М.В. Хачумов

Модели представления и кластеризации слабоструктурированной информации[1]

Аннотация. Предложена схема решения задачи бинарной кластеризации слабоструктурированной информации. Рассмотрены различные способы представления исходных данных задачи кластеризации. Рассмотрены методы последовательного сокращения и последовательного объединения кластеров, а также модель первоначального размещения кластеров. Даны оценки числа кластеров для решения задачи кластеризации. Предложен метод бинарной кластеризации точек, расположенных на окружности.

Ключевые слова: слабоструктурированные данные, геометрическая кластеризация, искусственная нейронная сеть, кластер, сетевая модель, классификация, задача коммивояжера.

Введение

Многие теоретические вопросы геометрической кластеризации и классификации слабоструктурированной информации остаются нерешенными до настоящего времени [1-6]. Слабоструктурированные данные (semi-structured data) - данные для которых точная структура заранее неизвестна и может меняться в потоке. К подобным данным обычно относят тексты и снимки, которые могут как порознь, так и одновременно присутствовать в документах, представленных в различных форматах. Для анализа таких данных необходимы математические модели представления информации и алгоритмы преобразования, выделения признаков и обработки. Возникающие при этом трудности преодолеваются наборами различных методов.

На Рис. 1 представлены основные этапы, модели и методы кластерного анализа слабоструктурированной информации.

Большое внимание уделяется метрикам, вопросам выбора числа и первоначального размещения кластеров. Формальные постановки ряда задач являются общими как при кластеризации текстов, так и изображений. Поэтому актуальной задачей является интеграция различных моделей в единой системе обработки.

1. Модели представления слабоструктурированной информации

Модели представления данных и знаний оказывают существенное влияние на выбор метода кластерного анализа. В случае многомодальной информации, характеризующейся наличием разнородных источников информации (текстов, изображений, звуков), и при наличии большого числа признаков к форме представления данных предъявляются повышенные требования. Можно выделить, на наш взгляд, четыре наиболее значимые системы представления исходных данных:

1) табличный способ, который широко применяется при решении задач классификации алгебраическим методом [1], нейронными сетями, деревьями решений и при построении опорного множества кластеров;

2) способ представления данных мультимножествами [3], эффективность которого показана на экономических задачах и исследуется

[1] Работа выполнена при поддержке проекта РФФИ №13-07-00025 А "Исследование методов анализа интегрированной текстовой, графической и речевой информации в системах интеллектуального управления динамическими объектами".

