Построение сложных классификаторов для объектов в многомерных пространствах

А.М. Крашенинников, Н.И. Гданский, М.Л. Рысин

Принцип линейной нормальной классификации объектов в многомерных пространствах признаков может быть использован для построения классификаторов в случае множеств сложной структуры, неразделимые в общем случае одной гиперплоскостью. В таких случаях предложено использовать совокупность иерархически связанных нормальных разделяющих гиперплоскостей, которая названа ***иерархическим нормальным классификатором (ИНК).***

Для каждого распознаваемого множества A_i ИНК содержит совокупность нормальных гиперплоскостей $\{\pi\}_i$, заданных множествами их коэффициентов $\bar{C} = (C_0, C_1, \dots, C_n)$. Все гиперплоскости $\{\pi\}_i$ разделены на слои. Число слоев обозначим L_i. Число гиперплоскостей в слое с номером s обозначим через N_s. Набор коэффициентов гиперплоскости из совокупности $\{\pi\}_i$ в слое с номером s, имеющей номер k, будем обозначать как $\bar{C}^i_{sk}$. Для упрощения выражений наряду с вектором координат точек $\bar{x}$ будем использовать однородные векторы $\bar{x}_p = (1, \bar{x})$, у которых на начальной позиции к $\bar{x}$ добавлена единица.

Алгоритм проверки включения заданной точки пространства $\bar{x}$ в множество A_i с использованием ИНК, содержащего L_i слоев, в каждом из которых (с номером s) хранится N_s гиперплоскостей $\bar{C}^i_{sk}$, заключается в том, что производится перебор по всем слоям s ИНК от 1 до L_i. Для каждого слоя s последовательно производится подстановка координат однородного вектора $\bar{x}_p = (1, \bar{x})$, во все уравнения плоскостей слоя. При получении первого же неотрицательного значения в скалярном произведении

$$\left(\bar{C}^i_{1k}, \bar{x}_p\right) \geq 0 \qquad (1)$$

делается вывод о вхождении точки $\bar{x}$ в множество A_i, выходим из алгоритма с ответом: $\bar{x} \in A_i$. Если же во всех скалярных произведениях для гиперплоскостей первого слоя выполняется условие $\left(\bar{C}_{1k}^{i}, \bar{x}_p\right) < 0$, то проверку необходимо продолжать в следующем слое. После подстановки в условие (1) коэффициентов гиперплоскостей последнего слоя L_{i} проверку завершаем. Если при этом ни одного неотрицательного значения в скалярных произведениях $\left(\bar{C}_{sk}^{i}, \bar{x}_p\right)$ не было получено, то отсюда следует, что: $\bar{x} \notin A_i$.

Применение ИНК позволяет решать задачу разделения множеств для совокупностей множеств любой структуры, имеющих сложное относительное расположение в пространстве признаков.

ИНК каждого множества A_i предложено определять путем его разделения с остальными множествами. Поскольку с точки зрения включения точек в A_i все другие множества одинаковы, то после объединения их можно считать одним множеством. Таким образом, для практического решения задачи построения ИНК отдельного множества достаточно разработать алгоритм только для пары множеств.

Для решения задачи построения ИНК отдельного множества в алгоритме для пары множеств предложено использовать две дополнительные операции по разделению множеств – отсечение и бинарную кластеризацию.

Если для пары множеств A_1 и A_2 не существует единой нормальной разделяющей гиперплоскости, то предлагается выполнить разделение A_1 и A_2 путем повторного применения принципа нормального разделения не к целым множествам, а к их частям.

Нормальную по отношению к межосевому вектору $\bar{C}_{12}$ гиперплоскость, которая отделяет все точки из A_1 и не содержит точек из A_2, назовем ***отсекающей*** для множества A_1, а подмножество A_{1o} - ***отсекаемым.*** Аналогично вводится отсекающая плоскость для множества A_2 ,. Практически построение отсекающих плоскостей производится перебором массива расстояний их точек до некоторой пробной нормальной плоскости.

Применение только одного нормального разделения и отсечения подмножеств в общем случае недостаточно для решения задач нормальной классификации множеств сложного вида – как для вложенных множеств, так и в тех случаях, когда отсекаемые множества пусты. Для преодоления данных затруднений предложено дополнительно применять близкое по назначению к кластеризации разбиение одного из множеств A_i и A_j на две части. Его задача - выделение пары максимально сгруппированных подмножеств. Назовем такой способ разбиения и получаемые подмножества для краткости ***бинарным***. Обозначим бинарные подмножества выделенного множества *A* через $\{A_1, A_2\}$.

