УДК 519.23

А.А.Моисеев

МОДИФИКАЦИЯ НЕКОТОРЫХ ПРОЦЕДУР АВТОМАТИЧЕСКОГО АНАЛИЗА ДАННЫХ

ГосНИИ химмотологии РФ

Проведено рассмотрение алгоритмов автоматического анализа данных, которое показало, что они имеют сравнительно простую основу. Генетическая оптимизация была сведена к двухшаговой версии случайного поиска экстремума, шагами в которой является предварительное смешивание результатов первичного поиска, аналогичное скрещиванию, и вторичный случайный поиск в выделенной области, соответствующий мутации. Метод потенциальных функций позволил сравнительно просто реализовать автоматическую кластеризацию входной выборки без ограничений на ее характер. В предложенном алгоритме обучения перцептронного классификатора обработка в ассоциативном нейроне была реализована в виде усреднение сигналов от подключенных рецепторов с вычитанием постоянной величины. Дополнительное использование условия нормировки адаптивных коэффициентов делает ее малосущественной при использовании выбора максимума в качестве решающего правила. Методически несложно реализована процедура обучения алгоритма нечеткого управления, базирующаяся на выравнивании частот реализации управляющих воздействий при использовании эквидистантной выборки входных состояний.

Ключевые слова: автоматический анализ данных, генетическая оптимизация, случайный поиск, скрещивание, мутация, потенрисциальные функции, кластеризация, перцептрон, классификатор, обучение, нечеткое управление

Наиболее значимой составляющей автоматизированной обработки информации является автоматический анализ данных. Наряду с классическими процедурами статистического анализа – факторного, дисперсионного, дискриминантного и др. [1] – он также включает ряд дополнительных процедур, не связанных напрямую со статистическим анализом. К ним, в частности относятся процедуры генетической оптимизации, классификации без учителя путем кластеризации исходной выборки данных, методы адаптации перцептронного классификатора по обучающей выборке, а также процедура нечеткого управления. В настоящее время предпринимаются значительные усилия по объединению подобных процедур в рамках единой интеллектуальной технологии [2, 3]. В рамках этих усилий здесь рассматриваются некоторые модификации указанных процедур путем их существенного упрощения – как идеологически, так и в части реализации.

Генетическая оптимизация представляет собой модификацию классических генетических алгоритмов поиска экстремума. Ее первым шагом является случайный поиск экстремума, например, максимума, в m – мерном пространстве факторов X. Совокупности случайных точек x_i, i = 1, …n, соответствует совокупность x_{ij} их координат, где j = 1, …m, а также совокупность y_i значений максимизируемой функции в этих точках. Пусть $\min = \min\limits_i y_i$ - минимальное значение указанной функции на введенной выборке. В качестве первого приближения точки максимума выберем точку с координатами $x'_j = \dfrac{\sum\limits_i x_{ij}(y_i - \min)^\alpha}{\sum\limits_i (y_i - \min)^\alpha}$, где α = 1, 2, …. – параметр элитарности [2], рост которого ведет к быстрому росту вероятности отбора точки максимума практически без изменения. Расчет точки максимума путем использования взвешенного среднего соответствует процедуре скрещивания генетического алгоритма. Точка x’ = (x’$_1$, …, x’$_m$) определяет центр области мутации (x’$_j$ - Δ, x’$_j$ + Δ), j= 1….m, в которой осуществляется дальнейший поиск максимума методом случайного поиска. При этом дополнительно вводится случайная выборка x’$_i$, i = 1,…,n’. Результатом этого поиска является $y' = \max\limits_{x\in\{x',x'_i\ldots,x'_{n'}\}} y(x)$, а точка x: y(x) = y’ соответствует результату мутации для этой итерации.

Полученная таким образом точка максимума заменяет точку минимума в исходной выборке, после чего осуществляется следующая итерация генетической оптимизации. Итерации продолжаются до останова процедуры при выполнении условия $\dfrac{\max_k - \min_k}{\max_0 - \min_0} < \varepsilon \in (0,1)$, где $\max_k$, $\min_k$ – значения максимума и минимума на k – той итерации, а $\max_0$, $\min_0$ - значения максимума и минимума при первичном случайном поиске.

