АЛГОРИТМ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ НАБОРОВ ИНФОРМАТИВНЫХ ПРИЗНАКОВ СОСТОЯНИЯ ПРОМЫШЛЕННОГО ОБОРУДОВАНИЯ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Каляда В.В.

Давыдов И.Г. – к.т.н.,доцент

Не так давно получил распространение термин «большие данные», обозначивший новую прикладную область — поиск способов автоматического быстрого анализа огромных объёмов разнородной информации. Самым перспективным подходом к анализу больших данных считается применение машинного обучения — набора методов, благодаря которым компьютер может находить в массивах изначально неизвестные взаимосвязи и закономерности. Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться, который находится на стыке математической статистики, методов оптимизации и классических математических дисциплин, но имеет также и собственную специфику, связанную с проблемами вычислительной эффективности и переобучения.

Кластерный анализ заключается в том, чтобы сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов т.е. — это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны. Отличие кластеризации от классификации в том, что кластеризация разбивает множество объектов на группы, которые определяются только ее результатом. Классификация относит каждый объект к одной из заранее определенных групп. [5]

Кластеризация данных включает в себя этапы:

1) Выделение характеристик
2) Определение метрики
3) Разбиение объектов на группы
4) Представление результатов

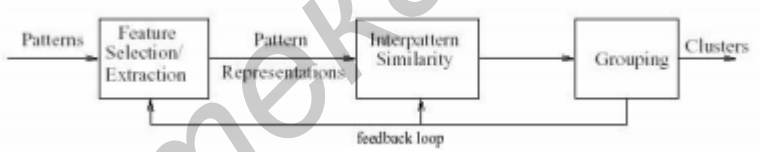

Рис.1 – Общая схема кластеризации

Постановка задачи кластеризации:

Дано:

X- пространство объектов;

$X^l = \{x_i\}_{i=1}^l$ - обучающая выборка;

$\rho : X \times x \to [0, \infty]$ - функция расстояния между объектами.

Найти:

Y - множество кластеров и

$\alpha : X \to Y$ -алгоритм кластеризации, такие что:

- каждый кластер состоит из близких объектов;
- объекты разных кластеров существенно различны.

В моей работе для обнаружения информационных признаков зарождающихся дефектов промышленного оборудования по виброакустическим сигналам применяется вейвлет-анализ и специальные базисные функции.

Далее используются алгоритмы кластеризации для автоматической оценки технического состояния промышленного оборудования.

Алгоритм кластеризации k-means

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

Алгоритм стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

Библиотека БГУИР

$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left(x_j - \mu_i\right)^2$, где k – число кластеров, S_j - полученные кластеры, i = 1,2,…,k и μ_i - центры масс векторов $x_j \in S_i$.

EM-алгоритм (англ. expectation-maximization)– алгоритм максимизации ожидания, используемый в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, в случае, когда модель зависит от некоторых скрытых переменных. Каждая итерация алгоритма состоит из двух шагов.

На E-шаге вычисляется ожидаемое значение функции правдоподобия, при этом скрытые переменные рассматриваются как наблюдаемые.

На M-шаге вычисляется оценка максимального правдоподобия, таким образом увеличивается ожидаемое правдоподобие, вычисляемое на E-шаге. Затем это значение используется для E-шага на следующей итерации. Алгоритм выполняется до сходимости.

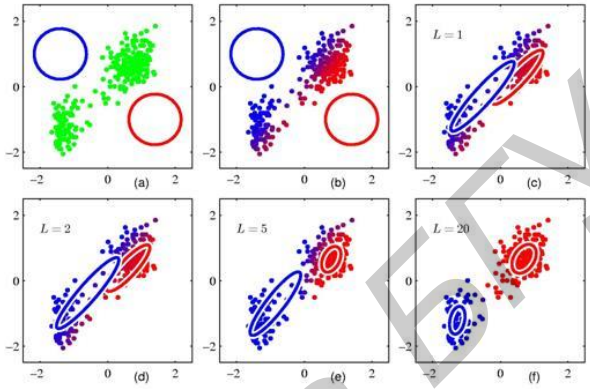

Рис.1 – EM-алгоритм

Самоорганизующаяся карта Кохонена (англ. Self-organizing map — SOM) — нейронная сеть с обучением без учителя, выполняющая задачу визуализации и кластеризации. Идея сети предложена финским учёным Т. Кохоненом.

Принцип работы:

- Инициализация карты, то есть первоначальное задание векторов веса для узлов.

Цикл:

- Выбор следующего наблюдения (вектора из множества входных данных).
- Нахождение для него лучшей единицы соответствия (best matching unit, BMU, или Winner) — узла на карте, вектор веса которого меньше всего отличается от наблюдения (в метрике, задаваемой аналитиком, чаще всего, евклидовой).
- Определение количества соседей BMU и обучение — изменение векторов веса BMU и его соседей с целью их приближения к наблюдению.
- Определение ошибки карты.

На входе:

X^l - обучающая выборка; η - темп обучения;

На выходе:

$w_{mh} \in R^n$ — векторы весов, m = 1…M, h = 1…H;

Два типа графиков — цветных карт M × H:

- Цвет узла (m, h) — локальная плотность в точке (m, h) — среднее расстояние до k ближайших точек выборки;
- По одной карте на каждый признак: цвет узла (m, h) — значение j-й компоненты вектора w_{mh}.

Библиотека БГУИР

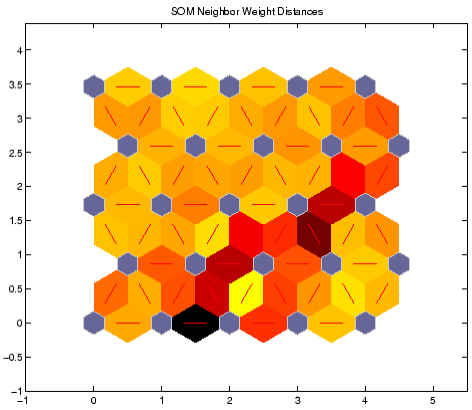

Рис.2 – Самоорганизующаяся карта Кохенена

Таким образом очевидно, что для достижения наибольшей эффективность предложенных алгоритмов для решения задачи состояния промышленного оборудования и обнаружения зарождающихся дефектов, необходимо экспериментировать с выбором мер расстояний и проводить многократное обучение алгоритмов на различных типах и видах оборудования. Никакого единого универсального решения данной задачи со 100% результатом не существует

Список использованных источников:

1. https://en.wikipedia.org/wiki/Cluster_analysis
2. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. 1999. Vol. 31, no. 3.
3. http://www.machinelearning.ru/wiki/index.php?title=Самоорганизующаяся_карта_Кохонена
4. http://www.machinelearning.ru/wiki/index.php?title=EM-алгоритм
5. Котов А., Красильников Н. Кластеризация данных. 2006. – 23с.

Библиотека БГУИР