Презентация на тему методы кластерного анализа.

Цель презентации:

1 ) Получить представление о истории кластерного анализа (и о самом кластерном анализе).

2 ) Получить представление о классификации методов кластерного анализа.

3 ) Получить представление о самих методах кластерного анализа.

4 ) Попытаться провести сравнительный анализ методов, описать плюсы и минусы.

5 ) Попытаться сделать вывод.

1 )

Целью кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные группы или кластеры.

Социально-экономические системы в их динамическом развитии характеризуются многомерными данными. При анализе информации о сложных системах необходимы обработка и анализ огромных объёмов разнородных данных, важно упорядочить их по иерархическим уровням и произвести группировку, разбиение множества исследуемых объектов и признаков на однородные группы или кластеры. Это позволяет сделать кластерный анализ, впервые применённый в социологии Трионом [25] в 1939г.

Встаёт проблема автоматизации процесса обработки и анализа информации, необходимости создания простого и доступного инструмента (программного пакета) для группировки и распознавания данных, произвольно структурированных, и типов данных. Остро возникает проблема представления данных в виде удобном и понятном для пользователя. Необходим программный код, который бы мог выполняться на различном HardWare, без его предварительного переноса на «родной язык» устройства. Актуально использование методов кластерного анализа для численных и нечетких данных.

На сегодняшний день актуальным является развитие методов кластерного анализа (см. [9],[10], [11], [12], [15] и др.), среди которых метод полных связей, метод максимального локального расстояния, метод Ворда, центроидный метод и др. Актуально их применение к решению экономических задач (см. [6], [17], [18], [20] и др.), в частности, сегментации рынка (модель Уинда и Кардозы и др.), их программная реализация.

На сегодняшний день актуально сочетание кластерного анализа с теорией нечетких множеств (см. [2], [3], [5], [7], [13], [21], [23] и др.), основы которой заложены в 60-е годы американским математиком Латфи Заде [5]. В настоящее время Японией и США активно разрабатываются электронные системы с нечеткими управляющими алгоритмами.

Существует программное обеспечение для кластерного анализа в узких предметных областях, например, «ClusterDelta» позволяет производить анализ данных на бирже, существуют программы для прогноза конъюнктуры рынка и др.

Существуют пакеты программ, реализующие набор методов кластерного анализа данных («STATISTICA Multivariate Exploratory Techniques», «MatLab» и др.) В программном пакете «STATISTICA Multivariate Exploratory Techniques» реализованы методы k-средних, иерархической кластеризации и двухвходового объединения. Используются различные метрики расстояний. Этот пакет не ориентирован на решение задач кластеризации в условиях неоднородных показателей, на обработку нечетких данных. В «MatLab» имеются встроенные функции, позволяющие реализовать алгоритмы четкой и нечеткой кластеризации. Эти пакеты рассчитаны на использование профессиональными математиками и в этом их ограничение.

Нами разработан программный продукт «Кластеризация ИТ-предприятий». Программный продукт позволяет группировать данные иерархическим центроидным методом кластерного анализа, неиерархическим FCM-методом, который предполагает возможность работы с нечеткими данными. Выполнена программная реализация методов создания, сохранения, загрузки, обработки кластеров (данных).

При анализе и прогнозировании социально-экономических явлений исследователь сталкивается с многомерностью их описания. Это  происходит при решении задачи сегментирования рынка, построении типологии стран по большому числу показателей, прогнозирования конъюнктуры рынка товаров, изучении и прогнозировании экономической депрессии и др.

Методы многомерного анализа – действенный количественный инструмент исследования социально-экономических процессов, описываемых большим  числом характеристик. К ним относятся кластерный анализ, таксономия, распознавание образов, факторный анализ и др. Кластерный анализ наиболее ярко отражает черты многомерного анализа в классификации, факторный анализ – в исследовании связи.

Первое применение кластерный анализ нашел в социологии. Название кластерный анализ происходит от английского слова cluster – гроздь, скопление. Впервые в 1939 был определен предмет кластерного анализа и сделано его описание исследователем Трионом. Главное назначение кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные группы или кластеры. Решается задача классификации данных и выявления соответствующей структуры в ней.

Большое достоинство кластерного анализа в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы. Это имеет большое значение, например, для прогнозирования конъюнктуры, когда показатели имеют разнообразный вид, затрудняющий применение традиционных эконометрических подходов.

Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы социально-экономической информации, делать их  компактными и наглядными.

Кластерный анализ можно использовать циклически. В этом случае исследование производится до тех пор, пока не будут достигнуты необходимые результаты. При этом каждый цикл здесь может давать информацию, которая способна сильно изменить направленность и подходы дальнейшего применения кластерного анализа. Этот процесс можно представить системой с обратной связью.