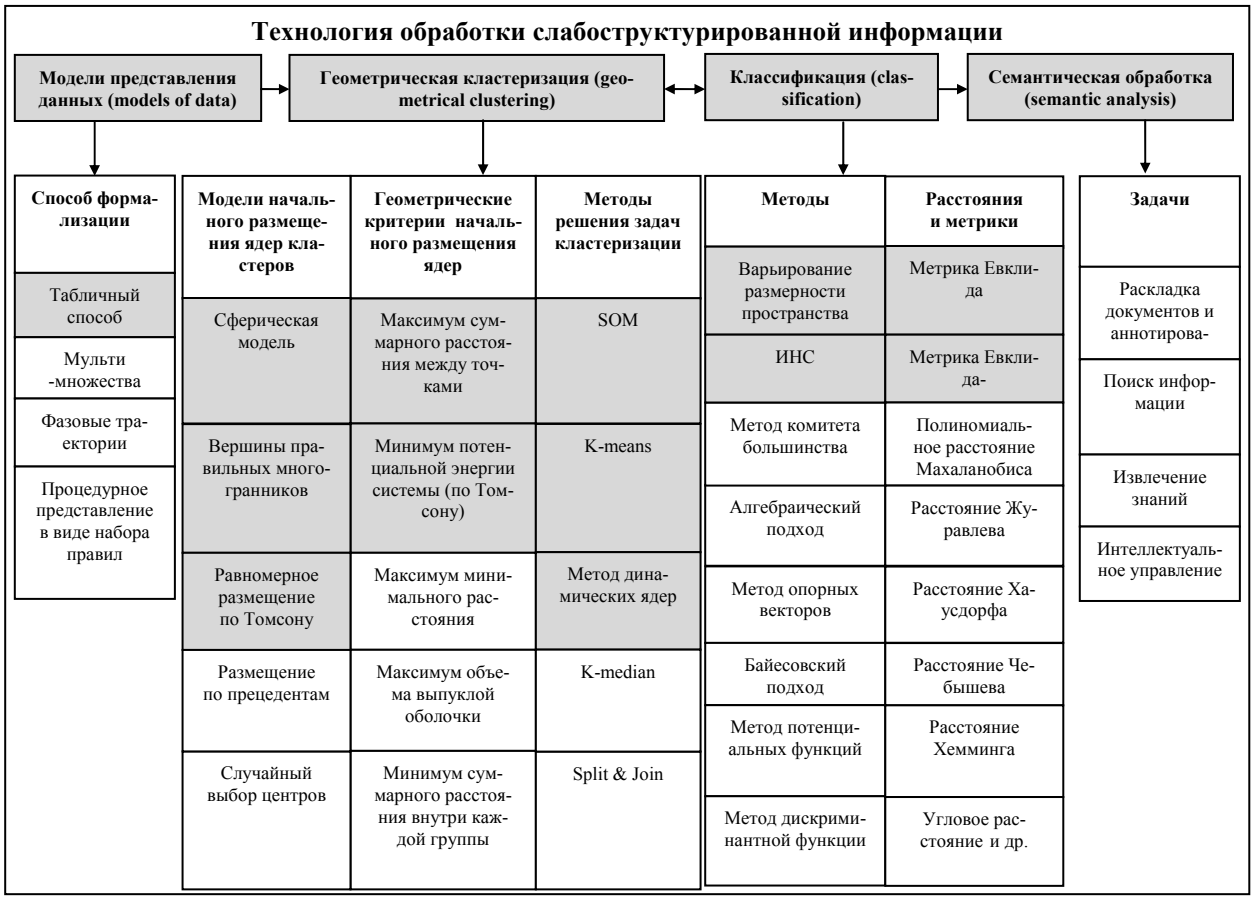


Рис. 1 . Составляющие технологии обработки слабоструктурированной информации

применительно к задачам распознавания графических образов, например печатных букв;

3) метод фазовых траекторий, предложенный для обработки мультимодальной информации [4], и пока недостаточно изученный на практике. Задача формирования траекторий в фазовом пространстве и интеллектуального управления ими на основе правил рассмотрена в работе [5];

4) процедурное представление, когда знания о конкретной проблемной области задаются в виде набора правил, и декларативное представление знаний, когда информация хранится в виде базы данных и/или базы знаний [6].

Рассмотрим некоторые модели представления данных, которые могут быть применены в задачах кластерного анализа [7-9].

1.1. Табличный способ задания исходных данных

Табличный способ задания, рассматриваемый как основной, был применен и исследован в работе [7] применительно к задаче бинарной классификации. Семантические модели относительно редко применяются в задачах кластеризации и классификации [6]. Обозначенные выше представления 2) и 3), рассматриваются как дополнение к табличному методу представления данных.

Пусть задана исходная информация для объектов $\omega_i, i = 1,..., m$ в виде обучающих выборок. Соответствующая структура данных представлена в Табл. 1. Классы Ω_1 и Ω_2 представлены матрицами значений признаков X_1 и X_2 размерности $(m_1 \times p)$ и $(m_2 \times p)$ соответственно, причем $m_1 + m_2 = m$.

Таким образом, Табл. 1 содержит данные об m объектах $\{\omega_1,..., \omega_m\}$, каждый из которых представлен вектором значений информационных признаков $x = \{x_1,..., x_n\}$ и отнесен экспертом к одному из двух классов $\{\Omega_1, \Omega_2\}$. Табличный способ является удобным представлением информации для построения решающих

Табл. 1. Представление обучающей выборки

Объекты	Признаки и их значения			Класс
	x_1	⋯	x_p	
ω_1	x_{11}	⋯	x_{1p}	Ω_1
…	…	…	…	…
ω_{m_1}	$x_{m_1 1}$	⋯	$x_{m_1 p}$	Ω_1
ω_{m_1+1}	$x_{(m_1+1)1}$	⋯	$x_{(m_1+1)p}$	Ω_2
…	…	…	…	…
ω_m	x_{m1}	⋯	x_{mp}	Ω_2

правил для задач классификации. В частности он хорошо подходит для решения задач классификации алгебраическим методом. Поскольку в задачах обработки документов происходит изменение размерности признакового пространства, то табличный метод должен работать совместно с методом варьирования признакового пространства [7].

1.2. Мультимножества для описания данных

В работе [3] предложен метод классификации совокупности многопризнаковых объектов, основанный на их представлении в виде точек метрического пространства мультимножеств. Мультимножество или множество с повторяющимися элементами служит удобной математической моделью для описания объектов, которые характеризуются многими разнородными (количественными и качественными) признаками и могут существовать в нескольких экземплярах с отличающимися, в частности, противоречивыми значениями признаков, свертка которых или невозможна, или математически некорректна. Мультимножеством A, порожденным обычным множеством $U=\{x_1,x_2,...\}$, все элементы которого различны, называется совокупность групп элементов вида $A=\{k_A(x)\bullet x \mid x\in U,\ k_A(x)\in Z+\}$. Здесь $k_A: U\to Z+=\{0,1,2...\}$ называется функцией числа экземпляров мультимножества, определяющей кратность вхождения элемента $x_i\in U$ в мультимножество A, что обозначено символом $\bullet$. Если $k_A(x)=\chi_A(x)$, где $\chi_A(x)=1$ при $x\in A$ и $\chi_A(x)=0$ при $x\notin A$, то мультимножество A становится обычным множеством. Если все мультимножества семейства $A=\{A_1,A_2,...\}$ образуются из элементов множества G, то G называется доменом для семейства A, а множество $SuppA=\{x \mid x\in G,\ \chi_{SuppA}(x)=\chi_A(x)\}$ – опорным множеством или носителем мультимножества A. Мощность мультимножества $|A|$ определяется как общее число экземпляров всех его элементов; размерность мультимножества $|SuppA|$ – как общее число различных элементов.

