, N_{cl})$ уровень близости к центру кластера $p > 1$.

Поскольку из всех внутренних точек наиболее удаленной от центра куба является внутренняя точка $\bar{O}$, то в качестве реперной точки $\bar{a}_{nov}(p)$, соответствующей значению p принимаем точку с координатами: $\bar{a}_{nov}(p) = (-\Delta_p, 0, \ldots, 0) = (-pn(q-1)/2, 0, \ldots, 0)$.

Для перехода от суммы расстояний $\Sigma\rho(\bar{O}, \bar{a}^u)$ (для начальной точки $\bar{O}$) к расстояниям $\Sigma\rho(\bar{a}_{nov}(p), \bar{a}^u)$ (для внешней реперной точки $\bar{a}_{nov}(p)$) в рассматриваемой метрике необходимо к каждому из q^n расстояний $\rho(\bar{O}, \bar{a}^u)$ в сумме $\Sigma\rho(\bar{O}, \bar{a}^u)$ прибавить величину $\Delta_p = pn(q-1)/2$. При этом:

$$\Sigma\rho(\bar{a}_{nov}(p), \bar{a}^u) = \Sigma\rho(\bar{O}, \bar{a}^u) + q^n \Delta_p = nq^n(q-1)/2 + pnq^n(q-1)/4 = nq^n(q-1)(2+p)/4$$

С учетом данной суммы сумма всех расстояний между всеми парами объектов $\{\bar{a}^v, \bar{a}^u\}$ множества $\{C^n_q \cup \bar{a}_{nov}(p)\}$ равна:

$$\Sigma\rho(\bar{a}^v, \bar{a}^u) = \Sigma(C^n_q)\rho(\bar{a}^v, \bar{a}^u) + 2\Sigma\rho(\bar{a}_{nov}(p), \bar{a}^u) = nq^{2n-1}(q^2-1)/3 + nq^n(q-1)(2+p)/2 = nq^n(q-1)[q^{n-1}(q+1)/3 + (2+p)/2].$$

Подстановка сумм в формулу (3) дает для класса $\{C^n_q \cup \bar{a}_{nov}(p)\}$ выражение критерия $F(\bar{a}_{nov}(p), A_f)$ средних расстояний через для внешней дополнительной точки $\bar{a}_{nov}(p)$ через величину уровня удаленности p для данной точки:

$F(\bar{a}_{nov}(p), A_f) = N_{cl} \cdot \Sigma\rho(\bar{a}^x, \bar{a}^u)/ \Sigma\rho(\bar{a}^v, \bar{a}^u) = N_{cl} \cdot [(2+p)/4]/[q^{n-1}(q+1)/3+(2+p)/2] = N_{cl}/[2+(4/3)q^{n-1}(q+1)/(2+p)]$.

В частности, для примера на рис.1 значения критерия F для пороговых значений $p = 1{,}5; 2{,}0; 2{,}5$ будут следующими:

$F(\bar{a}_{nov}(1.5), A_f) = 10/[2 + 16/3] = 15/11$; $F(\bar{a}_{nov}(2.0), A_f) = 10/[2 + 16/4] = 5/3$; $F(\bar{a}_{nov}(2.5), A_f) = 10/[2 + 16/5] = 25/13$.

Отметим, что полученная формула справедлива только для внешних по отношению к кубу C^n_q точек, для которых R>1. Для внутренних точек $\bar{a}_{nov}(p)$ куба (R≤1) не выполняется предложенный переход от $\bar{\rho}(\bar{O})$ к $\bar{\rho}(\bar{a}_{nov}(p))$. Поэтому для них полученная формула не верна.

Учитывая, что общее число элементов в множестве $\{C^n_q \cup \bar{a}_{nov}(p)\}$: $N_{cl} = V+1 = q^n + 1$, формула $F(\bar{a}_{nov}(p), A_f)$ для произвольных классов с параметрами $\{n, N_{cl}\}$ дает следующее выражение для теоретического значения критерия для заданного уровня близости p:

$$F_{cr}(p) = F(\bar{a}_{nov}(p), A_f) = N_{cl}/[2+(4/3)(N_{cl} -1+ (N_{cl} -1)^{(n-1)/n})/(2+p)]. \quad (5)$$

В частности, для пороговых значений уровня новизны $p = 1{,}5; 2{,}0; 2{,}5$ формулы принимают следующий вид:

$F_{cr}(1{,}5) = F(\bar{a}_{nov}(1{,}5), A_f) = N_{cl}/[2 + (8/21)(N_{cl} -1+ (N_{cl} -1)^{(n-1)/n})]$;

$F_{cr}(2{,}0) = F(\bar{a}_{nov}(2{,}0), A_f) = N_{cl}/[2 + (1/3)(N_{cl} -1+ (N_{cl} -1)^{(n-1)/n})]$;

$F_{cr}(2{,}5) = F(\bar{a}_{nov}(2{,}5), A_f) = N_{cl}/[2 + (8/27)(N_{cl} -1+ (N_{cl} -1)^{(n-1)/n})]$.

