УДК 681.5

С.А. Батуркин, Е.Ю. Батуркина, В.А. Зименко, И.В. Сигинов

СТАТИСТИЧЕСКИЕ АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ ДАННЫХ В АДАПТИВНЫХ ОБУЧАЮЩИХ СИСТЕМАХ

Приводятся основные положения теории статистической кластеризации, обобщенная классификация алгоритмов кластеризации. Описываются метод кластеризации к-средних (k-means) и варианты его применение для кластеризации обучаемых в адаптивных обучающих системах.

***Ключевые слова**: обучающие системы, адаптивный, алгоритм, кластер, статистический, стратифицирующий.*

Введение. Управление процессом обучение строится на основе принципов системного и процессного подходов. Данные, используемые в различных процессах необходимо структурировать, классифицировать, подвергать тщательному анализу. При организации обучения в группах учащихся необходимо создание аналитической системы, позволяющей группировать учащихся по различным категориям. Одним из видов дифференциации является кластеризация (от англ. cluster - скопление), то есть выделение объединений однородных элементов, которые могут рассматриваться как самостоятельные единицы, обладающие определёнными свойствами. Именно элементы кластеров, являющиеся в сущности многомерными векторами линейных пространств над полем признаков, удобнее всего рассматривать при моделировании учебного процесса. Моделирование учебного процесса позволяет значительно улучшить качество его анализа и тем самым создать один из надежнейших инструментов управления.

В случае организации обучения задача кластеризации решается зачастую методами математической статистики. Благодаря их применению появляется возможность автоматизации процесса кластеризации за счет применения электронно-вычислительных средств. Для адаптивных обучающих систем, реализованных на программном уровне, задачи кластеризации решаются только посредством применения аппарата математической статистики.

Среди наиболее известных методов кластеризации можно выделить следующие:

- *k*-средних (*k*-means);
- графовые алгоритмы кластеризации;
- статистические алгоритмы кластеризации;
- алгоритм ФОРЕЛЬ;
- иерархическую кластеризацию или таксономию;
- нейронную сеть Кохонена;
- ансамбль кластеризаторов [1].

Постановка задачи. Формальная постановка задачи кластеризации выглядит следующим образом: пусть X — множество объектов, Y — множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $p(x,x')$. Имеется конечная обучающая выборка объектов $\boldsymbol{X^m = \{x_1 ... x_m\} \subset X}$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике p, а объекты разных кластеров существенно отличались. При этом каждому объекту $\boldsymbol{x_i \in X^m}$ приписывается номер кластера y_i.

Алгоритм кластеризации — это функция $\boldsymbol{a : X \to Y}$, которая любому объекту $\boldsymbol{x \in X}$ ставит в соответствие номер кластера $\boldsymbol{y \in Y}$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров с точки зрения того или иного критерия качества кластеризации [2].

Кластеризация отличается от классификации тем, что метки исходных объектов y_i изначально не заданы и даже может быть неизвестно само множество Y. Последнее обстоятельство вводит следующие коррективы при применении кластерного анализа в обучающих системах:

- невозможность быстрого определения классов исследуемых объектов, в нашем случае – групп учащихся;
- невозможность применения стандартизированных методик;
- невозможность построения таксономий.

Однако, наряду с недостатками, статистическая кластеризация обладает и рядом достоинств:

- возможность задания заранее неизвестного класса объектов по начальным характеристикам;

- возможность кластеризации сколь угодно большого количества объектов в предельно короткие сроки;

- возможность введения в исследуемые совокупности стандартизованных индикаторов (среднестатистического профиля учащегося определенной категории).

Основной задачей приведенного исследования является поиск оптимального статистического алгоритма кластеризации, применимого в адаптивных обучающих системах. Подобный алгоритм улучшит качество управления учебным процессом за счет введения удобной математической модели обучаемого.

Статистические алгоритмы кластеризации. Статистические алгоритмы основаны на предположении, что кластеры описываются некоторым семейством вероятностных распределений, а сама задача кластеризации сводится к разделению смеси распределений по конечной выборке.

В основе кластерного анализа лежат две гипотезы байесовского подхода к разделению смесей вероятностных распределений [3].

Гипотеза 1 (о вероятностной природе данных). Объекты выборки X^ℓ появляются случайно и независимо, согласно вероятностному распределению, представляющему собой смесь распределений

$$p(x) = \sum_{y \in Y} w_y p_y(x), \qquad \sum_{y \in Y} w_y = 1, \quad (1)$$

где $p_y(x)$ - функция плотности распределения кластера y, w_y - неизвестная априорная вероятность появления объектов из кластера y. Конкретизируя вид распределений $p_y(x)$, чаще всего берут сферические гауссовские плотности. Это обычная практика - представлять кластеры в виде шаров.