Преимуществом данной процедуры является ее всеобъемлющий характер, позволяющий осуществлять поиск глобального максимума. Ее основной недостаток – низкая скорость сходимости за счет необходимости перебора всех точек исходной выборки. Вероятно, эту скорость удалось бы повысить, ограничившись случайным поиском в области мутации. Однако при этом существует опасность выплеснуть с водой и ребенка, пропустив точку максимума вне этой области. Поэтому вопрос о коррекции правила останова остается пока открытым.

Модифицированная кластеризация базируется на использовании метода потенциальных функций [4, 5]. Его первым шагом является расчет максимального расстояния Δ между крайними точками исследуемой выборки. В каждой из этих точек формируется потенциал вида $\varphi = \frac{1}{1+\alpha\left(\frac{r}{\Delta}\right)^2}$, где r – расстояние от крайней точки доданной, α – параметр потенциальной функции. К каждому из исходных кластеров относятся точки, удовлетворяющие условиям отнесения $\varphi_A(r) > 0.9$, $\varphi_B(r) > 0.9$. При добавлении точек в кластер соответствующие потенциалы трансформируются следующим образом:

$$\varphi_A \to \frac{1}{n_A}\sum_{r_i \in A}\varphi(r_i)$$
$$\varphi_B \to \frac{1}{n_B}\sum_{r_i \in B}\varphi(r_i) \qquad (1)$$

Добавление точек к кластерам A и B завершается, когда в выборке не остается точек, удовлетворяющим условиям $\varphi_A > 0.9$, $\varphi_B > 0.9$, где φ_A, φ_B определяются в соответствии с (1).

Из точек, не отнесенных к исходным кластерам, выберем точку C, сумма квадратов расстояний от которой до крайних точек максимальна. Эта точка интерпретируется как зародыш третьего кластера, в который входят точки, оставшиеся вне исходных кластеров и удовлетворяющие условию отнесения $\varphi_C(r) > 0.9$. Как и ранее, добавление точек в кластер будет приводить к трансформации его потенциала: $\varphi_C \to \frac{1}{n_C}\sum_{r_i \in C}\varphi(r_i)$. Отбор будет проводиться до тех пор, пока не останется точек, удовлетворяющих условию отнесения со скорректированным потенциалом. Эта ситуация отображена на Рисунке 1.

Отбор точек, максимально удаленных от крайних, в качестве зародышей кластеров и формирование этих кластеров по описанной методике продолжается до тех пор, пока не останется точек, не отнесенных к кластерам. Преимуществом построенной процедуры кластеризации является е общий характер, позволяющий осуществить кластеризацию без ограничений на кластеризуемую выборку. Недостатком является необходимость выбора параметра настройки α.

Принципиальная схема перцептронного классификатора [5] приведена на Рисунке 2. Его входами являются рецепторы, формирующие бинарные

сигналы $x_1 \ldots x_n$ с одинаковыми коэффициентами усиления $\frac{1}{n}$, линейные функции от которых формируются в ассоциативных нейронах $A_1 \ldots A_l$:

$$y_i = \frac{k}{n} \sum_{x_j \in A_i} x_j - \theta$$

Эти нейроны, таким образом, формируют среднее значение подключенных рецепторов за вычетом параметра классификатора $\theta \in (0,1)$. В свою очередь, выходы ассоциативных нейронов используются для формирования линейных комбинаций $z_i = \sum \lambda_{ij} y_j$ с адаптивными коэффициентами усиления λij, выбираемыми по результатам обучения с учителем.

В дальнейшем будем считать, что адаптивные коэффициенты λ_{ij} удовлетворяют условию нормировки $\sum_{j=1}^{k} \lambda_{ij} = 1$, i=1,…m. Их начальным приближением являются случайные величины, равномерно распределенные в интервале (0, 1) и затем скорректированные в соответствии с условием нормировки. Предположим, что для обучения предложен i – тый образ. Если $z_i \neq z_p = z_{max}$, осуществляется следующая коррекция адаптивных коэффициентов:

$$\begin{aligned} &\lambda_{ij} \to a\lambda_{ij}, y_j > 0 \\ &\lambda_{ij} \to \lambda_{ij}, y_j \le 0 \\ &\lambda_{pj} \to \lambda_{pj} / a, y_j > 0 \\ &\lambda_{pj} \to \lambda_{pj}, y_j \le 0 \\ &a > 1 \end{aligned} \qquad (2)$$

Затем скорректированные в соответствии с (2) коэффициенты дополнительно пересчитываются с учетом условия нормировки. Эта операция повторяется с предъявлением i – того образа, пока не будет выполнено условие $z_ш = z_{max}$. После этого для обучения предъявляется следующий образ и описанные выше операции повторяются.

