Нейский И.М.

Экспериментальные исследования адаптивной кластеризации фактографических данных

Большинство современных предприятий используют в своей деятельности информационные системы. Хранилища данных, в которых собираются данные по бизнес – процессам компании. Объемы накапливаемой информации увеличиваются с течением времени, поэтому актуальной задачей в развитии компании является переход от анализа тенденций текущих показателей деятельности предприятия к более комплексном подходу «извлечения знаний» из имеющихся данных в целях выявления закономерностей.

Изучением проблем и созданием решений в этой области активно занимаются направления Интеллектуальный анализ данных (Business Intelligence) и Управление знаниями (Knowledge Management), в рамках которых выделяются поднаправления Выявление знаний в базах данных (Knowledge Discovery in Databases), Анализ фактографических данных (Data Mining), Анализ неструктурированных данных (Text Mining) и др. Результаты исследований этих направлений положены в основу многих информационно-аналитических систем, которые используются, в основном, для персональной работы экспертов. Однако, современной тенденцией является применение указанных технологий и для централизованного управления организациями.

Для исследования структурированных массивов информации используется анализ фактографических данных, в котором выделены шесть различных задач: классификация, регрессия, кластеризация, выявление ассоциаций, выявление последовательностей, и прогнозирование.

Потребность в кластеризации возникает в тех областях/этапах деятельности, где есть необходимость в разбиении объектов (ситуаций) на непересекающиеся подмножества, называемыми кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Четкое разделение на кластеры возможно только в идеальных условиях и при сильно различающихся параметрах объектов кластеризации., поэтому для решения реальных задач все чаще применяются нечеткие методы, в которых разбиение объектов (ситуаций) выполняется на частично пересекающиеся подмножества.

Примерами практических задач, в которых используется, требуется или планируется применение кластерного анализа, являются следующие задачи:

- выделение групп клиентов брокерского обслуживания для формирования перечня предлагаемых сервисов;
- формирование потребительской корзины;
- принятие решения о выдаче потребительского кредита;
- сегментирование сферы деятельности с целью повышения эффективности производительности;
- обработка изображений;
- тематический анализ библиотеки документов;
- оптимизация использования складских помещений;
- выявление транзакций, проведенных по поддельным кредитным картам;
- выявление потенциальных болезней пациентов;
- построение показательной (репрезентативной) выборки и т.д.

На примере одного из крупных российских банков актуальной задачей, связанной со значительным ростом клиентской базы (приблизительно в 20 раз), стала задача разработки «удачных» тарифных политик для обслуживаемых клиентов в области брокерского обслуживания. Учитывая данные обстоятельства, использование ранее применяемых методов с привлечением только человеческих ресурсов стало невозможно, так как объем информации для анализа также возрос. Для перехода на использование машинных методов необходимо осуществить выбор метода или методов из существующих либо разработать собственный метод с учетом особенностей этой области.

Главная особенность анализа этой области в том, что его необходимо проводить на регулярной основе, чтобы сохранить конкурентные преимущества на рынке данного вида услуг, так как сложившаяся кризисная обстановка на всех мировых финансовых рынках ведет не только к спаду большинства показателей в различных отраслях экономики, но и является возможностью достичь более значимых результатов за счет повышения собственной эффективности. Поэтому появляется необходимость использовать дополнительные, ранее не используемые ресурсы, которые сосредоточены в компании - внутреннее информационное поле (аналитики, эксперты, накопленная информация об операциях, клиентах и т.д.). Сложностью проведения такого анализа является нетривиальность разыскиваемых закономерностей в силу большого количества информации и наличии НЕ-факторов.

