Тема моего научного исследования: методика адаптивной кластеризации фактографических данных.

Целью данной работы является разработка адаптивной методики кластеризации фактографических данных на основе дивизимных и итерационных методов.

Разработанная методика предназначена для бизнес – аналитиков, специалистов по анализу данных, разработчиков систем класса Data Mining.

Основными задачами работы являются:

1. Исследование методов и подходов интеллектуального анализа данных, используемых для кластеризации фактографических данных.

2. Разработка методики выбора существующих алгоритмов кластеризации.

3. Разработка методики адаптивной кластеризации фактографических данных.

4. Разработка методов докластеризации.

5. Разработка программного комплекса для автоматизации предложенной методики кластеризации.

6. Оценка эффективности предложенной методики с помощью экспериментальных исследований.

В исследуемой задаче выявлены две проблемы: выбор метода кластерного анализа, оценка полученных результатов.

Практическая задача, лежащая в основе научного исследования, связана с выделением групп клиентов брокерского обслуживания на основе интервальной информации об операциях клиента.

Брокерское обслуживание клиентов на фондовых рынках профессиональным участником торгов заключается в том, что клиенты, юридические и физические лица, заключают договор с брокером на осуществление комплекса услуг в интересах и за счет клиента. На основе полученной интервальной информации об операциях клиентов перед группой аналитиков ставится задача получения фактически сформированных групп клиентов для разработки коммерческого предложения по оказанию услуг с целью привлечения новых клиентов/групп клиентов.

Практическая задача, лежащая в основе научного исследования, имеет свои характерные особенности:

1. Количество исходных объектов – от 10 000 до 50 000 шт.

2. Количество значимых характеристик объектов – от 70 до 150 шт.

3. Типы характеристик – числовые, лингвистические.

4. Форма получаемых кластеров – сложная, с пересечениями.

5. Количество кластеров – результат исследования, от 5 до 30 шт.

6. Качество кластеризации – высокое.

7. Изменение объектов анализа – ежедневное.

Методика выбора метода кластеризации состоит из трех шагов: первый – выбор методики кластерного анализа, второй – настройка параметров выбранной методики (характеристических, итерационных, экспертных), третий – выполнение кластерного анализа и оценка полученного разбиения. В случае получения разбиения требуемого уровня качества поиск метода прекращается, в противном случае осуществляется переход на второй или первый шаг.

На основе литературных источников получена общая классификация методов кластерного анализа, в которых также предложены практические рекомендации по применению данных методов для решения практических задач, их достоинствах и недостатках. При анализе источников выделено восемь критериев выбора метода кластерного анализа: объем информации, размерность информации, типы атрибутов сущностей, чувствительность к равномерности информации и наличию отклонений, форма, количество, перекрываемость кластеров, качество кластеризации. Из исследованных методов выделены входные параметры, которые представлены в виде диаграммы. На диаграмме выделены параметры, имеющие одинаковую смысловую нагрузку в различных методах. На основе разработанных критериев предложен алгоритм выбора метода кластерного анализа.

После анализа существующих методик и алгоритмов для использования в адаптивной методике ADAKL были выбраны методики из различных классов, и используется теория графов и нечеткая логика. Определяющими факторами в выбранной комбинации являются:

— во-первых, использование теории графов в алгоритме позволяет осуществлять выделение кластеров произвольной формы и оптимальной структуры;

— во-вторых, при использовании математического аппарата нечеткой логики решается задача разделения объектов с лингвистическими атрибутами и позволяет сделать более полное разбиение исходного множества на кластеры, ликвидируя неопределенности, которые возникают при четком разбиении.

В основе разработанной методки ADAKL используется метод MST и идеи метода Fuzzy C-means. Данная методика состоит из пяти этапов: первый - нормализация значений числовых атрибутов, второй - вычисление матрицы взаимных расстояний между объектами, третий - построение минимального остовного дерева, четвертый - разделение объектов на кластеры и построение матрицы нечеткого разбиения, пятый - выбор наилучшего разбиения на основе оценки. Укрупнено данный алгоритм выполняет два

крупных действия – первичное размещение данных по кластерам, а затем – уточнение полученных центров кластеров и перераспределение объектов.