Гипотеза 2 (о форме кластеров). Объекты описываются n числовыми признаками $f_1(x),\ldots,f_n(x)$, $X=R^n$. Каждый кластер $y \in Y$ описывается n-мерной гауссовской плотностью $p_y(x)=N(x;\mu_y,\sum_y)$ с центром $\mu_y=(\mu_{y1},\ldots,\mu_{yn})$ и диагональной ковариационной матрицей $\sum_y = diag(\sigma^2_{y1},\ldots,\sigma^2_{yn})$.

При этих предположениях задача кластеризации совпадает с задачей разделения смеси вероятностных распределений и для её решения можно применить ЕМ-алгоритм. На Е-шаге по формуле Байеса вычисляются скрытые переменные g_{iy}. Значение g_{iy} равно вероятности того, что объект $x_i \in X^l$ принадлежит кластеру $y \in Y$. На М-шаге уточняются параметры каждого кластера $(\mu_y,\sum_y)$, при этом существенно используются скрытые переменные g_{iy}.

Однако ЕМ-алгоритм обладает тем недостатком, что в результате кластеризации каждый объект $x_i \in X^l$ приписывается к каждому кластеру с определенной вероятностью. С точки зрения педагогической практики в случаях, когда идет работа с большими массивами данных по каждому учащемуся, результат подобной кластеризации будет труден для понимания, и для большинства преподавателей недоступен для интерпретации [3].

Метод *k*-средних. Зачастую в практических целях используется метод k-средних, являющийся упрощением ЕМ-алгоритма. Главное отличие в том, что в ЕМ-алгоритме каждый объект x_i распределяется по всем кластерам с вероятностями $g_{iy} = P\{y_i = y\}$. В алгоритме k-средних (k-means) каждый объект жёстко приписывается только к одному кластеру.

Другое отличие в том, что в k-means форма кластеров не настраивается. Однако это отличие не столь принципиально. Можно предложить упрощённый вариант ЕМ, в котором форма кластеров также не будет настраиваться - для этого достаточно зафиксировать ковариационные матрицы $\sum_y$, $y \in Y$ [4].

С другой стороны, возможен и обобщённый вариант k-means, в котором будут определяться дисперсии кластеров вдоль координатных осей. Общий алгоритм кластеризации с использованием алгоритма к-средних можно представить в следующем виде:

1) произвольно сформировать начальное приближение центров всех кластеров $y \in Y$. В качестве центров можно взять наиболее удалённые друг от друга объекты выборки μ_y;

2) отнести каждый объект выборки к ближайшему центру:

$$y_i := \arg\min_{y \in Y} p(x_i, \mu_y), i = 1,\ldots,l; \quad (2)$$

3) вычислить новое положение центров:

$$\mu_{yi} := \frac{\sum_{i=1}^{l} [y_i = y] f_i(x_i)}{\sum_{i=1}^{l} [y_i = y]}, \quad (3)$$
$$y \in Y, i = 1,\ldots,n$$

4) Продолжать вычисления, пока y_i не перестанут изменяться.

Существует два канонических варианта алгоритма k-means. Вариант Бола – Холла, представленный выше, и вариант Маккина [5], отличающийся тем, что шаги 3 и 4 выполняются внутри одного цикла по объектам выборки. Когда находится объект, переходящий из одного

кластера в другой, центры обоих кластеров пересчитываются.

Алгоритм *k*-means крайне чувствителен к выбору начальных приближений центров. Случайная инициализация центров на шаге 1 может приводить к плохим кластеризациям. Для формирования начального приближения лучше выделить k наиболее удалённых точек выборки: первые две точки выделяются по максимуму всех попарных расстояний; каждая следующая точка выбирается так, чтобы расстояние от неё до ближайшей уже выделенной было максимально.

Кластеризация может оказаться неадекватной и в том случае, если изначально будет неверно угадано число кластеров. Тогда проводится кластеризация при различных значениях k и выбирается то, при котором достигается резкое улучшение качества кластеризации по заданному функционалу [6].

Алгоритм *k*-средних наиболее полно удовлетворяет требованиям простоты применения и удобства интерпретации результатов кластеризации в случае его применения в адаптивных обучающих системах. Большое количество программно-аппаратных средств, использующих алгоритм *k*–средних, позволяет реализовывать обучающие системы различного уровня сложности.

Применение кластеризации в адаптивных обучающих системах. При анализе учебной активности и уровня знаний каждый учащийся в обучающей системе может быть связан с n-мерным вектором. Данный вектор содержит в себе оценки различных параметров учащегося, например - общий логит знаний, академическую активность, креативность и т.п. В итоге каждый студент описывается вектором $x_i = \{x_{i1}, \ldots, x_{in}\}$, $x_i \in X^l$, где X^l – совокупность векторов, характеризующих некоторую группу учащихся или определенное направление. В дальнейшем, для краткости, будем называть эти векторы «стратифицирующими векторами обучаемого» или просто «стратифицирующими векторами».

