Нейский И.М.

Тема моей диссертационной работы: методика адаптивной кластеризации фактографических данных.

Целью работы является повышение эффективности процесса кластеризации фактографических данных с использованием методов интеллектуального анализа данных.

Экономическим назначением данной работы является оптимизация списка услуг брокерского обслуживания для выделенных групп клиентов, с технической точки зрения данная работа предназначена для выделения групп клиентов на основе данных об их операциях в рамках брокерского обслуживания.

Основными задачами работы является разработка методики проведения кластерного анализа фактографических данных с использованием имеющихся алгоритмов на основе статистических и характеристических данных об исходном наборе данных и разработка адаптивного алгоритма кластеризации фактографических данных смешанного типа.

Предметной областью для диссертационной работы выбрано брокерское обслуживание клиентов на фондовых рынках профессиональным участником торгов. Данный процесс состоит из того, что клиенты, юридические и физические лица, поступая на брокерское обслуживание, предоставляют свою персональную информацию, переводят денежные средства и ценные бумаги на счета брокерского обслуживания брокеру, оформляют договор. Далее процесс носит итеративный характер: клиент подает поручения на осуществление операций в соответствии с возможностями по оформленному договору, получает консультации, получает выписки и соответствующие документы о проводимых операциях, вводит/выводит денежные средства, вводит/выводит ценные бумаги. Брокер в свою очередь, в соответствии с поручениями, инициирует операции от имени клиента на биржевом, внебиржевом рынках, в депозитарном центре. Также данная работа найдет применение в следующих исследованиях: анализ клиентской базы (кредиты физических лиц), ранжирование существующих финансовых инструментов, проведение анализа отраслей с целью инвестирования ресурсов.

В поставленной задаче выявлены следующие проблемы: 1) выбор метода интеллектуального анализа данных; 2) большинство имеющихся на данный момент аналитических систем имеют узкую специализацию на предметной области; 3) при использовании одного инструмента интеллектуального анализа данных теряется большое число значимых закономерностей; 4) сложность оценки полученного разбиения.

Формализованная модель предметной области описывает процесс брокерского обслуживания клиентов в виде объектной модели с выделением основных категорий сущностей и поставленной задачи в математической форме.

Классификация методов интеллектуального анализа данных отражает составляющие компоненты ИАД и отмечает, что кластеризация является одним из этапов интеллектуального анализа данных. Также изображена многоступенчатость формирования результата при обработке фактографической информации.

В процессе работы над диссертацией была проведена классификация методов кластеризации: по способу обработки данных; по способу анализа данных; по возможности расширения обрабатываемого объема данных; по времени выполнения кластеризации; по количеству применений алгоритмов кластеризации.

Для принятия решения о выборе метода кластерного анализа при наличии исходного набора данных выделено восемь критериев: объем информации, размерность информации, типы атрибутов сущностей, чувствительность к равномерности информации и наличию отклонений, априорное (экспертное) представление о форме, количестве, перекрываемости кластеров, качество кластеризации.

Для адаптации существующих методов к исходному набору данных проведен анализ использующихся в исследованных методах параметров, выделены общие параметры, которые имеют одинаковую смысловую нагрузку и используются в более чем одном методе. Среди полученных параметров выделено три группы: характеристические,

итерационные и экспертные параметры. При выполнении кластеризации в большинстве случаев не всегда имеется априорное представление о характеристиках выделяемых кластеров и сложность процесса адаптации алгоритма заключается в том, что при адаптации требуется решить задачу нахождения приемлемого баланса между характеристиками предметной области и возможностями настройки алгоритма кластеризации. Адаптация алгоритма к предметной области носит итерационный характер и не всегда имеет конечное решение в виде итоговых значений параметров настройки алгоритма кластеризации.

После анализа отобранных методик кластеризации, выявления критериев выбора методов кластеризации, выделения параметров методов для адаптации к предметной области и на основании экспертных высказываний для проведения адаптивной кластеризации данных смешанного типа разработан алгоритм, основой которого является интеграция алгоритмов MST и Fuzzy C-means. Определяющими в выбранной комбинации были способность при использовании графов выделять кластеры произвольной формы и оптимальной структуры, при использовании математического аппарата нечеткой логики выполнить разделение объектов с лингвистическими атрибутами. Еще одним достоинством нечеткой кластеризации является то, что использование нечеткости при определении объектов по кластерам позволит сделать более полное разбиение исходного множества на кластеры, ликвидируя тем самым неопределенности, которые возникают при четком разбиении.

Данный алгоритм состоит из пяти этапов, часть из которых можно разбить на более мелкие части – подэтапные шаги. Разработанный алгоритм имеет конфигурационные параметры и состоит из следующих этапов: 1) нормализация значений числовых атрибутов; 2) вычисление матрицы взаимных расстояний между объектами; 3) построение минимального остовного дерева; 4) разделение объектов на кластеры и построение матрицы нечеткого разбиения; 5) выбор наилучшего разбиения на основе оценки.

Адаптивный алгоритм кластеризации представлен в математической форме на двух листах с описанием входных параметров, необходимых для проведения кластерного анализа. При адаптации алгоритма к работе в той или иной предметной области 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.

Весовой коэффициент позволяет выделять отдельную статистическую информация, которая играет наибольшую роль при выделении групп объектов. Размазанность кластеров позволяет делать более четкими или наоборот более размазанными границы между группами объектов. Также значительную роль при выделении групп объектов играет способ расчета информационного расстояния между объектами, а впоследствии и кластерами, с помощью которого можно сделать анализ «гладким», применив Евклидовы метрики, или сделать анализ «грубым», применив метрики с математическими функциями минимизации и максимизации.

