Матвеев Ю.Н., Даньшина А.Е., Стукалова Н.А., Туляков А.В. Обработка спектрофотометрической информации с использованием алгоритмов кластерного анализа. // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей XVI Междунар. научно-техн. конф. – Пенза: ПДЗ, 2016. – С. 121-124.

УДК 004.023

ОБРАБОТКА СПЕКТРОФОТОМЕТРИЧЕСКОЙ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА

Ю.Н. Матвеев, А.Е. Даньшина, Н.А. Стукалова, А.В. Туляков

SPECTROPHOTOMETRIC INFORMATION PROCESSING BY USING CLUSTER ANALYSIS ALGORITHMS

Yu.N. Matveev, A.E. Dan'shina, N.A. Stukalova, A.V. Tuljakov

Аннотация. Описан метод кластерного анализа, который позволяет обработать большие объёмы информации, приведены меры сходства объектов, на основании которых делаются выводы о «похожести» этих объектов. Примером метода кластеризации обработки информации служит метод k-средних.

Ключевые слова: фотометрия, спектрофотометрия, биологическиежидкости, кластерный анализ, метод k-средних, массив данных.

Abstract. There is given a reviewof a cluster analysis method which allows to process large volumes of information. The measurement of element similarities on the basis of whichconclusions about resembling of these objects are drawn is described. The method of k-averages is an example of a method of a clustering of information processing.

Keywords: photometry, spectophotometery, body liquids, cluster analysis, qualifier, k-means, multiple data.

Одним из методов исследования биологических жидкостей пациентов лечебных учреждений является спектрофотометрический анализ. Данный анализ основан на исследовании спектра излучения исследуемого препарата. Известно, что каждый химический элемент излучает определённый спектр оптических волн, по которому можно идентифицировать этот элемент. Подобно химическим элементам, молекулы также испускают определённый спектр излучения. Основываясь на этих данных, спектральный анализ позволяет определить химический состава вещества.

Измерительный прибор – спектрофотометр, с помощью которого определяют химический состав, улавливает излучения исследуемого раствора, преобразует их в электрические сигналы, которые и используются для расчёта необходимых параметров, причём эти параметры (световой поток, вязкость и прочее) могут быть различны в зависимости от вида и модели прибора. Однако какими бы ни были определяемые параметры, они образуют большой массив данных, требующий его упорядочивания.

Чтобы упорядочить большой объём информации, необходимо обратиться к методам классификации данных. Классификация – это система, в которой составляющие её объекты распределены на основании некоторых закономерностей, называемых классификаторами. Классификаторы основаны на использовании математических методов, позволяющих разбить исходные данные на классы.

Одним из таких методов является кластерный анализ. Он применяется, когда необходимо преобразовать большой массив данных в наглядные структуры –

группы. Кластером (от англ. cluster – гроздь, группа, скопление) является объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица, обладающая определёнными свойствами.

Для анализа данных используют меры сходства [1]. Выделяют четыре меры сходства:

1. Коэффициент корреляции – это показатель характера взаимного влияния изменения двух случайных величин. Коэффициент корреляции может принимать значения от -1 до +1. Если значение по модулю находится ближе к 1, то это означает наличие сильной связи, а если ближе к 0 – связь отсутствует. При коэффициенте корреляции, равном по модулю единице, говорят о линейной зависимости [2].

2. Мера расстояния устанавливает сходство или различие между объектами. Два объекта идентичны, если описывающие их переменные принимают одинаковые значения. В этом случае расстояние между ними равно нулю. Меры расстояния обычно не ограничены сверху и зависят от выбора шкалы измерений. Существует много различных мер расстояния, но наиболее часто используется евклидово расстояние [3].

3. Коэффициенты ассоциативности применяются, когда необходимо установить сходство между объектами, описываемыми бинарными переменными, причем 1 указывает на наличие переменной, а 0 – на ее отсутствие.

4. Вероятностные коэффициенты сходства – при образовании кластеров по этим мерам вычисляется информационный выигрыш от объединения двух объектов, а затем объекты с минимальным выигрышем рассматриваются как один.

Для определения сходства элементов системы необходимо составить вектор характеристик для каждого объекта. Известны алгоритмы, позволяющие работать как с числовыми, так и с качественными характеристиками [4].

После построения вектора характеристик проводят его нормализацию с целью одинакового вклада компонентов в расчет «расстояния». В процессе нормализации все значения приводятся к некоторому диапазону, например, [-1, 1] или [0, 1]. После этого можно переходить к измерению схожести, т.е. к измерению расстояния между компонентами. Стоит отметить, что выбор метрики полностью лежит на исследователе, поскольку результаты кластеризации могут существенно отличаться при использовании разных мер. После определения значений меры сходства между объектами необходимо применить метод кластерного анализа для создания групп сходных объектов (кластеров).

Подобно метрикам, существует множество методов кластеризации. Рассмотрим наиболее популярный – метод k-средних.

Этот алгоритм строит заданное число кластеров, расположенных как можно дальше друг от друга. Работа алгоритма делится на несколько этапов:

1. Случайно выбрать *k* точек, являющихся начальными «центрами масс» кластеров.
2. Отнести каждый объект к кластеру с ближайшим «центром масс».
3. Пересчитать «центры масс» кластеров согласно их текущему составу.
4. Если критерий остановки алгоритма не удовлетворен, вернуться к п. 2.

В качестве критерия остановки работы алгоритма обычно выбирают минимальное изменение среднеквадратической ошибки. Также возможно останавливать работу алгоритма, если на шаге 2 не было объектов, переместившихся из кластера в кластер.

К недостаткам данного алгоритма можно отнести необходимость задавать количество кластеров для разбиения.

Таким образом, необходимо отметить, что не существует универсального метода обработки данных, так как приходится экспериментировать с выбором мер расстояний, а иногда и с самим алгоритмом вычисления расстояний. Никакого единого решения этой задачи не существует. При работе с фотометрическими данными необходимо провести ряд вычислительных экспериментов, позволяющих выбрать наиболее наглядные результаты анализа.

Библиографический список

1. Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.

2. Гмурман В.Е. Теория вероятностей и математическая статистика: учеб. пособие для вузов. М.: Высшая школа, 2004. 479 с.

3. URL: http://www.aiportal.ru/articles/autoclassification/measuredistance.html

4. Матвеев Ю.Н., Стукалова Н.А., Михальцов Н.Г. Использование математического моделирования в задачах диагностики технологических систем // Проблемы информатики в образовании, управлении, экономике и технике: сб. статей XV Международной научно-технической конференции. Пенза, 2015. С. 160–164.

Матвеев Юрий Николаевич
Тверской государственный
технический университет,
г. Тверь, Россия
E-mail: matveev4700@mail.ru

MatveevYu.N.
Tver State Technical University,
Tver, Russia

Даньшина Анна Евгеньевна
Тверской государственный
технический университет,
г. Тверь, Россия

Dan'shina A.E.
Tver State Technical University,
Tver, Russia

Стукалова Наталия Александровна
Тверской государственный
технический университет,
г. Тверь, Россия

Stukalova N.A.
Tver State Technical University,
Tver, Russia

Туляков Андрей Вячеславович
Тверской государственный
технический университет,
г. Тверь, Россия

Tuljakov A.V.
Tver State Technical University,
Tver, Russia