Поскольку качество кластеризации повышается с уменьшением радиусов кластеров R_1, R_2 и увеличением межцентрового расстояния ρ_{12} между ними, то в качестве критерия сгруппированности подмножеств A_1 и A_2 предложено использовать ранее введенную степень разделимости подмножеств $\lambda(A_1, A_2)$, а в качестве меры его оптимальности - максимум. Условие оптимальности получаемого разбиения $\{A_1, A_2\}$ принимает вид:

$$\lambda(A_1, A_2) = \rho_{12} /(R_1 + R_2) \rightarrow \max(A_1, A_2),$$

(2)

В общем случае глобальный экстремум задачи (2) может достигаться не на единственной паре возможных подмножеств $\{A_1, A_2\}$. Точное ее решение можно найти перебором всех возможных вариантов разбиения множества на пары непустых подмножеств A_1, A_2 и вычислением для них значения $\lambda(A_1, A_2)$ с последующим сравнением полученного значения с текущим максимумом λ. Обозначим через ne число точек в исходном множестве $A(ne \geq 2)$.

Практически точный переборный алгоритм решения задачи (2) реализуется перебором всех кодовых чисел k из отрезка $[1; 2^{ne-1} - 1]$, описывающих все различные варианты разбиения A на подмножества A_1, A_2. По k формируются A_1, A_2 и производится вычисление значения критерия

$\lambda(A_1, A_2)$. В качестве оптимального принимается тот вариант разбиения, при котором достигается максимум значений $\lambda(A_1, A_2)$.

Принимая в качестве характерного параметра задачи число точек ne в разбиваемом множестве A, получим, что сложность точного переборного алгоритма равна $ne \cdot 2^{ne}$, т.е. является экспоненциальной. Поэтому использование точного алгоритма решения задачи бинарной кластеризации для обычных вычислительных устройств возможно только при относительно небольших разделяемых множествах, примерно для значений $ne < 15 - 18$.

Практически размер разделяемых множеств ne может быть достаточно большим. Также в процессе решения общей задачи классификации данный алгоритм может применять десятки раз. Поэтому основной задачей точного алгоритма является решение тестовых задач, а на практике для бинарной классификации необходимо использовать приближенные алгоритмы, сочетающие более низкую сложность с получением решений, достаточно близких значений критерия качества. У данных алгоритмов условие глобальной оптимальности заменяется локальной оптимальностью, при которой получаемое решение может быть лучшим только для ограниченного подмножества общей области поиска.

Изучение оптимальных решений задачи бинарной кластеризации множеств показывает, что в получаемых оптимальных подмножествах всегда присутствуют по одной точке из какой-либо или из нескольких пар максимально удаленных точек.

Поэтому построение бинарных подмножеств предложено начать с размещения в них точек, между которыми достигается максимальное расстояние. Представители выделенной пары максимально удаленных точек, размещаемые вначале для подмножеств A_1, A_2, обозначим через $\bar{a}_{11}$, $\bar{a}_{21}$ и назовем ***начальными***. Максимальная удаленность точек $\bar{a}_{11}$ и $\bar{a}_{21}$ позволяет сделать ряд заключений о местоположении всех других точек A и их возможном включении в подмножества A_1 и A_2. Они могут находиться только внутри пересечения сфер радиусов ρ_{max} с центрами в $\bar{a}_{11}$, и $\bar{a}_{21}$.

Наиболее простой вариант разделения реализуется с использованием нормальной гиперплоскости π_n, проходящей через среднюю точку $\bar{P}$ вектора $(\bar{a}_{11}, \bar{a}_{21},)$ (рис.1). Для точек этой гиперплоскости $(\bar{x} \in \pi_n)$ выполняется условие $\rho(\bar{x}, \bar{a}_{11}) = \rho(\bar{x}, \bar{a}_{21})$. Вводя для краткости прямое и обратное отношения $\delta_{12}(\bar{x}) = \rho(\bar{x}, \bar{a}_{11})/\rho(\bar{x}, \bar{a}_{21})$ и $\delta_{21}(\bar{x}) = \rho(\bar{x}, \bar{a}_{11})/\rho(\bar{x}, \bar{a}_{21})$ представим условие $(\bar{x} \in \pi_n)$ в виде:

