УДК 519.95

АНАЛИЗ ДАННЫХ И ПРИНЯТИЕ РЕШЕНИЙ С ПОМОЩЬЮ ЛОГИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ В ФОРМЕ ПОЛУПЛОСКОСТЕЙ

© 2017 Н.А. Игнатьев, Д.Ю. Саидов

Национальный университет Узбекистана имени Мирзо Улугбека, г. Ташкент, Узбекистан

Статья поступила в редакцию 29.09.2017

Рассматривается интеллектуальный анализ данных через решение задач распознавания с учителем. В качестве инструмента для извлечения новых знаний из баз данных предлагается использовать логические закономерности в форме полуплоскостей. Описано 3 способа анализа исходных и латентных признаков на основе: критерия Фишера; отношения внутриклассового сходства и межклассового различия, определяемого через функцию Лагранжа; критерия для вычисления оптимальной границы между значениями из разных классов. Предложена методика отбора информативных наборов признаков с учётом этих способов анализа. Рассматривалось отображение различных описаний объектов на числовую ось. Доказано, что использование оптимальной границы между классами на числовой оси в качестве порога для линейной решающей функции увеличивает обобщающую способность при распознавании. Этот эффект объясняется отказом от предположения о нормальном распределении данных выборки при выборе порога. Предложенная технология анализа данных востребована при разработке интеллектуальных систем.
*Ключевые слова:*логические закономерности в форме полуплоскостей, информативные признаки, обобщающая способность.

ВВЕДЕНИЕ

Линейные дискриминантные функции (ЛДФ) широко используются в задачах интеллектуального анализа данных. Низкие затраты вычислительных ресурсов, возможность содержательной интерпретации результатов распознавания в качестве новых знаний являются теми свойствами, которые находят применение их при моделировании процессов и явлений в слабо формализованных предметных областях. При компьютерной реализации ЛДФ не требуется таблицы прецедентов, достаточно хранить в памяти лишь веса признаков. В технических устройствах ЛДФ могут быть представлены в виде электронныхсхем и микросхем чипов.

Для показателей точности распознавания большое значение имеет такое свойство признакового пространства как линейная разделимость объектов классов. Одним из способов обнаружения этого свойства является использование нелинейных преобразований признаков.

Нелинейные преобразования признаков, как правило, приводят к описанию объектовв пространстве (обобщенном пространстве) более высокой размерности, чем исходное. В качестве примера можно привести обобщенные линейные дискриминантные функции представляемые с помощью произведений исходных признаков степени не выше 2 и называемыми квадратичными. Обобщенное признаковое пространство можно рассматривать как линейное или спрямляющее, но значительно большей размерности чем исходное.

Переход в спрямляющее пространство объясняется с позиций обобщающей способности алгоритмов распознавания. Теоретически такой подход приемлем, так как повышает меру статистического разнообразия (емкость) класса линейных решающих функций. Доказательство этого факта можно найти в работе В.Н. Вапника [1]. Утверждается, что выборку из m объектов в пространстве из n признаков при $n \geqslant m$ всегда можно с помощью ЛДФ разделить на два класса 2^m способами. В реальных прикладных задачах отношения между объектами выборок данных определяется скрытыми закономерностями и нет смысла рассматривать все возможные варианты разбиения на классы.

Обучение ЛДФ сводится к вычислению вектора весовых коэффициентов. Среди вычислительных методов безусловным лидером является линейный дискриминант Фишера. Лидерские качества демонстрируются в виде высоких относительно других методов показателях обобщающей способности к распознаванию.

Выводы о существовании признакового пространства с линейной разделимостью объектов классов скорее всего представляет теоретический интерес, но для практического использования, как правило, неприемлемы. Обобщенные функции, которые предлагаются для формиро-

Игнатьев Николай Александрович, доктор физико-математических наук, профессор кафедры «Алгоритмы и технологии программирования».
E-mail: n_ignatev@rambler.ru
Саидов Дониер Юсупович, старший научный сотрудник-исследователь. E-mail: doniyor_2286@mail.ru

вания признакового пространства, увеличивают сложность обучения и реализации ЛДФ на несколько порядков выше, чем на исходном признаковом пространстве.

