Ахмедов Руслан Эльдарович,
кандидат физико-математических наук, доцент,
Российский государственный университет
социальных технологий

ФУНКЦИОНАЛЬНЫЕ МОДЕЛИ АНАЛИЗА ДАННЫХ И ПРИЛОЖЕНИЯ К АКУСТИЧЕСКИМ СИСТЕМАМ

Аннотация. Рассматриваются методы анализа данных на основе кластерного подхода, дается формальное описание моделей систем обработки информации. Обсуждаются возможности применения указанных методов в системах акустического моделирования.

Abstract. The methods of data analysis based on the cluster approach are considered, formal models of information processing systems are described. The possibilities of the application of these methods in an acoustic modeling are discussed.

Ключевые слова: Кластеризация, нейронные сети, нечеткое моделирование, акустическая модель, обучение.

Keywords: Clustering, neural networks, fuzzy modeling, acoustic model, learning.

Введение

Необходимость решения вопросов, связанных с функционированием сложных объектов в различных областях деятельности, сопряжена с рядом проблем, решение которых нельзя, по-видимому, осуществить в рамках универсального математического или естественнонаучного аппарата. Одна из причин тому – невозможность во многих случаях построить формальную модель задачи, структурировать данные и факторы, имеющие значение для конкретного объекта или системы. Также важно учитывать то обстоятельство, что удачно построенная модель или классификация данных может существенно упростить решение, уменьшить затраты на техническую реализацию.

Прикладные задачи, связанные с обработкой больших объемов информации, требуют обеспечения наглядного и компактного хранения данных, а также их систематизации. В решении подобных задач мощным аппаратом являются модели кластерного анализа, которые позволяют проводить исследование в рамках междисциплинарной области знаний – интеллектуального анализа данных (Data Mining, [1], [2]). Здесь можно выделить два основных направления: Web Content Mining и Web Usage Mining, нацеленных, соответственно, на автоматизированный поиск информации и обнаружение скрытых закономерностей в действиях конкретных пользователей.

Виды и особенности моделей кластеризации

Наиболее известные модели кластерного анализа, используемые в современных исследованиях, можно условно разделить на 2 группы: плоские и иерархические.

Плоская кластеризация порождает совокупность кластеров, не имеющих явных взаимосвязей.

Иерархическая кластеризация создает иерархию кластеров. В общем случае задача плоской кластеризации допускает следующую формальную интерпретацию.

Дано:

1. Множество элементов $D = \{d_1,\ d_2,\ \dots,\ d_N\}$;
2. Желательное количество кластеров K;
3. Целевая функция, оценивающая качество кластеризации.

Необходимо определить соответствие $\gamma: D \to \{1,\ \dots,\ K\}$, которое должно обеспечить экстремум (минимум или максимум) целевой функции. Целевая функция определяется в

терминах сходства или расстояния между элементами. Сходство элементов выражается в виде одной из функций тематического сходства или в значениях на одних и тех же осях векторного пространства. Тематическое сходство определяется, как правило, с помощью меры сходства или евклидова расстояния в векторном пространстве. Если же упор будет сделан на другое сходство элементов, то можно выбрать другое представление.

Плоская кластеризация, порождающая совокупность кластеров, не имеющих явных взаимосвязей, эффективна и проста, но в результате создается простое неструктурированное множество кластеров, использующее количество кластеров как входной параметр.

Иерархическая кластеризация создает иерархию, то есть структурированное множество, которое является более информативным, чем неструктурированное множество кластеров. Для иерархической кластеризации не требуется заранее указывать количество желаемых кластеров, но эти преимущества в ряде случаев значительно снижают производительность.

Рассматривая модели кластеризации в целом, следует учитывать следующие их особенности:

1) универсальность модели, т. к. не требуется априорных представлений об исходных данных, в результате допускается возможность сравнения данных различных типов;

2) в отличие от задач классификации, кластерный подход предполагает обучение «без учителя», т. е. допускается анализ построенной модели без участия эксперта, осуществляющего разбиение на классы согласно заданному критерию;

3) входная информация для алгоритма кластеризации – метрика, изменение которой оказывает существенное влияние на результаты.

В то же время, поскольку многие современные системы достаточно сложны и слабо формализуемы, возникает необходимость построения моделей, наиболее полно отвечающих реальным условиям для изучаемых объектов. Один из таких подходов к моделированию использует понятия нечеткого множества и нечетких отношений, восходящие к исследованиям Л. Заде [3]. Базовыми характеристиками в подобных моделях служат нечеткие переменные, на основе которых строятся логические операции, обобщающие известные операции классической (булевой) логики. Далее вводится понятие лингвистической переменной, значениями которой являются нечёткие множества. Это дает возможность интерпретировать формальную нечетко-множественную модель в виде нейронных сетей. Нейронные сети находят широкое применение для решения различных задач защиты информации [4], [5].