$$\delta_{12}(\bar{x}) = \delta_{21}(\bar{x}) = 1.$$

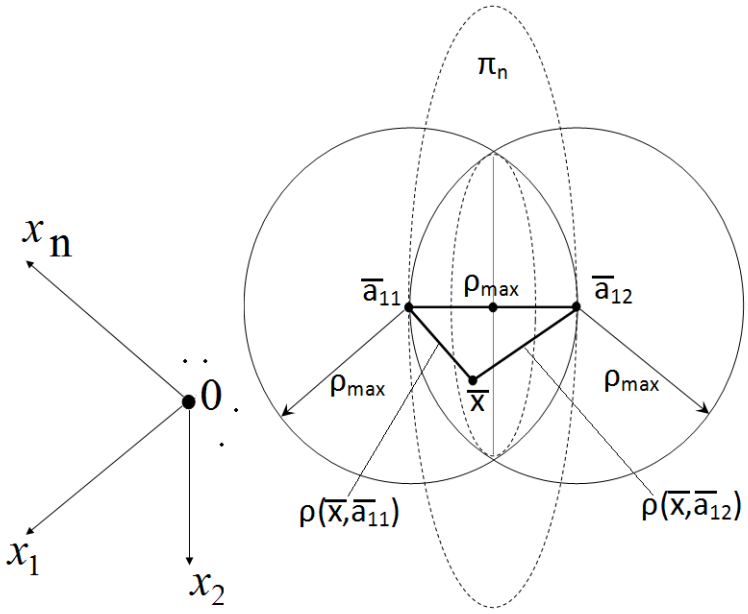


Рис. 1. Возможное местоположение точек множества *A*

К подмножеству A_1 относим те точки $\bar{x}$ множества *A*, которые лежат по одну сторону с его начальной точкой $\bar{a}_{11}$, в этом случае $\rho(\bar{x}, \bar{a}_{11}) < \rho(\bar{x}, \bar{a}_{21})$ (или $\delta_{12}(\bar{x}) < 1$.). К подмножеству A_2 относим те точки, которые ближе к $\bar{a}_{21}$, для них $\rho(\bar{x}, \bar{a}_{11}) > \rho(\bar{x}, \bar{a}_{21})$ (или $\delta_{12}(\bar{x}) > 1$).

Такой алгоритм разделения прост. Однако при его применении возникает неопределенность в отношении точек, лежащих на граничной плоскости π_n, у которых $\rho(\bar{x}, \bar{a}_{11}) = \rho(\bar{x}, \bar{a}_{21})$ $(\delta_{12}(\bar{x}) = 1)$. Также точки, лежащие достаточно близко границе π_n, могут быть не оптимально

включены в соответствующее подмножество из-за того, что они близки к другому подмножеству.

Для контроля подобных ситуаций предложено ввести предельную величину отклонения δb, которая позволяет априорно выделить:

а) множество точек, гарантированно входящих в оптимальное подмножество A_1 при выполнении условия: $\delta_{12}(\bar{x}) < \delta b$; (либо $\delta_{21}(\bar{x}) > 1/\delta b$) и

б) множество точек, гарантированно входящих в оптимальное подмножество A_2, для которых выполняется условие: $\delta_{21}(\bar{x}) < \delta b$; (либо $\delta_{12}(\bar{x}) > 1/\delta b$).

При введенном априорном пограничном значении δb возникает пограничный слой, точки которого удовлетворяют условиям:

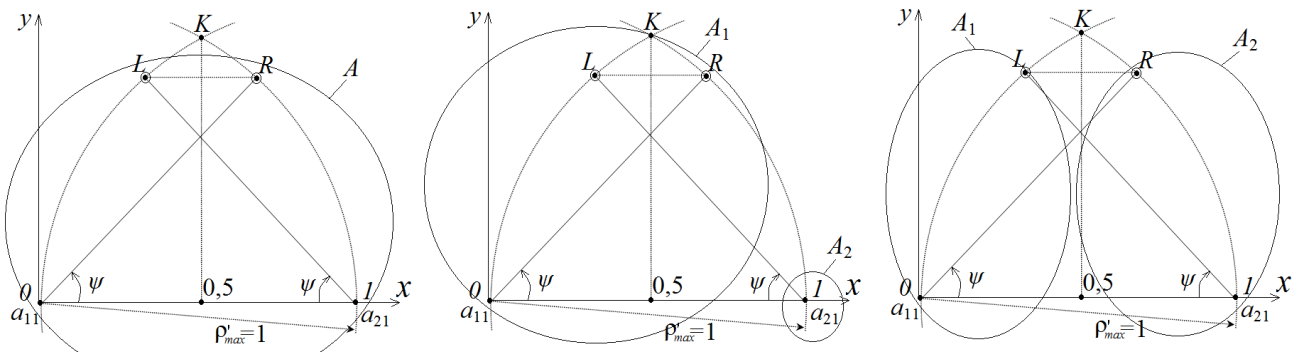
$$\delta b \leq \delta_{12}(\bar{x}) \leq 1/\delta b; \delta b \leq \delta_{21}(\bar{x}) \leq 1/\delta b.$$