Исследователи искали ответ на вопрос: с помощью каких нелинейных преобразований строить спрямляющее пространство? В методе SVM[2] такой выбор был сделан на использование ядерных функций. Несмотря на наличие теоретического обоснование метода никаких рекомендаций по выбору ядерных функций не разработано.

В [2] предлагалось решение проблемы линейной разделимости с помощью матриц попарного сходства объектов. Использование этих матриц рассматривалось в качестве одной разновидностей без признакового распознавания. Принцип без признакового распознавания распространяется на такие известные методы как ближайший сосед, *k* ближайших соседей, базовым свойством которых является локальная компактность по мере близости.

Проблемы поиска информативных признаков как в исходном, так и в расширенном (спрямляющем) пространстве оставалось открытой. Целью отбора была адаптация к той структуре признакового пространства, для которой существует линейная разделимость объектов классов.

Логические закономерности в форме полуплоскостей применяются в интеллектуальном анализе данных. Целью анализа является поиск скрытых закономерностей (новых знаний) из баз (хранилищ) данных.Результаты анализа востребованы при принятии решений в трудно формализуемых задачах.

Сложность формализации заключается в отсутствии единого критерия, для оптимизации которого можно использовать уже известные методы либо разрабатывать новые. Как правило, задачи принятия решения многокритериальные. Получить оптимальное решение по каждому критерию практически невозможно. Выбор критерия (критериев) остается за лицом принимающим решение (ЛПР).

Наиболее известный и широко применяемый на практике критерий Фишера [3] не претендует на полноту исследования структуры данных с помощью логических закономерностей в форме полуплоскостей. В работе предлагаются два новых критерия для решения этой проблемы и методология их использования для принятия решений. Описан эвристический метод отбора информативных наборов признаков в спрямляющем пространстве. Востребованность метода доказывается через вычисление показателей обобщающей способности алгоритмов распознавания, основанных на принципах разделения объектов поверхностями.

ПОСТАНОВКА ЗАДАЧИ

Рассматривается двухклассовая задача распознавания в стандартной постановке. Каждый из объектов выборки $E_0=\{S_1,\dots,S_m\}$ принадлежат одному из классов K_1 или K_2 ($E_0=K_1\cup K_2$) и описываются с помощью n количественных признаков $X(n)=(x_1,\dots,x_n)$. Для распознавания объектов на E_0 используется обобщенные линейные решающие функции вида $d(S)=w_1y_1+\dots+w_ry_r$, где $y_c=f_c(S)$, $f_c(S)\in\left\{x_{i_1}^{a_1}x_{i_2}^{a_2}\dots x_{i_t}^{a_t}\right\}$, $a_j\in\{0,1\}$, $j=1,\dots,t$, $t>1$.

Считается, что для оценки выбора информативного набора признаков $Y(p)=(y_1,\dots,y_p)$ используется функционал $F(E_0,Y(p))$. Требуется определить:

- критерии для оценки закономерностей в форме полуплоскостей;
- информативный набор признаков

$$Y(p)=\arg\max_{f_i(S)\in\Omega^t} F(E_0,Y(p)),$$

где Ω^t - множество обобщенных функций степени не выше t.

2. КРИТЕРИИ ДЛЯ ОЦЕНКИ ЗАКОНОМЕРНОСТЕЙ В ФОРМЕ ПОЛУПЛОСКОСТЕЙ

Закономерности в форме полуплоскостей представляют предикат вида $P(x)=\left[\sum_{i=1}^{n}w_ix_i\le w_0\right]\in\{0,1\}$. Геометрическим местом точек, равноудаленным от двух эталонов из разных классов, является гиперплоскость, значения весовых коэффициентов которой вычисляются через координаты эталонов [4]. В качестве таких эталонов в данной работе предлагается рассматривать векторы математических ожиданий M_1, M_2 значений признаков объектов по каждому из классов K_1 и K_2.