В задачах  социально-экономического прогнозирования перспективно сочетание кластерного анализа  с другими количественными методами (например, с регрессионным анализом).

Кластерный анализ имеет недостатки  и ограничения. В частности, состав  и количество кластеров зависит от  выбираемых критериев разбиения. При сведении исходного массива данных к более компактному виду могут возникать искажения, теряться индивидуальные черты отдельных объектов за счет  замены их характеристиками обобщенных значений параметров кластера. При проведении классификации часто игнорируется возможность отсутствия в рассматриваемой  совокупности каких-либо значений.

Существует достаточно много методов кластерного анализа. Например, метод полных связей, метод максимального локального расстояния, метод Ворда, центроидный метод (методом взвешенных групп) и др.

Алгоритмы кластерного анализа можно  подразделить на иерархические и неиерархические. Иерархические алгоритмы делятся на агломеративные, характеризуемые последовательным объединением  исходных элементов и соответствующим уменьшением числа кластеров, и на дивизимные (делимые), в которых число кластеров возрастает, начиная с одного, в результате чего образуется последовательность  расщепляющих групп.

Алгоритмы кластерного анализа имеют хорошую программную реализацию, которая позволяет решить задачи самой большой размерности.

Сочетание методов кластерного анализа с методами теории нечетких множеств открывает новые возможности для исследователей.

Основы нечеткой логики были заложены в конце 60-х лет в работах известного американского математика Латфи Заде [5]. Исследования такого рода было вызвано возрастающим неудовольствием экспертными системами. «Искусственный интеллект» легко справлялся с задачами управления сложными техническими комплексами, но был беспомощным при простейших высказываниях повседневной жизни.

Для создания действительно интеллектуальных систем, способных адекватно взаимодействовать с человеком, был необходим новый математический аппарат, который переводит неоднозначные жизненные утверждения в язык четких и формальных математических формул. Первым серьезным шагом в этом направлении стала теория нечетких множеств, разработанная Заде. Его работа «Fuzzy Sets», опубликованная в 1965 году в журнале «Information and Control», заложила основы моделирования интеллектуальной деятельности человека и стала начальным толчком к развитию новой математической теории. Он же дал и название для новой области науки –«fuzzy logic» (fuzzy - нечеткий, размытый, мягкий).

Дальнейшие работы профессора Латфи Заде и его последователей заложили фундамент новой теории и создали предпосылки для внедрения методов нечеткого управления в инженерную практику.

Аппарат теории нечетких множеств, продемонстрировав ряд многообещающих возможностей применения – от систем управления летательными аппаратами до прогнозирования итогов выборов, оказался вместе с тем сложным для практического воплощения. Нечеткая логика заняла свое место среди других специальных научных дисциплин – где-то посредине между экспертными системами и нейронными сетями.

Свое второе рождение теория нечеткой логики пережила в начале восьмидесятых годов, когда несколько групп исследователей (в основном, в США и Япони) всерьез занялись созданием электронных систем различного применения, использующих нечеткие управляющие алгоритмы. Теоретические основы для этого были заложены в ранних работах Коско и других ученых.

Третий период начался с конца 80-х годов и до сих пор. Этот период характеризуется бумом практического применения теории нечеткой логики в разных сферах науки и техники. До 90-ого года появилось около 40 патентов, относящихся к нечеткой логике (30 - японских). Сорок восемь японских компаний создают лабораторию LIFE (Laboratory for International Fuzzy Engineering), японское правительство финансирует 5-летнюю программу по нечеткой логике, которая включает 19 разных проектов - от систем оценки глобального загрязнения атмосферы и предвидения землетрясений до АСУ заводских цехов. Результатом выполнения этой программы было появление целого ряда новых массовых микрочипов, базирующихся на нечеткой логике. Сегодня их можно найти в стиральных машинах и видеокамерах, цехах заводов и моторных отсеках автомобилей, в системах управления складскими роботами и боевыми вертолетами.

В США развитие нечеткой логики идет по пути создания систем для большого бизнеса и военных. Нечеткая логика применяется при анализе новых рынков, биржевой игре, оценки политических рейтингов, выборе оптимальной ценовой стратегии и т.п. Появились и коммерческие системы массового применения.

Смещение центра исследований нечетких систем в сторону практических применений привело к постановке целого ряда проблем, в частности: новые архитектуры компьютеров для нечетких вычислений; элементная база нечетких компьютеров контроллеров; инструментальные средства разработки; разработка нечетких систем управления, и т.п.

**Задача кластерного анализа.** Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве , разбить множество объектов на ( – целое) кластеров (подмножеств) , , …, , так, чтобы каждый объект принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время, как объекты, принадлежащие разным кластерам были разнородными (см. [9],[10], [11], [12], [15] и др.).

