Литература

1. Веснин В.Р. Стратегическое управление: учеб. – ТК Велюи, Изд-во Проспект, 2006. - 328 с.

2. Adamus ,W. A new method of job evaluation. http://www.isahp.org/2009. Proceedings/Fina l_Papers/ 106_Adamus_REV_FIN.pdf, 2009.

Е.И. Чигарина, М.И. Шеремеев

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ В ЗАДАЧАХ ОБРАБОТКИ ДАННЫХ БОЛЬШОГО ОБЪЕМА

(Самарский национальный исследовательский университет
имени академика С.П. Королева)

В наше время особую актуальность приобрели «большие данные» (Big Data). Широкое распространение технологий автоматизированной обработки информации и накопление в компьютерных системах больших объёмов данных, сделали очень значимой задачу поиска неявных зависимостей, в имеющихся наборах данных. Для решения задачи структурирования данных используются методы математической статистики, теория баз данных, теория искусственного интеллекта, ставшая очень популярной в наше время, которая включает в себя машинное обучение и нейронные сети, а также различные вероятностные, визуальные, прогнозирующие, генетические алгоритмы. Задача структурирования большого количества данных разделяется на: распределение данных по известным классам (классификация) и распределение данных по неизвестным классам (кластеризация). В данной работе рассмотрены методы кластеризации. Задача кластеризации относится к статистической обработке данных, а также к широкому классу задач машинного обучения с учителем и без учителя.

Машинное обучение (англ. Machine Learning, ML) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Для решения задачи кластеризации можно выделить такие подходы как вероятностный подход, подходы на основе искусственного интеллекта, иерархический подход и другие. Для последующего анализа выбраны следующие алгоритмы кластеризации: алгоритм опорных векторов и алгоритм случайного леса, которые относятся к группе вероятностных подходов решения задачи кластеризации и иерархических алгоритм, который относится иерархическому подходу решения задачи кластеризации, а также алгоритм кластеризации, предлагаемый компанией Microsoft .

Основная идея метода опорных векторов – перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы.

Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этим параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора. То есть основными параметрами, влияющими на процесс обучения, являются: обучающая выборка и размер обучающей выборки.

Иерархическая кластеризация – комплекс алгоритмов, использующих разделение крупных кластеров на более мелкие или объединение мелких в более крупные. Соответственно, выделяют разделительную (дивизивную) и агломеративную (объединительную) кластеризации. В разделительной кластеризации всё исходное множество данных сначала рассматривается как один кластер, который расщепляется на два, тот в свою очередь ещё на два и т.д. В результате образуется иерархическое дерево кластеров. В агломеративной кластеризации также формируется иерархическое дерево, но путем объединения объектов в более крупные кластеры из более мелких. Сначала каждый объект рассматривается, как отдельный кластер, а затем производится объединение. Факторы, влияющие на процесс обучения – выбор вида дерева, а также количество элементов в расщепляющей выборке.

Алгоритм случайного леса (Random forest) – это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Чем больше деревьев, тем лучше качество, но время настройки и работы Random forest (RF) также пропорционально увеличиваются. Чем меньше глубина, тем быстрее строится и работает RF. При увеличении глубины резко возрастает качество на обучении, но и на контроле оно, как правило, увеличивается. Основными параметрами, влияющими на результат обучения являются: число деревьев, максимальная глубина дерева, число признаков для выбора расщепления, минимальное число объектов, по которым производится расщепление.

Алгоритм кластеризации Microsoft является алгоритмом сегментации или кластеризации, который выполняет итерацию вариантов в наборе данных, чтобы сгруппировать их в кластеры, содержащие подобные характеристики. Алгоритм кластеризации Microsoft сначала определяет связи в наборе данных и формирует ряд кластеров на основе этих связей. После первого определения кластеров алгоритм вычисляет, как кластеры представляют группирование точек, а затем пытается повторно определить группирования, чтобы создать кластеры, которые лучше представляют данные. Алгоритм последовательно выполняет этот процесс до тех пор, пока улучшить результаты, определяя кластеры, будет невозможно. При подготовке данных, предназначенных для использования в обучении модели кластеризации, следует учитывать требования к конкретному алгоритму, в том числе к объему необходимых данных, и то, как эти данные используются.

Для сравнения перечисленных алгоритмов определены такие критерии как время, затраченное на процесс обучения и выдачу результатов, а также

средняя ошибка кластеризации, равная усредненному значению вероятности непопадания в “нужный” кластер.

В качестве предметной области для исследования алгоритмов используется область медицинской диагностики, в частности, определение группы здоровья человека, учитывая его возраст, пол, ранее имеющихся диагнозов, региона проживания и других факторов.

И.У. Шарофутдинов

ЦИФРО-АНАЛОГОВАЯ ЛИНЕАРИЗАЦИЯ НА ОСНОВЕ АППРОКСИМАЦИИ НЕПРЕРЫВНЫМИ СПЛАЙНАМИ

(Ферганский государственный университет)

В настоящее время имеется ряд работ посвященных исследованию свойств сплайн - функций и их возможностей для технических приложений. Широкая популярность методов сплайн - аппроксимации объясняется тем, что они служат универсальным инструментом моделирования функций и по сравнению с другими математическими методами при равных с ними информационных и аппаратных затратах обеспечивают большую точность вычислений.

Актуальными являются задачи разработки методов, алгоритмов аппаратных и программных средств для быстрого поиска и выявления локальных особенностей сигналов. Анализ и восстановления сигналов составляет основу процессов решения задач обработки геофизических и сейсмических сигналов, обработки результатов стендовых испытаний, обработки изображений и других.

Требования высокой производительности вычислительных систем, применяемых в этих областях, могут быть удовлетворены как за счет разработки новых методов и алгоритмов цифровой обработки сигналов (ЦОС), так и с помощью многопроцессорных средств параллельно – конвейерных вычислений.

Для решения задач анализа и восстановления сигналов широко применяются обобщенные спектральные методы и методы сплайн-функций.

Сплайн-функция — гладкая кусочно-полиномиальная функция, используемая для выравнивания временных рядов. Применение сплайн-функция вместо обычных функций тренда эффективно, когда внутри анализируемого периода меняется тенденция, направление ряда. С.-ф. помогает выделить под периоды, внутри которых динамика показателя не претерпевает существенного изменения. Любой сплайн достаточной гладкости может быть представлен через базисные сплайны. В частности, при d=1 для разложения используются так называемые “нормализованные” базисные сплайны степени m (B - сплайны).