УДК 621.391:396

МОДЕЛИ И АЛГОРИТМЫ КЛАССИФИКАЦИИ МНОГОМЕРНЫХ ДАННЫХ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ С РАДИАЛЬНО-БАЗИСНЫМИ ФУНКЦИЯМИ

А. А. Сирота, А. В. Цуриков

Воронежский государственный университет

Поступила в редакцию 14.03.2013 г.

Аннотация. Рассматриваются модели классификации многомерных данных с использованием радиально-базисных функций, применительно к задаче создания контентно-зависимых цифровых водяных знаков. Исследуется вероятность ошибки классификации многомерных данных в зависимости от размерности признакового пространства.
Ключевые слова: цифровой водяной знак, классификация данных, нейронные сети, радиально-базисные функции.
Annotation. The paper determines approaches to creating high-dimensional data classifiers using radial-basis functions for developing content-dependent digital watermarks. It also examines the dependency between the data classification errors and the space dimension size.
Keywords: content-dependent digital watermark, data classification, neural networks, radial-basis function.

ВВЕДЕНИЕ И ПОСТАНОВКА ЗАДАЧИ

Одними из перспективных технологий, в которых находят применение методы и алгоритмы классификации данных, являются технологии цифровых водяных знаков (ЦВЗ), используемые для защиты авторских прав на объекты цифрового контента [1]. Наиболее часто ЦВЗ применяются для разметки мультимедиа файлов, служащими контейнерами для скрытного (стеганографического) встраивания идентификационных данных. Актуальными также являются задачи, связанные с защитой различных электронных документов, содержащие в своей основе текстовые данные. Существует несколько способов встраивания информации в текстовые контейнеры. Известны методы текстовой стеганографии, такие как метод выравнивания пробелами, метод чередования маркеров конца строки, двоичных нулей, знаков одинакового начертания и другие. Все эти методы объединяет одна особенность – они жестко привязаны к структуре расстановки элементов текста и в случае его переформатирования встроенная информация может быть потеряна. Поэтому возникает задача обоснования новых способов создания ЦВЗ, которые являются инвариантными формату представления текста и привязанными только к содержанию текста (контентно-зависимыми).

В работах [1-3] для создания ЦВЗ в объектах мультимедиа (цифровые изображения, аудио и видеофайлы данных) используется аппарат искусственных нейронных сетей, позволяющий реализовать функциональные модели преобразования данных. Показано, что при использовании ИНС процесс встраивания данных в файл-контейнер носит существенно менее прозрачный характер, что позволяет добиться большей скрытности и устойчивости создаваемых ЦВЗ. В связи с этим, представляет также интерес применение данного подхода для создания контентно-зависимых ЦВЗ для текстовых данных, представленных в различных форматах.

Будем исходить из следующей модели данных. В качестве ЦВЗ без ограничения общности будем рассматривать двоичную последовательность $D = \{d^{(p)}, p = 1, P\}$, при этом $d^{(p)} \in \{-1; 1\}, p = \overline{1, P}$ – скалярная величина ($m = 1$), которая несет в себе один бит информации в пределах встраиваемого сообщения. Пусть $z = (z_1, \ldots, z_n)^T$ случайный вектор, представляющий фрагмент текстового файла контейнера. Совокупность реализаций вектора

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 13-01-97507 р_центр_ а

$z \in R^n : z^{(p)}, p = \overline{1,P}$ представляет текст в целом и сопоставляется с элементами последовательности ЦВЗ. При формализации задачи необходимо представить каждый фрагмент контейнера в виде некоторого многомерного вектора данных $z \in R^n$, однозначно характеризующего этот фрагмент. Элементы вектора, а также его размерность, непосредственно зависят от выбора способа формирования этого вектора на основе исходного текста. К таким способам можно отнести способ, основанный на кодировании символов текста, способы, основанные на подсчете длины каждого слова в тексте, подсчете количестве слов в предложении и т.п. Каждый из этих способов дает возможность получения некоторых статистических характеристик текста.

Таким образом, преобразование текста в последовательность $B = \{z^{(p)}, p = \overline{1,P}\}$ можно осуществить различными способами, которые образуют конечное множество вариантов S. Тогда любой фрагмент текстового контейнера может быть представлен в виде вектора $z = z(s)$, где $s \in S, P = P(s)$.

Используя введенные обозначения можно свести исходную задачу к задаче нахождения отображения вида

$$\tilde{d}^{(p)} = F(d^{(p)}, z^{(p)}(s)),$$
$$p = \overline{1,P(s)}, s \in S,$$
$$E_d = \left\| d - \tilde{d} \right\| \to \min.$$

Это означает, что задача всегда сводится к задаче классификации произвольного множества элементов B на два класса, первый из которых соответствует значениям $d^{(p)} = 1$, а второй – значениям $d^{(p)} = -1$.