Способ нормализации является дополнительным механизмом улучшения качества кластеризации, который особенно актуален при наличии в данных каких – либо выбросов и отклонений. Степень удаленности элементов является регулятором идентичности элементов при определении объектов по группам, влияющим на количество итоговых групп.

Разработанный алгоритм обладает следующими достоинствами:

— двухэтапная кластеризация фактографических данных;

— способен работать с лингвистическими атрибутами объектов кластеризации с применением нечеткой логики и введением словарной системы для вычисления расстояний между объектами входного набора данных;

— использует весовые коэффициенты для анализируемых атрибутов объектов с целью повышения/понижения влияния атрибутов на результаты кластеризации и адаптации алгоритма к различным предметным областям;
— 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 предметной области: небольшое количество кластеров, наибольшая плотность, средняя удаленность объектов.

Анализ экспериментальных данных показывает большое количество клиентов с близкими параметрами по операциям с ценными бумагами и денежными средствами, но при этом среди этого множества клиентов можно выделить следующие группы, количество которых не является конечным: долгосрочные инвесторы, спекулянты, потребители, фонды управления активами пайщиков. Выявленный перечень групп не является конечным ввиду постоянного развития деятельности компании в этом направлении, что подтверждается постоянным ростом клиентской базы.

Для оценки эффективности разработанной методики проведено четыре исследования, 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 отражением следующих диаграмм: количество полученных кластеров, оценка разбиения, средняя оценка разбиения и итоговая оценка разбиения. Оценка разбиений проводилась на основе индекса истинности разбиения. 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.

Разработанное программное решение имеет двухзвенную архитектуру клиент – сервер и нацелено на использование данного программного обеспечения с СУБД, поддерживающими стандарт SQL92.

На укрупненной схеме обработки потоков бизнес-информации и архитектуры разрабатываемого программного комплекса отражены основные этапы работы программного решения и его место в общей схеме обработки информации. При обработке входного потока информации очистка данных может проводиться в несколько этапов: фильтрация незначащих элементов данных, фильтрация незначащих параметров данных, удаление противоречий. Также программное решение позволяет производить накопление полученных результатов кластеризации для дальнейшего использования при выполнении анализа.

На инфологической модели представлена схема базы данных, на которой можно выделить две группы сущностей: сущности, используемые для работы программного решения в части ведения справочников и перечня настраиваемых параметров,

используемых для анализа, а также сущности, позволяющие получить агрегированные показатели деятельности компании в исследуемой предметной области. Даталогическая модель базы данных получена на основе построенной инфологической модели. Полученная модель находится в первой нормальной форме.

При работе над диссертацией решены следующие из поставленных задач: проведено исследование методов интеллектуального анализа данных, разработаны критерии выбора методов кластеризации, разработана метода адаптации к предметной области, разработан алгоритм адаптивной кластеризации на основе интеграции методов интеллектуального анализа данных, разработана архитектура программного решения и программное решение, проведена апробация метода и анализ экспериментальных данных.

Для демонстрации работоспособности адаптивного алгоритма ADAKL проведем анализ данных об автомобилях с целью получения типов транспортных средств на основе их технических характеристик:

Для проведения анализа используем следующий исходный набор данных:

Идентификатор автомобиля	Максимальная скорость	Характерный цвет	Сопротивление воздушному потоку	Масса автомобиля	Ожидаемая группа
V1	220	red	0.30	1300	1
V2	230	black	0.32	1400	1
V3	260	red	0.29	1500	1
V4	140	grey	0.35	800	2
V5	155	blue	0.33	950	2
V6	130	white	0.40	600	2
V7	100	black	0.50	3000	3
V8	105	red	0.60	2500	3
V9	110	grey	0.55	3500	3

Для проведения эксперимента используем Евклидову метрику для определения расстояний между объектами и алгоритм Борувки для постоения минимального остовного дерева.

Эксперименты:

Номер экспери-мента	Весовые коэффициенты				Размазанность	Степень удаленности	Результат разбиения								
1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	0	1	1	2	1	1	1	1	1	1	1	1	1	1
3	1	0	1	1	2	0.45	1	1	1	1	1	1	2	2	2
4	1	0	1	1	2	0.35	1	1	1	2	2	2	3	3	3
5	1	1	1	1	2	0.35	1	1	1	1	2	1	2	2	3
6	1	0.5	1	1	2	0.35	1	1	1	1	1	1	2	2	3
7	1	0.13	1	1	2	0.35	1	1	1	2	2	2	3	3	3

Сводная таблица по оценке информационного расстояния между цветовыми оттенками:

	black	red	grey	blue	white
black	0.0000	0.3333	0.7451	0.3333	1.0000
red	0.3333	0.0000	0.5817	0.6667	0.6667
grey	0.7451	0.5817	0.0000	0.5817	0.2549
blue	0.3333	0.6667	0.5817	0.0000	0.6667
white	1.0000	0.6667	0.2549	0.6667	0.0000

Таким образом, применение в анализе символьных атрибутов объектов позволяет получать более обширное разбиение с использованием большинства параметров объектов, но данный процесс вносит необходимость экспертной или итерационной оценки качества разбиения при выявлении коэффициента влияния атрибута на информационное расстояние между объектами.