Для проведения валидного кластерного анализа совокупности стратифицирующих векторов необходимо задание первоначальных корректных условий кластеризации. Для выявления этих условий требуется первоначальная сортировка. Суть данной процедуры заключается в том, чтобы из совокупности X^l выделить необходимое количество устойчивых групп стратифицирующих векторов, в которых в дальнейшем будут выделяться векторы, имеющие среднестатистический для данной группы векторов набор параметров. Каждый такой вектор должен представлять из себя центр *n*–го кластера x_{in} при дальнейшем проведении кластеризации.

При проведении ряда процедур кластеризации по определенной дисциплине для определенной специальности или направления обучения формируются устойчивые стратифицирующие векторы – центры x_{in}. При задании первоначальной погрешности можно четко выделить эти устойчивые векторы, которые в дальнейшем будут применяться в качестве индикаторов при последующих процедурах кластеризации. Основная задача этих индикаторов - ускорить процесс кластеризации за счет того, что в качестве центров кластеров будут выбраны именно эти векторы. Таким образом, в структуре алгоритма *k*–средних будут убираться шаги 1 и 2.

Коррекцию индикаторов можно производить при желании и по необходимости в случаях существенного изменения программы обучения или содержания учебно-методических материалов курса.

Сами индикаторы будут представлять из себя не что иное, как наборы параметров, характеризующих определенную категорию учащихся.

Естественно, что указанные выше операции невозможно производить без программно-аппаратных средств автоматизации. Для написания корректного программного кода авторами статьи был разработан оригинальный алгоритм кластеризации, использующий в своей основе метод *k*–средних:

1) провести первоначальную сортировку, для чего из всей совокупности выделить необходимое (исходя из заданных условий) количество кластеров. Данная процедура полностью повторяет пп. 1-5 алгоритма *k*–средних;

2) выделить векторы - индикаторы. Для этого необходимо обработать несколько выборок исходных данных, используя алгоритм *k*–средних. В результате получаем несколько стратифицирующих устойчивых векторов x_{in}, являющихся индикаторами определенных категорий учащихся. Чем больше количество обработанных выборок, тем точнее получаются индикаторы;

3) использовать индикаторы x_{in} в качестве центров кластеров y_i при дальнейших операциях кластеризации, устраняя таким образом шаги 1,2,4,5 алгоритма *k*-средних.

Заключение. Для управления процессом обучения в адаптивных обучающих системах, необходимо валидное моделирование основных составляющих подпроцессов. Для принятия адекватных решений в вопросах управления, зачастую, необходимо применять математические методы анализа данных. Одним из наиболее понятных и распространенных методов явля-

ется кластеризация. Решение задачи кластеризации принципиально неоднозначно, и тому есть несколько причин:

- не существует однозначно наилучшего критерия качества кластеризации;

- число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием;

- результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом.

Разработанная авторами статьи методика, наряду с перечисленными, обладает еще одним недостатком: при проведении кластеризации требуется периодическая проверка валидности индикаторов, для чего с определенной периодичностью необходимо проверять отклонение параметров используемого индикатора от параметров векторов – центров кластера в одной из проверяемых совокупностей стратифицирующих векторов.

Однако, наряду с недостатками, методика обладает целым рядом достоинств:

- простота кластеризации, достигаемая за счет использования понятных алгоритмов;

- высокая скорость проведения анализа;

- простота интерпретации результатов кластеризации, достигаемая за счет того, что в шаге 3 разработанного алгоритма фактически используется иерархическая кластеризация, строятся таксономии [7];

- возможность использования алгоритма при обработке больших совокупностей данных.

Таким образом, каждый обучающийся в адаптивной обучающей системе может быть представлен посредством вектора, компонентами которого являются характеристики обучаемого, необходимые для оценки уровня его подготовки.

Дальнейшее развитие работы по данной тематике предполагает создание программного продукта, реализующего предложенную методику, и интеграцию его в существующую систему дистанционного обучения, с целью модернизации её до уровня адаптивной обучающей системы.

Библиографический список

1. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний.- Новосибирск: ИМ СО РАН, 1999.

2. *Загоруйко Н.Г., Ёлкина В.Н., Лбов Г.С.* Алгоритмы обнаружения эмпирических закономерностей. - Новосибирск: Наука, 1985.

3. *Кулаичев А.П.* Методы и средства комплексного анализа данных. – М.: ИНФРА- М, 2006.

4. *Лагутин М.Б.* Наглядная математическая статистика. М.: П-центр, 2003.

5. *Мандель И.Д.* Кластерный анализ. - М.: Финансы и статистика, 1988.

6. *Уиллиамс У.Т., Ланс Д.Н.* Методы иерархической классификации // Статистические методы для ЭВМ / под ред. М. Б. Малютова. - М.: Наука, 1986. - С. 269–301.

7. *Батуркин С.А., Гостин А.М., А.В. Пруцков и др.* Система внутреннего тестового контроля знаний РГРТУ: методические указания/ Рязан. гос. радиотехн. ун-т. – Рязань, 2007. – 68 с.