Санкт-Петербургский Государственный Университет

Математико-Механический факультет

Кафедра Системного Программирования

Оценка обоснованности кластеризации для данных высокой размерности

Дипломная работа студентки 544 группы

Сивоголовко Елены Владимировны

Допущена к защите

Заведующий кафедрой:
д.ф.м.н., профессор Терехов А. Н.

Научный руководитель:
д.ф.м.н., профессор Новиков Б. А.

Рецензент:
ассистент кафедры информатики Васильева Н. С.

Санкт-Петербург
2008 г.

Содержание

1 Введение

Кластеризация является одной из задач логического анализа данных. Ее предназначение формально можно определить как разбиение совокупности объектов на однородные по какому-то признаку группы – кластеры. Главное отличие кластеризации от классификации состоит в том, что первая из них не имеет никаких сведений о классах, которые ей предстоит выделить, за исключением, в лучшем случае, их числа или размера. Наибольшее распространение кластеризация получила в биологии, биоинформатике, маркетинговых исследованиях, информационном поиске и даже в психологии.

В настоящее время, разработано множество различных алгоритмов кластеризации. Их сложность зависит от размерности данных, объема кластеризуемого множества, области применения и так далее.

Проблема оценки обоснованности полученной структуры кластеров неотделима от задачи кластеризации. (Потому что каждый раз, когда вы получаете какой-либо результат, вам хочется знать на сколько он точен и достоверен.) Иногда степень обоснованности хочется назвать качеством кластеризации, но это не совсем верно, поскольку качество кластеров можно определить с двух разных сторон.

Первое определение рассматривает качество кластеризации как степень соответствия между полученным разбиением и реально существующими классами элементов целевого множества. Такая трактовка качества вполне уместна, поскольку практически у каждого метода кластеризации есть свои причины ошибаться в распределении объектов по классам. Поводом для получения некорректных результатов может послужить неправильно указанное число кластеров(в тех алгоритмах, где это необходимо), большая размерность кластеризуемых данных, чувствительность используемого метода к выбросам(отдельно стоящим элементам целевого множества), отсутствие гауссова распределения данных при работе с алгоритмами, предполагающими его наличие, неверно выбранная метрика схожести элементов и прочее.

Относительно сказанного выше, оценка качества кластеризации может проводится следующими способами

1. проверкой вручную. (Как ни странно, но это самый точный и часто используемый метод проверки)

2. установлением контрольных точек и проверка на полученных кла-

стерах.

3. определением стабильности кластеризации путем добавления в модель новых переменных.

4. созданием и сравнением кластеров, полученных с использованием различных методов.

Второй подход к трактовке качества принято называть обоснованностью кластеризации(в английской терминологии – cluster validity). Его основная идея заключается в том, что оценка качества кластеризации сводится к оценке свойств полученной структуры кластеров.

Важнейшими из этих свойств являются:

1. *Компактность кластеров* – Элементы каждого кластера должны быть расположены как можно ближе друг к другу.

2. *Отделимость кластеров* – Кластеры должны быть хорошо отделимы один от другого.

Очень часто их называют критериями обоснованности. Иногда ним добавляют еще и третий: близость элементов кластера к его центру [1].

Таким образом, кластеризация является обоснованной, а следовательно, качественной, если полученные кластеры, компактны и хорошо отделимы друг от друга. (Здесь следуте заметить, что отдельной проблемой является оценка отделимости в случае нечеткой кластеризации, когда целью является получение пересекающихся кластеров.)

Функции, предназначенные для оценки обоснованности кластеризации, называют метриками или индексами обоснованности.

Цель данной работы состоит в том, чтобы ввести новую метрику обоснованности, дающую правильную оценку компактности кластеров, предоставляющую возможность визуальной оценки и допускающую обобщение измерения отделимости на случай нечеткой кластеризации. Метрики, существующие на данный момент совокупностью указанных выше свойств не обладают.

2 Постановка задачи

Одной из многочисленных областей применения кластеризации является поиск изображений по содержанию. Существует активно разрабатываемая и на данный момент подтвержденная рядом экспериментов гипотеза

о том, что для различных семантических классов изображений будет полезно использовать различные алгоритмы поиска по содержанию [2–3], а выделение классов из множества данных – прямая задача кластеризации.

