Н.А. БАРЫШНИКОВА

МЕТОДИКА КЛАСТЕРИЗАЦИИ НА ОСНОВЕ ИЕРАРХИЧЕСКИХ И НЕИЕРАРХИЧЕСКИХ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА

В статье рассматривается необходимость проведения процедуры измерения показателей образовательных систем, которые должны быть открытыми, измеримыми для проведения различного рода аудитов (внутренних и внешних). Эксперты, проводящие измерения, анализирующие и принимающие решения по выявленным несоответствиям, должны быть обеспечены методиками мониторинга образовательных процессов и анализа показателей качества образовательных услуг. Описывается возможность использования для этих целей методики адаптивной кластеризации фактографических данных на основе дивизимных и итерационных методов.

Кластерный анализ, экспертный опрос, кластеризация, иерархические методы, агломеративные и дивизимные методы, неиерархические методы, итеративные методы.

В дискуссиях о новой системе образования периода инновационного развития России образовательные учреждения ориентируются на работу в открытой системе. Таким образом, работодатели, учредители и потребители новых знаний получат возможность реально влиять на положение дел в каждом отдельно взятом учебном заведении. Новая модель системы образования, очевидно, потребует и нового управления.

В современных условиях конкурентной борьбы важным условием выживания, развития образовательного учреждения является наличие у него системы менеджмента качества (СМК), соответствующей требованиям стандартов серии ISO 9000. Стандарт ISO 9001:2008 «Системы менеджмента качества. Требования» инициирует систему управления образовательного учреждения осуществлять мониторинг, измерение и анализ основных процессов образовательной деятельности [1, п.4.1]; определяет ориентацию руководства учебного заведения на потребителя образовательных услуг [1, п.5.2]; требует постоянного анализа СМК со стороны руководства [1, п.5.6], а также устанавливает требование наличия обратной связи от потребителей во входных данных для анализа СМК со стороны руководства [1, п.5.6.2].

Показатель удовлетворённости всех заинтересованных сторон становится в экономически развитых странах глобальным критерием совершенства (оптимальности) деятельности любой организации [2]. Поэтому приобретают актуальность постановка и решение задачи количественной оценки (и работающих методик измерения) этого показателя. Однако в настоящее время методы измерения показателя удовлетворённости заинтересованных сторон ещё недостаточно разработаны [3]. Одним из фактов постановки задачи количественной оценки удовлетворённости потребителей продукции или услуг является то, что удовлетворённость как таковая представляет собой нечёткое, размытое понятие, на значение которой влияют субъективные суждения, восприятие и эмоции человека. Значения удовлетворённости как аргумента выражаются в виде нечётких суждений через понятия естественного языка и плохо описаны методами математического анализа. Таким образом, возрос спрос на новые формализованные средства анализа проблем и оценки альтернативных вариантов решения выявленных проблем.

В ходе разработки методического сопровождения систем поддержки принятия решений при экспертной оценке качества альтернатив необходимо решить следующие задачи:

- разработать методику ранжирования многокритериальных альтернатив, учитывающую как числовые, так и лингвистические критерии, представленные в различных шкалах;
- выстроить методику корректировки весовых характеристик экспертов в соответствии с их квалификационным уровнем, опытом работы;
- разработать стратегию проведения общественной оценки потребителями образовательных услуг;
- разработать практические методы выбора управленческого решения, отвечающие предпочтениям экспертов и лиц, принимающих решения.

Регистрация величин, характеризующих субъективные значения количественных параметров при проведении экспертного опроса, обычно строится так, чтобы эксперт вполне однозначно отвечал на стандартный вопрос о предпочтении одного решения другому, о месте объекта в ранговом ряду, о числе, которое следует приписать объекту в заданной балльной шкале, и т.д.

В настоящее время в учреждениях, оказывающих образовательные услуги, активно разрабатываются и внедряются информационно-аналитические системы, которые направлены на мониторинг образовательных процессов и анализ ключевых показателей качества. Одна из важнейших задач в этой области – анализ и агрегирование многочисленных фактографических данных, часто решаемая с использованием методов кластеризации. На данный момент известно более 50 методов кластеризации, которые представлены в математической и алгоритмической форме, но при этом немногие имеют реализацию и рекомендации по использованию в сфере образования. Знание того, какие методы дают наилучший результат, может подсказать направление движения тем, кто планирует применять кластерный анализ для решения практических задач, создаёт новые алгоритмы или совершенствует существующие.