Таким образом, задача создания ЦВЗ для текстового контейнера сводится к построению алгоритма классификации многомерных данных, представляющих текст. Такие ЦВЗ являются контентно-зависимыми, так как фактического встраивания данных не происходит, а реализуется процедура «узнавания» кодированных фрагментов текста для извлечения последовательности элементов ЦВЗ. Подобные задачи с высокой степенью эффективности могут быть решены путем использования нейронных сетей на основе RBF (Radial Basis Functions). Поэтому представляется важным исследование потенциальных возможностей сетей с RBF на основе эталонной статистической модели данных для создания и исследования универсальных алгоритмов классификации фрагментов текста.

ЭТАЛОННАЯ МОДЕЛЬ КЛАССИФИКАЦИИ ДАННЫХ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ С RBF

При создании эталонной статистической модели необходимо учесть следующие соображения. Так как общая задача классификации текста сводится к определению принадлежности каждого его фрагмента одному из двух классов (1 и -1), в модели должно генерироваться два класса случайных векторов, образующих точки в многомерном пространстве произвольной размерности. Учитывая, что в процессе преобразования текст разбивается на фрагменты, каждый из которых содержит одинаковое количество структурных элементов (в зависимости от типа разбиения фрагментом может быть набор слов, абзацев, букв, а элементом – соответственно слово или буква), размерность пространства можно интерпретировать как количество элементов во фрагменте. Размерность пространства напрямую влияет на потенциальную разделимость данных при их классификации, так как ее увеличение фактически означает увеличение числа признаков, по которым можно классифицировать заданный фрагмент. Таким образом, для исследования потенциальных возможностей применения RBF при создании контентно-зависимых ЦВЗ можно предложить следующую статистическую модель.

В единичном гиперкубе I размерности n в соответствии с равномерным законом распределения случайным образом размещается P точек 2-х классов

$$\begin{gathered} B_+ = \{z^{(p)}, p = \overline{1,P_1}\}, \\ B_- = \{z^{(p)}, p = \overline{1,P_2}\}, \\ B = \{z^{(p)}, p = \overline{1,P}\} = B_+ \cup B_-, \\ P_1 + P_2 = P, \\ D_+ = \{d^{(p)} = 1,\ p = \overline{1,P_1}\}, \\ D_- = \{d^{(p)} = 0,\ p = \overline{1,P_2}\}, \\ D = \{d^{(p)}, p = \overline{1,P}\} = D_+ \cup D_-, \end{gathered} \quad (1)$$

Вероятность появления каждой точки первого класса обозначим Q_1. Вероятность появления точек второго класса – $Q_2 = 1 - Q_1$. Для ге-

нерации случайного количества точек с заданными вероятностями может использоваться алгоритм генерации биномиального распределения.

Основной задачей является разработка обучаемого классификатора совокупности данных $B = \{z^{(p)}, p = \overline{1,P}\}$ и оценка вероятности ошибки классификации точек двух типов, равномерно распределенных в гиперкубе I. Следует также учитывать, что после создания классификатора предъявляемые для классификации данные могут быть искажены (например, если на вход подается текст с опечатками), что также должно быть отражено в модели. Для этого вводится вероятность искажения каждого элемента (фрагмента) P_r, при этом искажение вносится в одну из компонент данного элемента, которая модифицируется в соответствии с равновероятным законом распределения. Синтезируемый классификатор должен быть устойчивым по отношению к данным искажениям.

Будем искать классификатор в виде нейронной сети, реализующей отображение вида

$$\tilde{d} = F(d,z) = sign\ \Phi(z) = \begin{cases} 1, & \Phi(z) \geq 0, \\ -1, & \Phi(z) < 0, \end{cases} \quad (2)$$

$$\Phi(z) = \sum_{i=1}^{K} w_i \varphi_i(z),$$

$$\varphi_i(z) = \varphi(\|z - u_i\|) = \exp\left[-\frac{(\|z - u_i\|)^2}{2\sigma_i^2}\right], \quad (3)$$

где u_i – центр i – ой радиально-базисной функции; σ_i – параметр влияния i – ой радиально-базисной функции; w_i – соответствующий весовой коэффициент этой функции; K – количество используемых функций. Можно заметить, что классификатор зависит от трех групп параметров – центров радиальных функции $U = (u_1,...,u_K)^T$, параметров влияния $\Sigma = (\sigma_1,...,\sigma_K)^T$, а также весовых коэффициентов $W = (w_1,...,w_K)^T$.

АЛГОРИТМ НАСТРОЙКИ ПАРАМЕТРОВ КЛАССИФИКАТОРА

Процесс построения классификатора разобьем на несколько этапов.