Введено понятие меры мультимножества, обладающей свойствами монотонности, симметричности, непрерывности и эластичности, а также понятие метрического пространства с тремя видами расстояний. Рассмотренный математический аппарат предоставляет широкие возможности для описания данных и выполнения операций над ними [8, 9]. Схема метода классификации с использованием предложенного описания заключается в построении решающих правил над небольшим количеством наиболее значимых (отобранных в результате селекции) признаков.

Все исходные мультимножества группируются (суммируются) в два мультимножества, представляющие два класса объектов. Мультимножества-суммы в свою очередь разбиваются на несколько мультимножеств-слагаемых по числу признаков, характеризующих объекты. По каждой группе признаков для каждой пары слагаемых мультимножеств генерируется пара новых мультимножеств, максимально удаленных друг от друга в метрическом пространстве. Граница между новыми слагаемыми в каждой паре определяется некоторым значением соответствующего признака. Различные комбинации таких «граничных» значений признаков дают обобщенные решающие правила для классификации объектов. Классификация осуществляется с помощью обобщенных решающих правил, составленных из разных «граничных» значений признаков, что напоминает метод решающих деревьев, обеспечивающих желаемый уровень точности классификации объектов.

1.3. Фазовые траектории для описания многомодальных данных

В работе [4] предложен метод классификации многомодальных данных для обработки

речи, изображений и текстов. Ассоциативность обращения к информации позволяет быстро получить нужную информацию, независимо от объемов выборки, а структурный подход к обработке информации – автоматически восстанавливать структуру и компактно хранить полученную информацию.

Любое ключевое слово можно представить в виде последовательности символов, а произнесенное вслух - в виде последовательности фонем. Изображение представляется для обработки в виде последовательности кодов его пикселей или последовательности чисел, характеризующих выделенные признаки (коэффициенты Фурье, вейвлет–коэффициенты, инварианты, описания иерархий и т.д.). Таким образом, независимо от модальности, информация о семантической единице информации (слове, изображении) представляется в виде последовательности двоичных чисел. Рассмотрено преобразование F двоичной последовательности в пространство R^n таким образом, что каждому n - членному фрагменту последовательности (т.е. вектору размерности n) соответствует точка $a(t)$ в R^n, с соответствующими координатами, а всей последовательности A соответствует последовательность точек - траектория: $\widetilde{A} = F(A)$. Здесь F - отображение в сигнальное пространство, основа для структурной обработки информации.

Преобразование F обладает свойством ассоциативности обращения к точкам траектории $\widetilde{A}$: любые n символов адресуют к соответствующей точке траектории. В общем случае среди n-членных фрагментов информационной последовательности может встретиться уже хранимый в памяти (возможно, эталонный) фрагмент, и траектории в этом случае пересекутся или совпадут. Одинаковые фрагменты последовательности преобразуются в одну и ту же траекторию. Под распознаванием здесь понимается процесс принятия решения о степени совпадения входной информации с запомненной ранее. В основе механизма распознавания лежит сравнение входной последовательности $\widetilde{A}$ и наиболее близкой к ней хранимой последовательности A с вычислением меры близости (по Хеммингу).

Решение о совпадении с заданной степенью точности принимается сравнением с порогом по распознаванию. В более простом случае при обучении в качестве несущей последовательности A используется последовательность, соответствующая запоминаемому событию, а в качестве информационной последовательности J - последовательность символов кода, соответствующего этому событию. Под распознаванием в этом случае понимается воспроизведение информационной последовательности J-кода события, которое инициирует входная последовательность A. Показано [4], что описанные процессы обработки информации: обучение, воспроизведение, формирование словаря, синтаксической последовательности, эффективны как в рамках одного ассоциативного процесса, так и в многоуровневых системах. Использование всех свойств ассоциативного процесса возможно лишь при включении его в иерархическую структуру, осуществляющую структурный анализ информации. К сожалению, автору этой интересной работы не удалось в полной мере реализовать на практике предложенный способ представления многомодальной информации в рамках одной информационной системы.

Рассмотренные модели представления информации дополняют друг друга, и могут быть, на наш взгляд, совместно использованы в системах кластеризации и классификации.

Задача кластерного анализа текстовой и графической информации заключается:

- в выделении информативных параметров, например, извлечении индекса текста, информативных параметров изображений (интегральных или/и инвариантных параметров). Данный этап можно назвать предобработкой информации;
- в кластеризации, т.е. автоматическом формировании рубрик и кластеров изображений;
- в классификации текста или изображения, создании реферата;
- в семантической обработке полученной информации с целью обобщения.

Модели целесообразно применять в том порядке, в каком они рассмотрены в этой работе, совместно с соответствующими им алгоритмами обработки информации.

