МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Юго-Западный государственный университет»
(ЮЗГУ)

Кафедра программной инженерии

Утверждаю:
проректор по учебной работе
_____ Локтионова О.Г.
«30» 06 2021г.

МЕТОДЫ КЛАСТЕРИЗАЦИИ ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ ФОРМИРОВАНИЯ ОБУЧАЮЩИХ ВЫБОРОК

Методические указания для выполнения лабораторной работы по дисциплине «Теория нейрокомпьютерных систем» для студентов направления подготовки 09.03.04 «Программная инженерия»

Курск 2021

УДК 004.932

Составитель: Р.А. Томакова

Рецензент
Кандидат технических наук, к.т.н., доцент А.В. Малышев

Методы кластеризации информативных признаков для формирования обучающих выборок: методические указания для проведения лабораторных работ и выполнения самостоятельной внеаудиторной работы по дисциплине «Теория нейрокомпьютерных систем» для студентов направления подготовки 09.03.04 «Программная инженерия» / Юго-Зап. гос. ун-т; сост. Р.А. Томакова. Курск, 2021. –14 с.

Рассмотрена методика изучения различных методов кластеризации информативных признаков, предназначенных для формирования обучающих выборок, применяемых при организации работы нейронных сетей различной архитектуры.

Методические указания предназначены для студентов всех форм обучения направления подготовки 09.03.04 «Программная инженерия»

Текст печатается в авторской редакции

Подписано в печать 2021 г. Формат 60×84 1/16.
Усл. печ. л. 0,7 . Уч.- изд. л. 0,6 Тираж экз. Заказ. 877 Бесплатно.
Юго-Западный государственный университет.
305040, г. Курск, ул. 50 лет Октября, 94.

МЕТОДЫ ПОСТРОЕНИЯ КЛАСТЕРНЫХ СИСТЕМ НА ОСНОВЕ СИНТЕЗА ОБРАЗОВ ИНФОРМАТИВНЫХ ПРИЗНАКОВ

Цель работы – изучение основных методов кластеризации информативных признаков и приобретение практических навыков для формирования обучающих выборок, применяемых при организации работы нейронных сетей различной архитектуры.

1. КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

1.1 Простой алгоритм выявления кластеров

Пусть задано множество X информативных признаков $\{x_1, x_2, ..., x_N\}$. Пусть также центр первого кластера z_1 совпадает с любым из заданных информативных признаков (образов) и определена произвольная неотрицательная пороговая величина T; для удобства можно положить, что $z_1 = x_1$. После этого вычисляется расстояние D_{21} между образом x_2 и центром кластера z_1 по формуле:

$$D_{i,j} = \left\| x_j - z_i \right\| = \sqrt{(x_j - z_i)' \cdot (x_j - z_i)} \qquad (1)$$

Если это расстояние больше значения пороговой величины T, то учреждается новый центр кластера $z_2 = x_2$. В противном случае образ x_2 включается в кластер, центром которого является z_1. Пусть условие $D_{21} > T$ выполнено, т. е. z_2 – центр нового кластера.

На следующем шаге вычисляются расстояния D_{31} и D_{32} от образа x_3 до центров кластеров z_1 и z_2.

Осуществляется проверка условия:

Если оба расстояния $D_{31} > T$ *и* $D_{32} > T$, то назначается новый центр кластера $z_3 = x_3$.

В противном случае образ x_3 зачисляется в тот кластер, центр которого к нему ближе.

Подобным же образом расстояния от каждого нового образа до каждого известного центра кластера вычисляются и сравниваются с пороговой величиной и далее проверяется условие: если все эти расстояния превосходят значение порога T, то учреждается новый центр кластера. В противном случае образ зачисляется в кластер с самым близким к нему центром.

Результаты описанной процедуры *определяются выбором центра первого кластера, порядком осмотра образов, значением пороговой величины T и, конечно, геометрическими характеристиками данных.*

На рисунке 1 представлены два различных варианта выбора центров кластеров для одних и тех же информативных признаков, возникшие в результате изменения величины порогового значения T и исходного центра кластера.