Решением задачи кластерного анализа являются разбиения, удовлетворяющие некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, который называют целевой функцией. Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонения:

,

где - представляет собой измерения -го объекта.

Для решения  задачи кластерного анализа необходимо определить понятие сходства и разнородности.

Объекты -ый и -ый попадут в один кластер, когда расстояние между точками и будет достаточно мало и попадут в разные кластеры, когда это расстояние будет достаточно большим. Попадание в один или разные кластеры объектов определяется понятием расстояния между и из , где – -мерное евклидово пространство.

Неотрицательная функция называется функцией расстояния (метрикой), если:

а) для всех и из ;

б) тогда и только тогда, когда ;

в) ;

г) ;

где – любые три вектора из .

Наиболее часто употребляются следующие функции расстояний:

евклидово расстояние        ,

– норма                            ,

супремум - норма              , ,

– норма                             .

Евклидова метрика является наиболее популярной. Метрика наиболее легкая для вычислений. Супремум-норма легко считается и включает в себя процедуру упорядочения, а - норма обобщает три первые функции расстояний.

Пусть n измерений , ,..., представлены в виде матрицы данных размером :

.

Тогда расстояние между парами векторов могут быть представлены в виде симметричной матрицы расстояний:

.

Важным понятием является понятие сходства между объектами и . Неотрицательная вещественная функция  называется мерой сходства, если:

1) для ;

2) ;

3) ;

Пары значений мер сходства можно объединить в матрицу сходства:

.

Величину называют коэффициентом сходства.

Естественной  мерой сходства характеристик объектов во многих задачах является коэффициент корреляции между ними



где , , , соответственно средние и среднеквадратичные отклонения для характеристик и . Мерой различия между характеристиками может служить величина  . В некоторых задачах  знак коэффициента корреляции несуществен и зависит лишь от  выбора единицы измерения. В этом случае в качестве меры различия  между характеристиками используется .

**Методы кластерного анализа.** Методы кластеризации могут быть иерархическими и неиерархическими. Иерархическая кластеризация (hierarchical clustering) характеризуется построением иерархической, или древовидной, структуры.

Иерархические методы могут быть агломеративными (объединительными) и дивизивными. Агломеративная кластеризация (agglomerative clustering) начинается с каждого объекта в отдельном кластере. Кластеры объединяют, группируя объекты каждый раз во все более и более крупные кластеры. Этот процесс продолжают до тех пор, пока все объекты не станут членами одного единственного кластера.

Разделяющая, или дивизивная, кластеризация (divisive clustering) начинается со всех объектов, сгруппированных в единственном кластере. Кластеры делят (расщепляют) до тех пор, пока каждый объект не окажется в отдельном кластере.

Обычно в маркетинговых исследованиях используют агломеративные методы, например методы связи, дисперсионные и центроидные методы. Методы связи (linkage methods) включают метод одиночной связи, метод полной связи и метод средней связи.

В основе **метода одиночной связи** (single method) лежит минимальное расстояние, или правило ближайшего соседа. При формировании кластера первыми объединяют два объекта, расстояние между которыми минимально. Далее определяют следующее по величине самое короткое расстояние, и в кластер с первыми двумя объектами вводят третий объект. На каждой стадии расстояние между двумя кластерами представляет собой расстояние между их ближайшими точками.

**Метод полных связей.** Суть данного метода в том, что два объекта, принадлежащих одной и той же группе (кластеру), имеют коэффициент сходства, который меньше некоторого порогового значения . В терминах евклидова расстояния это означает, что расстояние между двумя точками (объектами) кластера не должно превышать некоторого порогового значения . Таким образом, определяет максимально допустимый диаметр подмножества, образующего кластер.

Метод **средней связи** (average linkage) действует аналогично. Однако в этом методе расстояние между двумя кластерами определяют как среднее значение всех расстояний, измеренных между объектами двух кластеров, при этом в каждую пару входят объекты из разных кластеров.

**Метод Варда.** В этом методе в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений, которая есть ни что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров.

В **центроидных** методах (centroid method) расстояние между двумя кластерами представляет собой расстояние между их центроидами (средними для всех переменных). **Центроидный метод (centroid method)** Дисперсионный метод иерархической кластеризации, в котором расстояние между двумя кластерами представляет собой расстояние между их центроидами (средними для всех переменных). Каждый раз объекты группируют и вычисляют новый центроид. Изо всех иерархических методов методы средней связи и Варда показывают наилучшие результаты по сравнению с другими методами.

