Методы кластер-анализа, или автоматической классификации

Минск 2011

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ. ПРИКЛАДНАЯ СТАТИСТИКА 3

Развитие представлений о статистике 3
Прикладная статистика 4

1 ЗАДАЧА КЛАССИФИКАЦИИ 7

2 МЕТОДЫ КЛАССИФИКАЦИИ 11

3 ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ 15

3.1. Автоматическая классификация, основанная на описании классов «ядрами». Эвристические алгоритмы 16
3.1.1. Параллельные процедуры 17
Алгоритм k-эталонов 17
Алгоритм взаимного поглощения 18
3.1.2. Последовательные процедуры 19
3.2 Алгоритмы, использующие понятие центра тяжести 21
3.2.1 Параллельные процедуры 21
Алгоритм k-средних 21
3.3 Алгоритмы с управляющими параметрами, настраиваемыми в ходе классификации 32
3.3.1 Параллельные процедуры 32
Алгоритм ИСОМАД (Isodata) 32
Алгоритм Пульсар 35

БИБЛИОГРАФИЧЕСКИЙ УКАЗАТЕЛЬ 37

ВВЕДЕНИЕ. ПРИКЛАДНАЯ СТАТИСТИКА

Развитие представлений о статистике

Впервые термин «статистика» мы находим в художественной литературе – в «Гамлете» Шекспира (1602 г., акт 5, сцена 2). Смысл этого слова у Шекспира – знать, придворные. В течение следующих 400 лет термин «статистика» понимали и понимают по-разному. Вначале под статистикой понимали описание экономического и политического состояния государства или его части.

По Наполеону Бонапарту «Статистика – это бюджет вещей». Согласно формулировке 1833 г. «цель статистики заключается в представлении фактов в наиболее сжатой форме». Еще несколько значений этого термина. Статистика состоит в наблюдении явлений, которые могут быть подсчитаны или выражены посредством чисел (1895). Статистика – это численное представление фактов из любой области исследования в их взаимосвязи (1909).

В XX в. статистику часто рассматривают прежде всего как самостоятельную научную дисциплина. Статистика есть совокупность методов и принципов, согласно которым проводится сбор, анализ, сравнение, представление и интерпретация числовых данных (1925).Термин «статистика» употребляют еще в двух смыслах. Во-первых, в обиходе под «статистикой» часто понимают набор количественных данных о каком-либо явлении или процессе. Во-вторых, статистикой называют функцию от результатов наблюдений, используемую для оценивания характеристик и параметров распределений и проверки гипотез.

Прикладная статистика

Современный этап развития статистических методов можно отсчитывать с 1900 г., когда англичанин К. Пирсон основан журнал «Biometrika». Разработанная в первой трети XX в. теория анализа данных называется параметрической статистикой.Однако, позднее выяснилось, что это тупиковая ветвь теории статистики, поскольку реальные данные не подчиняются каким-либо параметрическим семействам, надо применять иные статистические методы. Параметрические методыкритиковал академик АН СССР С.Н.Бернштейн в 1927 г. в своем докладе на Всероссийском съезде математиков.

В нашей стране термин «прикладная статистика» вошел в широкое употребление в 1981 г. после выхода массовым тиражом (33940 экз.) сборника «Современные проблемы кибернетики (прикладная статистика)».

Прикладная статистика и математическая статистика – это две разные научные дисциплины.

Прикладная статистика - методическая дисциплина, являющаяся центром статистики. При применении методов прикладной статистики к конкретным областям знаний и отраслям народного хозяйства получаем научно-практические дисциплины типа "статистика в промышленности", "статистика в медицине" и др. Прикладная статистика нацелена на решение реальных задач. Поэтому в ней возникают новые постановки математических задач анализа статистических данных, развиваются и обосновываются новые методы.

В настоящее время статистическая обработка данных проводится, как правило, с помощью соответствующих программных продуктов. Разрыв между математической и прикладной статистикой проявляется, в частности, в том, что большинство методов, включенных в статистические пакеты программ (например, в заслуженные Statgraphics и

SPSS или в более новую систему Statistica), даже не упоминается в учебниках по математической статистике.

Методы прикладной статистики используются в зарубежных и отечественных экономических и технических исследованиях, работах по управлению, в медицине, социологии, психологии, истории, геологии и других областях. Например, в США - не менее 20 миллиардов долларов ежегодно только в области статистического контроля качества.

Статистическая дисциплина, занимающаяся прикладной реализации математико-статистических методов, проблемно- и методоориентированных систем автоматизированной статистической обработки данных, называется «анализом данных» или «прикладная статистика».

Наиболее актуальные направления исследований этой научной дисциплины:

а) Развитие методов анализа данных, не апеллирующих ких вероятностной природе, а также методов, нацеленных на выявление вероятностной и геометрической природы обрабатываемых данных в условиях отсутствия соответствующей априорной информации (кластер-анализ, многомерное шкалирование, томографические методы, целенаправленное проецирование многомерных данных и т. п.).

б) Формализация (математическая постановка) реальных задач статистического анализа данных в различных предметных областях и на базе этого опыта выработка типовых математических постановок задач.

в) Вычислительные вопросы компьютерной реализации методов статистического анализа данных.

г) Теория и практика генерирования на ЭВМ данных заданной природы и развитие на этой основе методов статистического анализа малых выборок.

д) Развитие прикладного программного обеспечения по методам статистического анализа данных с акцентом на создание

интеллектуализированных проблемно- и методо-ориентированных программных комплексов, способных обеспечить исследователя развитой системой машинного ассистирования.

1 ЗАДАЧА КЛАССИФИКАЦИИ

В прикладной статистике есть раздел многомерного статистического анализа, одной из задач которого является задача классификации. Необходимость анализа и формализации задач, связанных со сравнением и классификацией объектов, сознавали ученые далекого прошлого. Приведем четыре генеральные идеи и методологические принципы многомерного статистического анализа, на которых базируются, по существу, все основные разделы и подходы математического аппарата классификации.

1. Эффект существенной многомерности. Сущность этого принципа в том, что выводы, получаемые в результате анализа и классификации множества статистически обследованных (по ряду свойств) объектов, должны опираться одновременно на совокупность этих взаимосвязанных свойств с обязательным учетом структуры и характера их связей.
2. Возможность лаконичного объяснения природы анализируемых многомерных структур.
3. Максимальное использование «обучения» в настройке математических моделей классификации и снижения размерности.
4. Оптимизационная формулировка задач классификации и снижения размерности.

Под классификацией понимается система группировки множества объектов, составленная на основе учета общих признаков этих объектов и закономерных связей между ними.

Целью классификации является образование групп схожих между собой объектов, которые принято называть классами. При

геометрическом подходе в основе применения методов классификации лежит так называемая гипотеза компактности. Согласно ей, близким в содержательном смысле объектам в геометрическом пространстве признаков соответствуют обособленные множества точек, обладающие свойствами хорошей отделимости. А именно:

1. множества разных образов соприкасаются в сравнительно небольшом числе точек, либо вообще не соприкасаются и разделены точками, не принадлежащими ни одному из классов;
2. границы классов имеют сравнительно плавную форму – не изрезаны, и у классов отсутствуют глубокие выступы в пределы других классов.

В результате различные классы при выполнении гипотезы компактности могут быть разделены достаточно простыми гиперповерхностями.

Гипотеза компактности дает на практике хорошие результаты классификации, если есть достаточное соответствие между содержанием выделенных признаков и построенным геометрическим пространством.

Обобщением гипотезы компактности является гипотеза простой геометрической структуры. Она заключается в следующем: сходным в содержательном смысле объектам классификации соответствует простая структура в геометрическом пространстве признаков: расположенность вдоль прямой, на окружности, в сфере, по спирали, на решетке и т.п.

На основе гипотезы компактности разработано множество алгоритмов классификации.

