p-ISSN 1607-3274. Радіоелектроніка, інформатика, управління. 2015. № 1
e-ISSN 2313-688X. Radio Electronics, Computer Science, Control. 2015. № 1

НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

УДК 004.912:004.8

Бодянский Е. В.[1], Рябова Н. В.[2], Золотухин О. В.[3]

[1]Д-р техн. наук, профессор, профессор кафедры искусственного интеллекта Харьковского национального университета радиоэлектроники, Харьков, Украина
[2]Канд. техн. наук, доцент, и.о. зав. кафедрой искусственного интеллекта Харьковского национального университета радиоэлектроники, Харьков, Украина
[3]Ассистент кафедры искусственного интеллекта Харьковского национального университета радиоэлектроники, Харьков, Украина

МНОГОСЛОЙНАЯ АДАПТИВНАЯ НЕЧЕТКАЯ ВЕРОЯТНОСТНАЯ НЕЙРОННАЯ СЕТЬ В ЗАДАЧАХ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Рассмотрена задача классификации текстовых документов на основе нечеткой вероятностной нейронной сети в режиме реального времени. В массиве текстовых документов может быть выделено различное количество классов, к которым могут относиться данные документы. При этом предполагается, что данные классы могут иметь в n-мерном пространстве различную форму и взаимно перекрываться. Предложена архитектура многослойной адаптивной нечеткой вероятностной нейронной сети, которая позволяет решать задачу классификации в последовательном режиме по мере поступления новых данных. Предложен алгоритм обучения многослойной адаптивной нечеткой вероятностной нейронной сети, а также решена задача классификации на основе предложенной архитектуры в условиях пересекающихся классов, что позволяет относить один экземпляр текстового документа к разным классам с различной степенью вероятности. Архитектура классифицирующей нейронной сети отличается простотой численной реализацией и высокой скоростью обучения, и предназначена для обработки больших массивов данных, характеризующихся векторами признаков высокой размерности. Предлагаемая нейронная сеть и метод ее обучения предназначены для работы в условиях пересекающихся классов, отличающихся как формой, так и размерами.

Ключевые слова: классификация, адаптивная нечеткая вероятностная нейронная сеть, пересекающиеся классы, нейроны в точках данных.

НОМЕНКЛАТУРА

AFPNN – Adaptive Fuzzy Probabilistic Neural Network;
EPNN – Enhanced Probabilistic Neural Network;
FLVQ – Fuzzy LVQ;
FPNN – Fuzzy Probabilistic Neural Network;
PNN – Probabilistic Neural Network;
WTA – Winner Take All;
c_j – среднее арифметическое;
$\tilde{N}$ – количество нейронов в скрытом слое;
$w_l(N)$ – обучающая выборка;
η – параметр шага обучения;
D – евклидово расстояние (метрика);
j – нейрон-победитель;
j^* – индекс прототипа-победителя;
k – номер наблюдения;
l – ширина активационной функции;
m – число возможных классов;
N – объем обучающей выборки;
n – арность векторов;
N_X – количество векторов, относящихся к классу X;
$o_l^{[i]}$ – выходной сигнал скрытого слоя;
P – априорная вероятность;
p – относительная частота появления образов;
$p(x)$ – функция плотности вероятностей;
q^* – нейрон-победитель внутри блока;
u_p – уровень принадлежности;
w_{li} – синаптический вес;
w_l^T – транспонированный вектор синаптических весов;
$x(k, j)$ – сигнал вектора-образа, который участвовал в классификации;

DOI 10.15588/1607-3274-2015-1-5

$x_i(k, j)$ – сигнал вектора-образа, с известной классификацией;

$x_i(l)$ – синаптические веса в скалярной форме;

$x(k)$ – n-мерный вектор признаков с номером классифицируемого образа k;

σ – параметр ширины активационной функции.

ВВЕДЕНИЕ

На сегодняшний день классификация текста считается достаточно сложной проблемой. Классификация текста является деятельностью, которая становится все более значимой в наши дни. Это обусловлено огромным объемом доступной информации и проблемой поиска информации. К тому же большинство используемых баз данных являются политематическими с большим количеством категорий, которые превращают задачу классификации текста в более сложную.

Возникли новые проблемы, среди которых наиболее острой является информационная перегруженность и, как следствие, необходимость классификации последовательно поступающих документов в режиме реального времени. Эта задача весьма актуальна, например, для информационных агентств, разнообразных Интернет-издательств, которые должны постоянно классифицировать поток поступающих текстовых документов, в том числе новостных сообщений, аналитических обзоров, дайджестов, статей, докладов и т.п. При этом документы, подлежащие классификации, как правило, характеризуются разнородностью (политематичностью), т.е. затрагивают сразу несколько тем, как весьма различных, так и очень близких.

On-line классификация такого рода текстовых документов не является тривиальной задачей, поскольку в небольшом фрагменте текста может содержаться весьма ценная информация, и отнесение к соответствующему классу нельзя игнорировать, а близко расположенные классы могут пересекаться и/или сливаться. Поэтому желательно учесть принадлежность анализируемого документа к каждому из потенциально интересующих пользователя классов.