3. Два подхода к обнаружению новизны. Вспомогательные функции для анализа новизны. Алгоритм CUR_NOVELTY однократного определения новизны в заданном классе

Рассмотрим общую задачу определения и удаления новизны из класса A, содержащего N_{cl} объектов, при заданном пороговом значении уровня новизны p. В общем случае поиск новизны можно производить при числе объектов N_{cl}, не меньшем 3, поскольку иначе понятие новизны теряет смысл.

Исходя из специфики решаемой задачи удаления новизны в обособленных классах обучающих данных, возможны два основных подхода к обнаружению и удалению новизны: 1) однократный (слабый) и 2) итерационный (сильный).

При первом упрощенном (слабом) подходе новизна заданного уровня определяется и удаляется из класса A только один раз.

Однако, также, как и при устранении выбросов, удаление из класса A новизны уровня p (некоторого объекта $\{ \bar{a}_i \}$) может порождать в сокращенном множестве A\ $\{ \bar{a}_j \}$ другую новизну уровня p. Данная ситуация при удалении новизны может повторяться неоднократно. Поэтому гарантированное удаление новизны заданного уровня может обеспечить только итерационный (сильный) способ ее удаления. Рассмотрим программную реализацию обоих методов.

Входными данными задачи, в которой рассмотрены свойства $\boldsymbol{N_{cl}}$ объектов некоторого класса обучающей выборки в n –мерном пространстве значений признаков U, является массив OB[n][Ncl], содержащий численные характеристики свойств объектов.

Для сокращения объема вычислений и упрощения программной структуры основных алгоритмов использованы следующие вспомогательные функции.

3.1. Функция MAIN_DATA_CALC расчета основных данных, необходимых для определения новизны

При определении новизны в заданном классе A[Ncl] наряду с массивом OB[n][Ncl], в качестве основных данных рассмотрим также:

1) матрицу DIST[Ncl][Ncl] расстояний между объектами в A,

2) вектор SUM[Ncl] сумм SUM[i] ($\Sigma\rho(\bar{a}^i, \bar{a}^u)$, $\bar{a}^u \in A$) расстояний от выбранного объекта i ($0 \le i \le Ncl-1$) до всех остальных объектов в A.

Поскольку расчет матрицы DIST является самым трудоемким при определении новизны, то предложено только один раз вычислять эти данные по массиву OB[n][Ncl], а потом корректировать их, не повторяя численных расчетов. При значительных величинах n и Ncl такой подход существенно сокращает общий объем вычислений.

Предложено использовать полную матрицу расстояний, поскольку применение именно ее максимально ускоряет все последующие расчеты.

Вектор SUM[Ncl] удобен в качестве вспомогательного при коррекции данных.

Начальный расчет матрицы DIST[Ncl][Ncl] и вектора SUM[Ncl] по массиву OB[n][Ncl] производится функцией MAIN_DATA_CALC.

Входные данные функции MAIN_DATA_CALC:

1) n - размерность пространства U характеристик оьбъектов,

2) Ncl - число объектов в совокупности M,

3) OB[n][Ncl] – массив векторов значений характеристик для объектов из совокупности M,

4) RO(OB1,OB2,n) – функция расчета расстояния между объектами OB1,OB2 в заданной метрике пространства U.

Выходные данные функции MAIN_DATA_CALC:

1)матрица DIST[Ncl][Ncl],

2) вектор SUM[Ncl].