До разработки аппарата многомерного статистического анализа и, главное, до появления и развития достаточно мощной электронно-вычислительной базы главные проблемы теории и практики относились не к разработке методов и алгоритмов, а к полноте и тщательности отбора и теоретического анализа изучаемых объектов, характеризующих их признаков, смысла и числа градаций по каждому из этих признаков. Все

методы классификации сводились, по существу, к методу так называемой комбинационной группировки, когда все характеризующие объект признаки носят дискретный характер или сводятся к таковым, а два объекта относятся к одной группе только при точном совпадении зарегистрированных на них градаций одновременно по всем характеризующим их признакам. Однако по мере роста объемов перерабатываемой информации и, в частности, числа классифицируемых объектов возможность эффективной реализации подобной логики исследования становилась все менее реальной. Именно электронно-вычислительная техника стала тем главным инструментом, который позволил по-новому подойти к решению этой важной проблемы и, в частности, конструктивно воспользоваться разработанным к этому времени мощным аппаратом многомерного статистического анализа: методами распознавания образов «с учителем» (дискриминантный анализ) и «без учителя» (автоматическая классификация, или кластер-анализ) и т. д.

Итак, подведем итог и дадим окончательное, формальное определение задачи классификации:

> **Задача классификации** — формализованная задача, в которой имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется выборкой. Классовая принадлежность остальных объектов не известна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.
>
> **Классифицировать объект** — значит, указать номер (или наименование) класса, к которому относится данный объект.

Классификация объекта — номер или наименование класса, выдаваемый алгоритмом классификации в результате его применения к данному конкретному объекту.

Существуют также другие способы постановки эксперимента — т.н. обучение без учителя, но они используются для решения другой задачи — кластеризации или таксономии (ниже будут описаны методы классификации, основанные на обучении без учителя). В этих задачах разделение объектов обучающей выборки на классы не задаётся, и требуется классифицировать объекты только на основе их сходства друг с другом. В некоторых прикладных областях, и даже в самой математической статистике, из-за близости задач часто не различают задачи кластеризации от задач классификации.

2 МЕТОДЫ КЛАССИФИКАЦИИ

Выше мы определили, что классификация – это разбиение исследуемых объектов на классы подобных друг другу в некотором заданном априорно или найденном в процессе анализа объективно существующем отношении. В математическом плане задача классификации объектов формулируется как задача построения разбиения объектов исходного множества на некоторое определенное заранее или отыскиваемое в ходе анализа число непустых попарно непересекающихся подмножеств (классов) объектов.

Современные методы классификации могут быть структурированы следующим образом:

1. ***различают классификацию с учителем и классификацию без учителя*** или, по-другому, автоматическую классификацию. В первом случае, когда категории объектов установлены заранее, до начала анализа данных, говорят, что имеет место классификация с обучением (с учителем), и в результате проведения ее необходимо найти распределение объектов анализируемой таблицы "объекты-свойства" по известным классам. Во втором случае необходимо найти как структуру таблицы "объекты-свойства", т.е. совокупность возможно имеющихся в ней "естественных" классов, так и распределение объектов по этим классам; выявление структуры таблицы "объекты-свойства" при этом достигается при помощи тех или иных формальных параметрических и непараметрических критериев качества разбиения объектов на классы.
2. ***различают исключающие и не исключающие классификации***. В результате проведения не исключающей классификации один и тот же объект может быть одновременно отнесен сразу к нескольким классам как, например, в случае, когда в рабочем теле объекта

диагностирования обнаруживается сразу несколько различных дефектов. Не исключающие классификации не будут рассматриваться, поскольку они не являются иерархическими и не удовлетворяют гипотезе компактности; кроме того, исключающие классификации всегда могут быть организованы таким образом, чтобы они могли выявлять, в частности, как множественные, так и кратные дефекты.

3. ***различают внутренние и внешние классификации***. При проведении внутренней классификации все признаки априорного описания объектов признаются равноправными и используются при проведении классификации единообразно. При проведении внешней классификации, один из признаков априорного описания объектов выделяется в качестве внешнего, и цель анализа данных сводится к тому, чтобы на основе информации, доставляемой только остальными (внутренними) признаками, построить классификацию, наилучшим образом отражающую поведение признака, выделенного в качестве внешнего; по сути дела, при проведении внешней классификации выделенный признак используется в качестве "учителя".
4. ***классификации разделяют на иерархические и неиерархические***. Формально неиерархические классификации направлены на оптимизацию внутренних свойств выделяемых классов, а иерархические - на оптимизацию отношений между конкретными объектами и их совокупностью, представленной анализируемой таблицей "объекты-свойства", т.е. на оптимизацию структуры последней. Другими словами, при неиерархической классификации объекты агрегируются так, чтобы формируемые классы были по возможности наиболее однородными, а отношения между классами при этом не анализируются. Напротив, при иерархической классификации агрегирование объектов в один класс (а также двух

классов в один больший класс) производится только тогда, если такое объединение сопровождается минимальными приращениями неоднородности вновь получаемого класса.

5. ***классификации разделяются по методу, точнее направлению разбиения объектов на классы на восходящие (агломеративные) и нисходящие (дивизивные)***. При восходящей классификации агрегирование объектов выполняются путем объединения объектов в классы все возрастающего объема, пока не будет получен один класс, охватывающий все анализируемые объекты. Другими словами, восходящая (объединяющая) классификация начинается с такого исходного разбиения, в котором содержится ровно М классов по одному объекту в каждом, где М как обычно, число объектов в анализируемой таблице "объекты-свойства". При этом критерии объединения объектов и/или промежуточных классов на каждом очередном уровне иерархии классов (иерархии объединений) служит условие наименьшего приращения неоднородности образуемого класса по сравнению с уровнями неоднородностей всех других классов, которые могут быть в принципе образованы на данном уровне. При нисходящей классификации исходная совокупность анализируемых объектов разделяется путем последовательной дихотомии (деление пополам) до тех пор, пока не будет достигнут желаемый уровень однородности получаемых классов либо когда будет получено разбиение, в котором в каждом классе содержится ровно по одному объекту из числа анализируемых.
6. ***различают моно- и политетические классификации***. При монотетической классификации разбиение таблицы "объекты-свойства" на классы производится с использованием на каждом уровне иерархии единственного признака, который для данного уровня наиболее информативным. При политетической

классификации все признаки априорного описания объектов и на всех уровнях используются единообразно: при этом восходящие (агломеративные) классификации всегда являются политетическими.

3 ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ

Итак, если для всех объектов исходного набора известно, к какому классу они принадлежат, то такая постановка задачи называется классификацией с учителем (или с обучением). Обучение без учителя происходит тогда, когда принадлежность объектов в исходном наборе нам заранее не известна. В настоящей главе будут рассматриваться методы, относящиеся к автоматической классификации (АК), или обучении без учителя. К этому виду классификации, например, можно отнести «Метод главных компонент» или «Кластерный анализ». Рассмотрим подробнее методы АК, основанные на описании классов «ядрами».

«Ядерные» методы нацелены на выявление сгущений наблюдений в признаковом пространстве и ранее носили чисто эвристический характер, так как понятие компактности наблюдений в признаковом пространстве не было формализовано. Для ряда эвристических процедур с развитием теории были найдены функционалы качества разбиения наблюдений на группы и тем самым формализовано соответствующее им понятие компактности. В соответствии с этим алгоритмы классификации, основанные на описании классов ядрами, подразделяют на эвристические и оптимизационные. Кроме того, методы можно разделить по способу подачи наблюдений на вход алгоритма. Если наблюдения подаются по одному (последовательно), то соответствующие процедуры называются последовательными. Если на вход алгоритмов подаются сразу все наблюдения, то они называются параллельными. Преимуществом последовательных процедур является высокая скорость работы, параллельных — независимость получаемой классификации от порядка данных выборки в исходном множестве О.