В то же время большинство известных методов классификации относят текстовый документ к одному из четко различимых классов. Отсутствие возможности получить наиболее актуальную и полную информацию по конкретной теме делает бесполезной большую часть накопленных ресурсов. Поскольку исследование конкретной задачи требует все больших трудозатрат на непосредственный поиск и анализ информации по теме, многие решения принимаются на основе неполного представления о проблеме.

1 ПОСТАНОВКА ЗАДАЧИ

Пусть задан массив, содержащий N текстовых документов, описываемых n-мерными векторами-признаками, при этом часть документов является классифицируемыми, а часть нет. Предполагается также априорно, что в массиве может быть выделено m различных классов, к которым могут относиться данные документы. При этом предполагается также, что данные классы могут иметь в m-мерном пространстве различную форму и взаимно перекрываться. Необходимо создать классифицирующую нейро-фаззи систему, которая позволит производить простой и эффективный метод классификации при условии взаимно перекрывающихся классов и предложить архитектуру классифицирующей нечеткой вероятностной сети, которая позволит разбивать подающие на обработку документы как с точки зрения Байесовской, так и нечеткой классификации одновременно. Сеть должна быть простой в реализации и пригодной для обработки поступающих наблюдений в последовательном online режиме.

2 ОБЗОР ЛИТЕРАТУРЫ

Достаточно эффективным средством для решения задачи классификации текстовых документов являются вероятностные нейронные сети, введенные Д. Ф. Шпехтом [1], обучение которых производится по принципу «нейроны в точках данных», что делает его крайне простым и быстрым. В [2–4] были введены модификации PNN, предназначенные для обработки текстовой информации и отличающиеся наличием элементов конкуренции в процессе обучения и возможностью коррекции рецепторных полей ядерных активационных функций. В [5–7] были введены нечеткие модификации вероятностных сетей, в том числе и для обработки текстов [8], позволяющие решать задачу классификации в условиях пересекающихся классов. Вместе с тем, использование PNN и FPNN в задачах обработки текстов усложняется в случаях, когда объемы анализируемой информации велики, а векторы признаков (образы) имеют достаточно высокую размерность. Это затруднение объясняется тем, что как в PNN, так и в других нейронных сетях, обучаемых по принципу «нейроны в точках данных» [9], количество нейронов первого скрытого слоя (слоя образов) определяется числом векторов-образов обучающей выборки N, что, естественно, приводит к снижению быстродействия и требует хранения всех данных, использованных в процессе обучения сети, что естественно затрудняет работу в on-line режиме. Для преодоления этого недостатка в [10] была предложена улучшенная вероятностная нейронная сеть, где первый скрытый слой образован не образами, а прототипами классов, вычисленных с помощью обычного K-среднего (HCM) в пакетном режиме. Поскольку в задачах классификации число возможных классов m обычно существенно меньше объема обучающей выборки N, EPNN гораздо лучше приспособлена для решения реальных задач, чем стандартная PNN.

Вместе с тем, можно отметить такие основные недостатки EPNN, как возможность обучения только в пакетном режиме, когда обучающая выборка задана заранее, и четкий результат классификации (отнесение предъявляемого образа только к одному классу), в то время как при обработке текстовых документов достаточно часто возникает ситуация, когда анализируемый текст с различными уровнями принадлежности может одновременно относиться сразу к нескольким, возможно пересекающимся классам. В связи с этим в [11] была предложена нечеткая вероятностная сеть, где в первом скрытом слое производится адаптивное уточнение прототипов с помощью WTA-правила обучения Т. Кохонена [12], а выходной слой оценивает уровни принадлежности посту-

пающих на обработку образов к тем или иным классам с помощью процедуры нечетких *C*-средних (FCM) [13]. Такая сеть содержит минимально возможное количество нейронов, равное числу классов и потому характеризуется высоким быстродействием. Вместе с тем же сеть не учитывает ни размеры классов, ни частоту появления образов в каждом из этих классов, что естественно ограничивает ее возможности при обработке данных, чьи прототипы удалены друг от друга на различное расстояние, которые к тому же может изменяться с течением времени.

3 МАТЕРИАЛЫ И МЕТОДЫ

Классические вероятностные нейронные сети Д. Ф. Шпехта предназначены для решения задач байесовской классификации (распознавания образов на основе байесовского подхода), состоящего в том, что класс с наиболее плотным распределением в области неклассифицированного предъявляемого образа $x(k)$ будет иметь преимущество по сравнению с другими классами. Также будет иметь преимущество и класс с высокой априорной вероятностью. Так, для трех возможных классов A, B и C в соответствии с байесовским правилом выбирается класс A, если $P_A p_A(x) > P_B p_B(x)\ AND\ P_A p_A(x) > P_C p_C(x)$.

Стандартная PNN состоит из входного (рецепторного) слоя, первого скрытого, именуемого слоем образов, второго скрытого, называемого слоем суммирования, и выходного слоя, образованного компаратором, выделяющим максимальное значение на выходе второго скрытого слоя.