Для сокращения расчетов используется симметричность матрицы DIST. Алгоритм функции следующий.

```
{
 for i = 0 to Ncl-1 //Проход по всем компонентам вектора SUM и строкам матрицы расстояния DIST
  {
   SUM[i] = 0; // Инициализация очередной компоненты вектора SUM
   for j = 0 to i-1 do SUM[i] += DIST[j][i];//суммирование уже вычисленных известных значений расстояний
   DIST[i][i] = 0; // Задание значения диагональному элементу матрицы DIST.
   for j = i+1 to Ncl-1 // Проход по всем еще не определенным компонентам текущей строки матрицы DIST
  {
    DIST[i][j] = RO (OB[i], OB[j], n); //расчет очередного нового расстояния
    DIST[j][i] = DIST[i][j]; //формирование симметричного элемента матрицы
    SUM[i] + = DIST[i][j]; //наращивание компоненты SUM[i]
  }
 }//завершение прохода по компонентам вектора SUM и строкам матрицы DIST
}
```

3.2. Функция NOVELTY_DETECTION определения новизны в множестве объектов

Однократное определение новизны по заданному уровню p заключается в выявлении всех точек класса, для которых значение критерия средних расстояний превышает пороговую величину. Основные данные для точек класса (матрица DIST[Ncl][Ncl] и вектор SUM[Ncl]) при этом считаются уже определенными ранее.

Определение массива NumNov[NNov] объектов, отнесенных к новизне заданного уровня производится функцией NOVELTY_DETECTION.

Входные данные функции NOVELTY_DETECTION:

1) n - размерность пространства U характеристик объектов,

2) Ncl - число объектов в классе,

3) вектор SUM[Ncl] сумм SUM[i] расстояний от выбранного объекта i до всех остальных объектов,

4) p – уровень проверяемой новизны.

.Выходные данные функции NOVELTY_DETECTION:

1) NNov – число найденных объектов, отнесенных к новизне заданного типа,

2) NumNov[NNov] – массив номеров объектов, отнесенных к новизне.

Псевдокод функции NOVELTY_DETECTION имеет вид.

```
{
  NNov = 0; // инициализация числа объектов, являющихся новизной
  SUMW = 0; // инициализация суммы всех расстояний между объектами класса
  F = Ncl / (2 + (4/3) *(Ncl-1+pow ((Ncl -1), (n-1) /n) /(2+p)); // расчет критерия F по p,n и Ncl
  for  i = 0 to Ncl-1  SUMW += SUM[i]; //вычисление SUMW
  for i = 0 to Ncl-1 if(Ncl*SUM[i]/SUMW>=F) //определение новизны и ее фиксация
    {  NumNov[NNov] = i;   NNov +=1 };
}
```

3.3. Функция CORRECTION_NOVELTY_DATA коррекции данных по объектам класса с учетом удаления новизны

Определение новизны при помощи функции NOVELTY_DETECTION не изменяет данных исследуемого класса. Функция CORRECTION_NOVELTY_DATA моделирует удаление новизны из класса за счет выполнения следующих действий:

1) удаление информации о новизне из всех основных данных по объектам класса.

2) формирование полных данных по новизне.

Входные данные функции CORRECTION_NOVELTY_DATA:

1) n - размерность пространства U характеристик оьбъектов,

2) Ncl - число объектов в классе,

3) OB[n][Ncl] – массив векторов значений характеристик для объектов класса,

4) DIST[Ncl][Ncl] - матрица расстояний между объектами в классе,

5) SUM[Ncl] - вектор сумм расстояний от выбранного объекта до всех остальных объектов класса,

6) NNov – число найденных объектов, отнесенных к новизне заданного уровня,

7) NumNov[NNov] – массив номеров объектов, отнесенных к новизне заданного уровня.

Выходные данные функции CORRECTION_NOVELTY_DATA:

1) измененное число NclR объектов в классе,

2) скорректированный массив OBR[n][NclR] векторов значений характеристик для объектов класса,

3) скорректированный вектор SUMR сумм расстояний от объекта класса до остальных,

4) скорректированная матрица DISTR расстояний между оставшимися объектами

5) массив OBNov[n][NNov] векторов значений характеристик для объектов, являющихся новизной.