Под ядром класса подразумевается некоторая реально существующая или условное наиболее «представительное» наблюдение, весь комплекс характеристик которой является эталоном данного класса. Часто алгоритмы, основанные на описании классов «ядрами», используют процедуру классификации текущего выборочного элемента к ядрам по минимальности расстояний:

• задаться метрикой d;

• найти ядра классов;

• классифицировать все наблюдения к ядрам по минимальности расстояния до них.

Для нахождения ядер обычно используют обучающую выборку, по которой находят геометрические центры классов, или применяют специальные формальные процедуры.

Таким образом, понятие ядра имеет самый широкий смысл: ядро класса (т. е. группы точек) может быть подгруппой точек, центром тяжести, осью, случайной переменной и т. д.

3.1 Эвристические алгоритмы

Рассмотрим сначала методы АК, непосредственно опирающиеся на постановку задачи выделения в многомерном пространстве компактных групп точек. Такие методы и отвечающие им алгоритмы называются эвристическими, так как само понятие «компактная группа (облако) точек» не поддается строгой формализации. В прикладных задачах автоматической классификации они стали применяться одними из первых и до сих пор сохраняют большое значение в разведочном анализе данных благодаря наглядности интерпретации полученных результатов и, как правило, простоте реализации. Для ряда эвристических процедур с развитием теории АК были найдены функционалы качества разбиения на

группы и тем самым формализовано соответствующее им понятие «компактности».

Описанные ниже алгоритмы разделены на два класса: параллельные и последовательные процедуры.

3.1.1 Параллельные процедуры

Алгоритм k-эталонов

Приведем типичный пример эвристического алгоритма, основная идея которого заключается в том, что совокупность объектов, находящихся на одинаковом расстоянии от каждого из k эталонов (ядер), образует компактную группу.

Пусть для классификации имеется выборка O_1 ,..., O_n причем i-й объект O_i - характеризуется вектором признаков $X_i = (x_i^{(1)}, ..., x_i^{(p)})$. Рассмотрим p-мерное признаковое пространство X^p вместе с функцией $d(X_i, X_j)$, задающей в X^p расстояние (либо степень близости).

Формальная схема алгоритма

1. Выберем k эталонов X^*_i, ..., X^*_k и порог d_0.
2. Поставим в соответствие объекту O_i код из k двоичных символов $e_i = (e_i^{(1)}, ..., e_i^{(k)})$, где $e_i^{(L)} = 1$, если $d(X_i,X_L) < d_0$ и $e_i^{(L)} = 0$ в противном случае.
3. Разобьем выборку на классы, относя к одному классу объекты с одинаковым кодом.

В зависимости от порога d_0 и геометрии выборки число классов может варьироваться от 1 до 2^k. Анализируя полученное разбиение выборки на классы, исследователь может уточнить выбор эталонов и перейти к следующей итерации алгоритма. Если в качестве эталонов взять векторы признаков объектов из данной выборки, то на вход алгоритма достаточно подать матрицу взаимных попарных расстояний

$\{p_{ij} = d(X_i, X_j)\}$. Рекомендуется набор эталонов составлять из k случайно выбранных точек. Укажем наиболее важные модификации алгоритма, связанные со способом выбора эталонов:

а) эталоны строятся на основе представлений экспертов (или какой-либо другой априорной информации о типичных представителях классов);

б) эталоны берутся из векторов, соответствующих представителям заведомо разных классов, но не обязательно типичных в своем классе. При этом достоинством алгоритма является то, что число эталонов не обязано равняться числу классов;

в) эталоны могут пониматься в расширенном смысле, как ядра в методе динамических сгущений.

Следующая модификация рассматриваемой процедуры связана с возможностью варьировать порог сходства.

Алгоритм взаимного поглощения

Рассмотрим процедуру автоматической классификации, целесообразность которой обосновывается правилом формирования сплоченных коллективов людей по принципу взаимного интереса и симпатии. Ключевым словом здесь является «взаимный», т. е. подразумевается, что отношение «объект O_i близок (интересен) объекту O_j» не симметрично и объекты O_i и O_j объединяются, если не только O_i близок к O_j, но и наоборот.

Формальная схема алгоритма

Обозначения такие же, как и в алгоритме k-эталонов:

1. Объекту O_i поставим в соответствие порог $d_i > 0$, называемый его радиусом влияния, т. е. объект O_j считается близким к O_i (находится в сфере влияния объекта O_i), если $d(X_i,X_j) <= d_i$.

2. Объекту O_i поставим в соответствие код $e_i = (e_i^{(1)}, ..., e_i^{(k)})$, где $e_i^{(L)} = 1$, если $d(X_L, X_i) <= d_i$ и $e_i^{(L)} = 0$ в противном случае.

3. Выделим в выборке классы, относя набор объектов O_{i1} , ..., O_{im} к одному классу, если у их кодов e_{i1},..., e_{im} все координаты с номерами L = i1 ,..., im равны 1.

4. Выделим в выборке минимальное число классов, объединение которых дает всю выборку.

Ясно, что алгоритм дает в общем случае нечеткую классификацию выборки. Геометрически каждому объекту O_i в признаковом пространстве X^p отвечает шар S_i радиуса d_i с центром в точке X_i. Классу $\{O_{i1}, ..., O_{im}\}$ отвечает пересечение $\bigcap_{l=1}^{m} S_{il}$ шаров S_{iL} содержащее все центры $X_{i1}, ..., X_{im}$ и называемое областью взаимного поглощения данного класса. В ряде задач основной целью классификации является покрытие признакового пространства областями взаимного поглощения классов.

Ясно, что настройка рассматриваемого алгоритма на специфику решаемой задачи осуществляется выбором порогов d_i, $1 < i <n$. Приведем примеры такого выбора**:**

а) $d_i = \max_j d(X_i,X_j) - \delta$

б) $d_i = \min_j d(X_i,X_j) + \delta$

в) $d = \frac{\sum_{j=1}^{n} m_{ij}\, d(X_i,X_j)}{\sum_{j=1}^{n} m_{ij}}$

Здесь δ — некоторая константа; $m_{ij} > 0$ — весовые множители. Они задаются либо эвристически, либо на этапе разведочного анализа служат управляющими параметрами.

3.1.2 Последовательные процедуры

Простой последовательный алгоритм классификации

Приведем простой последовательный алгоритм классификации, в основе которого лежит предположение, что представители одного класса не могут быть удалены друг от друга более чем на заданную пороговую величину.

Пусть на классификацию объекты поступают последовательно, например по одному. Если исходная информация представлена в форме матрицы «объект — свойство», то параметрами алгоритма являются функция близости (расстояние) $d(O_i, O_j)$ между объектами и пороговое значение d_0; если исходная информация представлена матрицей взаимных расстояний, то единственным параметром является пороговое значение d_0.

Формальная схема алгоритма

1. Первый объект O_1 объявляется ядром e_1 первого класса.
2. Пусть на m-ом шаге выделено k классов с ядрами $e_1,...,..., e_k$.

Для поступившего объекта O_i:

если $d(O_m,e_1) < d_0$, то O_i относим к первому классу;

если $d(O_m, e_{L-1}) > d_0$ и $d(O_m, e_L) > d_0$, $2 < L < k$, то O_m относим к L-му классу;

если $d(O_m, e_L) > d_0$, $1 <L< k$, то O_m объявляется ядром e_{k+1} нового $(k + 1)$-го класса.

Если функция $d(O_i, O_j)$ удовлетворяет неравенству треугольника, как, например, когда $d(O_i, O_j)$ — метрика, то объекты, отнесенные алгоритмом к одному классу, удалены друг от друга не более чем на $2d_0$.