Исходной информацией для синтеза сети является обучающая выборка образов, образованная «пакетом» n-мерных векторов $x(1), x(2), ..., x(k), ..., x(N)$ с известной классификацией. Предполагается также, что N_A векторов относятся к классу A, N_B к классу B и N_C к классу C, т.е. $N_A + N_B + N_C = N$, а априорные вероятности могут быть рассчитаны с помощью элементарных соотношений:

$$P_A = \frac{N_A}{N}, P_B = \frac{N_B}{N}, P_C = \frac{N_C}{N}, P_A + P_B + P_C = 1.$$

Количество нейронов в слое образов сети Шпехта равно N (по одному нейрону на каждый образ), а их синаптические веса определяются значениями компонент этих образов по принципу «нейроны в точках данных» так, что $w_{li} = x_i(l), i = 1, 2, ..., n; l = 1, 2, ..., N$, или в векторной форме $w_l = x(l) = (x_1(l), x_2(l), ..., x_n(l))^T$.

Очевидно, что обучение в данном случае сводится к одноразовой установке весов, что делает его крайне простым.

Каждый из нейронов слоя образов имеет колоколообразную функцию активации, с помощью которой предъявляемый сети сигнал $x(k)$ преобразуется в скалярный выход нейрона $o_l^{[i]}(k) = \Phi(\|x(k) - w_l\|, \sigma)$ чаще всего на основе гауссиана

$$o_l^{[i]}(k) = \exp\left(-\frac{\|x(k) - w_l\|^2}{2\sigma^2}\right).$$

В [11] было показано, что в задачах нечеткой классификации более естественно использовать распределение Коши в виде

$$o_l^{[i]}(k) = \frac{1}{1 + \frac{\|x(k) - w_l\|^2}{2\sigma^2}},$$

где параметр σ задает ширину, $l = 1(A), 2(A), ..., N_A(A), N_A + 1(B), ..., N_A + N_B(B), N_A + N_B + 1(C), ..., N(C)$.

Для упрощения численной реализации входные векторы рекомендуется предварительно нормировать на гиперсферу [12] так, что $\|x(k)\| = \|w_l\| = 1$, что ведет к более простой форме активационной функции

$$o_l^{[i]}(k) = \frac{\sigma^2}{\sigma^2 + (1 + w_l^T x(k))}.$$

Слой суммирования образован обычными сумматорами, число которых равно числу классов (в рассматриваемом случае – три), которые просто суммируют выходы нейронов слоя образов, а выходной компаратор выделяет класс с максимальным выходным сигналом второго слоя.

Поскольку при работе с текстовыми документами N может быть велико, работа в online-режиме с помощью стандартной PNN весьма затруднительна. Именно по этой причине в [10] была введена крайне простая архитектура, число нейронов в которой равно числу классов (в нашем примере три), а классификация производится с помощью оценки расстояния до прототипов классов, вычисленных с помощью среднего арифметического

$$c_j = \frac{1}{N_j} \sum_{k=1}^{N_j} x(k, j),\ j = 1, 2, ..., m,$$

в нашем случае $m = 3, j = 1$ соответствует классу A, $j = 2 - B$ и $j = 3 - C$.

Понятно, что такая элементарная схема не способна оценить ни размеры классов, ни их взаимное перекрытие.

Для устранения указанных недостатков и предлагается многослойная адаптивная нечеткая вероятностная нейронная сеть, архитектура которой приведена на рис. 1.

Первый скрытый слой содержит m однотипных блоков (на рис.1 – A, B и C) по числу возможных классов, которое может изменяться в процессе online-обучения. Каждый из блоков содержит одинаковое число нейронов $\tilde{N} + 1\ (\tilde{N}_A = \tilde{N}_B = \tilde{N}_c = \tilde{N})$, при этом в каждом блоке $\tilde{N}$ нейронов (в нашем примере 3) обучаются по принципу «нейроны в точках даннных», а один нейрон $c_j(c_A, c_B, c_C)$ вычисляет прототипы классов. В каждом блоке между отдельными нейронами и между блоками в целом по внутриблочным и межблочным латеральным связям организуется процесс «конкуренции» по Кохонену, позволяющий оценить как центроиды (прототипы) классов, так и их размеры. Второй скрытый слой

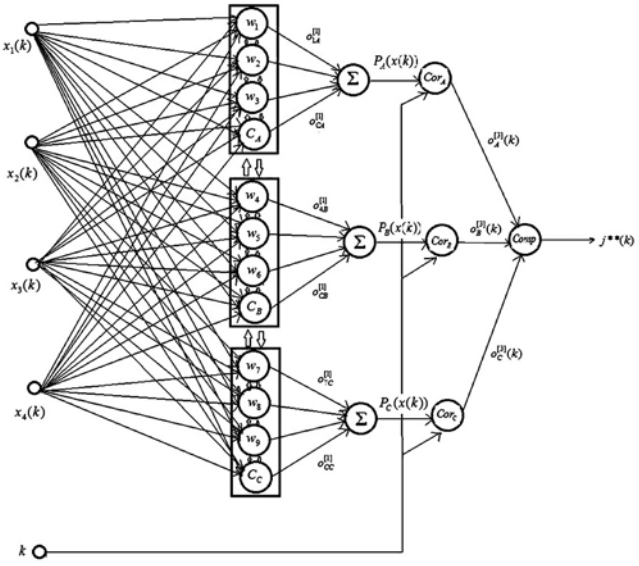

Рисунок 1 – Многослойная адаптивная нечеткая вероятностная сеть

сумматоров аналогичен слою в сети Шпехта, в третьем скрытом слое коррекции априорных вероятностей подсчитываются частоты появления образов в каждом из классов, а выходной слой-компаратор реализует собственно классификацию предъявленного образа.