Псевдокод функции CORRECTION_NOVELTY_DATA имеет следующий вид.

```
{
 for i = 0 to NNov -1    SUM[NumNov [i]]:=-1;//1.разметка сумм расстояний
```

```
for i = 0 to NNov-1 for j:=0 to Ncl-1 OBNov[i][j]:=OB[NumNov[i]][j];//2.формирование данных по объектам новизны
for i = 0 to Ncl-1   if(SUM[i] != -1)  //3. проход по сохраняющимся строкам матрицы DIST
 for j = 0 to Ncl-1   if(SUM[j] = = -1)    //коррекция сохраняющихся элементов SUM, разметка DIST
  {   SUM[i] - = DIST[i][j];    DIST[i][j] = -1;  }
for i = 0 to Ncl-1      if(SUM[i] != -1)    //4. Уплотнение остающихся строк матрицы DIST
   { ii = 0; j = 0; //проход по остающимся столбцам DIST, их перезапись
   while ((ii<Ncl-1)&&(j<Ncl))
    {   while (DIST[i][ii] == -1) ii += 1; //поиск очередного неотрицательного расстояния
      if(ii<Ncl) {DIST[i][j] = DIST[i][ii]; j += 1; ii += 1 };//перезапись неотрицательного расстояния
    }
   }
NclR = Ncl - NNov; // 5. коррекция Ncl
      ii=0; j=0;        // 6. перезапись SUM в SUMR, DIST в DISTR, OBR в OBR
  while ((ii<Ncl-1)AND(j<Ncl))
   {       while (SUM[ii]=-1) ii+=1;//поиск очередного неотрицательного расстояния
     if (ii<Ncl)
      { SUMR[j]=SUM[ii];
        for i= 0 to Ncl1 DISTR[j][i] = DIST[ii][i];
        for i= 0 to n-1 OBR[j][i] = OB[ii][i];
        j+=1; ii+=1;
      };
    }; //завершение перезаписи
  };
```

3.4. Алгоритм CUR_NOVELTY однократного определения текущей новизны в заданном классе

Функция реализует однократное определение новизны заданного уровня p в некотором классе A.

Применение вспомогательных функций дает возможность представить алгоритм CUR_NOVELTY, решающий данную задачу, в довольно простой форме.

Входные данные:

1) n - размерность пространства U характеристик оьбъектов,

2) Ncl - число объектов в совокупности M,

3) OB[n][Ncl] – массив векторов значений характеристик для объектов из совокупности M,

4) RO(OB1,OB2,n) – функция рассчета расстояния между объектами OB1,OB2 в заданной метрике пространства U.

5) p – уровень проверяемой новизны.

Выходные данные:

1) NNov – число найденных объектов, отнесенных к новизне заданного уровня,

2) NumNov[NNov] – массив номеров объектов, отнесенных к новизне заданного уровня,

3) NclR - скорректированное число объектов в классе

4) OBR - скорректированный массив векторов значений характеристик объектов в кластере,

5) OBNov - массив векторов значений характеристик объектов, отнесенных к новизне.

Вспомогательные данные:

1) матрица DIST[Ncl][Ncl] расстояний между объектами класса,

2) вектор SUM[Ncl] сумм расстояний от выбранного объекта до всех остальных объектов класса.

Алгоритм CUR_NOVELTY включает три этапа.

1. Расчет матрицы DIST[Ncl][Ncl] и вектора SUM[Ncl] по массиву OB[n][Ncl] при помощи функции MAIN_DATA_CALC.

2. Определение новизны уровня p в классе объектов A по матрице DIST и вектору SUM при помощи функции NOVELTY_DETECTION, которая выдает искомые число найденных объектов, отнесенных к новизне

заданного уровня NNov и массив номеров соответствующих объектов NumNov[NNov].

3. Если новизна уровня p и выше в классе A есть (NNov>1), то коррекция данных при помощи функции CORRECTION_DATA.

Пример 1. Рассмотрим однократное выделение новизны при помощи функции CUR_NOVELTY. n = 2, Ncl = 4. Изображающие точки объектов класса показаны на рис.2. Допустимый уровень новизны p = 1.5.

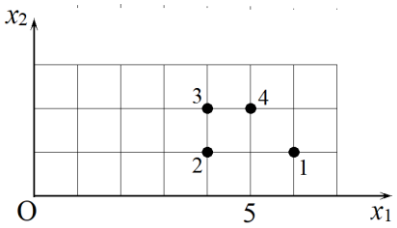

Рис.2. Точки, задающие объекты класса в примере 1

Функция CUR_NOVELTY дает следующие результаты:

1) предельное значение критерия: F_{lim} = F(n = 2, Ncl = 4, p = 1.5) =1.052;

2) суммы расстояний от заданной точки класса до всех остальных:

SUM[1] = 5.650; SUM[2] = 4.414; SUM[3] = 4.236; SUM[4] = 3.828;

3) сумма всех расстояний между точками внутри класса: SUMW =18.129;

4) значения критерия для объектов класса:

i=1) F = 1.247 > F_{lim}; i=2) F = 0.974 < F_{lim}; i=3) F = 0.935 < F_{lim}; i=4) F = 0.845 < F_{lim}.