Рассмотрим сначала задачу определения центров RBF-функций. Основная идея работы нейронной сети на основе радиальных базисных функциях состоит в нелинейном преобразовании входных многомерных данных в пространство большей размерности. Теоретическую основу такого подхода составляет *теорема Ковера о разделимости образов [4, 5], которая утверждает следующее: нелинейное преобразование сложной задачи классификации образов в пространство более высокой размерности повышает вероятность линейной разделимости образов в новом пространстве*. Известно, что задача классификации линейно-разделимых множеств относительно легко разрешима.

Рассмотрим множество B. Для каждого образа $z^{(p)} \in B,\ p = \overline{1,P}$ можно определить вектор, состоящий из множества действительных функций $\{\varphi_i(z^{(p)}) \mid i = 1,2,...,K\}$, вида

$$\varphi(z^{(p)}) = (\varphi_1(z^{(p)}), \varphi_2(z^{(p)}),..., \varphi_K(z^{(p)}))^T. \quad (4)$$

Так как образ $z^{(p)}$ является вектором в n – мерном входном пространстве, векторная функция $\varphi(z^{(p)})$ отображает точки n-мерного входного пространства в новое пространство размерности K.Функции $\varphi_i(z^{(p)})$ называются скрытыми, поскольку они играют роль скрытых элементов нейронных сетей прямого распространения. Соответственно пространство, образованное множеством скрытых функций $\{\varphi_i(z^{(p)})\}_{i=1}^{M}$, называется скрытым пространством или пространством признаков.

Дихотомия $\{B_+, B_-\}$ множества B называется φ – разделимой, если существует K – мерный вектор w, для которого можно записать

$$w^T\varphi(z^{(p)}) \geq 0, \quad z^{(p)} \in B_+,$$
$$w^T\varphi(z^{(p)}) < 0, \quad z^{(p)} \in B_-.$$

При этом гиперплоскость, задаваемая уравнением $z : w^T\varphi(z) = 0$, описывает разделяющую поверхность в φ – пространстве (т.е. в скрытом пространстве). Обратный образ этой поверхности определяет разделяющую поверхность во входном пространстве [5]. При этом доказано, что каждое исходной множество, случайным образом размещенное в многомерном пространстве является φ-разделимым с вероятностью 1 при условии соответственно большой размерности нового пространства K.

С учетом этого простейшим вариантом решения задачи назначения центров RBF может быть использование в качестве центра каждой радиальной функции одной из точек множества B: $u_i = z^{(i)}, i = \overline{1,P}$, сгенерированной в гиперкубе. При этом общее количество функций берется равным количеству классифицируемых точек $K = P$, а значение параметра влияния может быть подобрано одинаковым

для всех функций $\sigma_i = \sigma, \quad i = \overline{1,K}$. Использование всех точек в качестве центров радиальных функций означает, что во входном пространстве каждая точка будет окружена сферической разделяющей поверхностью одинакового радиуса и мы получим сферически разделяемую дихотомию.

Для окончательного построения классификатора при реализации данного подхода нужно подобрать значение параметра влияния, обеспечивающего отделение каждой точки $z^{(p)} \in B, \ p = \overline{1,P}$ от других и определить весовые коэффициенты путем решения системы линейных алгебраических уравнений (СЛАУ) вида

$$GW = d, \qquad (5)$$

$$G = \left\| g_{p,i} \right\|,$$

$$g_{p,i} = \left\| \varphi_i(z^{(p)}) \right\|,$$

$$p = \overline{1,P}, \ i = \overline{1,K}, \ P = K,$$

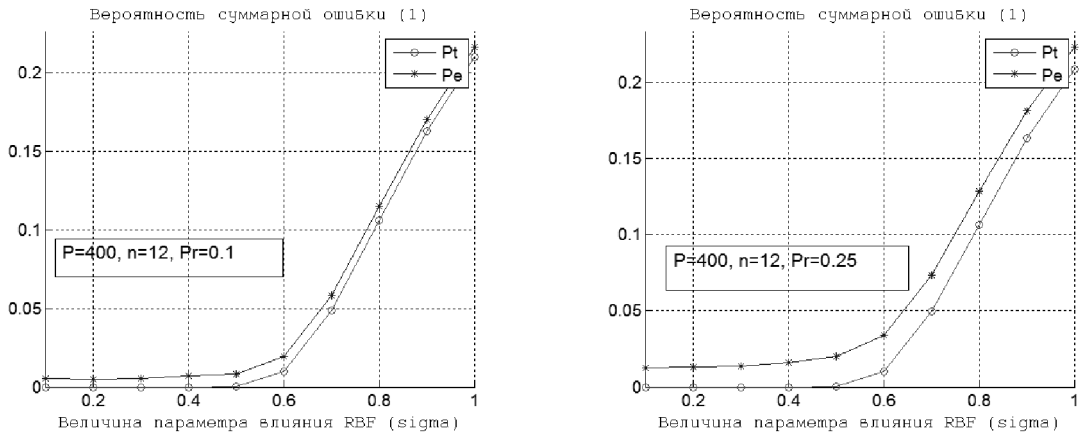
где G – матрица Грина, являющаяся в данном случае квадратной; $d = (d^{(1)}, \ldots, d^{(P)})^T$ – целевой вектор, определяемый исходя из исходного множества требуемой классификации данных $D = \{d^{(p)}, p = \overline{1,P}\} = D_+ \cup D_-$.