3.2 Алгоритмы, использующие понятие центра тяжести

При решении практических задач полезно иметь набор простых быстродействующих алгоритмов классификации для выработки первых представлений о структуре данных в признаковом пространстве. Таким алгоритмам посвящен этот параграф. Модификации алгоритмов, приведшие к ряду важных многопараметрических семейств их, ориентированных на проверку более сложных гипотез, возникающих уже в ходе исследования, описаны ниже в этой главе.

Пусть исходная информация о классифицируемых объектах представлена матрицей «объект — свойство», столбцы которой задают точки p-мерного евклидова пространства.

3.2.1 Параллельные процедуры

Опишем один из наиболее известных алгоритмов, модификациями которого являются многие важнейшие алгоритмы, приведенные далее.

Алгоритм k-средних

Единственным управляющим параметром является число классов, на которые проводится разбиение $S = (S_1, \dots, S_k)$ выборки X. В результатеполучается несмещенное разбиение $S_l^* = (S_1^*, \dots, S_k^*)$.

Формальная схема алгоритма k-средних

1. Выберем начальное разбиение $S^\circ = \left(S_1^\circ, \dots, S_k^\circ\right)$, где

$$S_l^\circ = \left\{X_{l1}^\circ, \dots, X_{ln_l}^\circ\right\}$$

$$\bigcup_1^k S_l^\circ = \boldsymbol{X}$$

$$S_l^{\circ} \cap S_{l'}^{\circ} = \emptyset, l \neq l'$$

2. Пусть построено m-е разбиение $S^m = (S_1^m, \dots, S_k^m)$. Вычислим набор средних $e^m = (e_1^m, \dots, e_k^m)$, где

$$e_l^m = \frac{1}{n_l} \sum_{j=1}^{n_l} X_{lj}^m$$

3. Построим минимальное дистанционное разбиение, порождаемое набором e^m и возьмем его в качестве $S^{m+1} = (S_1^{m+1}, \dots, S_k^{m+1})$, т. е.

$$S_1^{m+1} = \left\{ X \in \boldsymbol{X} : d(X,\ e_1^m) = \min_{1 \le l \le k} d(X,\ e_l^m) \right\}$$

……

$$S_l^{m+1} = \left\{ X \in \boldsymbol{X} \setminus \bigcup_{1}^{l-1} S_{l'}^{m+1} : d(X, e_l^m) = \min_{l \le l' \le k} d\left(X, e_{l'}^m\right) \right\}, 2 \le l \le k$$

где $d(X, e) = ||\boldsymbol{X} - e||$ — расстояние в R^p.

4. Если $S^{m+1} \neq S^m$, то переходим к п. 2, заменив m на m+1, если $S^{m+1} = S^m$, то полагаем $S^m = S^*$ изаканчиваем работу алгоритма.

Введем расстояние $d(\boldsymbol{X}, e)$ от точки $X \in R^p$ до множества$e = (e_1, \dots, e_k)$, где $e_l \in R^p$ по формуле $d(X, e) = \min_{1 \le l \le k} d(X,\ e_l^m)$. Тогда можно рассмотреть статистический разброс выборки $\boldsymbol{X}$ относительномножества $e = (e_1, \dots, e_k)$:

$$F(\boldsymbol{X}; e) = \sum_{X \in \boldsymbol{X}} d(X, e)^2$$

Определим статистический разброс разбиения $S = (S_1, \dots, S_k)$ выборкиX как разброс этой выборки относительно множества $e(S) = (e_1(S), \dots, e_k(S))$, где $e(S)$ — средний вектор класса S_l т. е. положим $F(S) = F(\boldsymbol{X}; e(S))$.

Непосредственно из построения минимального дистанционного разбиения следует формула

$$F(S) = \sum_{l=1}^{k} \sum_{X \in S^l} \left|\left|X - e_l(S)\right|\right|^2$$

Так как на последовательности разбиений S°, S^1,...,S^m,..., котораястроится в алгоритме k-средних, функционал $F(S)$ не возрастает, причем $F(S^m) = F(S^{m+1})$, толькоесли $S^m = S^{m+1}$, то для любогоначального разбиения S°алгоритм через конечное число шагов заканчивает работу.

Содержательно процедура алгоритма k-средних направлена на поиск разбиения S^* выборки $\boldsymbol{X}$ с минимальным разбросом.

В ряде случаев начальное разбиение S° задается как минимальное дистанционное разбиение, порожденное некоторым набором точек $e^{\circ} = \left(e_1^{\circ}, \dots, e_k^{\circ}\right)$. Результат классификации зависит от выбора e°. Обычно для проверки устойчивости результата рекомендуется варьировать выбор e°. В тех случаях, когда из априорных соображений нельзясразу выбрать число классов k, его находят либо перебором, либо вместо алгоритма k-средних используется алгоритм ИСОМАД (Isodata), в котором k является параметром,
настраиваемым в ходе классификации.

Демонстрация алгоритма

Имеется выборка наблюдений (рис. 3.2.1.1). В алгоритме k-средних необходимо заранее знать число классов (кластеров), на которое будет разбиваться множество наблюдений. Пусть число кластеров равно 3.

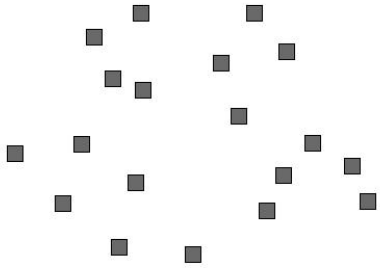

Рисунок 3.2.1.1

1. На первом шаге случайным образом выбираем из выборки 3 элемента и помечаем их как центры масс новых кластеров (рис. 3.2.1.2). Таким образом у нас 3 кластера, каждый из которых состоит из одного элемента.

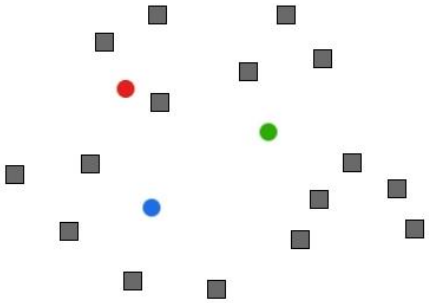

Рисунок 3.2.1.2

2. Далее пробегаемся по всем элементам, не состоящим ни в каком кластере, и добавляем текущий элемент в тот кластер, расстояние (вычисленное по выбранной метрике) от центра масс которого до этого элемента минимально. В результате этого все элементы будут разнесены по кластерам (рис. 3.2.1.3).

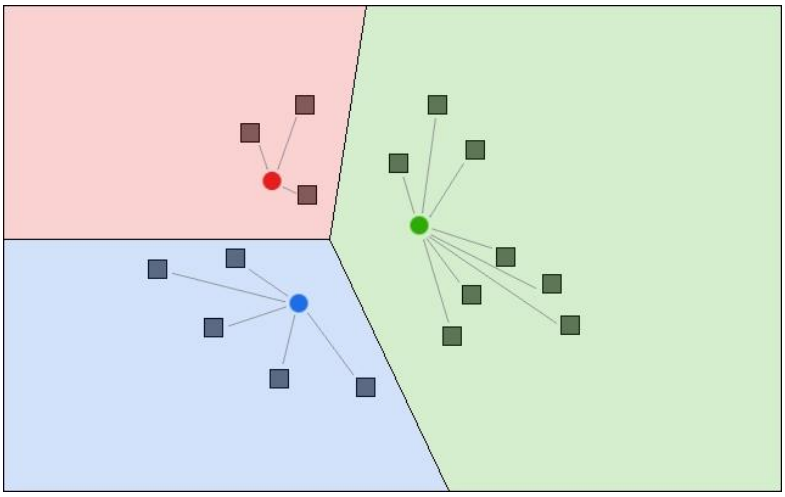

Рисунок 3.2.1.3

3. Вычисляем новый центр масс для каждого из кластеров (рис. 3.2.1.4).