Так как уровень нозизны превышает допустимый у объекта 1, но он является новизной е уровнtv новизны, превышающим p = 1.5. Все остальные точки класса не являются новизной.

4. Итерационное выделение новизны. Определение кластеров в классах. Алгоритм LEVEL_CLUSTER

В результате применения алгоритма CUR_NOVELTY к некоторому классу A при уровне новизны p после удаления новизны $\{\bar{a}_i\}$ уровня p среди множества оставшихся точек A\{ $\bar{a}_i$} может снова появиться новизна того же уровня. Такая же новизна может возникнуть и после повторного применения алгоритма CUR_NOVELTY и т.д. Итерационное удаление новизны будет продолжаться до тех пор, пока:

1) не будет получено подмножество, не содержащее новизны уровня p (кластер) или же

2) класс A не будет исчерпан – в нем останется менее 3 точек.

Рассматриваемые классы A_i, выделенные из обучающей выборки TE, представляют собой уже предварительно группированные совокупности точек. Поэтому для них введенное понятие уровня новизны p и статистический критерий F позволяют дать строгое определение кластера с заданной степенью p близости точек.

Кластером с предельным уровнем p близости точек для класса A_i назовем такое его подмножество $A_i(p) \subseteq A_i$ его точек, для которого критерий F относительной удаленности объектов не превышает порогового значения $F(p)$. Для всех точек $\bar{a}^y$ подмножества $A_i(p)$ выполняется условие:

$$F(\bar{a}^y, A_i) \le F_{cr}(p) = N_{cl}/[2+(4/3)(N_{cl}-1+(N_{cl}-1)^{(n-1)/n})/(2+p)].$$

Введенное понятие имеет смысл только для классов - предварительно сгруппированных подмножеств объектов, выделенных из их общей совокупности. В ином случае данное понятие не определено.

В общем случае число точек $n(A_i(p))$ в кластере $A_i(p)$ может изменяться от 2 до всего числа точек A_i (когда $A_i(p) = A_i$). Случае $n(A_i(p))$ 3 назовем вырожденным кластером, поскольку при одной и двух точках это понятие не определено.

Введенные уровни новизны позволяют также обоснованно ввести соответствующую им оценку плотности расположения объектов внутри кластеров.

1. Кластер первого типа - кластер с высокой равномерностью расположения объектов – это такой кластер, в котором нет близкой новизны, т.е. для всех его точек $\bar{a}^y$ выполняется условие: $F(\bar{a}^y, A_i) \le F_{cr}$ (1,5).

2. Кластер второго типа - кластер с средней равномерностью расположения объектов – это такой кластер, в котором, возможно, есть близкая новизна, но нет средней и дальней, т.е. для всех его точек $\bar{a}^y$ выполняется условие: $F(\bar{a}^y, A_i) \le F_{cr}$ (2,0).

3. Кластер третьего типа - кластер с низкой равномерностью расположения объектов – это такой кластер, в котором, возможно, есть близкая и средняя новизна, но нет дальней новизны, т.е. для всех его точек $\bar{a}^y$ выполняется условие: $F(\bar{a}^y, A_i) \le F_{cr}$ (2,5).

Очевидно, выделение кластера уровня p задает более высокие требования на искомое множество точек по сравнению с однократным удалением новизны такого же уровня p.

Однако применение функций MAIN_DATA_CALC, NOVELTY_DETECTION и CORRECTION_NOVELTY_DATA дает возможность построить алгоритм LEVEL_CLUSTER для выделения кластера заданного уровня p, имеющий довольно простую структуру.

Входные данные:

1) n - размерность пространства U характеристик объектов,
2) Ncl - число объектов класса,
3) OB[n][Ncl] – массив векторов значений характеристик для объектов класса,
4) RO(OB1,OB2,n) – функция расчета расстояния между объектами OB1,OB2 класса в заданной метрике пространства U,

5) *p* – уровень проверяемой новизны.

Выходные данные:

1) Ncluster - число точек в выделенном кластере уровня *p*,

2) OBC[n][Ncluster] – массив векторов значений характеристик для объектов выделенного кластера уровня *p*.

Вспомогательные данные:

1) матрица DIST[Ncl][Ncl] расстояний между объектами класса,

2) вектор SUM[Ncl] сумм расстояний от выбранного объекта i до всех остальных объектов класса,

3) массив текущей новизны Nov[NNov].