Кластеризация векторных представлений изображений сопряжена со многими трудностями. Во-первых, все вектора являются многомерными и, в зависимости от алгоритма построения вектора и его набора параметров, могут состоять из десятков или сотен компонент, а в некоторых случаях, представлением изображения может быть даже группа векторов высокой размерности. Во-вторых, в большей части цветовых пространств классическая евклидова метрика не пригодна для измерения схожести векторов изображений, поскольку не отражает реально существующие связи между картинками (это выведено опытным путем непосредственно в задаче поиска по содержанию). Поэтому для кластеризации приходится пользоваться функциями расстояния, дающими лучшие результаты для поиска, но в месте с тем не являющимися метриками по строгому математическому определению, это,например, метрика косинусов, взвешенная манхеттеновская метрика и многие другие. В-третьих, кластеры изображений, как правило, являются нечеткими, поскольку даже человеку не всегда просто определить к какому классу следует отнести ту или иную картинку.

Все вышеперечисленное делает оценку обоснованности кластеров изображений очень важной и интересной задачей. Как уже было упомянуто во введении, цель данной дипломной работы состоит в определений новой метрики обоснованности, сочетающей в себе возможность некоего "визуального"восприятия результатов кластеризации и верную оценку качества полученой кластерной структуры.

3 Обзор существующих подходов

3.1 Визуализация данных

Говоря об оценке обоснованности нельзя не упомянуть про визуализацию.

Под этим термином в логическом анализе данных чаще всего понимают любое визуальное представление кластеров и кластеризуемых объектов. Несмотря на то, что этот метод не дает метрических оценок, его

преимущества вполне понятны и объяснимы: графическая интерпретация информации воспринимается людьми гораздо лучше сухого набора цифр и атрибутов, предоставляемых метрическими индексами.

Визуализация является востребованной и динамично развивающейся областью. Одной из ее главных задач является корректное представление на двумерной или трехмерной плоскости данных многомерного пространства. Этот вопрос особенно актуален в таких предметных областях, как логический анализ текстовых данных и кластеризация генетического материала, где размерность элементов целевого множества может достигать нескольких тысяч.

3.2 Метрики оценки обоснованности

Способы оценки обоснованности кластеризации принято разделять на два больших класса, так называемые внутренние и внешние метрики. Их отличие состоит в том, что во внешних метриках используются данные, полученные априори. Чаще всего, это какие-то сведения о структуре классов в исходных данных. Внешние метрики широко используются для сравнения производительности различных методов кластеризации. Внутренние же метрики при подсчете степени обоснованности опираются только на то, что получено алгоритмом самостоятельно, без учета каких-либо внешних сведений.

Очевидно, что в большинстве случаев использовать внешние метрики не представляется возможным, поэтому здесь и далее в данной работе мы будем рассматривать исключительно внутренние метрики оценки обоснованности.

3.2.1 Статистические оценки

Методы математической статистики являются базовыми при измерении обоснованности результатов кластеризации. Иногда они используются сами по себе, иногда в качестве основы для более сложных метрик. Наиболее важные из статистических оценок могут быть перечисленых следующим образом:

1. Конечное значение целевой функции.

   Суть большинства алгоритмов сводится к нахождению экстремума некоторой целевой функции. Например, в алгоритмах к-средних,

это минимум суммы расстояний от каждого из объектов в кластере до центра этого кластера. По тому, насколько мало или велико полученное конечное значение, мы сможем сделать выводы о качестве работы алгоритма.

2. Среднее расстояние между элементами кластера.

   Метрика схожести, как и целевая функция, определяется в каждом алгоритме по-своему. Часто ее еще называют расстоянием между объектами кластера. Если продолжать примеры из к-средних, то там для этой цели обычно используется стандартная евклидова метрика. Среднее значение чаще всего вычисляется как среднее арифметическое: попарное расстояние между объектами в кластере суммируется и делится на общее число пар объектов.

3. Среднеквадратическое отклонение метрики схожести.

