И.С. Казанцева, О.П. Солдатова

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ ОБУЧЕНИЯ НЕЧЕТКОЙ НЕЙРОННОЙ СЕТИ ТАКАГИ-СУГЕНО-КАНГА ПРИ ПОМОЩИ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ДАННЫХ

(Самарский университет)

Целью данной работы является исследование алгоритмов кластеризации данных применительно к нечеткой нейронной продукционной сети Такаги-Сугено-Канга.

Первоначальная инициализация коэффициентов сети производится с помощью генератора псевдослучайных чисел. Подобный подход при генерации неудачных значений может привести к низкому качеству классификации. Для того чтобы уменьшить проблему случайных значений используют алгоритмы кластеризации.

Кластеризация – это автоматическое разбиение элементов некоторого множества на группы в зависимости от их схожести. Элементами множества могут быть данные или вектора характеристик. Данные группы называют кластерами. Нечёткая кластеризация подразумевает наличие пересекающихся множеств, элементы которых явно не могут принадлежать одному кластеру.

В данной работе рассматривается алгоритм нечёткой кластеризации C-Means, алгоритм нечеткой самоорганизации Густафсона-Кесселя и алгоритм C-эллипсоидов.

Алгоритм C-Means является итерационным, он позволяет разбить имеющееся множество точек на заданное число нечётких множеств. Алгоритм использует нечёткую матрицу принадлежности U с элементами u_{ij}, определяющими принадлежность каждого элемента исходного множества векторов x_t определенному кластеру, которые описываются своими центрами c_i.

В ходе решения задачи нечёткой кластеризации C-Means решается задача минимизации следующей целевой функции

$$E = \sum_{i=1}^{M}\sum_{t=1}^{P}(u_{it})^m \|c_i - x_t\|^2$$

Многократное повторение итерационной процедуры ведёт к достижению минимума функции E, который необязательно будет глобальным [1].

Алгоритм Густафсона-Кесселя повышает качество группирования за счет применения масштабирующей матрицы при расчете евклидового расстояния между вектором и центром кластера. В качестве масштабирующей обычно применяется симметричная положительно определенная матрица, т.е. матрица, у которой все собственные значения являются действительными и положительными. Множество собственных векторов, удовлетворяющих таким собственным значениям, образует в этом случае ортогональную базу многомерного пространства [1].

Основное отличие алгоритма С-эллипсоидов от алгоритма Густафсона-Кесселя заключается в расчете расстояний между множеством векторов x_i и центрами кластеров c_j:

$$d^2(x_i, c_j) = \|x_j - c_j\|^2 - \propto * \sum_{s=1}^{r} [S_{js}^T (x_j - c_j)]^2,$$

где $\|x_j - c_j\|^2$ - евклидово расстояние, S_{js} – собственный вектор ковариационной матрицы A_j – кластера j, $\propto = 1 - \frac{\lambda_2}{\lambda_1}$, где $\lambda_{1,2} -$ собственные значения матрицы A_j.

В настоящей работе рассматривается нечёткая нейронная продукционная сеть Такаги-Сугено-Канга, состоящая из 5 слоев. Первый слой сети состоит из элементов, которые выполняют фуззификацию входных переменных x_j. Элементы этого слоя вычисляют значения функций принадлежности $\mu_A(x_j)$, заданных гауссовыми функциями с параметрами. Второй слой, число элементов которого равно количеству правил в базе, выполняет агрегирование степеней истинности предпосылок соответствующих правил. В третьем слое генерируются значения функций $[p_{i0} + \sum_{j=1}^{N} p_{ij} x_j]$, которые умножаются на результат вычислений элементами предыдущего слоя. В четвертом слое первый элемент (сумматор) служит для активизации заключений правил в соответствие со значениями агрегированных в предыдущем слое степеней истинности предпосылок правил. Второй элемент (сумматор) проводит вспомогательные вычисления для последующей дефуззификации результата. Пятый слой, состоит из одного нормализующего элемента, выполняет дефуззификацию выходной переменной.

Выходной сигнал сети Такаги-Сугено-Канга можно представить формулой [2]:

$$y(x) = \frac{1}{\sum_{i=1}^{M} \left[\prod_{j=1}^{N} \mu_A^{(i)}(x_j)\right]} \sum_{i=1}^{M} \left[\prod_{j=1}^{N} \mu_A^{(i)}(x_j)\right] \left[p_{i0} + \sum_{j=1}^{N} p_{ij} x_j\right]$$

Тесты проводились с использованием модельных наборов данных, описывающих виды ирисов из базы UCI Machine Learning Repository [3].

Для определения качества классификации считалось отношение количества промахов классификации к общему числу векторов и суммарное среднеквадратичное отклонение (СКО).

Суммарное СКО определяется по формуле:

$$\text{СКО} = \sqrt{\frac{1}{p \cdot M - 1} \cdot \sum_{t=1}^{p} \sum_{i=1}^{M} (y_{i,t} - d_{i,t})^2},$$

где $y_{i,t}$ - реальное выходное значение t – номер обучающего примера (t= 1,2,...,p), а $d_{i,t}$ - ожидаемое выходное значение (i = 1,2,...,M).

В таблице 1 представлены результаты исследования зависимости СКО от количества кластеров для трёх алгоритмов кластеризации и алгоритма обратного распространения ошибки.

Таблица 1 – Зависимость СКО обучения от алгоритма

Количество кластеров	СКО			
	C-Means	Густафсона-Кесселя	С-эллипсоидов	Обратное распространение ошибки
3	0.459	0.424	0.440	0.703
4	0.445	0.410	0.420	0.534
5	0.387	0.404	0.345	0.524
6	0.201	0.395	0.214	0.519
7	0.279	0.292	0.232	0.477
8	0.280	0.412	0.331	0.530
9	0.359	0.423	0.447	0.653
10	0.362	0.425	0.454	0.687
11	0.415	0.493	0.468	0.712

В результате проведенного исследования можно сделать вывод, что приведенные выше алгоритмы кластеризации показывают примерно одинаковые результаты, однако применение алгоритмов кластеризации данных значительно понижает погрешность классификации, за счет исключения случайной составляющей в предпосылках правил.

Литература

1 Осовский, С. Нейронные сети для обработки информации [Текст] / С.Осовский – М.: Финансы и статистика, 2002 – 344 с.

2 Правила вывода [Электронный ресурс] // URL:http://fuzzy-group.narod.ru/files/Fuzzy_Modeling/Lection07.Fuzzy.logic.pdf

3 Ирисы Фишера [Электронный ресурс] // URL: https://archive.ics.uci.edu/ml/datasets/Iris

Ш.С. Каримов, Х.А. Бахриева, З.З. Нигматов

ПОСТРОЕНИЕ ИНФОРМАЦИОННОЙ СРЕДЫ ТЕХНОЛОГИЧЕСКОГО ПРОЦЕССА

(Ташкентский государственный технический университет)

Характерней особенностью современных производственных объектов является появление нескольких разрозненных автоматизированных систем, установленных на отдельных участках и зачастую реализованных на базе совершенно разных программно-аппаратных средств. Причинами этого может быть как отсутствие единой стратегии при выборе средств автоматизации, так и временная удаленность моментов создания различных автоматизированных систем. Следует отметить, что часто система автоматизации рассматривается как