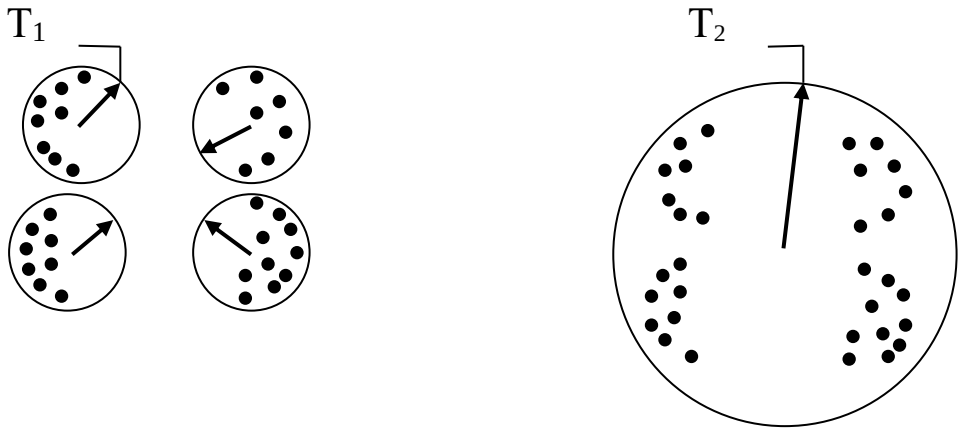

Рисунок 1. Влияние выбора величины порога и начального центра на количество кластеров в простой схеме кластеризации данных

Этот алгоритм позволяет просто и быстро получить приблизительные оценки основных характеристик заданного набора обучающих выборок. Кроме того, этот алгоритм привлекателен с вычислительной точки зрения, так как для выявления центров кластеров, соответствующих определенному значению порога T, *требуется реализация только одного просмотра выборочных элементов*. Практически же, для того чтобы хорошо понять

геометрию распределения образов с помощью такой процедуры, приходится проводить многочисленные эксперименты с различными значениями порога и различными исходными точками кластеризации. Поскольку изучаемые информативные признаки обычно имеют высокую размерность, визуальная интерпретация результатов исключается; поэтому необходимая информация добывается в основном при помощи сопоставления после каждого цикла просмотра данных расстояний, разделяющих центры кластеров, и количества образов, вошедших в различные кластеры.

Следует отметить, что важными характеристиками при этом являются ближайшая и наиболее удаленная от центра точки кластера, а также количество элементов, содержащихся в каждом кластере. Информацию, полученную таким образом после каждого цикла обработки данных, можно использовать для коррекции выбора нового значения порога T и нового исходного центра кластеризации в следующем цикле. Можно рассчитывать на получение с помощью подобной процедуры полезных результатов в тех случаях, когда в данных имеются характерные «гроздья», которые достаточно хорошо разделяются при соответствующем выборе порогового значения.

1.2 Метод выделения кластеров на основании алгоритма максиминного расстояния

Алгоритм максиминного $(\max\{\min\})$ расстояния, представляет собой итеративную процедуру выделения кластеров, в основе которой лежит нахождение евклидова расстояния между элементами информативных признаков. Реализацию этого алгоритма продемонстрируем на конкретном примере.

Рассмотрим выборку информативных признаков, состоящую из десяти двумерных образов:

$$X_1(0,0), X_2(3,8), X_3(2,2), X_4(1,1), X_5(5,3),$$

$$X_6(4,8), X_7(5,3), X_8(5,4), X_9(6,4), X_{10}(7,5)$$

Геометрическая интерпретация исследуемых информативных признаков представлена на рисунке 2.

На первом шаге алгоритма произвольным образом один из объектов выборки, например X_1, назначается центром первого кластера z_1.