Анализ обучаемости как живых систем, так и искусственных, построенных на базе нейронных сетей, приводит к выявлению общего свойства относительно ошибок обучения: средний уровень ошибки в обоих случаях постепенно снижается, начиная с некоторой итерации процесса обучения, причем возможен кратковременный резкий скачок ошибки обучения [6]. Вместе с тем, если эволюция живых систем характеризуется их способностью к стиранию памяти, следованию по «неверному» пути решения, отвлечению внимания, что отчасти объясняет поведение скорости обучения в определенных условиях, то для искусственно созданных нейронных систем подобные изменения являются неожиданным феноменом. Необходимо учесть, что скорость обучения естественных адаптивных систем на начальном этапе обучения обычно невысока, в то время как искусственно созданные нейронные системы характеризуются различным уровнем скорости обучения. Данное наблюдение позволяет сделать предположение, что различные типы искусственных адаптивных систем имеют некоторый скрытый фактор, который можно условно назвать «мгновенным переключателем» в процессе обучения системы. В случае подтверждения гипотезы, показывающей «квантованный» характер уровней обучения, открываются новые возможности для имитации и моделирования естественных систем.

Применение кластерного анализа в системах интеллектуальной обработки информации

В целях повышения эффективности обработки информации, сокращения времени использования человеческих ресурсов на рутинную работу, становится целесообразным построение подходящей модели, включающей в себя систему с обучением. Этот вопрос является актуальным, в частности, при распознавании потоков звуковой информации и выявлении ключевых слов и сочетаний, которые играют определяющую роль в конкретной задаче. Проблема моделирования систем с обучением – одна из составляющих в более общих задачах анализа неструктурированной речевой информации (АНРИ), которые имеют большое практическое значение, например, в бизнес-приложениях. С другой стороны, для расширения области применимости построенной модели необходимо провести эффективную классификацию отдельных элементов потока речевой информации с последующим определением шаблонных элементов в обучающей модели. С этой целью проводится анализ поступающей информации на фонетическом уровне и выделение простейших звуковых сочетаний – монофонов и трифонов. Для успешной реализации задачи моделирования предлагается использовать метод классификации по дереву регрессии (Classification and Regression Trees, или CART). Преимущество данного метода в том, что он успешно объединяет лингвистические знания (фонетический строй языка) и математический аппарат (метод минимизации среднеквадратической ошибки).

Алгоритм CART позволяет для каждой фонемы определить оптимальную последовательность вопросов путем выявления на каждом этапе ветвления вопроса, имеющего минимальное значение среднеквадратической ошибки. Пошаговая постановка вопроса необходима для определения характеристики образования звуков и звуковых сочетаний.

Вкратце алгоритм CART можно описать последовательностью действий, включающей вычисление значений суммарной среднеквадратической ошибки и взвешенной среднеквадратической ошибки на всех векторах, входящих в обучающую базу. Далее выполняется цикл алгоритма, в котором на каждом шаге ставится качественный вопрос относительно данного сочетания; таким образом, дерево регрессии разбивается на две «ветки», в зависимости от фонетических характеристик сочетания. Цель каждого разбиения состоит в том, чтобы минимизировать значение ошибки. Разность значений исходной взвешенной ошибки и суммы взвешенных ошибок на левой и на правой ветке определяет критерий разбиения, по которому производится оптимизация. В соответствии с построенным деревом регрессии формируются результирующие кластеры, которые затем передаются в модуль акустического моделирования для дальнейшей обработки.