   Среднеквадратичное(или стандартное) отклонение широко используется в статистике, его следует понимать, как выражение разброса данных относительно среднего значения. Вычисляют стандартное отклонение следующим образом: для начала, как и в предыдущем случае, считается среднее значение метрики схожести:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

   где N - число элементов в данном множестве, а $x_1 \ldots x_N$ - собственно элементы. Затем, непосредственно отклонение, находится как:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

4. Среднее расстояние между объектами, не лежащими в одном кластере.

   Следует заметить, что при больших объемах данных и разбиении на большое число кластеров эта метрика несколько сложна для вычисления. Ее суть заключается в следующем: для каждого из элементов данного кластера мы подсчитываем расстояния между ним и объектами, не вошедшими в этот кластер. Потом полученные

значений усредняются, чаще всего, тем же самым средним арифметическим.

Примером применения статистических методов для оценки обоснованности может служить кластеризационный алгоритм CLUTO [2].

Главным недостатком статистических методов является то, что их результаты сложно интерпретировать относительно двух критериев, приведенных выше, и если, исходя из среднеквадратического отклонения, можно составить какое-то представление о компактности, то отделимость в статистике никак не учитывается.

3.2.2 Индексы обоснованности

Индексы, представленные в этой части, уже в гораздо большей степени опираются на критерии обоснованности. Их используют в том случае, когда надо оценить степень "качества"одного кластеризационного алгоритма по сравнению с другим аналогичным, или же для определения наилучшего набора параметров кластеризации. Стоит заметить, что все они предназначены для случая отделимых, не пересекающихся кластеров.

Введем некоторые обозначения

n_c	число кластеров
d	размерность пространства
$d(x, y)$	расстояние между двумя элементами кластера
c_i	i кластер
v_i	центр i кластера
N	общее число элементов в кластеризуемом множестве

- индекс Данна

$$D = \min_{i=1\ldots n_c} \left\{ \min_{j=i+1\ldots n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1\ldots n_c} (diam(c_k))} \right) \right\}$$

здесь

$$d(c_i, c_j) = \min_{x \in c_i, y \in c_j} \{d(x, y)\} \text{ и } diam(c_i) = \max_{x,y \in c_i} \{d(x, y)\}$$

В данном случае функцию $d(c_i, c_j)$ следует понимать как расстояние между кластерами, а $diam(c_i)$ - как диаметр кластера. Предполагается, что чем лучше объекты полученного разбиения отделены

друг от друга, тем меньше будет диаметр кластера, по сравнению с межкластерным расстоянием. Главными недостатками индекса Данна являются большая вычислительная сложность и чувствительность к выбросам. (Чувствительность к выборосам очень легко объяснить тем, что выбросы могут серьезно исказить оценку кластерного диаметра, который, как указано выше, чаще всего определяется максимальным расстоянием между элементами одного кластера).

- SD индекс
  При определении обоснованнности кластеров SD индекс ориентируется на разброс элементов в кластере и на полную отделимость кластеров друг от друга. Разброс элементов подсчитывается с помощью диспресии в каждом кластере и в кластеризуемом множестве в целом.

  дисперсия множества дисперсия кластера

  $$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}\left(x_k^p - \overline{x_k^p}\right)^2 \quad \sigma_{v_i}^p = \frac{1}{\|c_i\|}\sum_{k=1}^{\|c_i\|}\left(x_k^p - \overline{x_k^p}\right)^2$$

  $$\sigma(x) = \begin{bmatrix}\sigma_x^1\\ \vdots \\ \sigma_x^d\end{bmatrix} \qquad \sigma(v_i) = \begin{bmatrix}\sigma_{v_i}^1\\ \vdots \\ \sigma_{v_i}^d\end{bmatrix}$$

  Уровень среднего разброса для множества кластеров определяется следующим образом

  $$Scatt = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{\|\sigma(v_i)\|}{\|\sigma(x)\|}$$

  Как несложно догадаться, параметер $Scatt$ служит мерой компактности кластеров. Для измерения отделимости используется расстояние между центрами кластеров:

  $$Dis = \frac{\max_{i,j=1\ldots n_c}\left(\|v_j - v_i\|\right)}{\min_{i,j=1\ldots n_c}\left(\|v_j - v_i\|\right)}\sum_{k=1}^{n_c}\left(\sum_{i=1,i\neq j}^{n_c}\|v_j - v_i\|\right)^{-1}$$

В итоге конечным значением индекса SD будет $SD = \alpha * Scatt + Dis$, где α некоторый весовой параметр. Как несложно заметить, чем ниже величины $Scatt$ и Dis, тем ниже SD и тем лучше получилась кластеризация.