Пусть $m_r^1\in M_1$, $m_r^2\in M_2$ – математическое ожидание (среднее-арифметическое) значений признака $y_r\in\Omega^t$ соответственно в классах K_1 и K_2. Внутриклассовое сходство и межклассовое различие признака $y_r\in\Omega^t$ по объектам $S_i\in E_0$ ($S_i=(y_{i1},\dots,y_{ip})$), $p>1$ вычислим соответственно как

$$\theta_r=\sum_{j=1}^{2}\sum_{S_i\in K_j}\left(y_{ir}-m_r^j\right)^2 \quad\text{и}\quad \gamma_r=\sum_{j=1}^{2}\sum_{S_i\in K_j}\left(y_{ir}-m_r^{3-j}\right)^2.$$

Для оценки веса (разделяющей способности) w_r признака $y_r\in\Omega^t$ по значениям θ_r, γ_r предлагается использовать функционал из [5]

$$J(w)=\frac{\sum_i w_i\theta_i}{\sum_i w_i\gamma_i}\to\min. \qquad (1)$$

При ограничении на веса в (1) $\sum_i w_i=1$, $w_i>0$ функция Лагранжа для решения задачи условной оптимизации имела вид

$$L(w)=\frac{\sum_i w_i\theta_i}{\sum_i w_i\gamma_i}+\lambda\left(\sum_i w_i-1\right),$$

а значения весов вычислялись как $w_i=\frac{\gamma_i-\theta_i}{\sum_j(\gamma_j-\theta_j)}$.

Согласно доказанной в [5] теоремы, необходимым и достаточным условием выбора признака $y_j\in Y(p)$ кандидатом на удаление из набора $Y(p)=(y_1,...,y_p)$ при ограничении $\sum_i w_i=1$, $w_i>0$ является $\frac{\theta_j}{\gamma_j}=\max_{y_i\in Y(p)}$. Соотношение

$$\frac{\theta_j}{\gamma_j} \tag{2}$$

дает возможность оценивать и упорядочивать признаки по плотности их распределения вокруг математических ожиданий классов. Чем выше плотность, тем меньше значение (2).

Также как и в (2) вычисление внутриклассового сходства по отдельному признаку используется в критерии Фишера [3]

$$\frac{|m_1-m_2|^2}{\tilde{s}_1+\tilde{s}_2}, \tag{3}$$

в котором сумма внутриклассового разброса $\tilde{s}_1+\tilde{s}_2=\theta_r$, а m_1-m_2 естьразность математических ожиданий классов K_1и K_2на числовой оси.

Отличный от (2) и (3) критерий рассчитан на анализ порядка расположения объектов классов на числовой оси [6]. Пусть

$$S_{r_1},S_{r_2},...,S_{r_m} \tag{4}$$

последовательность объектов, упорядоченная по невозрастанию значений признака $y_r\in Y(p)$. Упорядоченное множество значений (4) разделяется на два непересекающихся интервала $[c_1,c_2]$, $(c_2,c_3]$, каждый из которых рассматривается как градация номинального признака. Критерий для определения границыc_2основывается на проверке гипотезы (утверждения) о том, что каждый из двух интервалов содержит значения количественного признака объектов только одного класса.

Пусть u_i^1,u_i^2 – количество значений (4) некоторого количественного признака $y\in Y(p)$ класса K_i, i=1,2 соответственно в интервалах $[c_1,c_2]$, $(c_2,c_3]$, $|K_i|>1$, v – порядковый номер элемента упорядоченной по возрастанию последовательности (4) изE_0 определяющий границы интервалов как $c_1=S_{r_1}$, $c_2=S_{r_v}$, $c_3=S_{r_m}$. Критерий

$$w(y)=\left(\frac{\sum_{d=1}^{2}\sum_{i=1}^{2}u_i^d\left(u_i^d-1\right)}{\sum_{i=1}^{2}|K_i|\left(|K_i|-1\right)}\right)\left(\frac{\sum_{d=1}^{2}\sum_{i=1}^{2}u_i^d\left(|K_{3-i}|-u_{3-i}^d\right)}{2|K_1||K_2|}\right)\to\max_{c_1<c_2<c_3} \tag{5}$$

позволяет вычислять оптимальное значение границы между интервалами $[c_1,c_2]$, $(c_2,c_3]$. Выражение в левых скобках (5) представляет внутриклассовое сходство, в правых – межклассовое различие.

Значение $w_r=w(y_r)$ рассматривается как вес признака $y_r\in Y(p)$, а границы интервалов могут используются для нормирования значений признака объекта $S_i=(y_{i1},...,y_{ip})$ по формуле $\overline{y_{ir}}=\frac{y_{ir}-c_2}{c_3-c_1}$.