Процесс обучения сети начинается с установки начальных синаптических весов всех нейронов. Для архитектуры, приведенной на рис. 1, необходимо иметь девять $(\tilde{N}\cdot m)$ классифицированных образов по три на каждый класс A, B и C.

Так, например,

$$x(1,A)=w_1(0), x(2,A)=w_2(0), x(3,A)=w_3(0),$$

$$x(4,\mathrm{B})=w_4(0), x(5,\mathrm{B})=w_5(0), x(6,\mathrm{B})=w_6(0),$$

$$x(7,\mathrm{C})=w_7(0), x(8,\mathrm{C})=w_8(0), x(9,\mathrm{C})=w_9(0),$$

$$c_A(0)=\frac{1}{3}\sum_{k=1}^{3}x(k,A), c_B(0)=\frac{1}{3}\sum_{k=4}^{6}x(k,\mathrm{B}), c_C(0)=\frac{1}{3}\sum_{k=7}^{9}x(k,\mathrm{C}).$$

Далее векторы-образы, участвовавшие в формировании начальных условий, не используются и все последующие сигналы будут обозначаться $x(k,j)$, если они относятся к обучающей выборке $x(k)$, если они подлежали классификации.

Итак, пусть на вход сети подается первый образ $x(1,j)$, относительно которого известна его принадлежность к конкретному классу A или B или C. В результате межблочной конкуренции определяется прототип победитель j^* (при этом j не обязательно равно j^*), вектор параметров которого $c_{j^*}(0)$ в смысле принятой метрики (обычно евклидовой) наиболее близок к входному сигналу $x(1,j)$, т. е.

$$j^*=\arg\min_p\left(D(x(1,j),\mathrm{c}_p(0))\right)=\arg\min_p\left\|x(1,j)-c_p(0)\right\|^2=$$

$$=\arg\max_p x^T(1,j)\,\mathrm{c}_p(0)=\arg\max_p\cos(x(1,j),$$

$$c_p(0))\ \forall\ p=1,2,...,m,$$

при этом очевидно, что $-1\le\cos(x(1,j),c_p(0))=x^T(1,j)c_p(0)\le 1$ и $0\le\left\|x(1,j)-c_p(0)\right\|^2\le 4$.

При этом возможно возникновение двух взаимоисключающих ситуаций:

– входной вектор $x(1,j)$ и прототип-победитель $c_{j^*}(0)$ принадлежат одному классу, т. е. $j=j^*$;

– входной вектор $x(1,j)$ и победитель $c_{j^*}(0)$ принадлежат разным классам, т. е. $j=j^*$.

Далее производится настройка параметров нейронов и прототипов с помощью нечеткого LVQ правила обучения [14]

$$c_j(1)=\begin{cases}c_{j^*}(0)+\eta(1)(x(1,j)-c_{j^*}(0)), \text{если } j=j^*;\\ c_{j^*}(0)-\eta(1)(x(1,j)-c_{j^*}(0)), \text{если } j\ne j^*;\\ c_j(0), \text{если } j\text{-й нейрон не победил},\end{cases}$$

здесь $0<\eta(1)<1$ – параметр шага обучения, выбираемый обычно из эмпирических соображений.

Далее в случае пересекающихся классов несложно определить уровень принадлежности образа $x(1,j)$ к каждому из m имеющихся классов в виде [14]

$$u_p(1)=\frac{\left\|x(1,j)-c_p(1)\right\|^{-2}}{\sum_{l=1}^{m}\left\|x(1,j)-c_l(1)\right\|^{-2}},\ p=1,2,...,m.$$

На этом этап межблочной конкуренции заканчивается.

На этапе внутриблочной конкуренции в блоке, соответствующем классу j рассчитываются расстояния $D(c_j(1),w_q(0))=\left\|c_j(1)-w_q(0)\right\|^2$, где q пробегает все номера нейронов, соответствующих j-ому классу. Далее внутри j-го блока рассчитывается свой победитель $q^*(0)=\arg\min_q(D(c_j(1),w_q(0))$, ближайший к прототипу j-го класса и в случае, если выполняется условие $D(\mathrm{c}_j(1),w_{j^*}(0))<D(\mathrm{c}_j(1),x(1,j))$, вектор центра q^*-й функции активации заменяется на $x(1,j)$, увеличивая тем самым размеры класса, т.е. $w_{q^*}(1)=x(1,j)$. В противном случае все $w_q(0)$ остаются неизменными, увеличивая на единицу только свой индекс так, что $w_q(1)=w_q(0)$.

Таким образом, в процессе обучения по принципу «нейроны в точках данных» включаются только наблюдения, далеко отстоящие от текущего значения прототипа.