На основе существующих методов кластерного анализа построена классификация, которая разделяет методы по способу обработки данных на иерархические и неиерархические [5]. Иерархические методы в соответствии с классификацией делятся на агломеративные и дивизимные [5]. Агломеративная группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением количества кластеров [6]. Дивизимная группа методов характеризуется последовательным разделением исходных элементов и соответствующим увеличением количества кластеров [6]. Самую значимую часть неиерархических методов представляют итеративные методы [5], основанные на разделении набора данных на некоторое количество отдельных кластеров. Существуют два подхода для разделения данных. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве характеристик объектов, т.е. определение кластера там, где имеется большое «сгущение» объектов. Второй подход заключается в минимизации меры различия объектов [6].

Основная задача кластеризации – получение кластеров на основе множества исходных объектов. Существующие методы описаны в научных источниках, в которых, как правило, приводятся практические рекомендации по их использованию и описательные характеристики их возможностей. Адаптивность кластеризации означает возможность применения метода к выбранной предметной области после соответствующих настроек и выполнения этапа обучения. Следует отметить, что методы кластерного анализа являются контекстно-зависимыми. В данном направлении интеллектуального анализа данных выявлены две проблемы: 1) потеря значимых закономерностей при использовании одного инструмента анализа; 2) вычислительная сложность и большие временные затраты при применении инструментов на исходных данных.

Предлагаемая интеграция методов из двух разных классов – дивизимного и итеративного – направлена на устранение выявленных проблем [8]. Практическая задача, для решения которой используется описанная методика, может иметь следующие характеристики:

- количество исходных данных – более 10 000 объектов;
- количество значимых характеристик объектов – более 70 штук;
- типы характеристик – числовые, текстовые;
- форма получаемых кластеров – сложная, с пересечениями;
- количество кластеров – результат анализа, а не входной параметр;
- качество анализа – высокое.

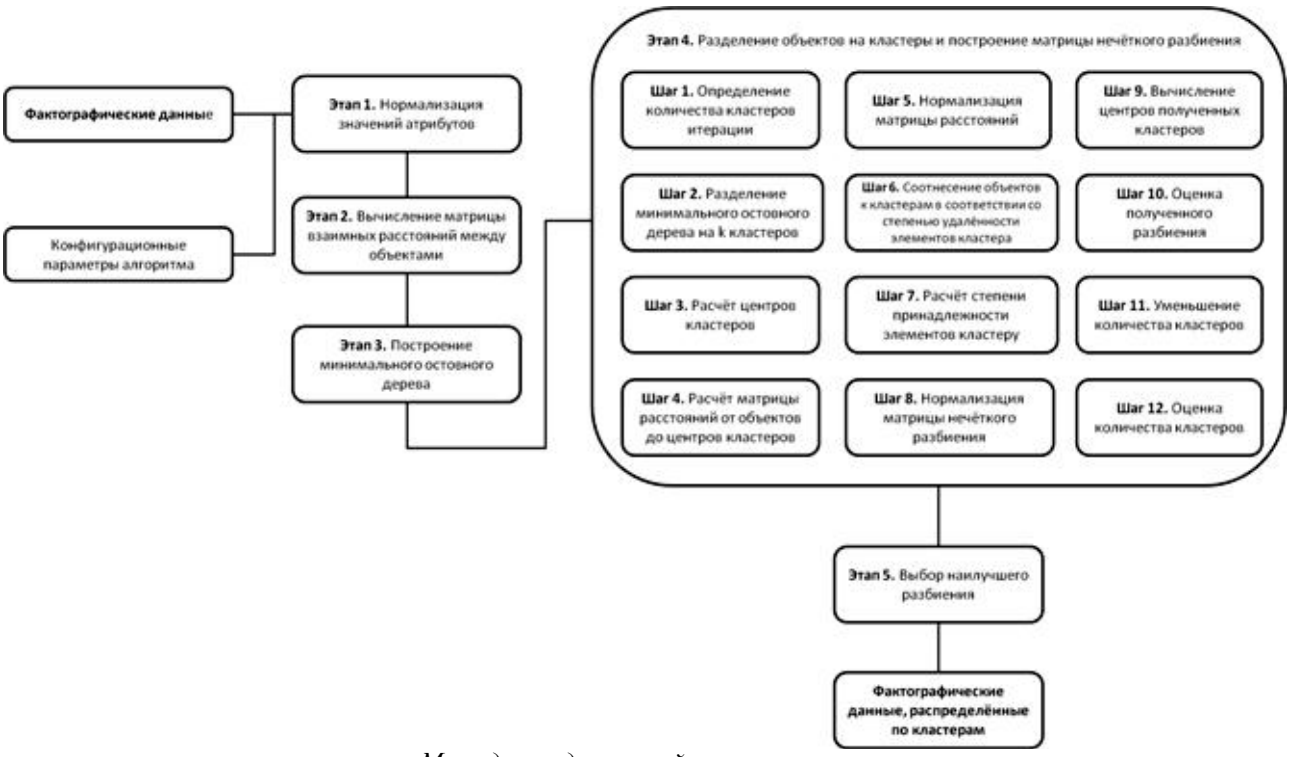
В рамках данной интеграции предлагается использование базовых принципов двух методов кластеризации: MST [7] и Fuzzy C-Means [9]. В результате интеграции получается методика с двухэтапной кластеризацией [8]. На первом этапе данной методики строится минимальное остовное дерево, образуя оптимизированную древовидную структуру из исходных элементов на основе характеристик кластеризуемых объектов. На втором этапе данной методики используется итеративный подход, с помощью которого сначала выделяются первичные кластерные центры на основе оптимизированной древовидной структуры, а потом центры кластеров и содержимое кластеров уточняются на основе вычисления степени принадлежности объекта кластеру и локального критерия остановки цикла. Сравнение достоинств и недостатков адаптивной методики представлено в таблице [8].