На рисунке 1 представлены результаты решения задачи подобным образом в виде зависимостей вероятности суммарной ошибки классификации данных от величины параметра влияния. Зависимость, обозначенная P_t, определяет вероятность ошибочной классификации исходных точек в многомерном пространстве, а зависимость, обозначенная как P_e, определяет вероятность ошибочной классификации точек, подвергнутых искажению с вероятностью P_r в соответствии с ранее описанной моделью. Представленные зависимости показывают, что путем простого перебора всегда можно найти значение параметра влияния, обеспечивающего однозначное разделение входных точек при условии $K = P$ и определении весовых коэффициентов в соответствии с (5).

При этом можно заметить, что при модификации входных данных в соответствии с моделью возмущений, разделяющая поверхность, построенная на основе исходных данных, перестает однозначно разделять точки в исходном пространстве, что приводит в основном к незначительному увеличению ошибки классификации.

Таким образом, использование подобной конфигурации нейронной сети достаточно эффективно с точки зрения минимизации ошибки классификации, но в то же время является существенно избыточным и приводит к усложнению нейронной сети. В связи с этим возникает задача синтеза и анализа классификатора в виде (2) с существенно уменьшенным количеством используемых функций $K < P$.

При $K < P$ вероятность ошибки возрастает, так как разделяющая поверхность в таком случае строится вокруг нескольких точек одного класса и в него могут попасть точки другого класса. Суть предлагаемого подхода к решению задачи в рассматриваемой постановке состоит в том, что отдельно определяются параметры RBF-функций, предназначенных для классификации точек первого класса (их количество P_1), и – параметры RBF-функций предназна-



а) б)

Рис. 1. Зависимости ошибки классификации от параметра влияния при различных значениях P_r

ченных для классификации точек второго класса (их количество P_2). Для этого в каждом множестве элементов B_+ и B_-, соответствующих точкам первого и второго классов, проводится независимое группирование точек по степени близости в многомерном пространстве. Пусть величина N_{me} определяет среднее количество точек, приходящихся на каждую группу. Тогда количество используемых радиальных функций и, соответственно, групп, используемых для отображения точек множества B_+ будет равно $K_1 = P_1 / N_{me}$, а для точек множества B_- – $K_2 = P_2 / N_{me}$. Обозначим полученные группы точек множества B_+ как $B_+^{(l)} = \{z^{(p,l)}, z^{(p,l)} \in B_+, p = \overline{1, N_l}\}, l = \overline{1, P_1}$ и группы точек множества B_- как $B_-^{(l)} = \{z^{(p,l)}, z^{(p,l)} \in B_-, p = \overline{1, N_l}\}, l = \overline{1, P_2}$. Для подобного группирования элементов, целесообразно использовать алгоритм кластеризации точек k-средних (k-means) [6], который выделяет заданное количество кластеров, основываясь на определении центров масс векторов (центроидов) и поиске наиболее близких к каждому центроиду элементов по критерию $\sum_{z \in B_{+,-}^{(l)}} (z^{(p,l)} - u_l)^2 \to \min$. Данные центроиды и назначаются в качестве центров RBF