Пусть к моменту поступления k-го наблюдения обучающей выборки сформированы все прототипы $c_j(k-1)$ и векторы параметров нейронов $w_l(k-1)$ общим числом $\tilde{N}m$. Тогда процесс обучения первого скрытого слоя может быть записан в виде следующей последовательности шагов:

– поступление на вход сети вектора-образа $x(k,j)$ с известной классификацией;

– определение прототипа-победителя $c_{j^*}(k-1)$ такого, что

$$j^*(k-1) = \arg\min_p (D(x(k,j), c_p(k-1))),\ p = 1,2,...,m;$$

– настройка параметра прототипа-победителя так, что

$$c_j(k) = \begin{cases} c_{j^*}(k-1) + \eta(k)(x(k,j) - c_{j^*}(k-1)), \text{если } j = j^*; \\ c_{j^*}(k-1) - \eta(k)(x(k,j) - c_{j^*}(k-1)), \text{если } j \neq j^*; \\ c_j(k-1), \text{если } j\text{-й нейрон не победил}, \end{cases}$$

– расчет уровней принадлежности

$$u_p(k) = \frac{\left\| x(k,j) - c_p(k) \right\|^{-2}}{\sum_{l=1}^{m} \left\| x(k,j) - c_l(k) \right\|^{-2}},\ p = 1,2,...,m;$$

– расчет внутриблочных расстояний в j-м классе $D(c_j(k), w_q(k-1))$, где q-все индексы нейронов j-блока;

– определение внутриблочного победителя $w_{q^*}(k-1)$ такого, что $q^*(k-1) = \arg\min_q (D(c_j(k), w_q(k-1))$,

– при выполнении условия $D(c_j(k), w_{j^*}(k-1)) < D(c_j(k), x(k,j))$ производится замена $w_{q^*}(k) = x(k,j)\ u\ w_q(k) = w_q(k-1)$.

Процесс обучения этого слоя производится до исчерпания обучающей выборки, т. е. завершается вычислением всех $c_j(N)$ и всех $\tilde{N}m$ весов $w_l(N)$.

Одновременно с этим в третьем скрытом слое происходит процесс подсчета относительных частот появления образов из разных классов

$$p_j = \frac{N_j}{N}.$$

На этом процесс обучения многослойной адаптивной нечеткой вероятностной нейронной сети завершается.

Пусть на вход обученной сети поступает некий неклассифицированный образ $x(k), k > N$. Этот сигнал поступает на все $\tilde{N}m$ нейронов сети, на выходах которых появляются значения

$$o_l^{[i]}(k) = \frac{\sigma^2}{\sigma^2 + (1 + w_l^T(N)x(k))}.$$

Здесь же в первом скрытом слое вычисляются уровни принадлежностей к каждому из возможных классов

$$u_j(k) = \frac{\left\| x(k) - c_j(N) \right\|^{-2}}{\sum_{l=1}^{m} \left\| x(k) - c_l(N) \right\|^{-2}}.$$

В принципе, можно говорить о принадлежности $x(k)$ к конкретному классу по максимальному значению принадлежности, однако в этом случае речь идет только о нечеткой классификации [14], а не байесовской.

Далее сумматоры второго скрытого слоя вычисляют плотности вероятностей

$$o_l^{[2]}(k) = \sum_q o_q^{[1]}(k),\ j = 1,2,...,m,$$

а q пробегает все номера нейронов (всего $\tilde{N}+1$), соответствующих j-му классу.

В третьем скрытом слое вычисляются произведение

$$o_l^{[3]}(k) = \frac{N_j}{N} o_j^{[2]}(k) = P_j p_j(x(k))$$

и, наконец, компаратор выходного слоя вычисляет класс победитель $j^{**}(k)$, которому с наибольшей вероятностью принадлежит предъявленный образ $x(k)$.

При предъявлении последующих образов $x(k+1), x(k+2),...$ классификация происходит аналогично предыдущему образу.

4 ЭКСПЕРИМЕНТЫ

В качестве экспериментальных данных использовалась выборка «20 Newsgroups», которая представляет собой набор из примерно 20000 новостных документов, разделенных на 20 различных групп. Этот текстовый корпус стал популярным набором данных для экспериментов в области интеллектуальной обработки текстовой информации. Одной из отличительных особенностей этой коллекции является значительный разброс в размерах документов, что осложняет задачу обработки информации. Исходная выборка данных была разделена на обучающую и тестирующую (60% и 40% соответственно).

Для эксперимента было выбрано 150 документов из различных категорий. После предварительной обработки было получено 61118 терминов для формирования вектора признаков для работы вероятностных нейронных сетей. Для оценки качества классификации использовались внешние меры полноты и точности).

5 РЕЗУЛЬТАТЫ

Результаты сравнения производительности простой FPNN и AFPNN для одинакового количества признаков представлены в табл. 1.

В ходе эксперимента рассматривалось, прежде всего, качество работы AFPNN. В табл. 2 представлен результат работы метода для значения параметра ширины активационной функции $\sigma = 0,05$. Показано, что в результате работы формируется набор значений вероятностей принадлежности входного текстового объекта к нескольким классам.