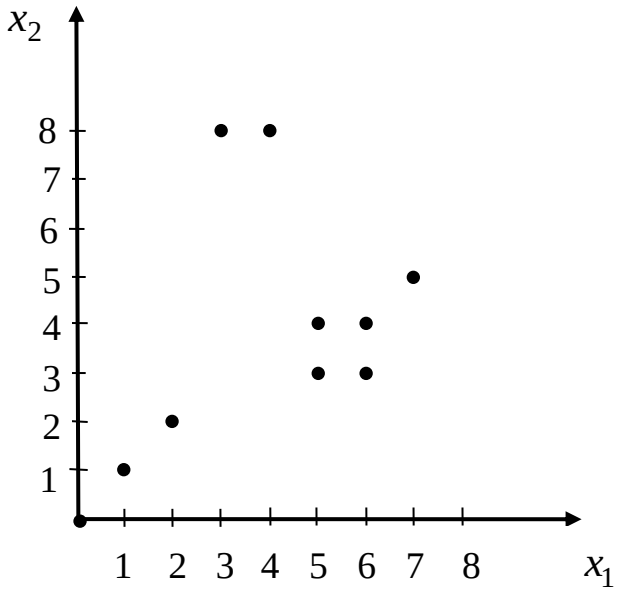

Рисунок 2. Выборка информативных признаков, предназначенная для реализации алгоритма максиминного расстояния

На рисунке 3 стрелками представлен порядковый номер шага, на котором производится выделение соответствующего центра кластера.

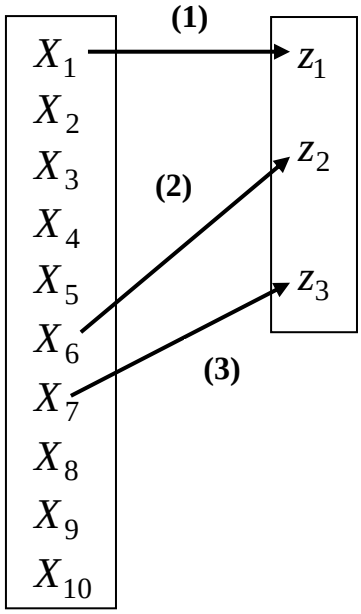

Рисунок 3. Выборка информативных признаков и центры кластеров соответственно

На втором шаге алгоритма определяется евклидово расстояние между центром первого кластера и всеми остальными элементами выборки $D_{1,j}, \quad j=\overline{1,10}$.

Затем отыскивается образ, отстоящий от образа X_1 на наибольшее расстояние $\max\{D_{1,j}\}$; в нашем случае это образ X_6, который и назначается центром второго кластера z_2.

На третьем шаге алгоритма производится вычисление расстояний между всеми образами выборки и центрами кластеров z_1 и z_2, т.е. определяются $D_{1,j}, D_{2,j}, \quad j=\overline{1,10}$. В каждой паре этих расстояний выделяется минимальное, т.е. $\min\limits_{j}\{D_{1,j}, D_{2,j}\}, \quad j=\overline{1,10}$. После этого находится *максимальное из этих найденных минимальных расстояний* $\max\limits_{i}\{\min\limits_{j}\{D_{1,j}, D_{2,j}\}\}, \quad j=\overline{1,10}$.

Если это расстояние составляет значительную часть расстояния между центрами кластеров z_1 и z_2 (например, больше либо равно среднего арифметического этого расстояния), то соответствующий образ назначается центром третьего кластера z_3.

В противном случае выполнение алгоритма прекращается. Если воспользоваться таким критерием, то легко убедиться в том, что центром кластера z_3 становится образ X_7.

На следующем шаге алгоритма вычисляются расстояния между тремя выделенными центрами кластеров и всеми остальными элементами выборки, т.е. $D_{1,j}, D_{2,j}, D_{3,j}, \quad j=\overline{1,10}$.

В каждой группе из трех расстояний выбирается минимальное, т.е. $\min\limits_{j}\{D_{1,j}, D_{2,j}, D_{3,j}\}, \quad j=\overline{1,10}$. После этого, как и на предыдущем шаге, находится максимальное из этих минимальных расстояний, т.е. $\max\limits_{i}\{\min\limits_{j}\{D_{1,j}, D_{2,j}, D_{3,j}\}\}, \quad j=\overline{1,10}$.

Если это расстояние больше либо равно среднего арифметического расстояния между центрами кластеров z_1, z_2, z_3, то соответствующий образ назначается центром кластера z_4. В противном случае выполнение алгоритма прекращается.

В общем случае описанная процедура повторяется до тех пор, пока на каком-либо шаге не будет получено максимальное расстояние, для которого условие, определяющее выделение нового кластера, не выполняется.

В этом простом примере были выделены три кластерных центра X_1, X_6 и X_7. Результаты кластеризации соответствуют геометрическим представлениям об этих данных.