3. ОТБОР ИНФОРМАТИВНЫХ НАБОРОВ ПРИЗНАКОВ

Задача поиска информативных наборов признаков для линейной разделимости объектов классов K_1иK_2 является *NP* полной. Из этого следует вывод, что кроме полного перебора других способов решить задачу поиска глобального экстремума функционала $F(E_0,Y(p))$ не существует. Используя некоторые эвристики, можно получить локальный экстремум функционала.

Смысл использования эвристик для решения проблемы линейной разделимости сводится к следующему. Пусть Ω^t – множество обобщенных функций степени не выше t. На множестве пар $(y_i,y_j)\subset Y(p)$ рассматривается сокращенный перебор с целью поиска экстремума по критерию Фишера

$$\Phi(w)=\frac{|m_1-m_2|^2}{\tilde{s}_1+\tilde{s}_2}\to\max. \tag{6}$$

Критерий (6) отличается от (3) тем, что для линейной проекции описаний объектов на числовую ось необходимо вычислять значения вектора весовw.Приемлемым считается результат, при котором точность распознавания на обучении по $(y_i,y_j)\subset Y(p)$ не ниже чем на исходном наборе $X(n)$.

Для вычисления коэффициентов дискриминантной функции $d(y)=w_1y_i+w_2y_j+w_0$ по паре $(y_i,y_j)\subset Y(p)$ сформируем матрицу ковариаций $Z=\begin{pmatrix}z_{11}&z_{12}\\z_{21}&z_{22}\end{pmatrix}$ и вектор-столбец разностей $\begin{pmatrix}m_i^1-m_i^2\\m_j^1-m_j^2\end{pmatrix}$, где m_i^1,m_j^1,m_i^2,m_j^2 математические ожидания по признакам y_i, y_j соответственно в классе K_1 и K_2. Решение системы линейных алгебраических уравнений

$$\begin{cases}w_1z_{11}+w_2z_{12}=m_i^1-m_i^2\\w_1z_{21}+w_2z_{22}=m_j^1-m_j^2\end{cases} \tag{7}$$

дает искомые значения весов w_1, w_2 дискриминантной функции.

Выбор коэффициентов линейного дискриминанта Фишера по (7) связан с предположением, что выборка данных распределена по

нормальному закону. Исходя из этого предположения выбор порога дискриминантный функции $d(y)$ производится как

$$w_0 = -(w_1(m_i^1|K_1| + m_i^2|K_2|) + w_2(m_j^1|K_1| + m_j^2|K_2|))/m. \quad (8)$$

Способ выбора порога без всяких предположений о природе среды впервые был предложен в [7]. Значение порога вычислялось по границе c_2 интервалов $[c_1,c_2]$, $(c_2,c_3]$ по (5) как

$$w_0 = \frac{c_2 + u(S)}{2}, \quad (9)$$

где $u(S)=w_1y_i+w_2y_j$, $S=(y_1,\ldots,y_p)$, $u(S)\in(c_2,c_3]$ – ближайщий к c_2 объект E_0 на числовой оси.

4. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Для вычислительного эксперимента были взяты 4 выборки данных из [6, 8, 9], содержащих представителей двух непересекающихся классов. Для описания объектов использовались признаки, измеренные в интервальных шкалах. Параметры выборок представлены в табл. 1.

Вычислительный эксперимент проводился на объектах выборок с описанием как в исходном так и в спрямляющем пространстве. Спрямляющее пространство было представлено обобщенными функциями степени не выше 2. Сравнительный анализ данных на выборке Chelust по критериям (2),(3),(5) представлен в табл. 2.

Нетрудно заметить, что между значениями критериев (см. табл. 2) нет линейной или квазилинейной зависимости. Многообразие отношений на множестве значений свидетельствует как о сложности структуры данных так сложности принятия решения по ним.