6 ОБСУЖДЕНИЕ

Предложенная авторами нейро-фаззи сеть позволяет решать задачу с точки зрения как нечеткой, так и вероятностной классификации, что обеспечивает ей преимущество по сравнению с классическими Байесовскими сетями и вероятностными нейронными сетями, все из которых не могут решать задачу в условиях перекрывающихся классов. Становится возможным определить более точные значения вероятностей принадлежности входящего текстового объекта к каждому из потенциально возможных классов. Данный метод предусматривает возможность обработки информации по мере ее поступления, характеризуется простотой реализации и высокой скоростью обработки информации.

Таблица 1 – Сравнительная характеристика качества классификации с использованием стандартной FPNN и AFPNN

Название класса	Количество документов	Точность		Отзыв	
		FPNN	AFPNN	FPNN	AFPNN
comp.graphics	100	0,73	0,81	0,78	0,83
comp.os.ms-windows.misc	80	0,65	0,69	0,71	0,79
comp.sys.ibm.pc.hardware	70	0,63	0,67	0,69	0,76
comp.sys.mac.hardware	20	0,60	0,61	0,64	0,72

Таблица 2 – Пример работы классификации с параметром ширины активационной функции $\sigma = 0,05$

№ входящего текстового объекта	Вероятность принадлежности к первому классу	Вероятность принадлежности ко второму классу	Вероятность принадлежности к третьему классу
1	1	$1{,}2412\ 10^{-144}$	$1{,}0302\ 10^{-132}$
2	0,24081	$9{,}1785\ 10^{-16}$	0,75919
3	$6{,}2498\ 10^{-81}$	1	$7{,}7081\ 10^{-68}$
4	$5{,}427\ 10^{-56}$	1	$2{,}3617\ 10^{-64}$
5	$3{,}6966\ 10^{-12}$	1	$2{,}5228\ 10^{-45}$

ВЫВОДЫ

Рассмотрена задача одновременной online нечеткой и вероятностной классификации текстовых документов, поступающих на обработку последовательно в реальном времени.

Введена архитектура классифицирующей нейронной сети, отличающаяся простотой численной реализации и высокой скоростью обучения и предназначенная для обработки больших массивов данных, характеризующихся векторами признаков высокой размерности. Предлагаемая нейронная сеть и метод ее обучения предназначены для работы в условиях пересекающихся классов, отличающихся как формой, так и размерами.

БЛАГОДАРНОСТИ

Работа выполнена в рамках госбюджетной научно-исследовательской темы Харьковского национального университета радиоэлектроники №265-1 «Методы создания общей онтологической базы социально-экономической образовательно-научной сети с целью интеграции в европейское пространство» при поддержке национального проекта TRUST: Towards Trust in Quality Assurance Systems программы «Tempus» Европейской комиссии (регистрационный номер 516935-TEMPUS-1-2011-1-FITEMPUS-SMGR).