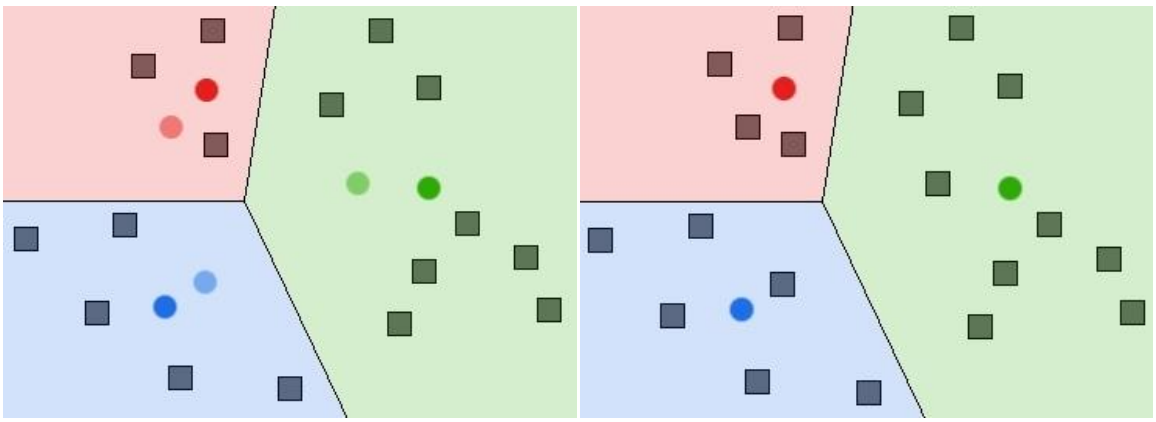

Рисунок 3.2.1.4

4. И снова производим перераспределение элементом по кластерам, основываясь на вычисленных центрах масс, производя процедуру, описанную выше в п.2 (рис. 3.2.1.5).

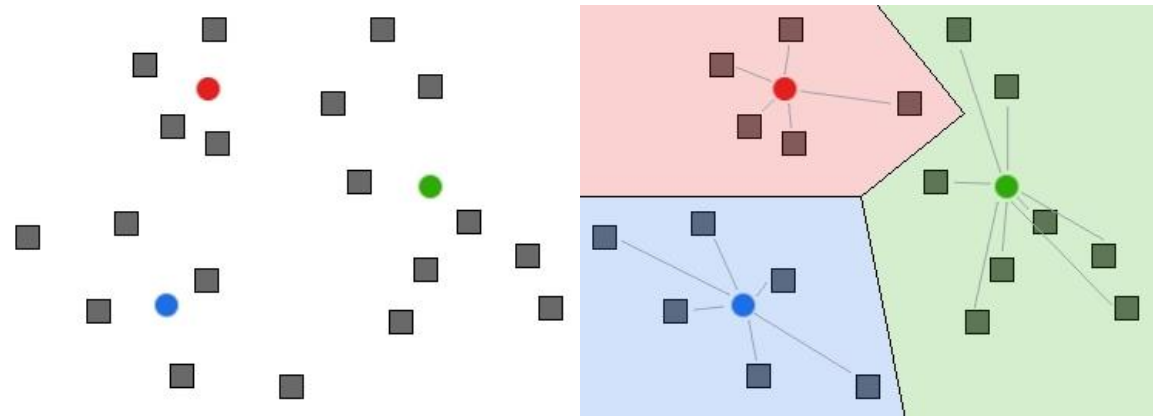

Рисунок 3.2.1.5

5. Повторяем шаги 3 и 4 до тех пор, пока центры масс всех кластеров не будут изменяться (центр масс кластера на текущем шаге равен центру масс на предыдущем) (рис. 3.2.1.6 и рис. 3.2.1.7).

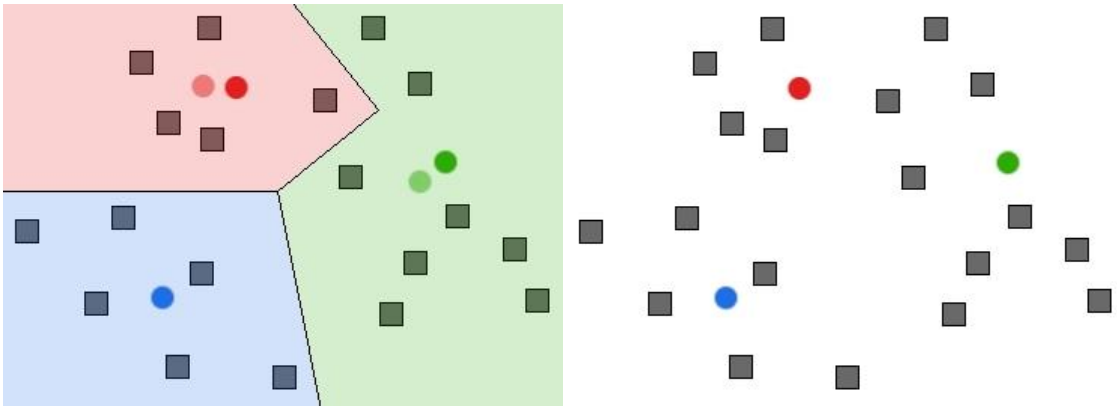

Рисунок 3.2.1.6

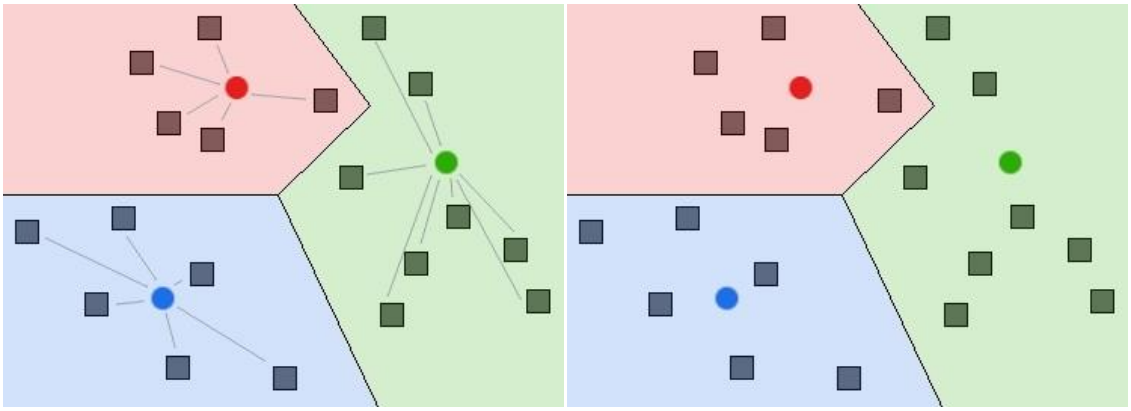

Рисунок 3.2.1.7

Алгоритм Форель

Первоначальное название ФОРЭЛ — ФОРмальный ЭЛемент. Единственным управляющим параметром является порог r — радиус шаров, которыми покрывается выборка $\boldsymbol{X}$. Пусть $D_r(e) \subset R^p$ — шар радиуса r с центром в точке e. Подвыборка $\boldsymbol{X}^1 = \boldsymbol{X} \cap D_r(e)$ называется несмещенной в $D_r(e)$, если ее средний вектор совпадает с e. Классификация при помощи алгоритма Форель разбивается на несколько последовательных этапов. На первом в выборке $\boldsymbol{X}$ выделяется несмещенная подвыборка $\boldsymbol{X}_1$ в некотором $D_r(e_1)$, которая объявляется

первым таксоном. На втором этапе та же процедура применяется к выборке $\boldsymbol{X} \setminus \boldsymbol{X}_1$. Таким образом, достаточно описать алгоритм только для первого этапа.