- RMSSDT и RS индексы.
  Эта группа индексов основана на среднеквадратичном отклонении и потому отделеность кластеров не учитывает. Физический смысл метрики RMSSDT заключается в измерении однородности кластеров, а поскольку кластеризация есть выделение однородных групп элементов, то чем ниже значение RMSSDT, тем лучше результат кластеризации.

$$RMSSTD = \sqrt{\frac{\sum_{i=1}^{n_c} \sum_{j=1}^{d} \sum_{k=1}^{n_{ij}} (x_k - \bar{x_j})^2}{\sum_{i=1}^{n_c} \sum_{j=1}^{d} (n_{ij} - 1)}}$$

  В данном случае n_{ij} это число элементов j размерности в i кластере.

  RS индекс трактуется как измерение непохожести кластеров. Принимаемые им значения расположены в промежутке от 0 до 1. Если в результате получилось 0 значит, что кластеры почти не отличимы друг от друга, 1 значит, что между кластерами существует значительное различие.

$$RS = \frac{SS_t - SS_w}{SS_t}$$

  где

$$SS_t = \sum_{j=1}^{d} \sum_{k=1}^{n_j} (x_k - \bar{x_j})^2 \text{ и } SS_w = \sum_{i=1}^{N} \sum_{j=1}^{d} \sum_{k=1}^{n_{ij}} (x_k - \bar{x_j})^2$$

4 Плотность кластерного ядра

В этой части будет описана введеная нами новая метрика обоснованности, для ее расчета нам понадобится определить следующие два понятия

4.1 Ядро кластера

Кластерным ядром, в общем смысле, называется часть элементов кластера, наиболее приближенная к его центру. Предполагается, что в этой области сосредоточено большинство данных кластера, чем дальше элемент от кластерного центра, тем больше вероятность что он окажется выбросом. Различия в точных определениях ядра в основном касаются того, какие именно элементы кластера следует считать близкими к центру и как эту близость подсчитывать. В данной работе мы определяем радиус кластерного ядра как

$$R_i = \bar{r}_i + \sqrt{2} * \sigma_i$$

здесь $\bar{r}_i$ это среднее расстояние между центром кластера и кластерными элементами, а σ_i среднеквадратичное отклонение вышеупомянутой величины.

Это определение неявным образом выведено из неравенства Чебышева: по меньшей мере 50% данных случайной выборки удалены не более чем на $\sqrt{2}$ среднеквадратичных отклонений от ее среднего значения. Таким образом, если рассматривать в качестве случайной выборки расстояния между центром кластера и его элементами, то можно утверждать, что не менее половины кластерных элементов попадет в гиперсферу, центр которой совпадает с центром кластера, а радиус равен R_i. Именно эту область кластера мы будем называть кластерным ядром. Стоит заметить, что наше определение, как впрочем и общее определение ядра предполагает, что элементы исходного множества имеют гауссово распределение. В общем случае, это конечно же неверно, но на практике данное допущение чаще всего работает.

4.2 Определение плотности

Обычно под плотностью кластера подразумевают дисперсию расстояний от каждого элемента до центра кластера. К сожалению, такой способ оценки, во-первых, чувствителен к выбросам, а во-вторых, может не давать адекватного результата даже на компактном, то есть вполне качественном с точке зрения обоснованности, кластере просто из-за его размера.

Мы предлагаем рассчитывать плотность кластера следующим обра-

зом

$$\rho_i = \frac{R_i}{\bar{N}_i}$$

В данном случае R_i это радиус кластерного ядра, определенный выше, а $\bar{N}_i$ число элементов, составляющих ядро.

Оценкой же для всего результата кластеризации будет являтся средняя плотность

$$\bar{\rho} = \frac{\rho_1 + \ldots + \rho_{n_c}}{n_c}$$

Таким образом мы получаем метрику обоснованности, которая легко вычислима, устойчива к выбросам и способна дать верное представление о компактности кластеров и близости элементов кластера к его центру. Плотность ядра никак не учитывает второй из критериев обоснованности, отделимость, но в данном случае это допустимо, поскольку кластеры изображений в большинстве своем не являются четкими. Проделанные эксперименты показывают, что не смотря на простоту введеная метрика способна давать адекватные результаты.

