文章编号:1001 - 9081(2006)10 - 2437 - 03

基于模糊角分类的神经网络用户兴趣模型分类算法

王秀丽[1],罗方芳[2],宁正元[1]
(1. 福建农林大学 计算机与信息学院,福建 福州 350002;
2. 福州大学 数学与计算机学院,福建 福州 350002)
(nzyfn@126.com)

摘 要:用户兴趣描述文件的快速分类是个性化搜索引擎的关键技术,提出了一种模糊角分类神经网络模型,该模型能接受用户兴趣描述文件的实向量输入,克服了角分类神经网络(CC4)对二进制输入的要求。模糊角分类神经网络模型根据用户信息所落入的 k 最近邻的样本泛化空间来进行分类,随着 k 值的增大,其分类效果趋近于贝叶斯分类算法。

关键词:模糊角分类;神经网络;分类算法
中图分类号: TP18 **文献标识码**:A

User interest profile classification algorithm based on FCC neural network

WANG Xiu-li[1], LUO Fang-fang[2], NING Zheng-yuan[1]
(1. *College of Computer and Information Technology, Agriculture and Forestry University of Fujian, Fuzhou Fujian* 350002, *China*;
2. *College of Mathematical and Computer, Fuzhou University, Fuzhou Fujian* 350002, *China*)

Abstract: Fast classification of user interest profile is a key technology for personalization search engine. A new kind of FCC neural network model was presented in this paper. It does not need the binary system input that is usually required by CC4, because it can accept the real vectors input. FCC neural network model works according to the k-nearest neighbor samples' generalization space which users´ information falls into. With the increasing of value k, the classification effect becomes close to that of Bayes classification algorithm.

Key words: Fuzzy Corner Classification(FCC); neural network; classification algorithm

0 引言

在当前主流的搜索引擎和未来一代搜索引擎的设计中,信息推荐是一个重要环节,信息检索的个性化是下一代搜索引擎的重要特征[1]。个性化信息推荐的实质是根据用户的特征判断用户所属类别并把相应类别的信息推荐给用户。对于信息的个性化推荐,一个最基本的要求是个性化推荐实现的时间应尽可能的减少。用户兴趣描述文件是根据用户的注册信息以及平时的网络行为(如:收藏、浏览、下载、删除等)动态变化的[2,3];相应的,用户所属的类别也不是一成不变的,会根据用户的兴趣爱好变更实时变化。根据用户的描述文件判定用户所属类别所花费的时间不能太长,否则无法有效地实现信息推荐。

目前,快速分类方法主要采用关联规则和神经网络技术[4]。基于关联规则的技术以频繁项目集或关联规则为分类模型的基础。但是,从大量数据中获取关联规则的时间复杂度较高,分类模型的建立与用户的检索过程必须分离。相比较而言,神经网络技术在这方面有着显著的优势。CC4 算法是角分类神经网络的代表,它以二进制向量为输入在对用户信息分类的处理中存在着严重的不足[5]。CC4 神经网络在对用户信息进行快速分类时,考虑到了用户信息之间的泛化距离 g。为获得用户信息之间的泛化距离 g,每个用户信息需要被表达成 0/1 向量[6~8]。用户信息向量的每个分量为介于[0,1]之间的实数值才能客观地描述用户的兴趣爱好,若按照 CC4 神经网络的要求过于简单地表达成 0/1 向量,将会对用户信息的正确分类产生严重的负面影响,降低了分类的准确率。为了提高分类的准确性,需要使 CC4 网络能够接受分量为介于[0,1]之间实数的向量的输入。

针对 CC4 存在的不足,本文提出了一种以实向量作为输入的神经网络模型以解决 CC4 神经网络输入的要求,称为模糊角分类(Fuzzy Corner Classification,FCC))神经网络。模糊角分类神经网络根据用户信息所落入的 k 最近邻的样本泛化空间来进行分类。随着 k 值的增大,其分类效果趋近于贝叶斯分类算法。

1 模糊角分类神经网络

1.1 FCC 神经网络结构

FCC 网络的结构如图 1 所示[9]。它是一个由输入层、隐层、规则库、输出层构成的全连接网络。输出层神经元的个数视具体问题而定。为了简便,图中仅给出了一个输出神经元,网络可以很容易地扩展成多输出神经元的网络[10]。

输入数据为[0,1]之间的数。输入向量 $\boldsymbol{X} = (x_1, x_2, \cdots, x_N)$,$N$ 的大小由具体问题决定。和 CC4 神经网络一样,隐层神经元数 H 等于训练样本的数目。

每个隐层神经元 $i(i = 1,2,\cdots,N)$ 的权重记为向量 $\boldsymbol{w}_i = (w_{i1}, w_{i2}, \cdots, w_{iN})$。$w_{ij}$ 代表着输入向量中的分量 x_j 到第 i 个隐

收稿日期:2005 - 12 - 30;**修订日期**:2006 - 01 - 18

作者简介:王秀丽(1963 -),女,汉族,河北保定人,副教授,主要研究方向:智能技术、文本挖掘; 罗方芳(1982 -),男,福建将乐人,硕士,主要研究方向:智能技术; 宁正元(1957 -),男,陕西武功人,教授,主要研究方向:智能计算、算法分析.