Значения 0,8001 по критерию (5)на признаках x_4x_5, x_5x_6 и x_6 указывает на то, что порядок

Таблица 1. Параметры выборок данных

№	Выборка данных	Количество	
		Объектов	признаков
1	Australian	690	14
2	Chelust	42	6
3	Gipertaniya	147	29
4	Seeds	140	7

Таблица 2. Сравнительный анализ данных Chelust по критериям (2), (3), (5)

Признаки (обобщенные функции)	Значения по критериям		
	(3)	(5)	(2)
x_1	0,0074	0,3781	0,7617
x_1x_2	0,0106	0,3781	0,6910
x_1x_3	0,0090	0,3620	0,7258
x_1x_4	0,0679	0,5757	0,2596
x_1x_5	0,0341	0,4196	0,4113
x_1x_6	0,0398	0,4355	0,3740
x_2	0,0134	0,3892	0,6391
x_2x_3	0,0158	0,3892	0,6014
x_2x_4	0,0811	0,6241	0,2270
x_2x_5	0,0439	0,4905	0,3517
x_2x_6	0,0499	0,4683	0,3230
x_3	0,0048	0,2884	0,8320
x_3x_4	0,0806	0,5312	0,2281
x_3x_5	0,0412	0,5180	0,3664
x_3x_6	0,0457	0,4743	0,3424
x_4	0,2669	0,8965	0,0819
x_4x_5	0,2278	0,9107	0,0946
x_4x_6	0,1797	0,8001	0,1170
x_5	0,1108	0,6254	0,1769
x_5x_6	0,1108	0,8001	0,1769
x_6	0,0787	0,8001	0,2322

расположения объектов одного класса относительно другого не изменился. Изменения есть у показателей плотности распределения объектов относительно математических ожиданий (центров) классов, вычисляемых по критериям(2) и (3).

Из максимального значения 0,9107 по критерию (5) следует вывод, что свойство линейной разделимости среди всех признаков из табл. 2 наиболее выражено у $x_4 x_5$. Несмотря на менее выраженное свойство линейной разделимости, показатель плотности распределения (среднеквадратичное отклонение от математических ожиданий классов) уx_4по (2) выше чем у $x_4 x_5$. Относительно малое значение отклонения (0,0819 относительно 0,0946) указывает на более высокую плотность распределения.

Влияние выбора порога дискриминантной функции w_0 с предположением о нормальности распределения выборки по критерию Фишера [3] и по критерию (5) по аналогии соответственно с (8) и (9) на исходных наборах признаков приводится в табл. 3.

Значение критерия (5), равное 1,0, по определению означает, что представители классов на числовой осине пересекаются между собой. Корректное (без ошибок) распознавание объектов на выборках Chelust и Seeds служат подтверждением этому определению.

Эффект от выбора значения порога по (8) или (9) в спрямляющем пространстве на данных Chelust демонстрируется в табл. 4 и рис. 1.

На рис. 1.a и 1.b показана последовательность расположения объектов классов по первой и второй паре признаков из табл. 4. В границах интервалов $[c_1, c_2]$, $(c_2, c_3]$ по (5) содержатся представители соответственно классов K_2 и K_1. При пороге, вычисленном по (8) (на рис. 1 указан жирной чертой), число ошибок равно соответственно 2 и 4.

Сравнивая результаты по данным Chelust из табл. 3 и табл. 4 отметим, что корректное распознавание в спрямляющем пространстве достигнуто за счет использования обобщенных функций не выше 2 степени. Для вычисления этих функций будет достаточно задать значения исходных признаков x_1, x_4, x_5 или x_2, x_4, x_5, x_6.

ЗАКЛЮЧЕНИЕ

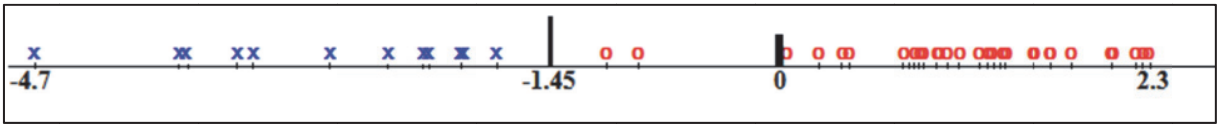
В работе рассмотрена проблема выбора решений в трудно формализуемых задачах путем анализа закономерностей в форме полуплоско-