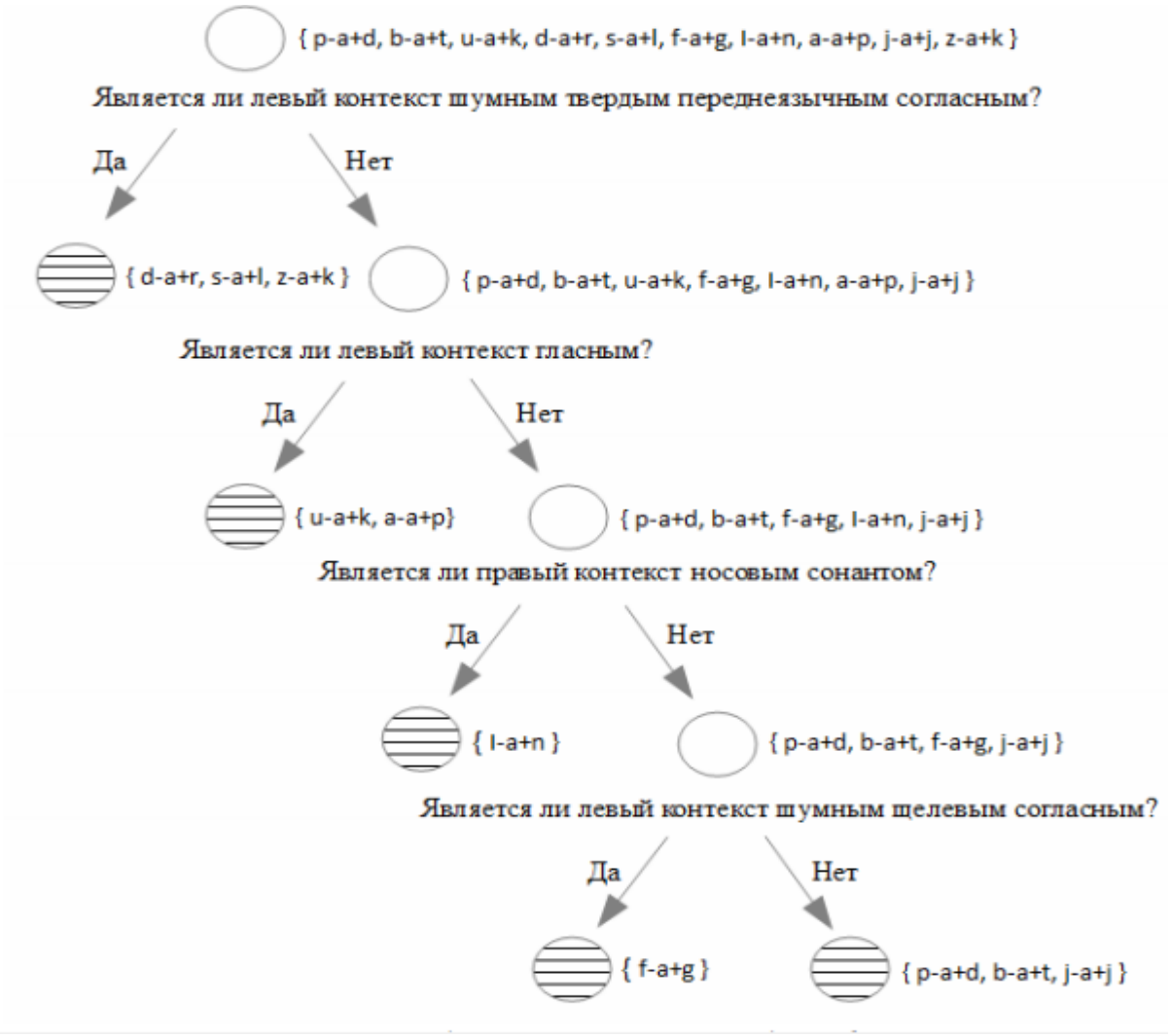

Рисунок 1. Пример дерева регрессии для фонемы «а»

Схема работы алгоритма для конкретного звукового сочетания дает классифицирующее дерево регрессии, листья которого являются кластерами, используемыми при дальнейшем акустическом моделировании. Левая ветка соответствует положительному ответу на вопрос, правая ветка соответствует отрицательному ответу на вопрос. Заштрихованный узел является терминальным (который не удалось разбить на две ветки по причине невыполнения критерия останова: либо вопрос не сокращает среднеквадратическую ошибку, либо отсутствует достаточное количество реализаций). Данные терминальные узлы и есть кластеры, в совокупности, составляющие акустическую модель.

Для улучшения пользовательских характеристик продуктов, построенных на базе модельной системы, целесообразно поставить вопрос об оптимизации работы алгоритма, с учетом варьирования условия завершения алгоритма, либо других параметров.

Пояснения. Среднеквадратическое отклонение точки от среднего в M-мерном пространстве признаков – $S_k^2 = \sum_{i=1}^{M} \left(X_i^{(k)} - \overline{X^{(k)}} \right)^2$;

суммарная ошибка – сумма среднеквадратических отклонений всех векторов, входящих в обучающую базу;

триф*о*н, монофо*н* – простейшие фонетические элементы (созвучия), по которым строятся шаблоны в акустической модели. Триф*о*н является сочетанием монофонов.

Список литературы:

1. Дюк В., Самойленко А. Data Mining: Учебный курс. // СПб.: Изд-во «Питер», 2001.– – 368 с.
2. Суркова А. С., Буденков С. С. Построение модели и алгоритма кластеризации в интеллектуальном анализе данных. // Вестник Нижегородского университета им. Н. И. Лобачевского, 2012.– № 2 (1)–с.198-202.
3. Zadeh L. A. From computing with numbers to computing with words – from manipulation of measurements to manipulation of perceptions. //Int. J. Appl. Math. Comput. Sci.– 2002.– Vol.12. №3. – p. 307-324
4. Котенко И. В. Интеллектуальные механизмы управления кибер- безопасностью /Управление рисками и безопасностью.// Труды ИСА РАН.– 2009.– Т.41.– Москва, URSS. – с.74-103.
5. ФСТЭК России. ГОСТ Р 52633.0-2006– [Электронный ресурс] – Режим доступа. – URL:http://www.posoh.ru/auto_ident/metki/doc/52633.0-2006.doc (дата обращения 20.08.2015)
6. Лоренц В. А., Гавриков В. Л., Хлебопрос Р. Г. Анализ обучения нейронной сети задачам, содержащим скрытую закономерность.// Вестник КрасГАУ. –2012. –Т. 5. –с. 88-92.