Алгоритм LEVEL_CLUSTER для выделения кластера с заданным предельным уровнем *p* близости точек.

```
{
MAIN_DATA_CALC (n, Ncl, OB, RO, DIST, SUM); // 1. Расчет матрицы DIST[Ncl][Ncl] и вектора SUM[Ncl]
Nnov =1; // 2. Инициализация текущего количества новизны
  while ((Nnov>0)&&(Ncl>2)) //3. Цикл while Условие выполнения итераций
  {
        NOVELTY_DETECTION (n, Ncl,OB, p, DIST, SUM, NNov, NumNov);
        if (Nnov>0)
         {

CORRECTION_DATA(n,Ncl,SUM,DIST,NNov,Nov,NclR,SUMR,DISTR,OBR, OBRNov);//найдена новизна
            Ncl = Ncl1; SUM = SUM1; DIST = DIST1; OB=OBR;
  }//завершение цикла while ((Nnov>0)&&(Ncl>2))
  Ncluster = Ncl;
  OBC=OB;
```

}//завершение алгоритма LEVEL_CLUSTER

Пример 2. Рассмотрим итерационное выделение новизны при помощи функции LEVEL_CLUSTER. n = 2, N_{cl} = 5. Изображающие точки объектов класса показаны на рис.3. Допустимый уровень новизны p = 3.5.

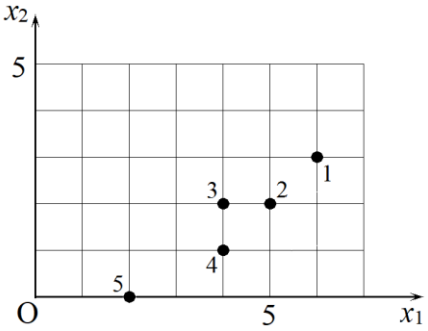

Рис.3. Точки, задающие объекты класса в примере 2

Итерация 1. Функция LEVEL_CLUSTER дает следующие результаты:

1) предельное значение критерия: F_{lim} = F(n = 2, Ncl = 5, p = 3.5) =1.447;

2) суммы расстояний от заданной точки класса до всех остальных:

SUM[1] = 11.479; SUM[2] = 7.434; SUM[3] = 7.064; SUM[4] = 7.479; SUM[4] = 13.670;

3) сумма всех расстояний между точками внутри класса SUMW = 47.126;

4) значения критерия для объектов класса:

i=1) F = 1.218 < F_{lim}; i=2) F = 0.789 < F_{lim}; i=3) F = 0.750 < F_{lim}; i=4) F = 0.793 < F_{limm}; i=5) F = 1.450 > F_{limm}.

Уровень новизны у объекта 5 превышает допустимый, он является новизной и удаляется из класса.

Итерация 2: 1) F_{lim} = F(n = 2, Ncl = 4, p = 3.5) =1.176;

2) SUM[1] = 5.236; SUM[2] = 3.414; SUM[3] = 4.0; SUM[4] = 4.650;

3) SUMW = 17.301;

4) i=1) F = 1.211 > F_{lim}; i=2) F = 0.790 < F_{lim}; i=3) F = 0.925 < F_{lim}; i=4) F = 1.075 < F_{limm}.

Уровень новизны превышает допустимый у объекта 1, он является новизной и удаляется из класса.

Итерация 3: 1) F_{lim} = F(n = 2, Ncl = 3, p = 3.5) =1.0609;

2) SUM[1] = 2.414; SUM[2] = 2.000; SUM[3] = 2.414;

3) SUMW = 6.828;

4) i=1) F = 1.0606 < F_{lim}; i=2) F = 0.879 < F_{lim}; i=3) F = 1.0606 < F_{lim}.

Новизна уровня выше p = 3.5 отсутствует. Следовательно, кластер с предельным уровнем новизны p = 3.5 в исходном классе существует и его составляют точки {2,3,4}.

Рассмотренные алгоритмы CUR_NOVELTY (однократное определение текущей новизны заданного уровня p) и LEVEL_CLUSTER (определения кластера уровня p) задают слабый и сильный виды удаления новизны из классов обучающей выборки.

Выбор метода устранения новизны определяется общими требованиями к обучающей выборке и классификатору, который строится по ней.

Длину входа задачи определения новизны определяет массив OB[n][Ncl], в котором заданы векторы значений характеристик для объектов из анализируемого класса. Длина массива пропорциональна n·Ncl.