$$u_l^{(-)} = \sum_{z \in B_-^{(l)}} z^{(p,l)}, l = \overline{1, P_2}.$$

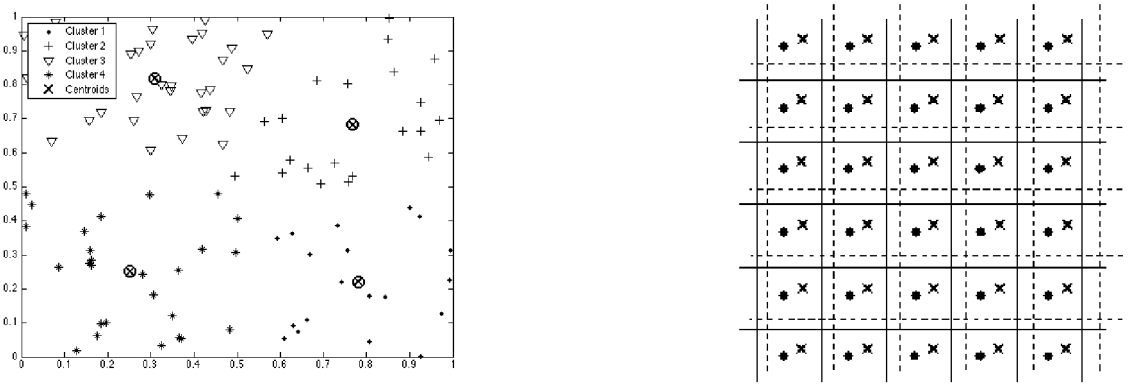
На рис. 2,а можно увидеть пример работы данного алгоритма при разбиении множества B_+ на четыре кластера.

Таким образом, алгоритм k-средних последовательно применяется к каждому из множеств B_+ и B_-, что приводит к разбиению общего гиперкуба, соответственно, на K_1 и K_2 ячеек, как это схематично представлено на рис. 2,б. В пространстве данных происходит образование «решетки», получаемой при разбиении на кластеры точек первого класса (x), при этом поверх нее независимо накладывается решетка, получаемая при разбиении точек второго класса точек (•). Так как каждая ячейка будет классифицироваться своей функцией RBF, то, очевидно, что наложение ячеек с точками разных классов будут приводить к ошибкам классификации.

Помимо центров радиальных функций, необходимо назначить параметры влияния для RBF-функций. В предлагаемом алгоритме рассматривается два способа задания этого параметра. Первый способ предполагает задание общего значения σ для всех функций путем последовательного перебора в заданном диапазоне от 0 до 1. Второй способ основан на определении радиуса каждого полученного кластера (расстояния от центра до наиболее удаленной точки) и задании параметра влияния как

$$\begin{aligned} \sigma_l^{(+)} &= \rho R^{(p,l)}, \\ R^{(p,l)} &= \max_{z \in B_+^{(l)}} \left(\left\| z^{(p,l)} - u_l \right\| \right), \\ l &= \overline{1, P_1}, \\ \sigma_l^{(-)} &= \rho R^{(p,l)}, \\ R^{(p,l)} &= \max_{z \in B_-^{(l)}} \left(\left\| z^{(p,l)} - u_l \right\| \right), \\ l &= \overline{1, P_2}, \end{aligned} \tag{6}$$

где ρ – эмпирически подбираемый коэффициент, обычно равный 0,3...0,5. На заключительном этапе построения классификатора необходимо вычисление весовых коэффициентов. В данном случае ($K < P$) СЛАУ для нахождения



а) б)

Рис. 2. Результат работы процедуры кластеризации точек k-means и его интерпретация

весовых коэффициентов является переопределенной. Для ее решения могут быть использованы различные методы, в том числе метод псевдоинверсии Мура-Пенроуза (нормальное псевдорешение), метод основанный на разложении SVD (Singular Value Decomposition), метод регуляризации по А. Н. Тихонову [5]. Как показали следования, последний вариант позволяет сформировать наиболее устойчивые решения в виде

$$\begin{gathered} GW = d, \\ w = w^{(a)} + (G^T G + \alpha I)^{-1} G^T (d - G w^{(a)}), \\ G = \left\| g_{p,i} \right\|, \\ \underline{g_{p,i}} = \left\| \varphi_i (z^{(p)}) \right\|, \\ p = \overline{1,P}, i = \overline{1,K}, \ K < P, \end{gathered} \tag{7}$$

где $w^{(a)}$ – априорное решение, которое в данном случае целесообразно выбрать следующим образом: $w_l^{(a)} = 1, \quad l = \overline{1,K_1}, \qquad w_l^{(a)} = -1,$ $l = \overline{K_1 + 1, K_2}$, α – параметр регуляризации, выбираемый одним из стандартных методов.

В соответствии с описанной последовательностью действий, можно привести общую блок-схему алгоритма обучения классификатора (рис. 3).

ИССЛЕДОВАНИЕ АЛГОРИТМА

При проведении исследований рассматривались различные варианты организации обработки данных. Первоначально моделировался алгоритм, в котором для всех RBF-функций проводилось назначение общего параметра влияния в режиме прямого перебора. Пример соответствующих зависимостей для вероятностей ошибок P_t, P_e от σ при $n = 20, N_{me} = 8$ представлены на рис. 4 а, б, в, г. Зависимости даны для вариантов решения СЛАУ для нахождения весовых коэффициентов методом Мура-Пенроуза (MP), методом Singular Value Decomposition (SVD) методом регуляризации по Тихонову для двух значений $\alpha = 0.1$ и $\alpha = 0.001$. При данном способе задания σ для RBF можно увидеть, что ошибка имеет определенный минимум, причем наибольший минимум имеет ошибка, получаемая при использовании метода регуляризации.