1.3 Метод выделения кластеров на основании алгоритма К-внутригрупповых средних

Алгоритм, основан на вычислении K внутригрупповых средних, состоит из последовательности шагов.

Шаг I. Задаются произвольным образом K исходных центров кластеров $z_1(1)$, $z_2(1)$, …, $z_K(1)$. Так как выбор центров кластеров осуществляется произвольно, то чаще всего, в качестве исходных центров используются, например, первые K результатов выборки из заданного множества информативных признаков.

Шаг 2. На k-м шаге итерации заданное множество информативных признаков {X} распределяется по K кластерам по следующему правилу:

$x \in S_j(k)$, если $\|x - z_j(k)\| < \|x - z_i(k)\|$ для всех $i = \overline{1, K}$, $i \neq j$, (2)

где $S_j(k)$- множество информативных признаков, входящих в кластер с центром $z_j(k)$.

В случае равенства решение принимается произвольным образом.

Шаг 3. На основе результатов шага 2 определяются новые центры кластеров $z_j(k+1)$, $j = \overline{1, K}$, исходя из условия, что сумма квадратов расстояний между всеми образами, принадлежащими множеству $S_j(k)$ и новым центром кластера должна быть минимальной.

Новые центры кластеров определяются по формуле:

$$z_j(k+1) = \frac{1}{N_j} \sum_{x \in S_j(k)} x, \quad j = \overline{1, K}, \tag{3}$$

где N_j – число выборочных образов, входящих в множество $S_j(k)$. Очевидно, что название алгоритма «К-внутригрупповых средних» определяется способом, принятым для последовательной коррекции назначения центров кластеров.

Шаг 4. Осуществляется проверка равенства

$$z_j(k+1) = z(k), \ j = \overline{1, K}. \qquad (4)$$

Если условие (4) достигается, то выполнение алгоритма заканчивается.

Иначе реализация алгоритма повторяется с шага 2.

Следует отметить, что качество работы алгоритма, основанного на вычислении K внутригрупповых средних, существенно зависит от числа выбираемых центров кластеров, от назначения исходных центров кластеров, от последовательности осмотра информативных признаков и, естественно, от геометрических особенностей выборочных данных.

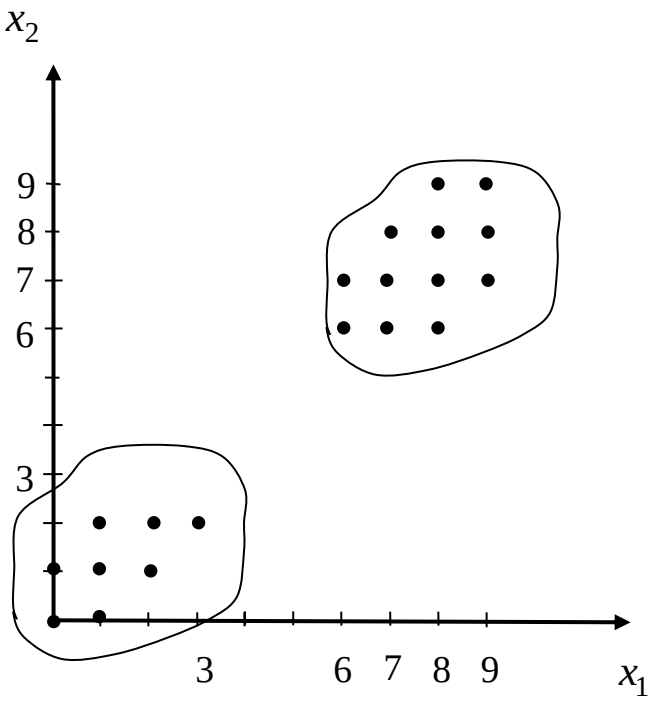

Рисунок 4. Выборка информативных признаков, использованная для иллюстрации работы алгоритма *K* средних

Пример. Для реализации алгоритма K-внутригрупповых средних рассмотрим множество информативных признаков X, представленные на рисунке 4.

$$X = \{X_1(0,0), X_2(2,0), X_3(0,3), X_4(1,1), X_5(2,1), X_6(1,2), X_7(2,2),$$
$$X_8(3,2), X_9(6,6), X_{10}(7,6), X_{11}(8,6), X_{12}(6,7), X_{13}(7,7), X_{14}(8,7),$$
$$X_{15}(9,7), X_{16}(7,8), X_{17}(8,8), X_{18}(9,8), X_{19}(8,9), X_{20}(9,9)\}$$

Шаг 1. Задается число исходных центров кластеров $K = 2$, и выбираются начальные центры кластеров, например,

$z_1(1) = x_1 = (0, 0)'$, $z_2(1) = = x_2 = (1,0)'$.