СПИСОК ЛИТЕРАТУРЫ

1. Specht D. F. Probabilistic neural networks / D. F. Specht // Neural Networks. – 1990. – Vol. 3 (1). – P. 109–118.
2. Бодянский Е. В. Семантическое аннотирование текстовых документов с использованием модифицированной вероятностной нейроной сети / Е. В. Бодянский, О. В. Шубкина // Системные технологии. – Днепропетровск, 2011. – Вып. 4 (75). – С. 48–55.
3. Bodyanskiy Ye. Semantic annotation of text documents using modified probabilistic neural network / Ye. Bodyanskiy, O. Shubkina// Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications: 6th IEEE International Conferences, Prague, 15–17 September 2011: – Prague: Czech Technical University In Prague, 2011. – P. 328–331.
4. Bodyanskiy Ye. Semantic annotation of text documents using evolving neural network based on principle «Neurons at Data Points» / Ye. Bodyanskiy, O. Shubkina // Workshop on Inductive Modelling «IWIM 2011» : 4th Interational Conference, Zhukyn-Kyiv, 4–10 July 2011: Kyiv: IRTC ITS, 2011. – P. 31–37.
5. Bodyanskiy Ye. A learning probabilistic neural network with fuzzy inference / Ye. Bodyanskiy, Ye. Gorshkov, V. Kolodyazhniy, J. Wernstedt // Artificial Neural Nets and Genetic Algorithms «ICANNGA 2003» : 6th International Conference, Roanne, France April 23-25 April 2003 : proceedings. – Wien : Springer-Verlag, 2003. – P. 13–17.
6. Bodyanskiy Ye. Resource-allocating probabilistic neuro-fuzzy network / Ye. Bodyanskiy, Ye. Gorshkov, V. Kolodyazhniy // European Union Society for Fuzzy Logic and Technology «EUSFLAT 2003» : 3rd Internetional Conference, Zittau : proceedings. – Zittau : University of Applied Sciences at Zittau/ Goerlitz, 2003. – P. 392-395.
7. Bodyanskiy Ye. Probabilistic neuro-fuzzy network with non-conventional activation functions / Ye. Bodyanskiy, Ye. Gorshkov, V. Kolodyazhniy, J. Wernstedt // Knowledge-Based Intelligent Information and Engineering Systems : 7th International Conference KES 2003, Oxford, 3–5 September 2003 : proceedings. – Berlin-Heidelberg-New York : Springer, 2003. – P. 973–979. – (Lecture Notes in Computer Science, Vol. 2774).
8. Бодянский Е. В. Классификация текстовых документов с помощью нечеткой вероятностной нейронной сети / Е. В. Бодянский, Н. В. Рябова, О. В. Золотухин // Восточно-европейский журнал передовых технологий – 2011. – № 6/2 (54). – С.16–18.
9. Zahirniak D. R. Pattern recognition using radial basis function network / D. R. Zahirniak, R. Chapman, S. K. Rogers, B. W. Suter, M. Kabriski, V. Pyatti // Aerospace Application of Artificial Intelligence: 6 International Conference, 5–8 June 1990 : proceedings. – Dayton : Ohio, 1990. – P. 249–260.
10. Ciarelli P. M. An enhanced probabilistic neural network approach applied to text classification / P. M. Ciarelli, E. Oliveira // Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications : 14th Iberoamerican Conference CIARP 2009, Jalisco, 15–18 November 2009 : proceedings. – Berlin-Heidelberg : Springer-Verlag, 2009. – P. 661–668. – (Lecture Notes in Computer Science, Vol. 5856).
11. Bodyanskiy Ye. Modified probabilistic neuro-fuzzy network for text documents processing / Ye. Bodyanskiy, I. Pliss, V. Volkova // International Journal Computing. – 2012. – 11. – № 4. – P. 391–396.
12. Kohonen T. Self-Organizing Maps / T. Kohonen. – Berlin: Springer, 1995. – 362 p.
13. Bezdek J. C. Convergence theory for fuzzy c-means: Counterexamples and repairs / J. C. Bezdek, R. J. Hathaway, M. J. Sabin, W. T. Tucker // IEEE Transaction on Systems, Man, and Cybernetics.–1987. – SMC-17. – № 5.– P. 873–877.
14. Bezdek J. C. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing / J. C. Bezdek, J. Keller, R. Krishnapuram, N. R. Pal // N.Y.: Springer Science + Business Media, Inc., 2005. – 776 p.

Статья поступила в редакцию 17.10.2014.
После доработки 22.12.2014.

Бодянський Є. В.[1], Рябова Н. В.[2], Золотухін О. В.[3]

[1]Д-р техн. наук, професор, професор кафедри штучного інтелекту Харківського національного університету радіоелектроніки, Харків, Україна

[2]Канд. техн. наук, доцент, в.о. зав. кафедрою штучного інтелекту Харківського національного університету радіоелектроніки, Харків, Україна

[3]Асистент кафедри штучного інтелекту Харківського національного університету радіоелектроніки, Харків, Україна

БАГАТОШАРОВА АДАПТИВНА НЕЧІТКА ЙМОВІРНІСНА НЕЙРОННА МЕРЕЖА В ЗАДАЧАХ КЛАСИФІКАЦІЇ ТЕКСТОВИХ ДОКУМЕНТІВ

Розглянуто задачу класифікації текстових документів на основі нечіткої імовірнісної нейронної мережі в режимі реального часу. У масиві текстових документів може бути виділено різну кількість класів, до яких можуть відноситися дані документи. При цьому передбачається що дані класи можуть мати в *n*-вимірному просторі різну форму і взаємно перекриватися. Запропонована архітектура багатошарової адаптивної нечіткої імовірнісної нейронної мережі, яка дозволяє вирішувати задачу класифікації в послідовному режимі по мірі надходження нових даних. Запропонований алгоритм навчання багатошарової адаптивної нечіткої імовірнісної нейронної мережі, а також вирішена задача класифікації на основі запропонованої архітектури в умовах пересічних класів, що дозволяє відносити один екземпляр текстового документа до різних класів з різним ступенем імовірності. Архітектура класифікуючої нейронної мережі відрізняється простотою чисельної реалізацією і високою швидкістю навчання, і призначена для обробки великих масивів даних, що характеризуються векторами ознак високої розмірності. Пропонована нейронна мережа і метод її навчання призначені для роботи в умовах пересічних класів, що відрізняються як формою, так і розмірами.

Ключові слова: класифікація, адаптивна нечітка імовірнісна нейронна мережа, класи, що перетинаються, нейрони в точках даних.