5 Эксперименты

5.1 Синтетические данные

В качестве синтетических данных используется классическая модель: множество точек на двумерной плоскости.

В данном случае, целевое множество содержит порядка 1500 элементов. Отметим что, в нем присутствуют нечетко отделимые кластеры и кластеры с не гауссовым распределением данных, что создает дополнительные трудности алгоритму кластеризации, в качестве которого выбран классический метод к-средних. Для определения расстояния между двумя элементами используется стандартная евклидова метрика: $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$. Результаты разбиений на различное число кластеров представлены на Рис. 1-4. Значения средней плотности ядра для каждого из этих результатов кластеризации показано на Рис. 5. Нетрудно заметить, что самому корректному разбиению(Рис. 2) соответствует максимальное значение плотности.

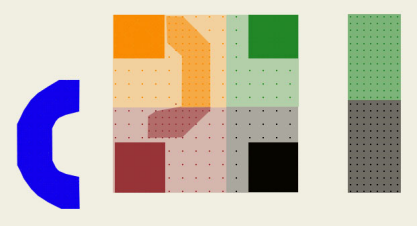


Рис. 1: разбиение на пять кластеров

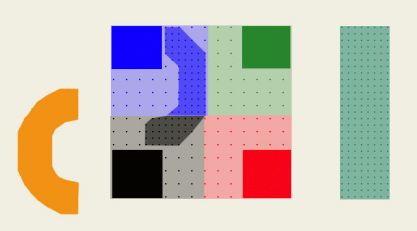


Рис. 2: разбиение на шесть кластеров(правильное)

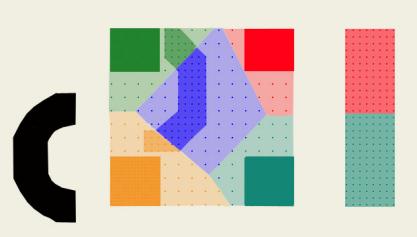


Рис. 3: разбиение на шесть кластеров(неправильное)

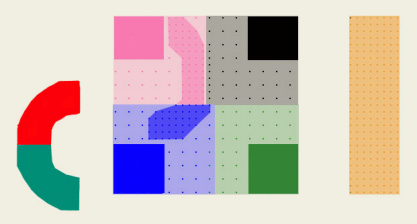


Рис. 4: разбиение на семь кластеров

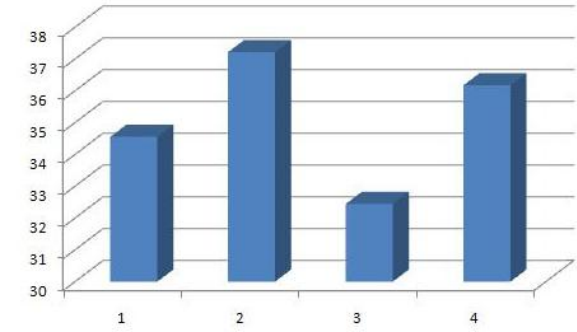


Рис. 5: средняя плотность ядра

5.2 Реальные данные

В качестве реальных данных используются две базы изображений, предназначенных для тестирования различных алгоритмов поиска картинок по содержанию в группе теории баз данных. Первая из них – ImageDBCorel – содержит 650 элементов, в ней вручную выделены девять семантических классов, вторая – CorelSmall-100 – состоит из 100 изображении, которые разбиты на шестнадцать различных семантических классов. Фотографии в этих базах взяты из коллекции CorelPhotoSet.