Адаптивный алгоритм кластеризации представлен в математической форме на двух листах с описанием входных параметров, необходимых для проведения кластерного анализа. При методики к работе с практической задачей наибольшее влияние, кроме самих значений атрибутов, оказывают следующие входные параметры алгоритма: весовой коэффициент влияния атрибута объекта, размазанность кластеров, степень удаленности элементов, способ определения расстояния между объектами, способ проведения нормализации значений числовых атрибутов.

Для разработанного алгоритма выполнена аналитическая оценка сложности, которая показывает, что разработанный алгоритм имеет:

— линейную зависимость аналитической сложности от количества входных атрибутов вне зависимости от их типа, от общего количества кластеров.

Разработанный алгоритм обладает следующими достоинствами:

— двухэтапная кластеризация фактографических данных;

— способен работать с лингвистическими атрибутами объектов кластеризации;

— использует весовые коэффициенты для анализируемых атрибутов объектов с целью повышения/понижения влияния атрибутов на результаты кластеризации;

— использует степень удаленности объектов/элементов для соотнесения объектов в кластеры при разделении;

— использует размазанность кластера, для определения нечеткости отнесения объекта к кластеру;

— использует способ определения расстояния между объектами, разработанный на основе базовых метрик: Евклидово расстояние, Квадрат Евклидова расстояния, расстояние Чебышева, с введением в функцию вычисления расстояния весовых коэффициентов и логики по вычислению расстояний между значениями лингвистических атрибутов;

— предлагает три способа построения минимального остовного дерева, результат работы которых одинаков, но все три способа отличаются разной вычислительной сложностью, что является определяющим на больших объемах данных;

— использует критерий оценки разбиения на кластеры с учетом специфики решаемой практической задачи: небольшое количество кластеров, наибольшая плотность, средняя удаленность объектов.

Предложенный алгоритм обладает следующими недостатком: квадратичная зависимость аналитической сложности алгоритма от количества исходных данных по объектам кластеризации.

Проверка оценки аналитической сложности с помощью эмпирических данных отражает незначительные расхождения с результатами анализа фактических данных.

Для оценки эффективности разработанной методики проведено четыре серии исследований по пятьдесят испытаний, три из которых лежат в области деятельности кредитной организации на фондовых рынках, а четвертое – в области анализа транспортных средств с целью выделения характерных групп. Оценка эффективности приведена в графическом виде с использованием средних величин и диаграмм. Оценка разбиений проводилась на основе индекса истинности разбиения с использованием контрольного примера. При сравнении эффективности методик использовались следующие алгоритмы: карты Кохонена, метод k-средних, разработанный алгоритм (ADAKL). По результатам средней оценки разбиений алгоритмы расположились в следующем порядке: ADAKL, карты Кохонена, метод k-средних. Превосходство разработанной методики перед картами Кохонена достигается использованием математического аппарата нечеткой логики и внутренних словарей системы при определении информационных расстояний между объектами.

Ввиду того, что одним из самых ресурсоемких этапов является третий этап, то для расширения исходных данных в процессе проведения анализа необходимо произвести докластеризацию добавляемых данных и расширить алгоритм проведения анализа исходных данных ADAKL еще шестью этапами исследования, на которых выполняется нормализация значений числовых атрибутов, расчет оценочной функции обоих наборов данных, определение разницы оценки между полученными значениями оценочных функций, поиск ближайшего элемента из исходного множества и распределение добавленных объектов по полученным от начального разбиения кластерам.

Анализ эмпирических данных подтверждает незначительность по сравнению с временем работы основного алгоритма времени работы докластеризации дополнительных данных. Время работы докластеризации сравнимо с временем работы второго этапа основной методики.

На укрупненной схеме архитектуры разработанного программного комплекса отражены основные этапы работы программного решения и его место в общей схеме обработки информации.