На рисунке 5 представлены зависимости вероятностей ошибок от N_{me} при реализации переборного и адаптивного способа назначения параметра влияния для каждого кластера в соответствии с (6). В качестве методов решения системы уравнений здесь представлены только метод SVD и метод регуляризации, так как метод Мура-Пенроуза во всех случаях дает примерно одинаковые результаты с методом SVD. При этом графики, представленные на рис. 5 а,б, даны для случая использования для всех RBF-функций общего значения параметра влияния $\sigma = 0.1$. Графики, представленные на рис. 5 в,г

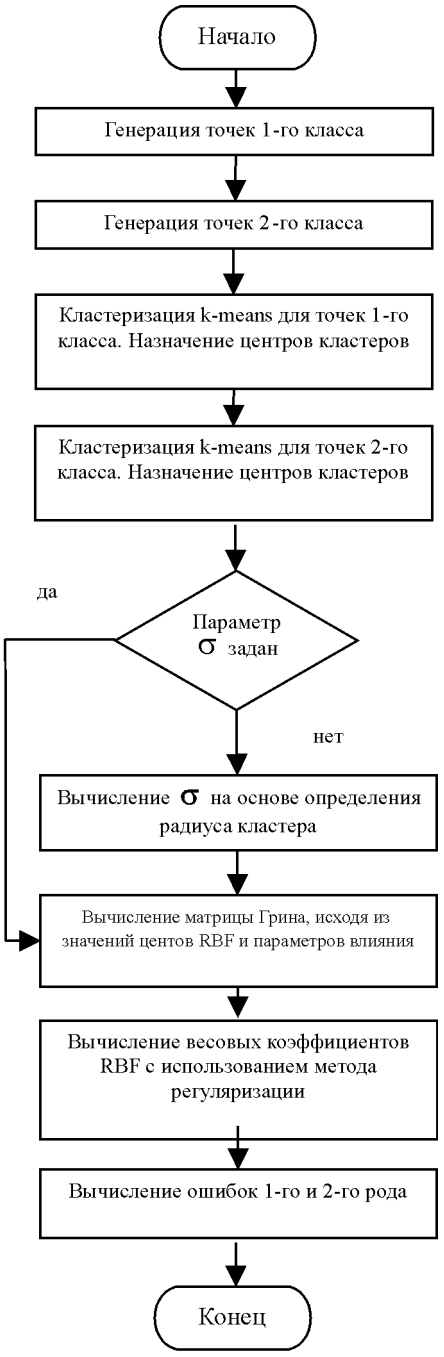


Рис. 3. Блок-схема алгоритма обучения классификатора

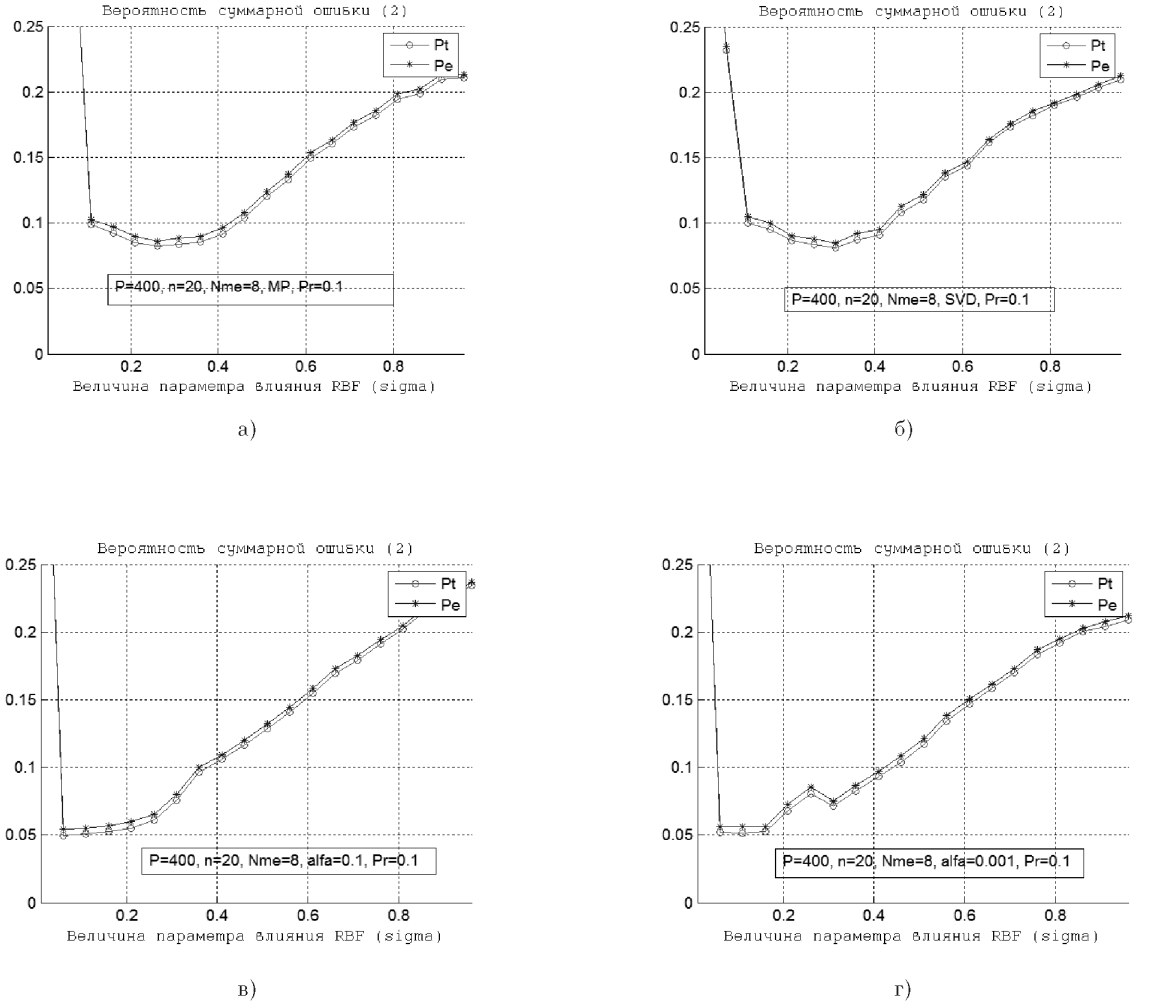


Рис. 4. Зависимости для вероятности ошибки классификации от величины параметра влияния

даны для случая использования адаптивного способа назначения параметра влияния индивидуально для каждого кластера. Из рис. 5 а,б следует, что метод регуляризации в данном случае позволяет получить ощутимый выигрыш с точки зрения вероятности ошибки при классификации данных. При реализации адаптивного способа задания параметра влияния вероятность ошибки больше почти в два раза и мало зависит от выбранного метода решения СЛАУ.

В целом на основании полученных в ходе исследования результатов можно сделать вывод, что наиболее целесообразным для работы алгоритма является использование количества RBF, обеспечивающее отображение 4...8 точек в кластере. При этом общее классифицируемое количество элементов может иметь порядок 10^2...10^3, а размерность признакового пространства порядка 10...20. Предложенная модель данных достаточно полно представляет основные закономерности классификации образов в многомерном пространстве и может быть использована при определении параметров алгоритмов создания контентно-зависимых цифровых водяных знаков.

Работа выполнена при поддержке гранта Минобрнауки РФ по программе «Развитие кооперации российских вузов и производственных предприятий» (Постановление Правительства № 218 от 09.04.2010 г. – 3 очередь, № 02. G25.31.0002)

СПИСОК ЛИТЕРАТУРЫ

1. *Сирота А.А.* Нейросетевые модели и алгоритмы стеганографического скрытия информации / А. А. Сирота, М. А. Дрюченко // Труды Российского научно-технического общества радиотехники, электроники и связи имени А. С. Попова. – Москва, 2010. – Т. 2. – С. 335-338.

2. *Дрюченко М.А.* Математическое и программное обеспечение для реализации новых технологий со-