Кластеризация будет производится в пространстве цветовых моментов, построенном по представлению цвета в цветовом пространстве L*a*b, векторное представление картинки в нем получается следующим образом: из изображения выделяется пять регионов определенного цвета, и для каждого региона подсчитывается размер в пикселях и девять моментов – $\mu_a, \mu_b, \mu_L, \sigma_{aa}, \sigma_{ab}, \sigma_{bb}, \sigma_{La}, \sigma_{Lb}, \sigma_{LL}$. Таким образом получившиися вектор содержит 50 компонент(или 45, если не учитывать размер регионов). Заметим, что на выбраном пространстве признаков евклидова метрика показывает не лучшие результаты, и потому мы будем использовать следующие модификации метода к-средних:

- *сферический алгоритм к-средних*
  В качестве меры близости элементов используется косинусная метрика [10]: $d(x,y) = \dfrac{x^\dagger y}{\|x\|_2 \|y\|_2}$

- *алгоритм к-средних со взвешенной манхеттенской метрикой*
  В этом случае близость картинок оценивается следующим образом: для начала изображения разбиваются на регионы, обычно их бывает пять, и затем схожесть региона l изображения Q и региона k изображения K считается как

  $$d_{region}(Q_l, K_k) = \sum_{i \in L,a,b} \mid \frac{\mu_i(Q_l) - \mu_i(K_k)}{\alpha(\mu_i)} \mid + \sum_{i,j \in L,a,b} \mid \frac{\sigma_{ij}(Q_l) - \sigma_{ij}(K_k)}{\alpha(\sigma_{ij})} \mid$$

  здесь μ_i и σ_{ij} некоторые цветовые характиристики регионов, а $\alpha(\mu_i)$ и $\alpha(\sigma_{ij})$ – среднеквадратичные отклонения этих величин.
  Расстояние же между картинками в целом, находится по формуле

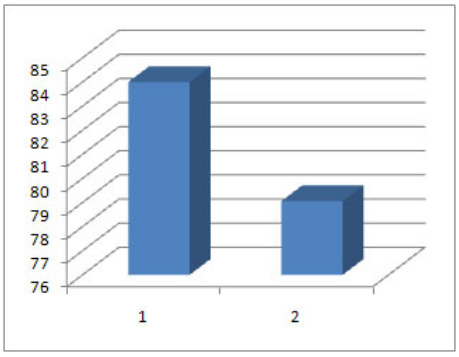
  $$d_{image}(Q,K) = N_Q \left(\frac{d_{region}(Q_0, K_0)}{N_{Q_0}} + \min_{f \in T_{90}} \sum_{l=1}^{4} \frac{d_{region}(Q_l, f(K)_l)}{N_{Q_l}} \right)$$

здесь N_Q - размер изображения в пикселях и N_{Q_l} - размер отдельного региона в пикселях. Физический смысл функции $f(K)$ - это поворот изображения K на 0, 90, 180 и 270 градусов. Данная метрика была впервые введена в статье [8].

Поскольку кластеры изображений являются нечетко отделимыми кластерами, то мы будем сравнивать обоснованность кластеризации не для различного числа кластеров, как в случае с синтетическими данными, а для разных распределений по кластерам, полученных с фиксированными параметрами кластеризации. Другими словами, мы будем разбивать имеющиеся изображения на число кластеров равное числу семантических классов и проверять, как изменяется значение плотности кластерного ядра в зависимости от того, похоже ли полученное распределение по кластерам, на распределение полученное вручную. В данном случае, нам потребуется определение сравнимости отделных кластеров, потому что без него мы не сможем оценить корректность среднего показателя плотности для кластеризации в целом. Поэтому, в контексте данной работы будет использоваться следующее определение сравнимости: кластер A сравним с кластером B по семантическому классу C, если оба кластера содержат по меньшей мере половину изображений из класса C. Примеры результатов сравнения отдельных кластеров на приведены на Рис. 6, Рис. 7 и Рис. 8

верная кластеризация

неверная кластеризация

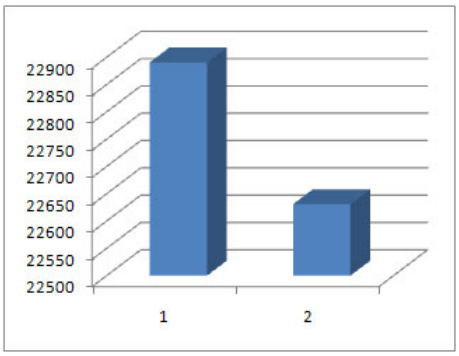


плотность ядра

Рис. 6: База данных ImageDBCorel. Алгоритм: сферический к-средних. Сравнение по семантическому классу "Облака".