Максимальную сложность из вспомогательных функций, определяющих сложность обоих алгоритмов, имеет функция MAIN_DATA_CALC, в которой производится расчет матрицы расстояний DIST[Ncl][Ncl] и вектор сумм расстояний SUM[Ncl].

При расчете одного расстояния между объектами, например, в евклидовой метрике, затрачивается 1) n вычитаний, 2) n возведений в квадрат, 3) (n-1) сложений и 4) 1 извлечение квадратного корня. Во всей функции MAIN_DATA_CALC такие расстояния вычисляются ровно N (N - 1)/2 раз.

Порядок трех основных операций совпадает и равен n. Поэтому сложность функции MAIN_DATA_CALC равна $O(n \cdot N^2)$. Так как (n·Ncl) <

$(n \cdot N^2) < (n \cdot Ncl)^2$, то сложность рассмотренных алгоритмов выше линейной, но меньше квадратичной по длине входа задачи.

Заключение

В статье рассмотрена актуальная частная задача определения и удаления новизны в классах обучающих данных, используемых при построении классификаторов при обучении с учителем. От эффективности ее решения во многом зависит качество классификаторов, получаемых на таких выборках.

Предложена геометрическая интерпретация новизны. На ее основе введено общее понятие уровня новизны и три ее качественные градации. Это дает возможность придать новизне объекта четкий пространственный смысл и оценить допустимые для нее пределы в конкретных решаемых задачах. Однако из-за сложности вычисления уровня новизны для практических расчетов предложено использовать связанный с ней статистический критерий средних расстояний. При помощи критерия можно косвенно оценивать уровень близости объектов ко всем объектам класса на основании довольно простых вычислений.

Использование уровня новизны позволило дать четкую геометрическую интерпретацию новизны и в частности, обоснованно выбирать пороговые значения при ее поиске. Применение статистического критерия F позволяет существенно упростить поиск новизны. Зависимость F(p), найденная на *n*–мерных кубах, позволила связать геометрическую оценку новизны с ее статистической оценкой.

Найдена верхняя оценка критерия F(p) для дополнительных точек, попадающих внутрь основного кластера класса, а также величина критерия для внешних точек кластера. Получена формула для значений критерия на внешних точках произвольных классов с параметрами $\{n, N_{cl}\}$, а также

формулы для пороговых значений критерия F при трех введенных градациях новизны, которым соответствуют значения уровня p=1.5; 2.0;2.5.

Рассмотрен обычный алгоритм однократного обнаружения и удаления новизны, который выявляет только текущую новизну заданного уровня, но не гарантирующий ее отсутствие в получаемом сокращенном классе.

С использованием геометрической аналогии введено понятие кластера с предельным уровнем p близости точек, гарантирующих отсутствие в итоговом множестве объектов новизны данного уровня. Также дано определение кластеров первого, второго и третьего типов, соответствующих отсутствию в них близкой, средней и дальней новизны.

Для определения кластера с заданным предельным уровнем p близости точек, требуется применять итерационный (сильный) метод удаления новизны соответствующего уровня из класса. Рассмотрен алгоритм, реализующий данный метод полного удаления новизны заданного уровня.

Показано, что оба алгоритма имеют сложность выше линейной по длине входа задачи, но меньше квадратичной.

Предложенные алгоритмы обнаружения и удаления новизны в классах обучающих данных вместе с алгоритмами удаления выбросов обеспечивают эффективное удаление шумов в этих данных. Применение таких исправленных обучающих данных обеспечивает построение на них качественных классификаторов.

References

1. Mottl V., Seredin O., Krasotkina O. Compactness Hypothesis, Potential Functions, and Rectifying Linear Space in Machine Learning. URL: link.springer.com/chapter/10.1007/978-3-319-99492-5_3.

2. Santoyo Sergio. A Brief Overview of Outlier Detection Techniques. What are outliers and how to deal with them? 2017. URL:

towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561.

3. Chen D., Jain R. A robust backpropagation learning algorithm for function approximation. IEEE Trans Neural Netw 5(3). 1994. pp. 467–479.

4. Liano K. Robust error measure for supervised neural network learning with outliers. IEEE Trans Neural Netw 7(1). 1996. pp. 246– 250.