2. Схема решения задачи бинарной кластеризации

В основе предлагаемой технологии лежит сетевая модель анализа, которая позволяет в

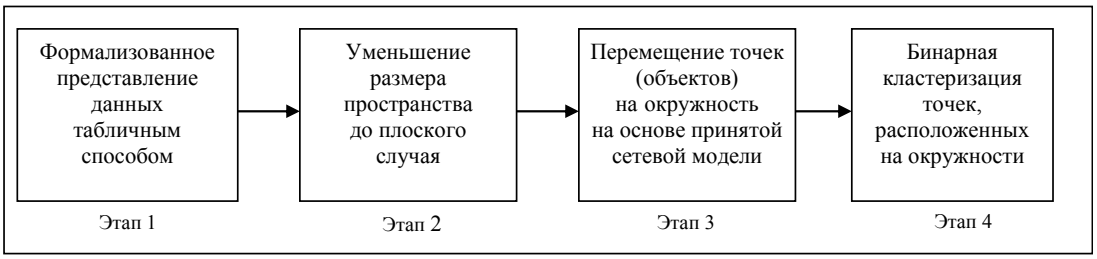


Рис. 2. Схема решения задачи бинарной кластеризации

комплексе решать задачи: коммивояжера на основе построения контура минимальной длины, кластеризации на основе метода динамических ядер, классификации при фиксированных ядрах с использованием набора метрик [10-13].

На Рис. 2 рассмотрена схема совместного применения различных методов кластерного анализа для решения задачи бинарной кластеризации, содержащая следующие этапы:

1) дается формализованное представление данных табличным способом, в котором объекты описаны в виде векторов признакового пространства;

2) происходит уменьшение размерности признакового пространства до плоского случая. Для этого используется метод варьирования признакового пространства [7];

3) решается задача коммивояжера для нахождения минимальной длины обхода точек. Предлагается итерационная процедура на основе нейронной сети Кохонена, с помощью которой точки передвигаются на окружность, что может сократить время и повысить качество дальнейшей кластеризации;

4) выполняется бинарная кластеризация на основе предложенного в статье метода.

3. Постановка задач кластеризации

Целью кластеризации является объединение объектов в непересекающиеся группы (кластеры, классы, множества) «близких» объектов. Задача *геометрической кластеризации* формулируется следующим образом: разбить множество точек $X \subset R^p$, $|X| = n$ на k $(k > 1)$ – подмножеств (кластеров, классов) $C_1, C_2, ..., C_k$, $C_i \cap C_j = \varnothing \quad \forall i, j, \quad i \neq j, \quad X = \bigcup_{i=1}^{k} C_i$ так, чтобы выполнялось условие:

$$H = \sum_{i=1}^{k} \sum_{x \in C_i} d(a_i, x) \to \min, \qquad (1)$$

где $a_i \in R^p$ – ядро кластера C_i, $d(a_i, x)$ – расстояние между ядром a_i и любой точкой x из множества X.

Один из известных методов решения задачи – метод динамических ядер использует в качестве меры квадрат расстояния Евклида

$$D_E(a_i, x) = \|x - a_i\|^2, \qquad (2)$$

причем $\sum_{x \in C_i} \|x - a_i\|^2$ определяет внутрикластерное расстояние для кластера C_i.

Метод заключается в следующем. Сначала задаются начальные значения ядер. Затем выполняют следующие шаги:

• разбиение на классы C_i при фиксированных значениях ядер:

$$C_i = \{x : D_E(a_i, x) \le D_E(a_j, x)\}; \qquad (3)$$

• оптимизация значений ядер при фиксированном разбиении на классы:

$$\sum_{x \in C_i} d(a_i, x) \to \min. \qquad (4)$$

Процедура (3)-(4) останавливается, если после очередного выполнения разбиения на классы не изменился состав ни одного класса. Характеристики кластеров представлены на Рис. 3.

Для изучения полученного разбиения объектов на однородные группы применяют следующие математические характеристики кластеров:

- *центр кластера* – среднее геометрическое место точек в пространстве переменных;
- *дисперсия кластера* – мера рассеяния точек в пространстве относительно центра кластера;