Для них невозможно сразу же сделать заключение о принадлежности к оптимальным множествам A_1 и A_2. Рассмотрим оценку возможной величины априорного отклонения δb. Максимальные значения данного отклонения достигаются в модельной ситуации, когда:

- разделяемые точки множества *A* лежат в одной гиперплоскости (рис.2 а),
- есть две промежуточных группы с центрами $\bar{L}$ и $\bar{R}$ и довольно большими числами точек $N >> 1$ на границах возможной области, симметрично расположенные слева и справа относительно крайней точки области $\bar{K}$, угловые отклонения точек $\bar{L}$ и $\bar{R}$ соответственно от точек a_{11} и a_{21}обозначим через ψ.

Перейдем для сокращения обозначений к масштабированным координатам, значения которых разделены на величину ρ_{max} и введем в рассмотренной плоскости вспомогательную систему координат с центром в точке a_{11} и осью *x*, проходящей через точку a_{21}. В ней координаты точек $\bar{L}$ и $\bar{R}$ следующие:

$$\bar{L} = (1 - \cos\psi; \sin\psi); \bar{R} = (\cos\psi; \sin\psi).$$



а б в

Рис.2

При угле ψ, близком к $\pi/3$, и приближении точек $\bar{L}$ и $\bar{R}$ к $\bar{K}$ оптимальным вариантом разбиения будет присоединение обеих промежуточных групп к одной из начальных точек, например, к a_{11} (рис.2 б). В этом случае параметры получаемых множеств A_1, A_2 и величина критерия будут следующие:

$\overline{C_1} \approx (\bar{L}+\bar{R})/2 = (0{,}5; \sin\psi)$; $\overline{C_2} \approx (1;0)$; $\rho_{12} \approx (0{,}25+\sin^2\psi)^{0{,}5}$;

$R_1 \approx 1/2$; $R_2 \approx 0$; $\lambda^{(1)} = \rho_{12}/(R_1+R_2) \approx 2(0{,}25+\sin^2\psi)^{0{,}5}$.

При меньшем угле ψ и более удаленном взаимном расположении точек $\bar{L}$ и $\bar{R}$ оптимальным вариантом разбиения будет присоединение точки $\bar{L}$ к a_{11}, а $\bar{R}$ к a_{21}, (рис.2 в). При этом получим следующие параметры множеств A_1, A_2 и величину критерия:

$\overline{C_1} \approx \bar{L} = (1-\cos\psi; \sin\psi)$; $\overline{C_2} \approx \bar{R} = (\cos\psi; \sin\psi)$; $\rho_{12} \approx 2\cos\psi - 1$;

$R_1 \approx R_2 \approx ((1-\cos\psi)^2+\sin^2\psi)^2/2 = (2(1-\cos\psi))^{0{,}5}/2$;

$\lambda^{(2)} = \rho_{12}/(R_1+R_2) \approx (2\cos\psi-1)/(2(1-\cos\psi))^{0{,}5}$.

При пороговом положении точек $\bar{L}$ и $\bar{R}$ выполняется равенство: $\lambda^{(1)} = \lambda^{(2)}$. Отсюда следует условие для порогового значения угла ψ_π:

$(2\cos\psi_\pi - 1)/(2(1-\cos\psi_\pi))^{0{,}5} = 2(0{,}25+\sin^2\psi)^{0{,}5}$.

Перейдем для упрощения вида выражения к новой переменной :

$t = 2(1-\cos\psi_\pi)$; $0 \le t \le 4$.