Формальная схема алгоритма

1. Выберем начальное разбиение $S^\circ = (S_1^\circ, S_2^\circ)$ выборки $\boldsymbol{X}$.

2. Пусть построено m-е разбиение $S^m = (S_1^m, S_2^m)$. Вычислим средний вектор e^m класса S_1^m.

3. Построим разбиение $S^{m+1} = (S_1^{m+1}, S_2^{m+1})$, где $S_1^{m+1} = \{X \in \boldsymbol{X}: d(X, e^m) \leq r\}$; $S_2^{m+1} = \boldsymbol{X} \setminus S_1^{m+1}$.

4. Если $S^{m+1} \neq S^m$, то переходим к п. 2, заменив m на m +1, если $S^{m+1} = S^m$, то полагаем $S^m = \boldsymbol{X}_1$ и заканчиваем работу первого этапа алгоритма.

Пополним пространство R^p «точкой» $*$, такой, что $d(X,*) \leq r$ для всех $X \in R^p$. Тогда статистический разброс выборки $\boldsymbol{X}$ относительно множества $\boldsymbol{e} = (e,*)$, где $e \in R^p$, запишется в виде

$$F(\boldsymbol{X}, \boldsymbol{e}) = \sum_{X \in \boldsymbol{X}} d(X, \boldsymbol{e})^2 = \sum_{X \in S_1} \left||X - e|\right|^2 + r^2|\overline{S_1}|$$

где $S_1 = \left\{X \in \boldsymbol{X} \cdot d(X, e) = \left||X - e|\right| \leq r\right\}$, $|\overline{S_1}|$ — число элементов в множестве $\boldsymbol{X} \setminus S_1$. При помощи функционала $F(\boldsymbol{X}, \boldsymbol{e})$ показано, что последовательность разбиений S°, S^1,..., S^m,..., которая строится на первом этапе алгоритма Форель, стабилизируется и алгоритм через конечное число

шагов заканчивает работу.

Применение алгоритма Форель для ряда последовательных значений $r_\nu - r_0 - \nu\Delta$, где $\Delta = \frac{r_0}{N}$, $\nu = 1, 2, \ldots, N-1$, позволяет оценить наиболее предпочтительное число классов для данной выборки. При этом основанием для выбора числа классов может служить многократное повторение одного и того же числа классов для нескольких последовательных значений r_ν и его резкое возрастание для следующего

шага по ν. На основе алгоритма первого этапа Форели строится целое семейство алгоритмов, целью которых является разбиение выборки на заданное число классов, покрытие выборки $\boldsymbol{X}$ областями более сложной формы, чем шары, и т. п. Имеется модификация алгоритма первого этапа Форели, в которой порог r является параметром, настраиваемым в ходе поиска первого сгустка $\boldsymbol{X}_1$ (см. алгоритм Пульсар в п. 3.3.1).

Демонстрация алгоритма

Допустим, было дано некоторое множество классифицируемых объектов (рис 3.2.1.8).

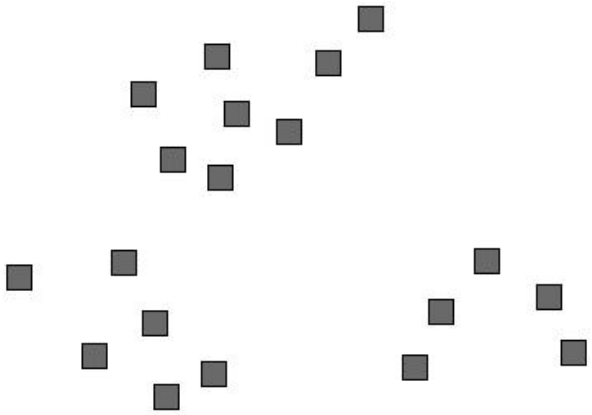

Рисунок 3.2.1.8

1. Построить гиперсферу радиуса R_0 охватывающую все множество точек (рис 3.2.1.9):

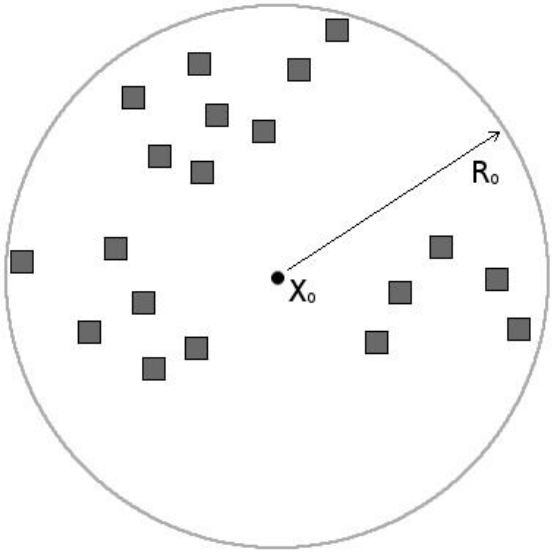

Рисунок 3.2.1.9

2. Установить радиус гиперсферы $R_1 = 0.9R_0$ и перенести центр сферы в любую из внутренних точек (расстояние до которых меньше радиуса) (рис 3.2.1.10):

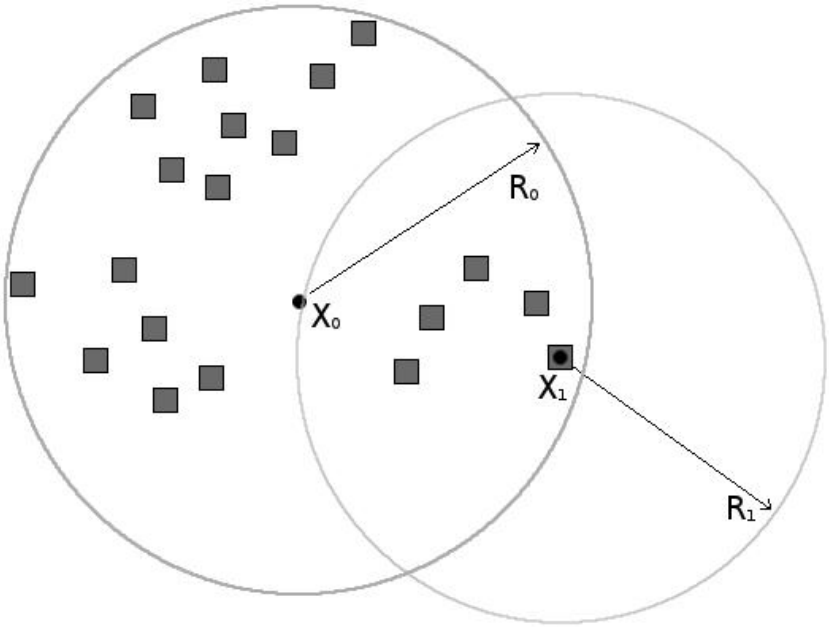

Рисунок 3.2.1.10

3. Вычислить новый центр масс и перенести в него центр сферы R_1 (рис 3.2.1.11):