Шаг 2. Выполняется проверка условия: если

$\|x_1 - z_1(1)\| < \|x_1 - z_i(1)\|$ и $\|x_3 - z_1(1)\| < \|x_3 - z_i(1)\|$, $i = 2$,

то $S_1(1) = \{x_1, x_3\}$.

Аналогично устанавливается, что остальные информативные признаки расположены ближе к центру кластера $z_2(1)$, и поэтому $S_2(1) = \{x_2, x_4, x_5, \ldots, x_{20}\}$.

Шаг 3. Осуществляется коррекция назначенных центров кластеров:

$$z_1(2) = \frac{1}{N_1} \sum_{x \in S_1(1)} x = \frac{1}{2}(x_1 + x_3) = \begin{pmatrix} 0{,}0 \\ 0{,}5 \end{pmatrix};$$

$$z_2(2) = \frac{1}{N_2} \sum_{x \in S_2(1)} x = \frac{1}{18}(x_2 + x_4 + \ldots + x_{20}) = \begin{pmatrix} 5{,}67 \\ 5{,}33 \end{pmatrix}.$$

Шаг 4. Выполняется проверка условия:

так как $z_j(2) \neq z_j(1)$, $j = 1,2$, то осуществляется возврат к шагу 2.

Шаг 2. Выбор новых центров кластеров приводит к выполнению неравенств $\|x_\ell - z_1(2)\| < \|x_\ell - z_2(2)\|$ для $\ell = 1,2,\ldots,8$

и выполнению неравенств $\|x_\ell - z_2(2)\| < \|x_\ell - z_1(2)\|$ для $\ell = 9,10,\ldots,20$.

Следовательно, $S_1(2) = \{x_1, x_2, \ldots, x_8\}$ и $S_2(2) = \{x_9, x_{10}, \ldots, x_{20}\}$.

Шаг 3. Коррекция назначения центров кластеров:

$$z_1(3) = \frac{1}{N_1} \sum_{x \in S_1(2)} x = \frac{1}{8}(x_1 + x_2 + \ldots + x_8) = \begin{pmatrix} 1{,}25 \\ 1{,}13 \end{pmatrix}$$

$$z_2(3) = \frac{1}{N_2} \sum_{x \in S_2(2)} x = \frac{1}{12}(x_9 + x_{10} + \ldots + x_{20}) = \begin{pmatrix} 7{,}67 \\ 7{,}33 \end{pmatrix}$$

Шаг 4. Так как $z_j(3) \neq z_j(2)$, $j = 1{,}2$, то выполняется возврат к шагу 2.

Шаг 2. Получаем те же результаты, что и на предыдущей итерации: $z_1(4) = z_1(3)$ и $z_2(4) = z_2(3)$.

Шаг 3. Также получаем идентичные результаты.

Шаг 4. Так как $z_j(3) = z_j(2)$, $j = 1{,}2$ алгоритм сошелся, в результате получены следующие центры кластеров:

$$z_1 = \begin{pmatrix} 1{,}25 \\ 1{,}13 \end{pmatrix},\ z_2 = \begin{pmatrix} 7{,}67 \\ 7{,}33 \end{pmatrix}.$$

Центры найденных кластеров достаточно хорошо согласуются с геометрической интерпретацией информативных признаков, представленных на рисунке 4.

2. ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОГО ВЫПОЛНЕНИЯ

1. Методом, указанным преподавателем, реализовать один из алгоритмов кластеризации информативных признаков.

2. Сгенерировать обучающую выборку информативных признаков, состоящую из 10 двумерных объектов.

3. Вывести на экран количество найденных кластеров, указать координаты их центров, сформировать кластеры, перечислив элементы выборки, входящие в каждый из них.

4. Вывести на экран изображения всех элементов информативных признаков. Указать величину порогового значения, в случае реализации простого алгоритма выявления кластеров.