верная кластеризация

неверная кластеризация

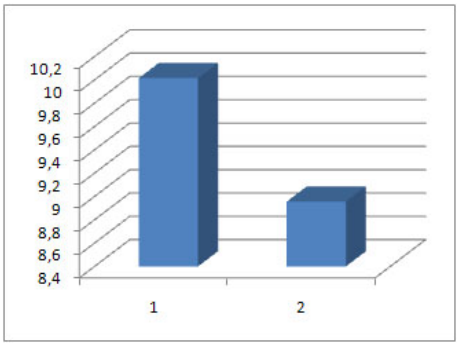


плотность ядра

Рис.7: База данных ImageDBCorel. Алгоритм: к-средних с манхеттенской метрикой. Сравнение по семантическому классу "Деревья".

верная кластеризация

неверная кластеризация



плотность ядра

Рис. 8: База данных CorelSmall-100. Алгоритм: сферический к-средних. Сравнение по семантическому классу "Кошки".

Таким образом, кластеры из более схожих изображений имеют лучшую оценку по плотности, что подтверждается экспериментами. На Рис. 9 представлена зависимость плотности от точности распределения картинок по кластерам для некоторых семантических классов базы CorelSmall-100. В качестве алгоритма кластеризации использовался сферический к-средних, для измерения точности кластеризации выбрана F-мера, которая рассчитывается следующим образом.

$$Precision(c_i) = \frac{p_+}{p_+ + n_+}$$

$$Recall(c_i) = \frac{p_+}{p_+ + p_-}$$

$$F1(c_i) = \frac{2 * Precision(c_i) * Recall(c_i)}{Precision(c_i) + Recall(c_i)}$$

В данных уравнениях p_+ обозначает число элементов семантического класса, попавших в кластер c_i, p_- – число элементов из того же семантического класса, но не попавших в c_i, а n_+ – оставшиеся элементы кластера c_i, принадлежащие другим семантическим классам.

Рис. 9: Зависимость плотности от точности распределения по кластерам

для некоторых семантических классов базы данных CorelSmall-100.

Как можно заметить, чем ближе результат кластеризации к реальному распределению по семантическим классам, тем лучше средний показатель плотности.

6 Реализация тестовой среды

Для проверки значимости введенной метрики была создана специальная тестовая среда, включающая в себя получение реальных или синтетических данных, непосредственно кластеризацию и оценку результатов. Ее ядром являются алгоритмы кластеризации, построенные на основе метода к-средних. Этот метод был выбран нами, потому что, во-первых, его достаточно просто реализовать, во-вторых, он дает возможность работать с различными метриками и, в-третьих, из-за случайного выбора начальных центров кластеров он способен давать результаты различного качества, что очень полезно при проверки точности метрики. Общая схема алгоритмов семейства к-средних пояснена на вкладке Algorithm 1.

Algorithm 1 K-Means

1: Проинициальзировать начальные центров кластеров C
2: Для каждого элемента x_i вычислить степень его принадлежности к кластеру с центром c_j, то есть функцию принадлежности$m(c_j \mid x_i)$ и вес элемента $w(x_i)$
3: Для всех кластерных центров пересчитать их положение

$$c_j = \frac{\sum_{i=1}^{N} m(c_j \mid x_i) w(x_i) x_i}{\sum_{i=1}^{N} m(c_j \mid x_i) w(x_i)}$$

4: Повторять шаги 2 и 3, до тех пор, пока кластеры не станут устойчивыми или не кончится лимит итераций