Таблица 3. Точность распознавания в исходном пространстве признаков

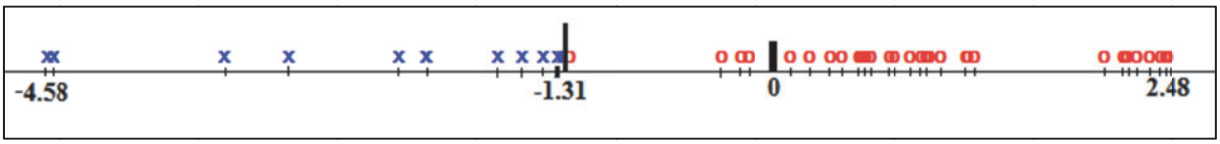
№	Выборка данных	Значение критерия		Число ошибок при выборе порога по критерию	
		(3)	(5)	(6)	(5)
1	Australian	0,0088	0,6176	96	84
2	Chelust	0,5446	1,0000	3	0
3	Gipertaniya	0,2180	0,672	8	1
4	Seeds	0,1616	1,0000	2	0

Таблица 4. Точность распознавания в спрямляющем пространстве

№	Комбинации из пар признаков	Значение критерия		Число ошибок при выборе порога по критерию	
		(3)	(5)	(8)	(9)
1	x_1x_5, x_4x_5	0,5426	1,0	2	0
2	x_2x_6, x_4x_5	0,2870	1,0	4	0



a)



b)

Рис. 1. Последовательность расположения объектов классов и границы пороговна числовой оси

стей. Предложены критерии для анализа и методология их использования, которая востребована для разработки и управления технических устройств на основе систем искусственного интеллекта.

СПИСОК ЛИТЕРАТУРЫ

1. *Вапник В.Н.* Восстановление закономерностей по эмпирическим данным. М.: Наука. 1979. 447 с.
2. *Середин О.С.* Линейные методы распознавания образов на множестве объектов произвольной природы, представленные попарными сравнениями. Общий случай // Известия Тульского государственного университета. Естественные науки. 2012. Вып. 1. С. 141-152.
3. *Дуда Р., Харт П.* Распознавание образов и анализ сцен. Мир. 1976. 512 с.
4. *Ту Дж., Гонсалес Р.* Принципы распознавания образов. М: Мир, 1978.416 с.
5. *Игнатьев Н.А.* Выбор минимальной конфигурации нейронных сетей // Вычислительные технологии.2001.Т.6. №1. С. 23-28.
6. *Игнатьев Н.А.* Вычисление обобщенных показателей и интеллектуальный анализ данных // Автоматика и телемеханика. 2011. № 5. С.183-190.
7. *Игнатьев Н.А., Нуржонов Ш.Ю.* Выбор параметров регуляризации для повышения обобщающей способности дискриминантных функций // Узбекистон Республикаси Курол Кучлари академияси нинг хабарлари. 2014. Т. 1. № 1(14).С. 81-87.
8. Knowledge discovering from clinical data based on classification tasks solving / *N.A. Ignat'ev, F.T. Adilova, G.R. Matlatipov, P.P. Chernysh* // Medinfo. Amsterdam: ios press. 2001. pp. 1354-1358.
9. Index of /ml/machine-learning-databases. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/ (дата обращения 14.03.2017).

DATA ANALYSIS AND DECISION-MAKING WITH LOGICAL REGULARITIES IN THE FORM OF HALF-PLANES

© 2017 N.A. Ignatyev, D.Y. Saidov

National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan

The intellectual analysis of data through solving recognition problems with the teacher is considered. As a tool for extracting new knowledge from databases, it is proposed to use logical regularities in the form of half-planes. Three methods for analyzing the initial and latent features are described on the basis of: Fisher's criterion; The relationship of intra-class similarity and the interclass difference defined through the Lagrange function; Criterion for calculating the optimal boundary between values from different classes. A methodology is proposed for selecting informative sets of features taking into account these methods of analysis. The mapping of various descriptions of objects onto the numerical axis was considered. It is proved that using the optimal boundary between classes on the numerical axis as a threshold for a linear decision function increases the generalizing ability in recognition. This effect is explained by the rejection of the assumption of a normal distribution of sample data when choosing a threshold. The proposed technology for data analysis is in demand in the development of intelligent systems.

Keywords: Logical regularities in the form of half-planes, informative features, generalizing ability.

Nikolay Ignatyev, Doctor of Physics and Mathematics, Professor at the Algorithms and Program Technologies Department. E-mail: n_ignatev@rambler.ru
Doniyor Saidov, Senior Researcher.
E-mail: doniyor_2286@mail.ru