Условие принимает вид:

$(1-t)/t^{0,5} = (1+t(4-t))^{0,5}$.

Умножая обе части на знаменатель левой части и возводя обе части в квадрат, получим:

$(1-2t+t^2) = t(1+4t-t^2)$.

Перенося все слагаемые в левую части и приводя подобные слагаемые, получим кубическое уравнение относительно t:

$t^3 - 3t^2 - 3t + 1 = 0$.

Подстановкой несложно проверить, что одним из корней будет значение t = -1. Данное значение на входит в допустимый отрезок [0;4]. Разделив уравнение на (t +1), получим квадратное уравнение относительно t:

$t^2 - 4t + 1 = 0$.

Его корни: $t_{1,2} = 2 \pm (3)^{0,5}$. Условию $0 \le t \le 4$ удовлетворяет корень $t_2 = 2 - (3)^{0,5}$.

Подставляя выражение для t, получим:

$2(1-\cos\psi_\pi) = 2 - (3)^{0,5}$; $\cos\psi_\pi = (3)^{0,5}/2$; $\psi_\pi = \arccos((3)^{0,5}/2) = \pi/6$.

При данном значении угла $\rho(\bar{x}, \bar{a}_{21}) = 2\sin(\psi_\pi/2) \approx 0{,}52$. Ему соответствует теоретическое значение предельной величины отклонения δb: $\delta b = \delta b_{21}(\bar{R}) = 0{,}52/1 = 0{,}52$. Поскольку данная величина найдена для предельных, в действительности не реализуемых вариантов подмножеств точек в A, то для практических расчетов принята величина априорного отклонения δb = 0,6. При этом условия априорного включения точки $\bar{x}$ из множества A в подмножества A_1 и A_2 имеют вид, соответственно:

$0 < \delta_{12}(\bar{x}) \le 0{,}6;\ 0 < \delta_{21}(\bar{x}) \le 0{,}6;$.

Данное правило также предложено применить для последующего после априорного расширения подмножеств A_1 и A_2. Только для тех точек, к которым данное правило уже не применимо, применяется переборный принцип разделения.

Рассмотрим приближенный алгоритм решения задачи.

1. *Исходные данные*:

1) n - размерность пространства U,

2) ne - число точек в множестве A, ($n_1 \geq 2$),

3) $A[1:ne][1:n]$ - массив координат точек множества A.

2. *Решаемые задачи*:

1) определение чисел элементов n_1, n_2 и массивов координат точек в квазиоптимальной паре бинарных подмножеств A_1 и A_2, у которых значение критерия $\lambda(A_1,A_2)$ близко к глобальному максимуму λ_{max};

2) определение центров тяжести $C1,C2$ найденных квазиоптимальных бинарных подмножеств A_1 и A_2.

Приближенный алгоритм бинарной кластеризации (ПАБК).

Шаг 1. Предварительный анализ относительного положения точек A. Построение матрицы расстояний.Определение min и max расстояний. Формирование списка PR[1:P] всех пар максимально удаленных точек. Введение начального значения критерия текущего оптимального разбиения: $KR\_MIN := 2\rho_{max}$.

Шаг 2. Перебор всех P пар максимально удаленных точек. Цикл по параметру s ($1 \leq s \leq$ P) по всем парам максимально удаленных точек.

Шаг 2.1. Начальные присваивания:

а) номера очередных максимально удаленных точек: $m1$:= PR[s][1]; $m2$:= PR[s][2];

б) засылка точек $m1$ и $m2$ в текущие множества $A_{1т}$ и $A_{2т}$ и центры тяжести $\bar{C}_{1т}$ и $\bar{C}_{2т}$;

в) формирование начального списка координат точек невключенных вершин AN, а также списков расстояний RC1 и RC2 точек $m1$ и $m2$ до точек из AN.

Шаг 2.2. Циклическое наращивание текущих множеств $A_{1т}$ и $A_{2т}$ за счет включения в них близких точек. Во нвутреннем цикле просмариваются все невключенные точки. Для них определяется соотношение D12=RC1[i]/RC2[i]. Если D12<=0.6, то точка из AN включается в $A_{1т}$; если D12>=1.67, то точка из AN включается в $A_{2т}$. Иначе точка остается в множестве AN. Если произошло включение новых точек в множество $A_{1т}$, то

корректируется его центр тяжести $\bar{C}_{1т}$ и список расстояний RC1. Аналогично, если произошло включение новых точек в множество $A_{2т}$, то корректируется его центр тяжести $\bar{C}_{2T}$ и список расстояний RC2.