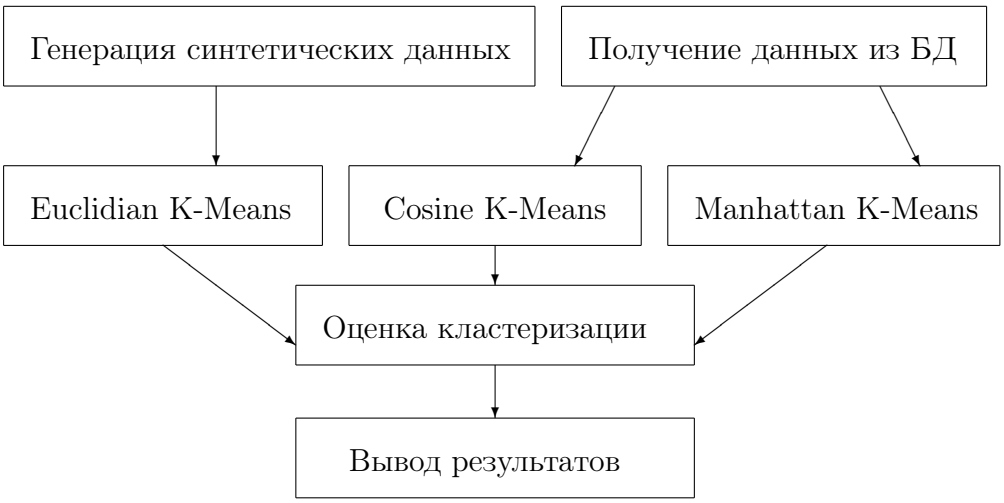
В нашем случае мы пользуемся к-средних с жесткой функцией рас-

пределения и постоянными весами

$$m(c_l \mid x_i) = \begin{cases} 1, \text{ если } l = argmin_j d(x_x, c_j) \\ 0, \text{ в противном случае} \end{cases}$$
$$w(x_i) = 1$$

здесь $d(x_x, c_j)$ – это метрика схожести. Другие варианты алгоритма более подробно описаны в [9].

Как уже упоминалось выше, для синтетических данных применяется исключительно евклидова метрика, а для реальных – косинусная и взвешенная манхетеннская.
Структура приложения в целом



В качестве языка реализации использовался C♯. Векторные представления картинок хранились в базе данных MSSQL 2005, доступ к серверу осуществлялся по технологии ADO.NET.

7 Заключение

7.1 Результаты

В данной работе мы ввели новую метрику обоснованности кластеризации – плотность кластерного ядра,которая позволяет определить компакт-

ность кластеров и составить представление о их внутренней структуре. Эффективность метрики подтверждается рядом экспериментов, как на синтетических, так и на реальных данных.

7.2 Дальнейшие направления исследования

Одним из недостатков данного определения плотности ядра является неявная зависимость от гауссова распределения данных. Безусловно, при дальнейшем использовании метрики нам придется обобщить ее до случая других распределений. Другим важным направлением исседований является применимость плотности к оценке нечетких кластеризаций, поскольку, как уже упоминалось раньше, кластеры изображений трудно отделить один от другого. Еще одним возможным обобщением метрики является ее расширения до случая неметрических пространств, поскольку неметрическая кластеризация тоже весьма полезна для поиска изображений по содержанию.

Список литературы

[1] Reginald E. Hammah and John H. Curran: Validity Measures for the Fuzzy Cluster Analysis of Orientations. IEEE Transacions on pattern analysis and machine intelligence, vol. 22, no. 12, December 2000

[2] George Karypis: CLUTO A Clustering Toolkit.

[3] Moses Charikar, Chandra Chekuri, Tomas Feder, Rajeev Motwani: Incremental Clustering and Dynamic Information Retrieval.

[4] Chen Li, Edward Y. Chang, Hector Garcia-Molina, Gio Wiederhold: Clustering for Approximate Similarity Search in High-Dimensional Spaces.

[5] Ferenc Kovacs, Csaba Legany, Attila Babos: Cluster Validity Measurement Techniques.

[6] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis: Cluster Validity Methods : Part I, SIGMOD Rec., Vol. 31, No. 2, pp. 40-45, 2002

[7] M. Halkidi, Y. Batistakis and M. Vazirgiannis: Cluster validity methods: part II, SIGMOD Rec., Vol. 31, No. 3, pp. 19-27, 2002

[8] Markus Stricker, Alexander Dimai: Spectral Covariance and Fuzzy Regions for Image Indexing, "Machine Vision and Applications 10:66-73, 1997

[9] Greg Hamerly, Charles Elkan: Alternatives to the k-means algorithm that find better clusterings, CIKM '02, November 4-9, 2002, McLean, Virginia, USA, pp. 600-607

[10] Shi Zhong: Efficient Online Spherical K-means Clustering