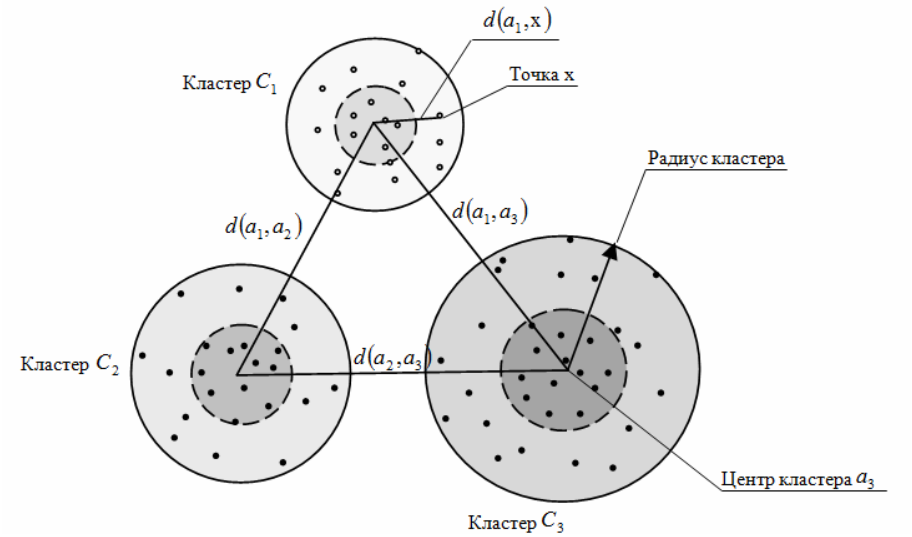


Рис. 3. Геометрическая интерпретация кластеров

- *среднеквадратичное отклонение (СКО)* объектов относительно центра кластера;
- *радиус кластера* - максимальное расстояние точек от центра кластера.

4. Методы выбора первоначального числа кластеров

Выбор методов определения первоначального числа кластеров весьма ограничен. Число классов часто задается экспертами исходя из целесообразности с учетом особенностей предметной области. Если таких соображений нет, то число классов определяется экспериментально. Наиболее распространенным является метод, основанный на последовательном парном объединении наиболее близких объектов ограниченной выборки с усреднением их характеристик [2]. Он позволяет автоматизировать выбор числа кластеров, причем окончательное решение принимает пользователь-эксперт.

4.1. Метод последовательного сокращения числа кластеров

Принцип построения метода, называемого также методом «отжига» [14], состоит в том, что на основе критерия качества класса принимается решение об удалении этого класса или о слиянии этого класса с другим. Алгоритм является аналогом схемы DEL [15], предназначенной для выбора оптимального набора признаков в задачах распознавания. Заметим, что в некоторых случаях, напротив, целесообразно разбиение класса на два. Пусть вначале имеется m классов (кластеров), которое необходимо уменьшить до величины n. Тогда количество тестируемых систем кластеров или признаков составит:

$$N = m + (m-1) + (m-2) + ... + (n+1) = \sum_{i=0}^{m-n} (m-i).$$

Величина N при этом оказывается существенно меньше, чем C_m^n. Процесс останавливается, когда получают минимальную по размеру систему признаков с приемлемым качеством.

В качестве критериев качества класса используются:

1) *Количественный критерий*. Класс, в котором менее N экземпляров (точек в пространстве признаков) считается пустым и полежит удалению. Порог выбирается экспертом в зависимости из смысла задачи и вида меры близости.

2) *Критерий сферической разделимости*. Два класса считаются сферически разделимыми, если сумма радиусов двух классов меньше расстояния между ядрами (центрами) этих классов. Если классы сферически неразделимы, то они сливаются в один. Решение о разбиении классов на два может служить расширением данного метода. В этом случае используют критерий равномерности.

3) *Средняя мера близости* точек класса от ядра должна быть не менее половины или трети от максимума меры близости точек от ядра (радиуса класса). Если это не так, то класс раз-

бивается на два (порождается еще одно ядро вблизи исходного).

4.2. Метод последовательного увеличения числа кластеров

Альтернативой служит метод последовательного увеличения числа кластеров [14]. Идея метода состоит в том, чтобы, начав с малого числа классов, увеличивать его до тех пор, пока не будет получена «хорошая» классификация. Метод является аналогом общей схемы алгоритма ADD [15].

В качестве критерия качества используют такое понятие, как часто воспроизводящийся класс. Проводится достаточно большая серия классификаций с различным начальным выбором классов. Определяются классы, которые возникают в различных классификациях. Считаются частоты появления таких классов. Критерием получения «истинного» числа классов может служить снижение числа часто повторяющихся классов. Т.е. при числе классов k число часто повторяющихся классов заметно меньше чем при числе классов $k-1$ и $k+1$. Начинать следует с двух классов. Трудоемкость метода ADD приблизительно такая же, как и у алгоритма DEL. Метод хорошо работает при небольшом числе классов. При достаточно большом числе классов и большом объеме экземпляров, которые необходимо разбить на классы, такая процедура подбора становится слишком медленной.

Оба описанных алгоритма крайне трудны в реализации из-за сложности оценки реального качества получаемых решений. Поэтому на практике лучше воспользоваться более простыми и доступными средствами автоматизированного выбора числа кластеров.

5. Проблема и модели начального расположения кластеров

Универсальным способом задания начального положения ядер является задание начального разбиения представительной части объектов на классы. При этом в начальном разбиении могут участвовать не все объекты. Далее, решая задачу (1), получают начальные значения кластеров. Однако могут представлять интерес и другие способы или модели начального представления данных.

Рассмотрим некоторые модели выбора начального расположения кластеров. Как правило, эффективность расположения может быть проверена только экспериментальным путем в результате решения конкретных задач и их оценки экспертом.