Анализ существующих решений и методов показал, что на текущий момент не существует специализированных или успешно примененных универсальных методов для решения описанной задачи. На данный момент известно более 100 методов кластеризации, поэтому для проведения конкретного практического исследования всегда возникает задача обоснованного выбора наиболее подходящего метода. Еще одной проблемой в данной области является оценка качества получаемого результата и выбор количества групп – кластеров, которое является входным параметром для большинства алгоритмов. Таким образом, задача построения эффективной тарифной политики разбивается на две подзадачи: техническую и экономическую. Решение экономической задачи состоит в оценке стоимости внедрения и сопровождения предлагаемых продуктов, а решением технической задачи является сбор и анализ имеющейся информации с помощью одного или совокупности методов фактографического анализа.

В связи с тем, что на данный момент не существует достаточного количества практических рекомендаций по применению существующих методов в данной предметной области и количество методов достаточно велико, была разработана методика адаптивной кластеризации, которая направлена на решение этой задачи. Данная методика, состоящая из четырех этапов, позволяет осуществить выбор метода кластерного анализа и получить конечное разбиение множества исходных объектов на кластеры. На основе методики получено, что для решения задачи разбиения клиентов брокерского обслуживания необходимо разработать новый метод адаптивной кластеризации, в котором количество кластеров является результатом исследования.

После проведенного анализа для решения поставленной задачи из инструментов выполнения кластеризации были выбраны: теория графов и нечеткая логика. Определяющими факторами в выбранной комбинации является способность при использовании графов выделять кластеры произвольной формы и оптимальной структуры, а при использовании математического аппарата нечеткой логики решается задача разделения объектов с лингвистическими атрибутами. За основу для нового метода в части первичного разделения объектов на кластеры взята идея метода MST, использующего минимальные остовные деревья, и идея метода Fuzzy C-means. На базе этих методов разработан метод ADAKL, который является двухэтапным и использует оценочную функцию разбиения, повышающую качество проводимой кластеризации. Вычисление глобального критерия делает алгоритм кластеризации во много раз быстрее, чем при использовании локального критерия при парном сравнении объектов.

Совокупность использованных методов и их модификация позволили преодолеть недостатки каждого из них: для MST - применение нечеткости позволяется сделать более плавное разбиение, помещая объекты в разные кластеры с разной степенью принадлежности, для Fuzzy C-Means - предварительное использование MST и модифицированного критерия оптимальности позволяет сократить количество итераций исследования входного набора данных, а следовательно, и снизить временные, человеческие и технические затраты на проведение исследований.

Вместе с тем предложенный метод обладает квадратичной зависимостью аналитической сложности алгоритма от количества исходных данных по объектам кластеризации, что существенно увеличивает временные затраты при регулярном появлении новых данных и повторной кластеризации.

Частично преодолеть этот недостаток можно за счет специальной процедуры докластеризации, которая определяет необходимость повторного запуска исследования полного массива данных и, в случае отсутствия признаков появления новых значимых групп объектов, осуществляет распределение новых (расширяющих) объектов по имеющимся кластерам. Для расширения исходных данных в процессе проведения анализа необходимо произвести дополнительное исследование добавляемых данных.

Для оценки работоспособности ADAKL в сравнении с другими алгоритмами были проведены три основных и одна дополнительная серии экспериментов. Исследование производилось на трех методах (самоорганизующиеся карты Кохонена, алгоритм k-средних и метод ADAKL), по результатам которого наилучшее разбиение на исследованных массивах по сериям экспериментов получено с применением метода ADAKL.

Проведенные эксперименты подтвердили, что использование интеграции методов кластеризации (многоэтапная кластеризация) улучшает качество выявления знаний в сравнении с одноэтапными методами, а также то, что превосходство разработанного метода достигается использованием математического аппарата нечеткой логики и внутренних словарей системы при определении информационных расстояний между объектами.

На основе метода ADAKL было разработано программное решение, с помощью которого выполнялось выделение групп клиентов и определение их доли от общего количества клиентов. Последующий анализ экономических показателей полученных групп объектов позволил дать названия кластерам, и разработать более целевую, направленную на конкретную клиентскую группу тарифную политику, а также предложить им более выгодные условия по совершаемым видам операций, увеличив количество этих операций и объем комиссионных сборов.