5. При реализации простого алгоритма выявления кластеров предусмотреть возможность изменения величины порогового значения T. Произвести сравнительный анализ влияния выбора начального центра кластеризации и величины порогового значения T на количество полученных кластеров.

6. Вывести на экран изображения элементов, принадлежащих каждому из отысканных кластеров. Выделить цветом найденные центры соответствующих кластеров.

7. Предоставить отчет. Содержание отчета:

1. Титульный лист.
2. Тема и цель работы.
3. Формальная постановка задачи кластеризации информативных признаков.
4. Основные теоретические сведения.
5. Задание, учитывая свой вариант.
6. Ход выполнения, обучения и тестирования.
7. Выводы.

3. КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какие методы для выделения кластеров информативных признаков существуют?
2. В чем заключается идея простого алгоритма построения кластеров?
3. Что такое пороговое значение, в чем заключается смысл?
4. Как влияет выбор величины порогового значения на количество кластеров?
5. Сформулируйте, от чего зависит работа простого алгоритма построения кластеров?
6. В чем заключается идея алгоритма максиминного расстояния для выделения кластеров?
7. Какой критерий оценки расстояний используется для анализа работы алгоритма максиминного расстояния?
8. В чем заключается идея алгоритма K-внутригрупповых средних для построения кластеров и выделения их центров?
9. Как осуществляется коррекция назначенных центров кластеров алгоритма K-внутригрупповых средних?
10. Сформулировать критерий окончания процесса итераций алгоритма K-внутригрупповых средних для построения кластеров и выделения их центров.

РЕКОМЕНДАТЕЛЬНЫЙ СПИСОК ЛИТЕРАТУРЫ

1.Борисовский, С.А. Нейросетевые модели с иерархическим пространством информативных признаков для сегментации плохоструктурированных изображений/ С.А. Борисовский, А.Н. Брежнева, Р.А. Томакова // Биомедицинская радиоэлектроника, 2010. – № 2. – С. 49-53.

2. Дж Ту, Гонсалес, Р. Принципы распознавания образов / Р. Гонсалес, Дж Ту – М.: Мир, 1978. – 411 с.

3. Кореневский, Н.А. Нейронные сети с макрослоями для классификации и прогнозирования патологий сетчатки глаза/ Н.А. Кореневский, Р.А. Томакова, С.П. Серегин, А.Ф. Рыбочкин // Медицинская техника, 2013. – № 4. – С. 16-18.

4. Томакова Р.А. Теоретические основы и методы обработки и анализа микроскопических изображений биоматериалов: монография / Р.А. Томакова, С.Г. Емельянов, С.А. Филист. – Курск, Юго-Зап. гос. ун-т, 2011. –202с.

5. Томакова Р.А. Интеллектуальные технологии сегментации и классификации биомедицинских изображений: монография / Р.А. Томакова, С.Г. Емельянов, С.А. Филист. – Курск, Юго-Зап. гос. ун-т, 2012. – 222с.

6. Томакова, Р.А. Структурно-функциональные решения нечетких нейронных сетей для интеллектуальных систем анализа разнотипных признаков/ Р.А. Томакова, С.А. Филист, В.В. Жилин, С.А. Горбатенко //Фундаментальные и прикладные проблемы техники и технологии, 2011. - № 1. –С. 85-91.

7. Томакова, Р.А. Универсальные сетевые модели для задач классификации биомедицинских данных/ Р.А. Томакова, С.А. Филист, Яа Зар До// Известия Юго-Западного государственного университета, 2012. – № 4-2(43). – С.44-50.

8. Томакова, Р.А. Метод обработки и анализа сложноструктурируемых изображений на основе встроенных функций среды MATLAB / Р.А. Томакова, С.А. Филист // Вестник Читинского государственного университета. – 2012. – № 1 (80). – С. 3-9.

9. Томакова, Р.А. Нейросетевые модели принятия решений для диагностики заболеваний легких на основе анализа флюорограмм

грудной клетки/ Р.А. Томакова, М.В. Дюдин, М.В. Томаков // Биомедицинская радиоэлектроника. 2014. №9. С. 12-15.

10. Хайкин, С. Нейронные сети: полный курс / С. Хайкин. – М.: ООО «И.Д. Вильямс», 2006. – 1104с.