Bodyanskiy Ye. V.[1], Ryabova N. V.[2], Zolotukhin O. V.[3]

[1]Dr.Sc., Professor, Professor of Department of Artificial Intelligence, Kharkiv National University of Radioelectronics, Kharkiv, Ukraine

[2]PhD., Associate Professor, Acting Head of Department of Artificial Intelligence, Kharkiv National University of Radioelectronics, Kharkiv, Ukraine

[3]Assistant of the Department of Artificial Intelligence, Kharkiv National University of Radioelectronics, Kharkiv, Ukraine

MULTILAYER ADAPTIVE FUZZY PROBABILISTIC NEURAL NETWORK IN CLASSIFICATION PROBLEMS OF TEXT DOCUMENTS

The problem of text documents classification based on fuzzy probabilistic neural network in real time mode is considered. A different number of classes, which may include such documents, can be allocated in an array of text documents. It is assumed that the data classes can have an n-dimensional space of different shape and mutually overlap. The architecture of the multlayer adaptive fuzzy probabilistic neural network, which allow to solve the problem of classification in sequential mode as new data become available, is.proposed. An algorithm for training the multilayer adaptive fuzzy probabilistic neural network is proposed, and the problem of classification is solved on the basis of the proposed architecture in terms of intersecting classes, which allows to determine the belonging a single instance of a text document to different classes with varying degrees of probability. Classifying neural network architecture characterized by simple numerical implementation and high speed training, and is designed to handle large data sets, characterized by the feature vectors of high dimension. The proposed neural network and its learning method designed to work in conditions of overlapping classes, differing both the form and size.

Keywords: classification, adaptive fuzzy probabilistic neural network, overlapping classes, neurons in the data points.

REFERENCES

1. Specht D. F. Probabilistic neural networks, *Neural Networks*, 1990, Vol. 3 (1), pp. 109–118.
2. Bodyanskiy Ye. V., Shubkina O. V. Semanticheskoe annotirovanie tekstovyh dokumentov s ispol'zovaniem modificirovannoj verojatnostnoj nejronoj seti, *Sistemnye tehnologii*. Dnepropetrovsk, 2011, Vyp.4 (75), pp. 48–55.
3. Bodyanskiy Ye., Shubkina O. Semantic annotation of text documents using modified probabilistic neural network, *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications: 6th IEEE International Conferences*, Prague, 15–17 September 2011. Prague, Czech Technical University In Prague, 2011, pp. 328–331.
4. Bodyanskiy Ye., Shubkina O. Semantic annotation of text documents using evolving neural network based on principle «Neurons at Data Points», *Workshop on Inductive Modelling «IWIM 2011», 4th Interational Conference, Zhukyn-Kyiv, 4–10 July 2011*. Kyiv, IRTC ITS, 2011, pp. 31–37.
5. Bodyanskiy Ye., Gorshkov Ye., Kolodyazhniy V., Wernstedt J. A learning probabilistic neural network with fuzzy inference, *Artificial Neural Nets and Genetic Algorithms «ICANNGA 2003», 6th International Conference, Roanne, France April 23–25 April 2003, proceedings.* Wien, Springer-Verlag, 2003, pp. 13–17.
6. Bodyanskiy Ye., Gorshkov Ye., Kolodyazhniy V. Resource-allocating probabilistic neuro-fuzzy network, *European Union Society for Fuzzy Logic and Technology «EUSFLAT 2003», 3rd Internetional Conference, Zittau, proceedings*. Zittau, University of Applied Sciences at Zittau/Goerlitz, 2003, pp. 392–395.
7. Bodyanskiy Ye., Gorshkov Ye., Kolodyazhniy V., Wernstedt J. Probabilistic neuro-fuzzy network with non-conventional activation functions, *Knowledge-Based Intelligent Information and Engineering Systems, 7th International Conference KES 2003, Oxford, 3–5 September 2003, proceedings*. Berlin-Heidelberg-New York, Springer, 2003, pp. 973-979. (Lecture Notes in Computer Science, Vol. 2774)
8. Bodyanskiy Ye. V., Ryabova N. V., Zolotukhin O. V. Klassifikacija tekstovyh dokumentov s pomoshh'ju nechetkoj verojatnostnoj nejronnoj seti / Ye.V. Bodyanskiy, // *Vostochno-evropejskij zhurnal peredovyh tehnologij*, 2011, №6/2 (54), pp. 16–18
9. Zahirniak D. R., Chapman R., Rogers S. K., Suter B. W., Kabriski M., Pyatti V. Pattern recognition using radial basis function network, *Aerospace Application of Artificial Intelligence, 6 International Conference, 5–8 June 1990, proceedings*. Dayton, Ohio, 1990, pp. 249–260.
10. Ciarelli P. M., Oliveira E. An enhanced probabilistic neural network approach applied to text classification, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 14th Iberoamerican Conference CIARP 2009, Jalisco, 15–18 November 2009, proceedings.* Berlin-Heidelberg, Springer-Verlag, 2009, pp. 661–668. – (Lecture Notes in Computer Science, Vol. 5856)
11. Bodyanskiy Ye., Pliss I., Volkova V. Modified probabilistic neuro-fuzzy network for text documents processing, *International Journal Computing*, 2012, 11, No.4, pp. 391–396.
12. Kohonen T. Self-Organizing Maps. Berlin, Springer, 1995, 362 p.
13. Bezdek J. C., Hathaway R.J., Sabin M. J., Tucker W. T. Convergence theory for fuzzy c-means: Counterexamples and repairs, *IEEE Transaction on Systems, Man, and Cybernetics*, 1987, SMC-17, No. 5, pp. 873–877.
14. Bezdek J. C., Keller J., Krishnapuram R., Pal N. R. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. N.Y, Springer Science + Business Media, Inc., 2005, 776 p.