Шаг 2.3. Оценка результатов наращивания текущих множеств $A_{1т}$ и $A_{2т}$ за счет включения в них близких точек.

Если все точки из AN включены в $A_{1т}$ и $A_{2т}$ (решение задачт бинарной кластеризации получено), то запись полученных данных и выход из алгоритма.

Если не все точки из AN включены в $A_{1т}$ и $A_{2т}$, то разделение оставшихся выполняется путем перебора вариантов по аналогии с точным решением.

Завершение работы алгоритма.

Моделирование точного и приближенного алгоритмов производилось на широком наборе множеств. Как правило, результат работы приближенного алгоритма совпадает с разбиением, полученным по точному алгоритму. В специальных модельных случаях значения критерия у приближенного метода хуже, чем у точного примерно на 15 %.

В частности, для модельного множества A ={{0;0}; {1;0};{1;1};{0;1};{0;0,8}; {0,2;1}; {0,25;0,25}; {1,00;0,5}} (рис.3а) в двумерном пространстве признаков точное решение (рис.3 б) дает значение критерия, равное λ_{max}=1.097.

Решение: n_1 = 5, A_1 = {{1.0,0.0}; {1,00;0,5};{0.0,0.0};{1.0,1.0}; {0.25,0.25}; n_2 = 3, A_2 = {{0.0,1.0},{0.0,0.8},{0.2,1.0}}, полученное по приближенному методу, дает значение критерия λ_{max}=0.930, что на 15% хуже, чем глобально оптимальное значение.

Применение дополнительных операций отсечения и бинарной кластеризации позволяет построить общий алгоритм разделения множеств произвольной структуры со сложным относительным пространственным положением путем построения иерархических нормальных классификаторов соответствующих множеств.

Литература:

1. Н.И. Гданский, А.М. Крашенинников. Бинарная кластеризация объектов в многомерных пространствах признаков [Текст] // Труды Социологического конгресса. РГСУ. 2012. – 94-98 с.

2. Н.И. Гданский, М.Л. Рысин, А.М. Крашенинников, Линейная классификация объектов с использованием нормальных гиперплоскостей [Электронный ресурс] // «Инженерный вестник Дона», 2012, №4 – Режим доступа: http://ivdon.ru/magazine/archive/n4p1y2012/1324 (доступ свободный) - Загл. С экрана. – Яз. рус.

3. Н.И. Гданский, А.В. Карпов, А.А. Бугаенко. Оптимальное интерполирование типовых динамик в задаче управления с прогнозированием [Электронный ресурс] // «Инженерный вестник Дона», 2012, №3 – Режим доступа: http://ivdon.ru/magazine/archive/n3y2012/936 (доступ свободный) - Загл. С экрана. – Яз. рус.

4. Л. Г. Комарцова, А. В. Максимов. Нейрокомпьютеры // Изд-во МГТУ им. Н.Э. Баумана, 2002. — С. 320.

5. Н.И. Гданский, А.М. Крашенинников. Сравнение эффективности методов бинарной кластеризации множество точек-прецендентов [Текст] // Математический методы и приложения: Труды двадцать вторых математических чтений РГСУ. АПКиППРО. 2013. – 59-67 с.

6. Л.Н. Ясницкий. Введение в искусственный интеллект. — 1-е. // Издательский центр «Академия», 2005. — С. 176.

7. Н.И. Гданский, М.Л. Рысин, А.М. Крашенинников. Применение современных информационных технологий в учебном процессе высшей школы [Текст]: монография // Изд-во РГСУ, 2012, ISBN 978-5-905675-31-7. – С.150.

8. Н.И. Гданский, А.М. Крашенинников, Е. А. Слюсарев. Использование геометрического подхода при построении классификаторов в системах искусственного интеллекта [Текст] // Математическое моделирование в

проблемах рационального природопользования. Сборник научных трудов Всероссийской молодежной конференции. РГСУ. 2012. – с.36-43.

9. Structure of Decision. The Cognitive Maps of Political Elites // Ed. by R. Axelrod. – Princeton: Princeton University Press, 1976. - 405 p.

10. Shapiro M.J., Bonham G.M. Cognitive processes and foreign policy decision-making // International Studies Quarterly. 1973. – P. 147–174.