Предъявление образов осуществляется в циклическом порядке, пока персептрон не начнет различать их безошибочно. Это является признаком останова процедуры обучения. Величина θ при описанной организации обучения оказывается не столь существенной , поскольку при выполнении условия нормировки для адаптивных коэффициентов она дает лишь постоянный сдвиг, не влияющий на выбор максимума. Изначально ее можно выбрать, например, равной $\theta = \frac{1}{n}$, а используя ее вариацию в

интервале (0, 1), можно добиться линейной сепарабельности распознавания, если последняя имеет место.

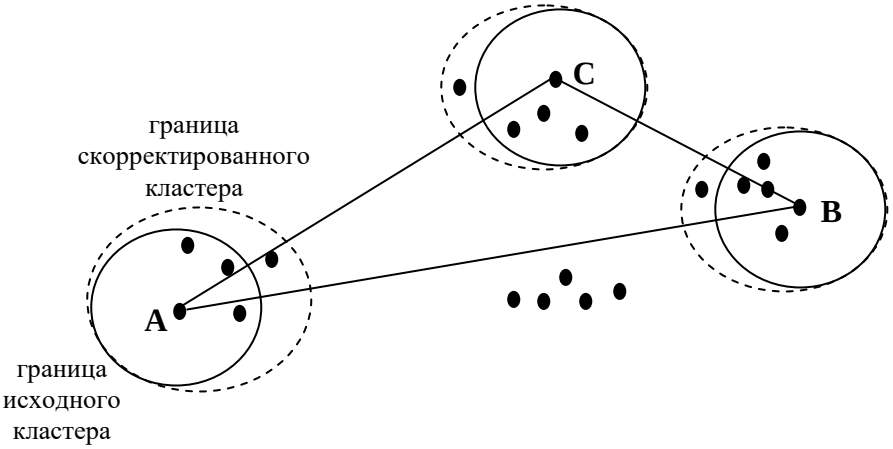

Рисунок 1- Кластеризация выборки

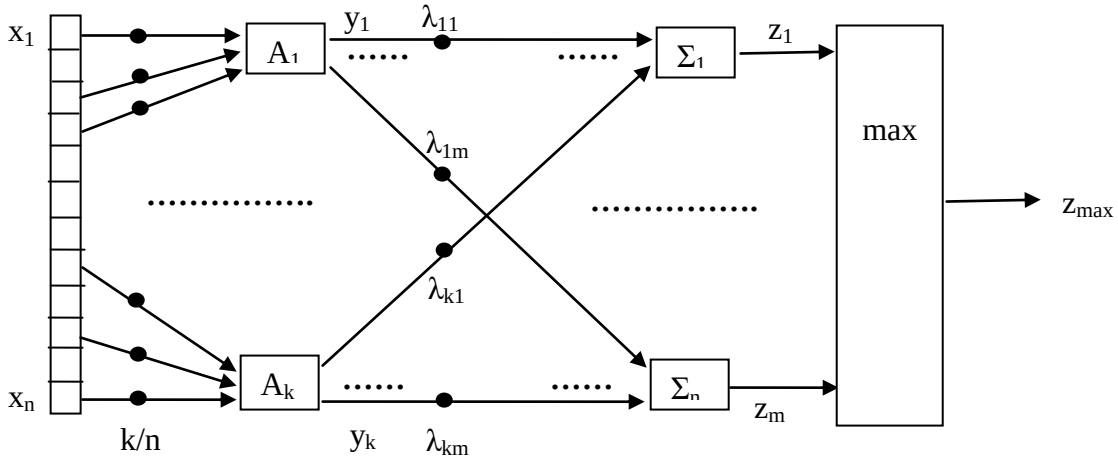

Рисунок 2 - Перцептронный классификатор

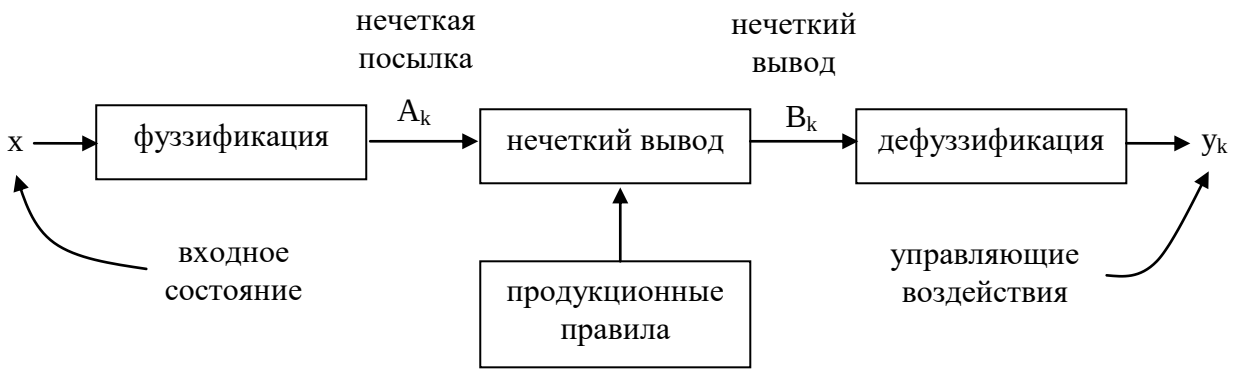

Рисунок 3- Нечеткое управление

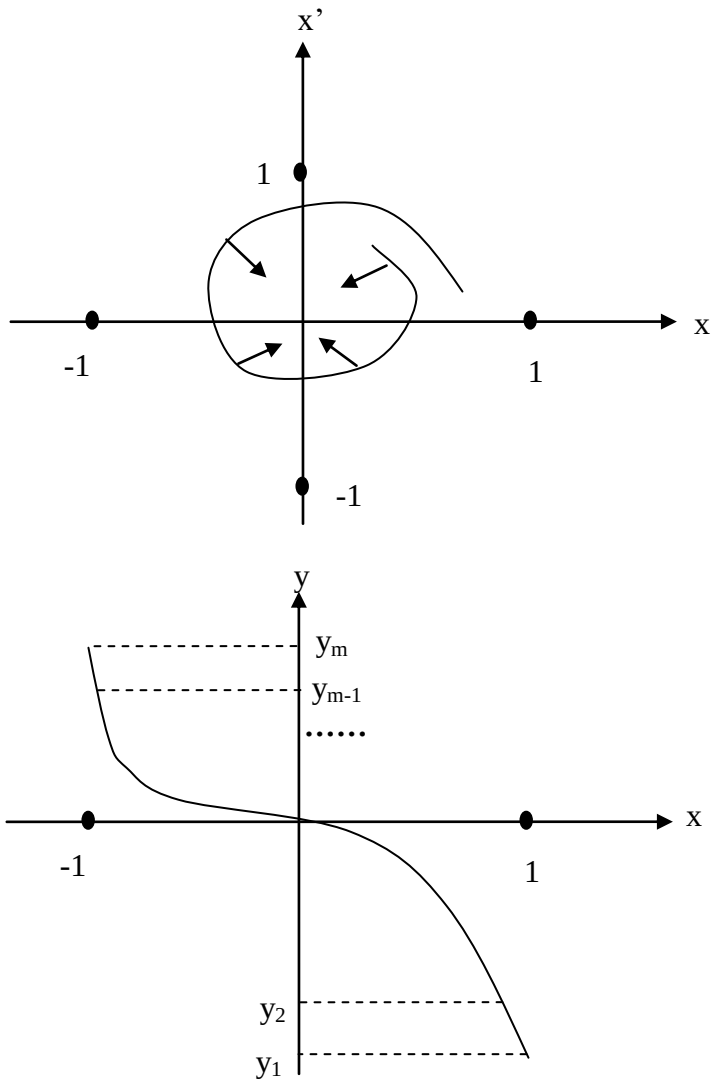

Рисунок 4 - Динамика регулирования и управляющие воздействия

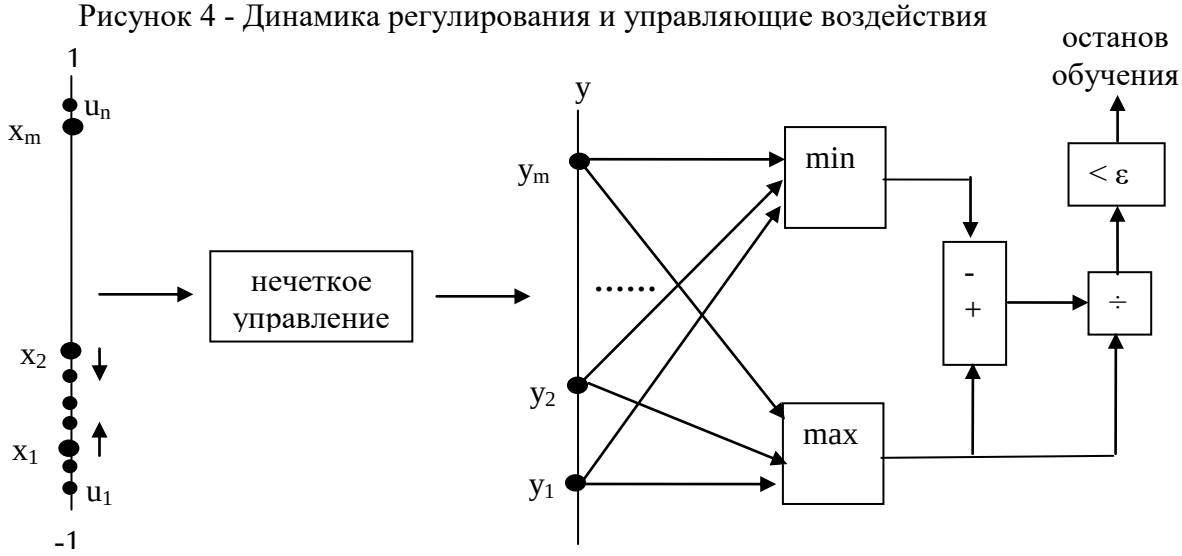

Рисунок 5- Обучение нечеткого управления