5.1. Модель случайного размещения кластеров

При автоматизированном подходе кластеры может задать сам пользователь исходя из требований задачи. Основной схемой при автоматическом подходе является выбор случайных объектов в качестве кластеров. В этом случае ядра (центры, центроиды) кластеров являются точками того же пространства, что и объекты. При случайном назначении ядер алгоритм оказывается чувствительным к возможным выбросам и результаты, получаемые для одной и той же выборки документов, будут отличаться. Случайный выбор экземпляров не представляется, безусловно, лучшим решением, и может привести к потерям времени на последующих этапах.

5.2. Модель размещения кластеров на границе p-шара

Геометрической моделью размещения кластеров может служить распределение точек на m-сфере. Рассматривается следующая задача. Расположить n точек в p - шаре радиуса r так, чтобы сумма расстояний между ними была максимальна [16]. Решение получено для плоского случая [17] и для частного трехмерного случая [16]. Доказано, что решение сводится к задаче размещения n точек на поверхности шара. Т.е. для максимизации суммарного расстояния между кластерами, они должны располагаться на сфере [16]. Этот вывод был использован в предложенной схеме решения задачи бинарной кластеризации.

6. Оценки необходимого числа кластеров в задаче кластеризации

Получим оценку числа кластеров, исходя, например, из минимума времени решения задачи. Пусть n - количество пунктов, m - количество кластеров ($n > m$), тогда $\left(\frac{n}{m}\right)$ - среднее чис-

ло пунктов, приходящихся на один кластер. Временная сложность алгоритма кластеризации оценивается в разных источниках как $O(m^2)$ или $O(m^3)$. Запишем аналитическую зависимость средней продолжительности решения задачи при кубической сложности в виде [18]:

$$T(m) = am^3 + am\left(\frac{n}{m}+1\right)^3 = am^3 + a\frac{(n+m)^3}{m^2}. \quad (5)$$

Решая оптимизационную задачу, получим:

$$3am^2 + a\frac{3(n+m)^2m^2 - 2m(n+m)^3}{m^4} = 0,$$

$$3m^5 + m^3 - 3n^2m - 2n^3 = 0. \quad (6)$$

Результаты численного решения уравнения (6) содержатся в Табл. 2 (результаты округлены до целых чисел).

Табл. 2. Результаты численного решения уравнения

n	10	50	100	200	300	400	500
m	4	10	15	22	29	34	39

Заметим, что приведенные оценки являются достаточно грубыми и требуют своего уточнения.

7. Бинарная кластеризация

Предлагается решение задачи бинарной кластеризации точек, расположенных на окружности. Пусть имеются *n* точек (объектов), упорядоченных на границе круга L. Без потери общности разобьем множество точек, расположенных на окружности, на два подмножества (кластера) C_1 и C_2 так, чтобы сумма внутриклассовых расстояний была больше межклассового расстояния. Для этого выполним следующие действия.

Пронумеруем точки на окружности от 1 до n. Пусть пара (l,k) характеризует кортеж, образуемый последовательностью точек, где l $(l = 1,...,n)$ - номер начальной точки кластера, k $(2 \le k \le n-2)$ – число точек кластера, расположенных на окружности.

Для каждого l построим дискретную функцию $F(l,k)$, характеризующую качество разбиения, в зависимости от числа элементов в кластере *k*. Лучшее локальное решение:

$$F(l,k) = \frac{\textit{расстояние между классами}}{\textit{сумма внутриклассовых расстояний}} \to \max_k \quad (7)$$

Оптимальное решение соответствует глобальному максимуму среди всех l $F(l,k) \to \max\limits_{l,k}$

В качестве расстояния будем использовать хорошо зарекомендовавшую себя метрику Евклида-Махаланобиса d_{E-M}, которая учитывает статистические свойства кластеров.

Определим составляющие функции качества.

Внутриклассовое расстояние. Подсчитываем для каждого класса $C_1 = C_1(l,k)$ и $C_2 = C_2(l,k)$ матрицы ковариаций $S_1 = S_1(l,k)$ и $S_2 = S_2(l,k)$, а также внутриклассовые расстояния как суммы расстояний от каждой точки класса $x \in C_i$ до центра соответствующего класса $\bar{x}_i$:

$$\sum_{x,\in C_1} d_{E-M}(x, C_1) = \sum_{x,\in C_1} \sqrt{(x-\bar{x}_1)^T S_1^{-1}(x-\bar{x}_1)},$$

$$\sum_{x\in C_2} d_{E-M}(x, C_2) = \sum_{x\in C_2} \sqrt{(x-\bar{x}_2)^T S_2^{-1}(x-\bar{x}_2)}.$$

Расстояние между классами. Для каждого варианта (l,k) подсчитываем межклассовое расстояние, т.е. расстояние между классами C_1 и C_2. Для этого строим объединенную матрицу ковариаций по формуле:

$$S_{1,2} = \frac{1}{n-2}(S_1 + S_2).$$

Расстояние между классами подсчитывается как расстояние между центрами классов с объединенной матрицей:

$$d_{E-M}(C_1, C_2) = \sqrt{(\bar{x}_1 - \bar{x}_2)^T S_{1,2}^{-1}(\bar{x}_1 - \bar{x}_2)}.$$

Оценка качества:

$$F(l,k) = \frac{d_{E-M}(C_1, C_2)}{\sum\limits_{x\in C_1} d_{E-M}(x, C_1) + \sum\limits_{x\in C_2} d_{E-M}(x, C_2)} \to \max_{l,k}. \quad (8)$$

Процедура имеет временную сложность (по числу итераций) $O(n(n-3)/2)$.