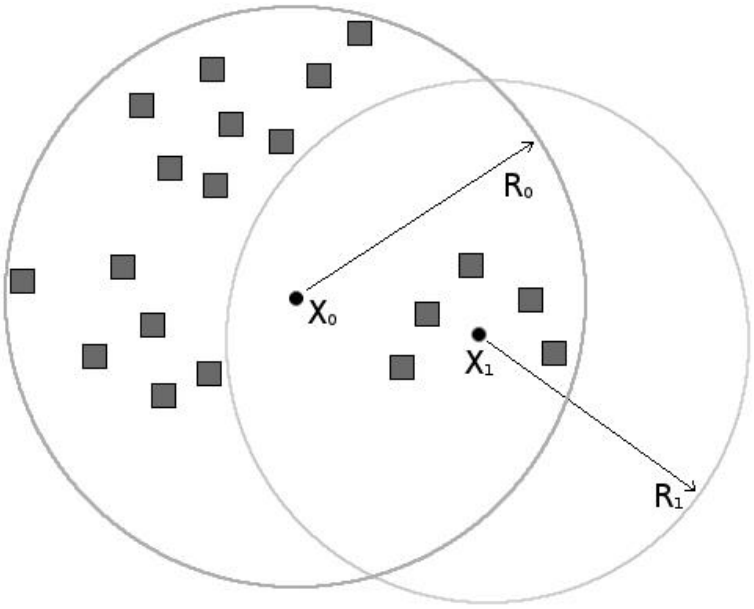

Рисунок 3.2.1.11

4. Если новый центр масс отличается от предыдущего необходимо вернуться к шагу 2 и повторить цикл. Цикл будет повторяться до тех пор пока центр тяжести не перестанет смещаться. Таким образом, центр сферы перемещается в область локального сгущения точек. В предложенном примере центр сферы $X_1 \neq X_0$, поэтому: необходимо установить новый радиус сферы $R_2 = 0.9R_1$ и перенести центр сферы в произвольную внутреннюю точку (рис 3.2.1.12):

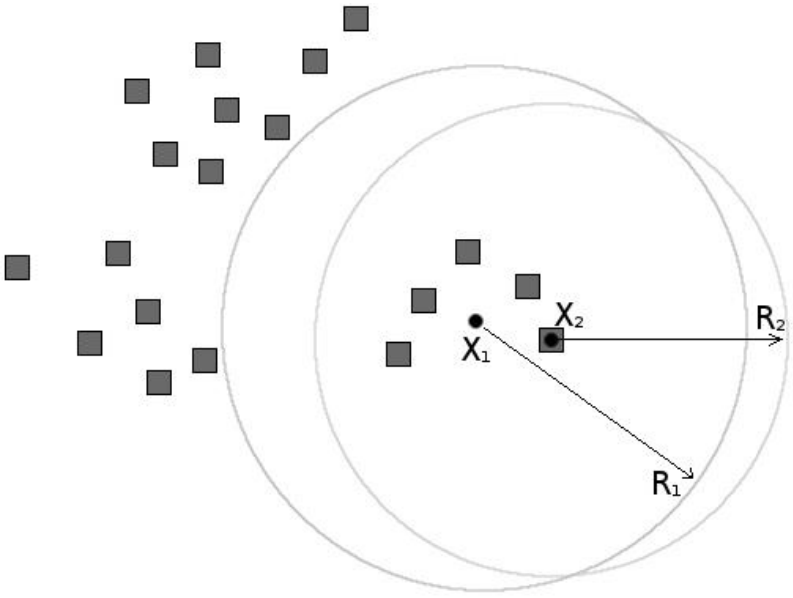

Рисунок 3.2.1.12

5. Вычислить новый центр тяжести и перенести в него центр сферы. Новый центр тяжести $X_2 = X_1$, поэтому внутренние точки текущей сферы объединяются в кластер (рис 3.2.1.13):

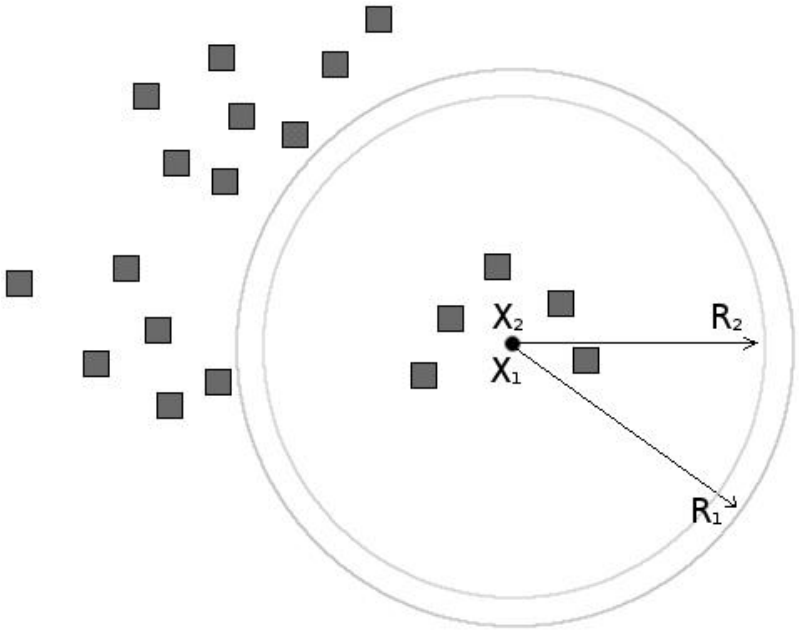

Рисунок 3.2.1.13

6. Точки, принадлежащие новому кластеру, исключаются из анализа, и работа алгоритма повторяется с п.1. И так до тех пор, пока все точки не будут исключены из анализа.

3.3 Алгоритмы с управляющими параметрами, настраиваемыми в ходе классификации

3.3.1 Параллельные процедуры

Рассматриваемые ниже алгоритмы ИСОМАД и Пульсар являются модификациями соответственно алгоритмов k-средних параллельного типа и Форель.

Алгоритм ИСОМАД (Isodata)

ИСОМАД — Итеративный Самоорганизующийся Метод Анализа Данных.

Основной процедурой в этом алгоритме, как и в алгоритме k-средних, является минимальное дистанционное разбиение, порожденное набором центров. Число классов заранее не фиксируется, а определяется в ходе классификации. Для этого используется ряд вспомогательных эвристических процедур, параметрами которых регулируются характеристики межклассовой и внутриклассовой структуры выборки на этапах классификации. Конфигурация (схема) ИСОМАД не является фиксированной, ее развитие отражает богатый опыт практического применения этого алгоритма.

Опишем наиболее распространенный вариант.

Параметры, определяющие процедуру классификации:

k — предполагаемое число классов;

k_0 — начальное (разведочное) число классов;

θ_n — минимально допустимое число элементов в классе (функция от n, где n — число элементов во всей выборке);

θ_s — порог внутриклассового разброса;

θ_c — порог межклассового разброса;

Q — максимально допустимое количество пар центров классов, которые можно объединить;

I — допустимое число циклов итерации.

Конкретные значения параметров задаются на основе априорной информации либо на этапе разведочного анализа выбираются из общих соображений, а затем корректируются от итерации к итерации.

Пусть на классификацию поступила выборка X — $\{X_1, ..., X_n\}$, где $X_i \in R^p$. Выберем начальный набор центров с $e^0 = (e^0_1, \ldots, e^0_{k0})$.

Формальная схема алгоритма

1. Выбираются значения параметров.

2. Строится минимальное дистанционное разбиение S — $(S_1, ..., S_{k0})$ выборки X, порожденное набором центров.

3. Пусть n_L — число элементов в классе S_L. Составляется $\bar{S} = (\bar{S}_1, ..., \bar{S}_{km})$ из классов S_L разбиения S, у которых $n_L \geq \theta_n$, где k_m— полученное (текущее) число классов. $\bar{S}$ присваивается обозначение $S = (S_1, ..., S_{km})$.

4. Вычисляется набор центров $e = (e_1, \ldots, e_{km})$ из средних векторов классов, входящих в разбиение S.

5. Вычисляется вектор $D = (D_1, \ldots, D_{km})$, где

$$D_L = \frac{1}{n_L} \sum_{X \in S_L} \|X - e_L\|, L = 1, \ldots, k_m$$

6. Вычисляется

$$\overline{D} = \frac{1}{n} \sum_{L=1}^{k_m} n_L D_L$$

7.