К другому типу процедур кластеризации относятся **неиерахические методы кластеризации** (nonhierarchical clustering), часто называемые методом -средних. Эти методы включают последовательный пороговый метод, параллельный пороговый метод и оптимизирующее распределение. В последовательном пороговом методе (sequential threshold method) выбирают центр кластера и все объекты, находящиеся в пределах заданного от центра порогового значения, группируют вместе. Затем выбирают новый кластерный центр, и процесс повторяют для несгруппированных точек. После того как объект помещен в кластер с этим новым центром, его уже не рассматривают как объект для дальнейшей кластеризации.

Аналогично работает **параллельный пороговый метод** (parallel threshold method), за исключением того, что одновременно выбирают несколько кластерных центров и объекты в пределах порогового уровня группируют с ближайшим центром.

Метод **оптимизирующего распределения** (optimizing partitioning method) отличается от двух изложенных выше пороговых методов тем, что объекты можно впоследствии поставить в соответствие другим кластерам, чтобы оптимизировать суммарный критерий, такой как среднее внутри кластерное расстояние для данного числа кластеров.

Недостатки неиерархических методов состоят в том, что число кластеров определяется заранее и выбор кластерных центров происходит независимо. Результаты кластеризации могут зависеть от выбранных центров. Многие неиерархические процедуры выбирают первые случаев *( –* число кластеров), не пропуская никаких значений в качестве начальных кластерных центров. Таким образом, результаты кластеризации зависят от порядка наблюдений в данных.

Неиерархическая кластеризация быстрее иерархических методов, и ее выгодно использовать при большом числе объектов или наблюдений.

Возможно использовать иерархические и неиерархические методы в тандеме. Во-первых, первоначальное решение по кластеризации получают, используя такие иерархические методы, как метод средней связи или метод Варда. Полученное этими методами число кластеров и кластерных центроидов используют в качестве исходных данных в методе оптимизирующего распределения.

**Данные.** Кластерный анализ можно применять к интервальным данным, частотам, бинарными данным. Важно, чтобы переменные изменялись в сравнимых шкалах.

Неоднородность единиц измерения и вытекающая отсюда невозможность обоснованного выражения значений различных показателей в одном масштабе приводит к тому, что величина расстояний между точками, отражающими положение объектов в пространстве их свойств, оказывается зависящей от произвольно избираемого масштаба. Чтобы устранить неоднородность измерения исходных данных, все их значения предварительно нормируются, т.е. выражаются через отношение этих значений  к некоторой величине, отражающей определенные свойства данного показателя. Нормирование исходных данных для кластерного анализа иногда проводится посредством деления исходных величин на среднеквадратичное отклонение соответствующих показателей.

Другой способ сводиться к вычислению, так называемого, стандартизованного вклада. Его еще называют -вкладом. -вклад показывает, сколько стандартных отклонений отделяет данное наблюдение от среднего значения:

,

где – значение данного наблюдения, – среднее, – стандартное отклонение. Среднее для -вкладов является нулевым и стандартное отклонение равно 1.

Стандартизация позволяет сравнивать наблюдения из различных распределений. Если распределение переменной является нормальным (или близким к нормальному), и средняя и дисперсия известны или оцениваются по большим выборным, то -вклад для наблюдения обеспечивает более специфическую информацию о его расположении.

Заметим, что методы нормирования означают признание всех признаков равноценными с точки зрения выяснения сходства рассматриваемых объектов. Уже отмечалось, что применительно к экономике признание равноценности различных показателей оправдано не всегда. Желательно наряду с нормированием придать каждому из показателей вес, отражающий его значимость в ходе установления сходств и различий объектов.

Экспертные оценки дают известное основание для определения важности индикаторов, входящих в ту или иную группу показателей. Умножение нормированных значений показателей на коэффициент, соответствующий среднему баллу оценки, позволяет рассчитывать расстояния между точками с учетом неодинакового веса их признаков.

Довольно часто при решении  подобных задач используют не один, а два расчета: первый, в котором все признаки считаются равнозначными, второй, где им придаются различные веса в соответствии со средними значениями экспертных оценок.

**Принятие решения о количестве кластеров.** Главный вопрос кластерного анализа – вопрос о количестве кластеров. Здесь нет твердых правил, позволяющих быстро принять решение, но можно руководствоваться следующим.

При определении количества кластеров руководствуются теоретическими и практическими соображениями. Например, если цель кластеризации – выявление сегментов рынка, то менеджмент может захотеть получить конкретное число кластеров.

В иерархической кластеризации в качестве критерия можно использовать расстояния, при которых объединяют кластеры.

В неиерархической кластеризации чертят график зависимости отношения суммарной внутригрупповой дисперсии к межгрупповой дисперсии от числа кластеров. Точка, в которой наблюдается изгиб или резкий поворот, указывает на приемлемое количество кластеров. Увеличение числа кластеров за эту точку обычно безрезультативно.

Относительные размеры кластеров должны быть достаточно выразительными.

2 )

3 )

4 )

5 )

-