Полезным применением процедуры обучения является адаптивная настройка алгоритма нечеткого управления [2, 6]. Его традиционная схема отображена на Рисунке 3 в предположении, что формируемые управляющие воздействия y_k образуют конечное множество, упорядоченное по возрастанию. Эти воздействия представляют собой выходы процедуры дефуззификации нечетких выводов B_k (термов выходной лингвистической переменной), функции принадлежности которых аппроксимируются гауссовыми формами вида $\mu_{Bk}(y) = \exp\left(-\left(\frac{y-y_k}{\sigma_{yk}}\right)^2\right)$ (3). Величины σ_{yk} выбираются при этом так, чтобы значения «чужих» функций принадлежности в точках y_k были бы $<<1$. Этому условию удовлетворяет, например выбор σ_{yk} в виде:

$$\sigma_{yk} = \min\left(\frac{y_k - y_{k-1}}{3}, \frac{y_{k+1} - y_k}{3}\right) \qquad (4)$$

Таким образом, функции принадлежности нечетких выводов B_k можно считать заданными соотношениями (3, 4). В этом случае обучение сводится к выбору функций принадлежности термов A_k входной лингвистической переменной. В предположении, что эти функции также аппроксимируются гауссовыми формами

$$\mu_{Ak}(y) = \exp\left(-\left(\frac{x-x_k}{\sigma_{xk}}\right)^2\right) \qquad (5),$$

задача их выбора оказывается эквивалентной выбору параметров x_k, σ_{xk} в соотношении (5).

Этот выбор подчиним следующим соображениям. Предположим, что входное состояние управляемого объекта финитно и будем в дальнейшем интерпретировать x как относительное отклонение этого состояния от 0, соответствующего уставке регулирования. Поскольку скорость объекта в этой ситуации также ограничена, будем интерпретировать x’ как относительную скорость. Типичное поведение x, x’ в ситуации регулирования отображено при этом на фазовой диаграмме, приведенной на Рисунке 4. Из нее видно, что управляющее воздействие определяется по существу только величиной отклонения x. Форма зависимости y(x) управляющих воздействий от состояния также приведена на Рисунке 4. Априорно она неизвестна, и именно это обуславливает необходимость обучения.