层神经元的连接权值。

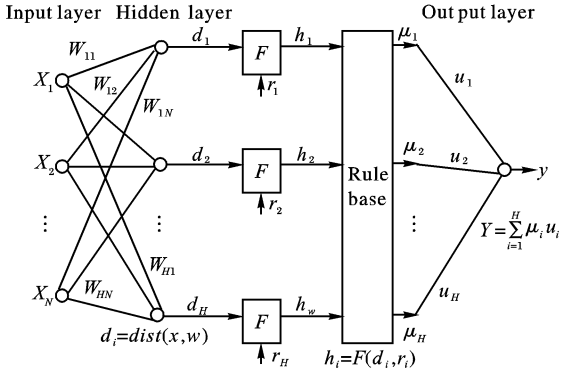


图1 FCC 结构图

对每个隐层神经元 i 首先要计算测试向量 $\boldsymbol{X}$ 和它的权值向量 $\boldsymbol{w}_i$ 之间的欧氏距离 d_i。r_i 为该隐层神经元 i 的泛化半径，由 d_i、r_i 构成激活函数 F 的输入，来计算隐层神经元 i 的输出 l_i。如图2所示，各个隐层神经元的输出合在一起构成了距离向量 $\boldsymbol{L}=(l_1,l_2,\cdots,l_H)$，通过 $\boldsymbol{L}$ 可以衡量测试向量与每个训练样本的相似性。

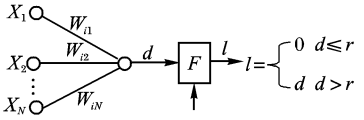


图2 FCC 隐层神经元

CC4 神经网络中 r 应用于整个神经网络，而 FCC 神经网络中每个隐层神经元都有其自身的 r，这样可以保证每个训练向量的泛化空间不重叠，每个隐层神经元（训练样本）都有一个以 w 为中心，r 为泛化半径的泛化空间。若测试向量 $\boldsymbol{X}$ 与 w 的距离小于或等于 r，则可视 $\boldsymbol{X}$ 与 w 所对应的训练样本不可区分，可以被划入该训练样本所代表的类中。

规则库使得 FCC 神经网络相对于 CC4 神经网络更具有一般性。它由 IF－THEN 集组成，距离向量 $\boldsymbol{L}$ 通过 IF－THEN 规则产生关于测试向量隶属于各个训练样本所属类的隶属度向量 $\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_H)$。通过输出层的权重 $u=(u_1,u_2,\cdots,u_H)$ 和 $\boldsymbol{\mu}$ 可以得到 FCC 神经网络的输出 $y=\sum_{i=1}^{H}\mu_i u_i$。

规则库的作用是判定测试向量 $\boldsymbol{X}$ 的模糊隶属度。但在 CC4 神经网络中，对于某一类测试样本要么属于它，要么不属于它，简言之 CC4 网络中测试向量的隶属度不是 0 就是 1。然而，FCC 神经网络中隶属度的值可以取[0,1]区间中的任一值。

规则库中包含两条确定测试向量模糊隶属度的 IF－THEN 规则。IF－THEN 规则是通过隐层神经元的输出 l 来判定的。记 m 为隐层神经元输出 $l_i=0$ 的个数。两条 IF－THEN 规则如下：

Rule1：IF $m=1$ THEN 使用 1NN 来确定 μ_i。

Rule2：IF $m=0$ THEN 使用 kNN 来确定 μ_i。

1.2 FCC 神经网络的训练过程

FCC 神经网络的训练过程包含两个步骤：

1）通过训练样本的输入／输出来确定输入层、输出层的权重。w_{ij} 为从输入神经元 i 到隐层神经元 j 之间的连接权值，当第 j 个训练样本提交给网络时，x_{ij} 为该训练样本的第 i 个输入神经元的输入，$w_{ij}=x_{ij}$，$i=1,2,\cdots,N$，$j=1,2,\cdots,H$。输出层的权重 u_i 等于训练样本的输出 y_i，$u_i=y_i$，$i=1,2,\cdots,H$。

2）即为训练过程，各隐层神经元泛化半径的确定。记 $d_{\min}$ 为训练向量 $\boldsymbol{i}$ 到其他训练向量 $\boldsymbol{j}(\boldsymbol{j}\neq\boldsymbol{i})$ 的非零距离中的最小距离，则第 i 个隐层神经元的泛化半径 $r_i=d_{\min}/2$，这能确保隐层神经元 i 的泛化空间不会和其他隐层神经元的泛化空间重叠。