Для устранения чувствительности к выборам в дополнение к адаптивной методике предлагается использование предобработки исходных данных в виде фильтрации незначащих компонентов, нормализации данных и т.п.

Таким образом, методику адаптивной кластеризации можно представить оптимизированной древовидной структурой из исходных элементов на основе характеристик кластеризуемых объектов – на первом этапе. На втором этапе данной методики сначала необходимо выделить первичные кластерные центры на основе оптимизированной древовидной структуры, а затем центры и содержимое кластеров уточняются на основе вычисления принадлежности объекта кластеру (см. рисунок).

Сравнение методов кластеризации MST и Fuzzy C-Means

Метод	Достоинства	Недостатки
MST	Простота использования; высокая скорость; понятность и прозрачность алгоритма	Алгоритм слишком чувствителен к выбросам; требуется задание количества кластеров
Fuzzy C-Means	Возможность частичного отнесения объекта к нескольким кластерам	Высокая вычислительная сложность; требуется задание количества кластеров; неопределенность с объектами, которые удалены от центров всех кластеров
Адаптивная методика	Понятность и прозрачность алгоритма; двухэтапная кластеризация; нечеткость при определении объекта в кластер; возможность использования объектов с разными типами атрибутов (числовые и текстовые); количество кластеров определяется в результате анализа; приемлемое время работы и конечность результата	Нелинейная зависимость времени анализа от количества исходных объектов; чувствительность к выбросам



Методика адаптивной кластеризации

Количество объектов исследования с целью разбиения на кластеры постоянно растёт во время проведения анализа массива фактически зарегистрированных данных, увеличивая время проведения исследования, поэтому может возникнуть задача оптимизации времени анализа исходных данных в виде докластеризации новых объектов, которыми могут быть данные при проведении внутренних аудитов СМК в образовательных учреждениях. Количество запланированных аудитов может расти в связи с внедрением и реализацией федеральных государственных образовательных стандартов третьего поколения в учреждениях профессионального образования, одной из целей которых является привлечение потенциальных потребителей – будущих работодателей к участию в формировании и реализации профессиональных образовательных программ.

ЛИТЕРАТУРА

1. ИСО 2001:2008. Системы менеджмента качества. Требования. М., 2008. 63с.
2. *Конти Т.* Система заинтересованных сторон: Стратегическая ценность//Методы менеджмента качества. 2003. №1.
3. *Швец В.Е.* Измерение процессов в системе менеджмента качества: опора на стратегию и структуру // Сертификация. 2003. №1.
4. *Мирошников В.В., Борбаць Н.М.* Методика оценки удовлетворённости заинтересованных сторон организации // Информационые технологии. 2007. №3.
5. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP/ А.А. Баргесян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. 2-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2008.
6. *Чубукова И.А.* Data Mining: учебное пособие. М.: Интернет-Университет Информационных Технологий; БИНОМ; Лаборатория знаний, 2006.
7. *He H., Singh A.* Efficient Algorithms for Mining Significant Substructures in Graphs with Quality Guarantees. – Department of Computer Science University of California. Santa Barbara, 2004.
8. *Нейский И.М.* Адаптивная кластеризация на основе дивизимных и итерационных методов // Информационные технологии в образовании, науке и производстве: сборник трудов третьей международной научно-практической конференции /под ред. Ю.А. Романенко. М., 2009.
9. *Штовба С.Д.* Введение в теорию нечетких множеств и нечеткую логику [Электронный ресурс]. URL:http://matlab.exponenta.ru/fuzzylogic/book1/index.php.