8. Пример решения задачи бинарной кластеризации

Рассмотрим пример совместного применения метода варьирования пространства признаков [7] и метода, основанного на сетевой модели.

1) Исходная таблица прецедентов (Табл.3).

Табл. 3. Исходные прецеденты

	x_1	x_2	x_3	x_4	x_5	x_6	Класс
W_1^1	1	2	0	1	-1	0	1
W_2^1	-1	0	-1	-2	-2	1	1
W_3^1	1	-1	2	-1	-1	0	1
W_1^2	0	2	1	-1	0	1	2
W_2^2	2	-1	0	-1	-1	0	2

2) Построим матрицу ковариаций C.

$$C = \begin{pmatrix} 1.3 & -0.55 & 0.45 & 0.60 & 0.25 & -0.55 \\ -0.55 & 2.30 & -0.20 & 0.90 & 0.50 & 0.30 \\ 0.45 & -0.20 & 1.30 & 0.15 & 0.50 & -0.20 \\ 0.60 & 0.90 & 0.15 & 1.20 & 0.25 & -0.35 \\ 0.25 & 0.50 & 0.50 & 0.25 & 0.50 & 0.00 \\ -0.55 & 0.30 & -0.20 & -0.35 & 0.00 & 0.30 \end{pmatrix} \quad (9)$$

3) Найдем собственные векторы и собственные числа для матрицы ковариаций (9), решая определитель (Табл.4):

$$\begin{vmatrix} c_{11}-\lambda_1 & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ c_{21} & c_{22}-\lambda_2 & c_{23} & c_{24} & c_{25} & c_{26} \\ c_{31} & c_{32} & c_{33}-\lambda_3 & c_{34} & c_{35} & c_{36} \\ c_{41} & c_{42} & c_{43} & c_{44}-\lambda_4 & c_{45} & c_{46} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55}-\lambda_5 & c_{56} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55} & c_{56}-\lambda_6 \end{vmatrix} = 0 \quad (10)$$

Табл. 4. Вектор собственных чисел

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
0	2.96	1.18	2.45	0	0.29

4) Результирующая Табл,5 векторов в системе признаков $Y = (y_1,...,y_n)$.

Табл. 5. Векторы в новой системе координат

Векторы в новой системе.	y_1	y_2	y_3	y_4	y_5	y_6	Класс
V_1^1	-0.03	**0.27**	-0.01	**-0.43**	-0.02	-0.85	1
V_2^1	-0.16	**0.87**	-0.06	**0.39**	0.18	0.08	1
V_3^1	-0.32	**0.07**	0.75	**-0.38**	0.34	0.2	1
V_1^2	0.64	**-0.06**	0.47	**0.45**	0.24	-0.29	2
V_2^2	-0.45	**-0.36**	-0.22	**0.31**	0.66	-0.26	2

Для дальнейшего хода исследований нам важно использовать наибольшие собственные числа $\lambda_2 = 2.96$, $\lambda_4 = 2.45$ и соответствующие им собственные векторы. После понижения размерности исходные объекты разместились на плоскости так, как показано на Рис. 4, а).

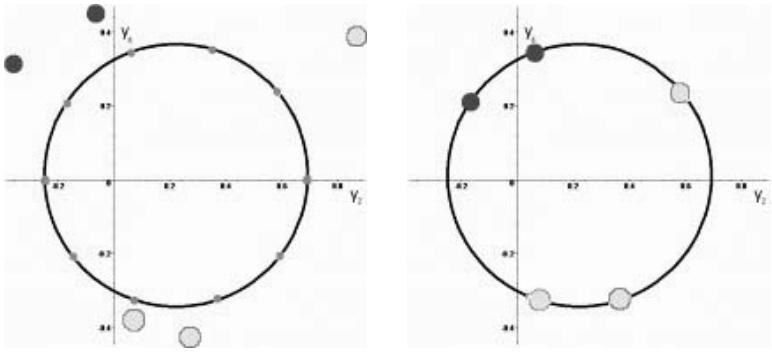
На этом же рисунке показаны равномерно размещенные на окружности точки-нейроны. Далее решается задача кластеризации на основе сетевой модели. При этом сетевая модель (на основе нейронной сети Кохонена) необходима для итерационного «притягивания» объектов к окружности. После чего и решается упрощенная задача бинарной кластеризации (7).