Примем в первом приближении, что центры входных термов $x_1 \ldots x_m$ разделяет область $x \in (-1, 1)$ на равные части. Входные состояния также зададим эквидистантной выборкой $u_1 \ldots u_n$, где $n >> m$. Качественно обработка этой выборки отображена на Рисунке 5. При правильном выборе $x_1 \ldots x_m$ эквидистантная входная выборка должна порождать равновероятные выборки откликов y. Согласно рисунку этой равновероятности соответствует условие $\frac{\max - \min}{\max} < \varepsilon \in (0,1)$, где max, min – максимальная и минимальная частота одинаковых откликов. Если это условие не выполняется, то интервал на x, соответствующий частотному максимуму, должен сжиматься, а соответствующий частотному минимуму – расширяться. Признаком останова процедуры обучения является приближенное выполнение условии равночастотности откликов.

Проведенное рассмотрение показывает, что рассмотренные алгоритмы автоматического анализа имеют сравнительно простую основу. Так, генетическая оптимизация представляет собой двухшаговую версию случайного поиска экстремума, шагами в которой является предварительное смешивание результатов первичного поиска, аналогичное скрещиванию, и вторичный случайный поиск в выделенной области, соответствующий мутации. Метод потенциальных функций позволяет сравнительно просто реализовать автоматическую кластеризацию входной выборки без ограничений на ее характер. В Предложенном алгоритме обучения перцептронного классификатора обработка в ассоциативном нейроне представляет собой усреднение сигналов от подключенных рецепторов и вычитание постоянной величины. Дополнительное использование условия нормировки адаптивных коэффициентов воспроизводит эту величину на выходе, что малосущественно при использовании выбора максимума в качестве решающего правила. Методически несложной является и процедура обучения алгоритма нечеткого управления, базирующаяся на выравнивании частот реализации управляющих воздействий при использовании эквидистантной выборки входных состояний.

ЛИТЕРАТУРА

1. Статистические методы для ЭВМ, п/ред. Энслейна К., Рэлстона Р., Уилфа Г., М., Наука, 1986, 464 с.
2. Рутковская Д., Пилиньский Р., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы, М., Телеком, 2006, 452 с.

3. Ротштейн А.П., Интеллектуальные технологии идентификации, Винница, Универсум, 1999, 320 с.
4. Айзерман М.А., Браверман Э.М., Розеноэр Л.И. Метод потенциальных функций в теории обучения машин, М., Наука, 1970, 384 с.
5. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения), М., Наука, 1974, 416 с.
6. Цыпкин Я.З. Адаптация и обучение в автоматических системах, М., Наука, 1968, 400 с.

A.A.Moiseev

SOME PROCEDURES MODIFICATION OF DATA ANALYSIS

State Research Institute of Chimmotology

Performed some algorithms consideration of data analysis, that's shown their base simplicity. Genetic optimization were transformed to two – step version of stochastic search, whose steps are preliminary mixing of primary search results (interpreted as crossing) and secondary stochastic search (interpreted as mutation). Potential function method allowed implementing the simple procedure of clasterization without any additional requirements to input sample. Learning algorithm of perceptron's classifier was used the preliminary averaging in secondary neurons with any constant subtraction. Additional adaptive coefficients normalizing do it insufficient at maximization used as decisive function. Fuzzy control learning were developed that's based on control transactions frequencies equalization at equidistant sample of input states.

Keywords: data analysis, genetic optimization, stochastic search, crossing, mutation, potential functions, clasterization, perceptron, classifier, learning, fuzzy control

REFERENCES

1. Statistical methods for digital computers, ed. by Enslein K. ea, NY., Wiley, 1977, 464 p.
2. Rutkowskaya D. ea Neyronnye seti, geneticheskiye algoritmy i nechetkiye sistemy (Neuron's nets, genetic algorithms and fuzzy systems), publisher "Telecom", 2006, 452 p.
3. Rotshtein A. Intellektualnye technologii identifikatsii, Vinnitsa, publisher "Universum", 1999, 320 p.
4. Aizerman M. ea Metod potentsialnych functsij v teorii obucheniya mashin (Potential functions method in mashine learning theory), M., publisher "Nauka", 1970, 384 p.
5. Vapnik V. ea Teoriya raspoznavaniya obrazov (statisticheskiye problem obucheniya), (Image recognition theory (statistical learning problems)), M., publisher "Nauka", 1974, 416 p.
6. Tsypkin J. Adaptatsiya i obucheniye v avtimaticheskich sistemach (Adaptation and learning in control systems), M., publisher "Nauka", 1968, 400 p