а) Если текущий цикл итерации последний, то переход к 11;

б) если $k_m \leq k/2$, то переход к 8;

в) если текущий цикл итерации имеет четный порядковый номер или $k_m \geq 2k$, то переход к 11; в противном случае переход к 8.

8. Для каждого класса S_L вычисляется вектор $б_L = (б^1_{L,\ldots,} б^p_L)$, где

$$б^j_L = \sqrt{\frac{1}{n_L}\sum_{X \in S_L}(X^j - e_L^j)^2}, \text{ где } j = 1,\ldots,p;\ L = 1,\ldots,k_m;$$

9. В каждом векторе a, отыскивается координата

$$б^{jL}_L = max_{1 \le j \le p}(\sigma_L^j),\ 1 \le L \le k_m$$

10. Если $б^{jL}_L > \theta_s$ для некоторого L, причем

а) $D_L > \overline{D}$ и $n_L > 2(\theta_n + 1)$

или

б) $k_m \le k/2$,

то класс S_L с центром e_L расщепляется на два новых класса S_L^+, S_L^- с центрами e_L^+, e_L^-, где соответственно $e_L^{\pm} = e_L \pm \varepsilon_L$, $\varepsilon_L = (\varepsilon_L^1,\ldots, \varepsilon_L^p)$ и $\varepsilon_L^j = 0$, если $j \ne j_L$, $\varepsilon_L^{jL} = \gamma б^{jL}_L$, $0 < \gamma \le 1$.

Если расщепление класса на этом шаге происходит, то переход к 2 с набором центров

$$(e_1,\ldots,e_{L-1}, e_L^+, e_L^-, e_{L+1,\ldots,} e_{km}),$$

в противном случае переход к 11.

11. Вычисляется матрица (d_{ij}) взаимных расстояний между центрами классов

$$d_{ij} = \|e_i - e_j\|$$

12. Расстояния d_{ij}, где $i < j$, сравниваются с порогом θ_c. Пусть $d_{i_1 j_1} \le d_{i_2 j_2} \le \cdots \le d_{i_{Q1} j_{Q1}}$ — упорядоченная последовательность тех из них, которые меньше θ_c. Вычеркнем из этой последовательности $d_{i_{Q1} j_{Q1}}$, если и только если в наборе $i_1, i_2, \ldots i_{Q1-1}$ встречается индекс i_{Q1} либо в наборе $j_1, j_2, \ldots j_{Q1-1}$ встречается индекс j_{Q1}. Проделаем аналогичную операцию с $d_{i_{Q1-1} j_{Q1-1}}$ и так далее до $d_{i_2 j_2}$.

Пусть $d_{l_1 t_1} \le d_{l_2 t_2} \le \cdots \le d_{l_{Q2} t_{Q2}}$ — полученная в результате последовательность. Заметим, что по построению $i_1 = l_1$, $j_1 = t_1$. Положим $q = \min(Q, Q_2)$.

13. Слияние классов. Для каждой пары (l_i, t_i), 1 < i < q классы S_{l_i}. и S_{t_i} сливаются в класс $S_{l_i} \cup S_{t_i}$. Непосредственно из 12 следует, что, если на предыдущем шаге было k_m

классов, то теперь остается k_m - q классов совокупности, которым переиндексацией присваивается обозначение S = (S_1, ..., $S_{k_m-q_n}$). Вычисляется набор центров e = (e_1, ..., e_{k_m-q}) средних векторов классов, входящих в S.

14. Если текущий цикл итерации последний, то алгоритм заканчивает работу. В противном случае переход к 1, если пользователь решил изменить какой-либо из параметров алгоритма, либо переход к 2, если в очередном цикле итерации параметры не меняются. Завершением цикла итерации считается каждый переход к 1 либо к 2.

Алгоритм Пульсар

Этот алгоритм, как и алгоритм Форель, состоит из последовательности одинаковых этапов, на каждом из которых выделяется один компактный класс (сгусток). Но радиус шара (величина окна просмотра) не фиксируется, а меняется (пульсирует) в ходе классификации. Для этого в алгоритм включены управляющие параметры, позволяющие поиск окончательного радиуса реализовать в виде процедуры стохастической аппроксимации.

Опишем этап выделения одного сгустка.

Параметры, определяющие процедуру классификации:

r_{min}, r^{max} — минимальный и максимальный радиусы;

n_{min}, n^{max} — минимальное и максимальное число элементов в классе;

$v_{доп}$ — допустимое число колебаний радиуса. (Говорят, что произошло колебание радиуса, если $\Delta r_m \times \Delta r_{m+1} < 0$, где Δr_m = r_m – r_{m-1}, r_m — значение радиуса на m-м шаге);

δ - порог, регулирующий скорость изменения радиуса.

Формальная схема алгоритма

1. Выберем начальный центр e^0 и значения параметров.

2. Для радиуса $r_0 = \frac{r_{min}+r^{max}}{2}$ построим класс $S^0 = \{ X \in X: \|X - e^0\| \leq r_0\}$, вычислим число элементов n_0 в классе S^0 и присвоим v (числу колебаний радиуса) значение $v_0 = 0$.

3. Пусть на m-м шаге для центра e^m выбран радиус r_m,

построен класс $S^m = \{ X \in X: \|X - e^m\| \leq r_m\}$, подсчитано число его элементов n_m и значение $v = v_m$.

Положим

$$e^{m+1} = \frac{1}{n_m} \sum_{X \in S^m} X$$

$$r_{m+1} = \begin{cases} \min(r + \gamma\delta, r^{max}), \text{если } n_m \leq n_{min} \\ \max(r - \gamma\delta, r_{min}), \text{если } n_m > n^{max} \text{или } e^{m+1} \neq e^m, \text{причем } v_m < v_{\text{доп}} \\ r_m - \text{ в остальных случаях} \end{cases}$$

Здесь $\gamma = (1 + v_m)^{-1}$, Порог $v_{доп}$ учитывается при выборе радиуса r_{m+1} только тогда, когда $e^{m+1} = e^m$ и одновременно $n_m > n^{max}$.

Далее положим

$v_1 = v_0 = 0$ и для $m \geq 1$

$$v_{m+1} = \begin{cases} v_m, \Delta r_m * \Delta r_{m+1} \geq 0; \\ v_m + 1, \Delta r_m * \Delta r_{m+1} < 0; \end{cases}$$

4. Если $e_{m+1} = e_m$, $r_{m+1} = r_m$, то алгоритм заканчивает работу, в противном случае переходим к 3, заменив m на m+1.

ВЫВОДЫ

1. Описаны наиболее известные и хорошо зарекомендовавшие себя при решении прикладных задач алгоритмы разбиения исследуемой совокупности объектов на классы как при известном, так и при неизвестном заранее числе классов.

2. Общим для всех рассмотренных алгоритмов является то, что в них распределение объектов по классам (классификация) осуществляется при помощи сформированного в ходе классификации набора «ядер» классов.

3. Алгоритмы различаются правилами распределения объектов по классам, типом ядер, тем, является ли набор управляющих параметров фиксированным или настраиваемым в ходе классификации, а также тем, как поступают объекты на классификацию: вся совокупность сразу или порциями по одному, по нескольку.

БИБЛИОГРАФИЧЕСКИЙ УКАЗАТЕЛЬ

1. Прикладная статистика: Классификации и снижение размерности. Справ.изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; Под ред. С. А. Айвазяна.— М.: Финансы и статистика, 1989.— 607 с: ил. ISBN 5—279—00054—X.
2. Прикладная статистика. А.И. Орлов. — М.: Экзамен, 2004.— 483 с: ISBN 5-472-01122-1.
3. Геоинформатика: Учеб. для студ. вузов / Е.Г. Капралов, А.В. Кошкарев, В.С. Тикунов и др; Под ред. В.С. Тикунова — М: Издательский центр «Академия», 2005. — 480 с.: цв. ил.