Как видно из Рис.4, б), объекты в результате расположились на окружности так, как их классифицировали эксперты в исходной таблице.

Таким образом, совместное применение методов позволяет решать задачу бинарной кластеризации.

Заключение

За основу представления слабоструктурированных исходных данных задачи кластеризации целесообразно принять табличный способ, который широко применяется при решении задач классификации алгебраическим методом, нейронными сетями, деревьями решений и при построении опорного множества кластеров. Рассмотренные методы выбора первоначального числа кластеров и модель их размещения позволяют решать задачу бинарной кластеризации. Причем решение задачи размещения точек на окружности на основе сетевой модели рассматривается как шаг, предшествующий кластеризации.



а) размещение исходных объектов б) результат кластеризации

Рис. 4. Пример решения задачи кластеризации

Показана целесообразность совместного применения различных методов кластерного анализа на примере решения задачи бинарной кластеризации.

Литература

1. Журавлев Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. — М.: Фазис, 2005. —159 с.
2. Хачумов М.В. Задача кластеризации текстовых документов. — Информационные технологии и вычислительные системы, № 2, 2010, с.42-49.
3. Петровский А.Б. Пространства множеств и мультимножеств. – М.: УРСС, 2003. – 248 с. - http://www.raai.org/about/persons/petrovsky/pages/Petrovsky_2003.pdf
4. Харламов А.А. Нейросетевой подход к интегрированному представлению и обработке информации в интеллектуальных системах. – Автореферат докторской диссертации.-М.: 2009. – 31 с.
5. Osipov G. Strategies for Stabilization Behaviour of Intelligent Dynamic Systems. – Proc. of 20th European Meeting on Cybernetics and Systems, Vienna, 2010, pp.195-197.
6. Люгер Д.Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. – М.: Издательский дом «Вильямс», 2003 – 864 с.
7. Толмачев И.Л., Хачумов М.В. Бинарная классификация на основе варьирования размерности пространства признаков и выбора эффективной метрики. – Искусственный интеллект и принятие решений, № 2, 2010, с.3-10.
8. Хачумов М.В. О проблеме интеграции анализа текстовых данных и графических образов для задач поиска и кластеризации документов. — Сборник трудов Девятой международная научно-практическая конференция «Исследование, разработка и применение высоких технологий в промышленности» (Санкт-Петербург, 22-23 апреля 2010). – СПб: Издательство Политехнического университета, 2010, Т.3, с 157-160.
9. Хачумов М.В. Модели представления данных в задачах распознавания образов и кластеризации. – Тезисы докладов Всероссийской конференции с международным участием «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологических систем» (18-22 апреля 2011 Москва). – М.: РУДН, 2011, с.146-148.
10. Ежов А.А., Шумский С.А. Нейрокомпьютинг и его применения в экономике и бизнесе. Оптимизация с помощью сети Кохонена. http://www.intuit.ru/department/expert/neurocomputing/6/4.html
11. Ежов А.А., Шумский С.А. Нейрокомпьютинг и его применения в экономике и бизнесе. Оптимизация и сеть Хопфилда. –http://www.intuit.ru/department/expert/neurocomputing/6/2.html
12. Хачумов М.В. Нейронная сеть Кохонена с универсальной метрикой. – Материалы XVI Международной конференции по вычислительной механике и современным прикладным программным системам (ВМСППС'2009, 25-31 мая 2009. Алушта), Изд-во МАИ – ПРИНТ, 2009, с.742-745.
13. Тищенко И.П. Алгоритмическое и программное обеспечение мультипроцессорных систем для распознавания графических образов на основе нейросетевого подхода. – Автореферат канд. дисс.: Переславль-Залесский, 2009. – 24 с.
14. Миркес Е.М. Нейроинформатика. Учебное пособие. – Красноярск: Издательство Красноярского государственного технического университета, 2003. – http://www.softcraft.ru/neuro/ni/p00.shtml.
15. Загоруйко Н.Н. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Института математики, 1999. – 270 с.
16. Хачумов М.В. Расстояния, метрики и кластерный анализ. – Искусственный интеллект и принятие решений, № 1, 2012, с. 81-89.
17. Атаманов В.В., Козачок М.А., Трушков В.В., Хачумов М.В. Выбор первоначального расположения кластеров в нейронной сети Кохонена. – Нейрокомпьютеры: разработка и применение, №1, 2009, с.73-76.
18. Серая О.В., Дунаевская О.И. Многошаговая кластеризация в задаче коммивояжера высокой размерности. – Математика и кибернетика - фундаментальные и прикладные аспекты, № 5/5(35), 2008, с. 37-40.

Хачумов Михаил Вячеславович. Научный сотрудник Института системного анализа РАН. Окончил Российский университет дружбы народов в 2009 году. Кандидат физико-математических наук. Автор 19 печатных работ. Область научных интересов: интеллектуальный анализ данных, интеллектуальные системы управления, обработка изображений, кластерный анализ слабоструктурированных данных. E-mail: khmike@inbox.ru