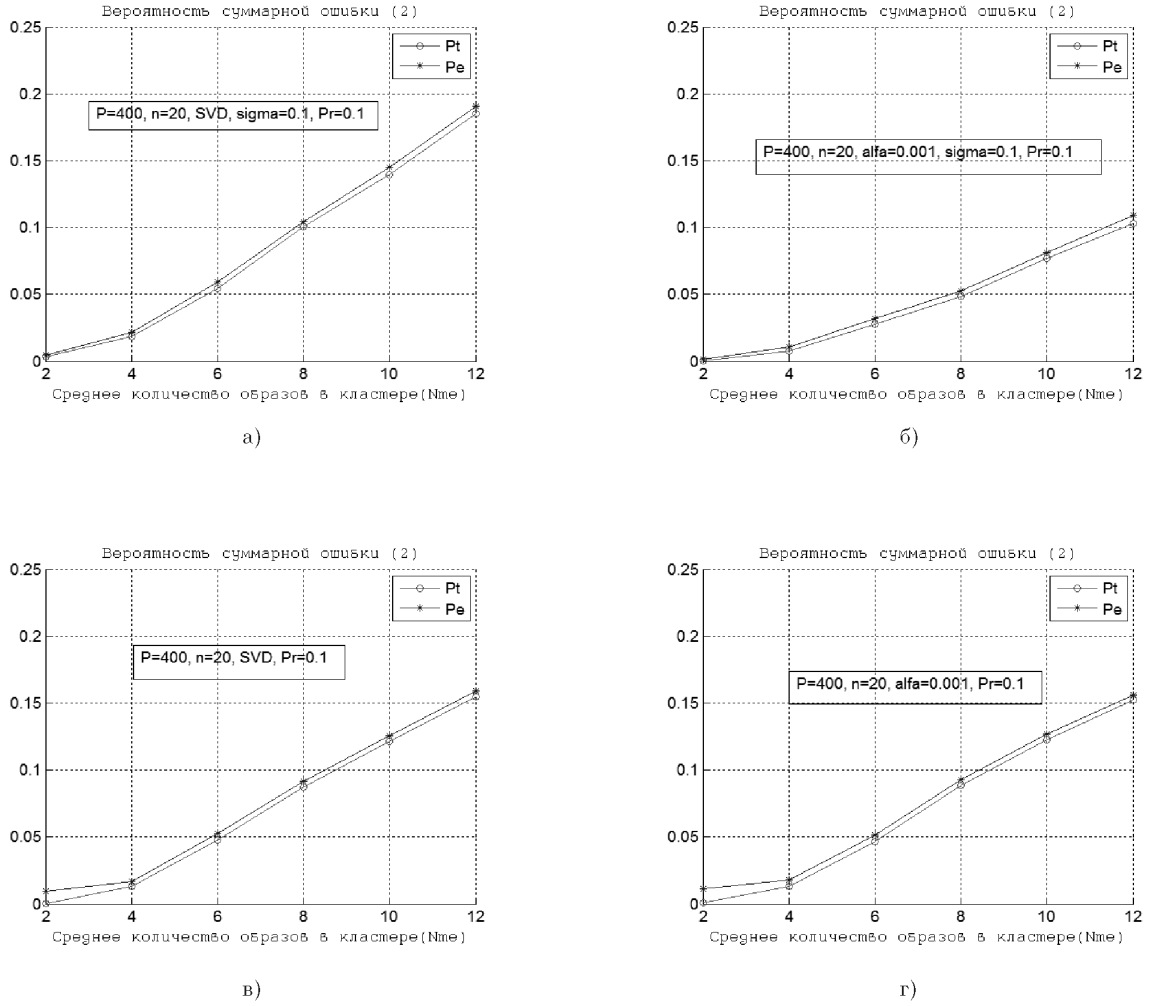


Рис. 5. Зависимости для вероятности ошибки классификации от среднего количества образов в кластере

здания цифровых водяных знаков в интересах защиты авторских прав на цифровые объекты интеллектуальной собственности / М. А. Дрюченко // Инновационные технологии на базе фундаментальных научных разработок - прорыв в будущее : сборник докладов научной конференции студентов, аспирантов и молодых ученых, 16-17 апреля 2012 г. — Воронеж, 2012. — С. 46-48.

3. *Алгазинов Э.К.* Математическое и программное обеспечение для создания цифровых водяных знаков с использованием искусственных нейронных сетей / Э. К. Алгазинов, М. А. Дрюченко, Е. Ю. Митрофанова, А. А. Сирота // Информационные технологии. — 2012. — № 9. — С. 60—66.

4. *Осовский С.* Нейронные сети для обработки информации / пер с польского И. Д. Рудинского. – М.: Финансы и статистика, 2002. – 344 с.

5. *Хайкин С.* Нейронные сети: полный курс, 2-е изд., испр. – М.: ООО “И.Д. Вильямс”, 2006. – 1104 с.

6. *Rajaraman A.* Mining of Massive Datasets / Anand Rajaraman, Jeffrey D. Ullman // Cambridge: Cambridge University Press, 2012. – 326 с.

Сирота Александр Анатольевич – доктор технических наук, профессор, профессор кафедры информационных систем ВГУ. E-mail: sir@cs.vsu.ru

Цуриков Андрей Владимирович – аспирант кафедры информационных систем ВГУ. E-mail: andrew.tsurikov@gmail.com

Sirota A. A. – Doctor of Technical Sciencies, Professor of Information Systems Department, Voronezh State University. E-mail: sir@cs.vsu.ru

Tsurikov A. V. – PhD student of Information Systems Department, Voronezh State University. E-mail: andrew.tsurikov@gmail.com.