5. J. S. Anstey, D. K. Peters and C. Dawson. Discovering Novelty in Time Series Data, Proc. Newfoundland Electrical and Computer Engineering Conference, IEEE, Newfoundland and Labrador Section, November 2005. URL: researchgate.net/publication/250381613_Discovering_Novelty_in_Time_Series_Data

6. Chandola V., Banerjee A., Kumar V. Outlier Detection: A Survey, Univ. of Minnesota TR 07-017, August 2007. URL: researchgate.net/publication/242403027_Outlier_Detection_A_Survey

7. Miljković D. Review of Novelty Detection Methods. Hrvatska elektroprivreda. Conference MIPRO proceedings of the 33rd International Convention At Opatija, Croatia. May 2010. URL: researchgate.net/publication/261424710_Review_of_novelty_detection_methods.

8. S. Marsland. Novety Detection in Learning Systems, Robotics and Autonomous Systems, 51(2-3). 2005. pp. 191-206.

9. Martinez D. Neural tree density estimation for novelty detection. IEEE Transactions on Neural Networks, Volume: 9 Issue: 2. 1998. pp. 330-338.

10. Lee H., Hwang B., Cho S. Analysis of Novelty Detection Properties of Autoassociative MLP. Journal of the Korean Institute of Industrial Engineering. Vol.28, No2. 2002. URL: researchgate.net/publication/263358725_Analysis_of_Novelty_Detection_Properties_of_Autoassociative_MLP

11. Rowland B., Maida A. S. Spatiotemporal Novelty Detection Using Resonance Networks. Proceedings of the 17th Annual Florida AI Research Society Conference. May 2004. pp. 676-681.

12. Tanaka T., Weitzenfeld A. Adaptive Resonance Theory in Neural Simulation Language. The MIT Press. 2002. pp. 157-169.

13. Oliveira A.L.I., Neto F.B.L., Meira S.R.L. Combining MLP and RBF Neural Networks for Novelty Detection in Short Time Series. Proc. of MICAI. 2004.URL:researchgate.net/publication/296580021_Combining_MLP_and_RBF_neural_networks_for_novelty_detection_in_short_time_series,

14. Markou M., Singh S. Novelty detection: A review - Part 2: Neural network based approaches. Signal Processing 83(12). pp. 2499-2521.

15. Schölkopf B., Williamson R., Smolax A., Shawe-Taylory J., Platt J. Support Vector Method for Novelty Detection. Advances in Neural Information Processing Systems 12. URL: proceedings.neurips.cc/paper/1999/file/8725fb777f25776ffa9076e44fcfd776-Paper.pdf

16. Banerjee A., Chandola V., Kumar V., Lazarević A. Anomaly Detection: A Tutorial, Proc. of SIAM Data Mining Conference. April 2008. 60 p.

17. Hautamäki V. Outlier Detection Using k-Nearest Neighbour Graph, 17th International Conference on Pattern Recognition Vol. 3, Cambridge. 2004. pp.80-84.

18. Miljković D. Novelty Detection In Machine Vibration Data Based On Cluster Intraset Distance. Proc. CTS, MIPRO. 2008. pp. 59-66.

19. Hallgrímsson, B., Jamniczky, H. A., Young, N. M. The generation of variation and the developmental basis for evolutionary novelty. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 2012. pp. 501-517.

20. Hughes, A. L. The evolution of functionally novel proteins after gene duplication. Proceedings of the Royal Society of London B: Biological Sciences. 1994. pp.119-124.

21. Kailing K., Kriegel H-P., Kröger P. Density-Connected Subspace Clustering for High-Dimensional Data. In: Proc. SIAM Int. Conf. on Data Mining (SDM'04), pp. 246–257.

22. Agrawal R., Gehrke J., Gunopulos D., Raghavan P. Automatic Subspace Clustering of High Dimensional Data. Data Mining and Knowledge Discovery. 2005, pp. 5-33.

23. Kriegel H-P., Kröger P., Zimek A. Subspace clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012. pp. 351-364.

24. Wolfson, H.J., Rigoutsos, I. Geometric Hashing: An Overview. IEEE Computational Science and Engineering, 1997. pp.10-21.

25. Mian A.S., Bennamoun M., Owens R. Three-dimensional model-based object recognition and segmentation in cluttered scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, Oct. 2006, pp. 1584-601.

26. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (PDF) (Second ed.), New York 2008. pp.129-135.

27. Hall P., Park B.U., Samworth R.J. Choice of neighbor order in nearest-neighbor classification. 2008. 36p.

28. Bremner D., Demaine E., Erickson J., Iacono J., Langerman S., Morin P., Toussaint G.T. Output-sensitive algorithms for computing nearest-neighbor decision boundaries. Discrete and Computational Geometry. 2005. p. 593-604.