1.3 模糊隶属度的泛化

完成了 FCC 网络的训练之后，FCC 神经网络即可部署展开了。输入测试向量，得到隐层的距离向量 $\boldsymbol{L}=(l_1,l_2,\cdots,l_H)$，通过 $\boldsymbol{L}$ 衡量测试向量和已训练向量之间的相似性。规则库作用于距离向量 $\boldsymbol{L}$，根据 1NN 或 kNN 算法，产生一个隶属度向量 $\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_H)$，利用输出层权重 u，$u=(u_1,u_2,\cdots,u_H)$ 和 μ 产生输出 $y=\sum_{i=1}^{H}\mu_i u_i$。

1）1NN 算法　当一个距离向量 $\boldsymbol{L}$ 的分量有一个为 0 时，使用 1NN 算法即测试向量落入某一训练向量的泛化空间，对于该训练向量所在类，测试向量的隶属度为 1，对于其他类的隶属度为 0。因此，如果 $l_j=0$，则隶属度向量 $\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_H)$ 为：

$$\mu_i=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases}\quad i=1,2,\cdots,H$$

2）kNN 算法　当距离向量 $\boldsymbol{L}$ 的任一分量均不为 0 时，即测试向量 $\boldsymbol{X}$ 并不完全落入某一测试向量的泛化空间时，采用 kNN 算法，测试向量隶属于样本向量的 k 近邻训练样本类，对非 k 近邻的样本类，其隶属度为 0。下面先从 $k=2$ 讨论起，进而将 k 一般化。

在给出了点 $\boldsymbol{X}$（代表一个测试向量 $\boldsymbol{x}$）后，取两个点 A 和 B 代表两个训练样本 a 和 b，它们是 x 的最近邻。它们之间的关系函数 μ(close to a) 和 μ(close to b) 由 x 和 a 与 b 的距离决定，当 x 完全落入 a 的泛化空间时 μ(close to a)$=1$，μ(close to b)$=0$；反之，μ(close to a)$=0$，μ(close to b)$=1$，这时 kNN 退化成 1NN 算法。除了这两种极端情况，x 的隶属于 a 和 b 的隶属度可记为 $\mu(x,a)$ 和 $\mu(x,b)$。

$\mu(x,a)=b/(a+b)$；

$\mu(x,b)=a/(a+b)$。

这是 $k=2$ 的情况，对于 k 取更大的值，上述公式可以更一般化为：

$\mu(x,a)=b/(a+b)=(1/a)/(1/a+1/b)$

$\mu(x,b)=a/(a+b)=(1/b)/(1/a+1/b)$

当 $k=3$ 时，c 也是 x 的 k 近邻样本，则隶属度可表示为：

$\mu(x,a)=(1/a)/(1/a+1/b+1/c)$

$\mu(x,b)=(1/b)/(1/a+1/b+1/c)$

$\mu(x,c)=(1/c)/(1/a+1/b+1/c)$

易证 $\sum_{i=1}^{k}\mu_i=1$。

2 FCC 的用户多隶属度判定

2.1 用户描述文件作为输入

用户描述文件主要是通过交互法生成的。所谓交互法是指根据与用户的交互操作来抽取用户的特征信息。在第一次生成用户描述文件时让用户回答一系列问题，然后根据用户选择的答案，启发式地转到下一个问题。这样根据用户的回答对用户的兴趣进行评价[1]。通过这种交互法得到对既定分类的兴趣的用户描述文件。例如，设有兴趣为“计算机、体育、数学”，某用户描述文件为(0.7,0.2,0.5)体现了该用户对计算机类较为兴趣，体育兴趣不大，对数学方面的内容一般

关心。当有“计算机”和“数学”方面的新网页信息或者新的文档,则可推荐给该用户。

若既定划分好的兴趣有 N 类,模糊角分类(FCC)神经网络的输入层设 N 个神经元。搜索引擎预先把用户划分成 H 类,每一类预先设定一个用户描述文件作为训练样本,则 FCC 神经网络的隐层神经元个数为 H,每个隐层神经元对应一个训练样本。

当新的用户描述文件生成时,将其输入 FCC 神经网络,通过已训练好的 FCC 神经网络权值,计算出该用户描述文件所属的类别。特别的,一个用户描述文件可以属于若干类别,对各类别有自己的隶属度。

2.2　实验分析

将用户的兴趣爱好划分为 10 类:军事、体育、财经、计算机、健康、文学、旅游、游戏、影视娱乐、数码电子。给出 10 组用户样本数据,每一组用户为一种兴趣领域的代表。表 1 给出了这 10 组用户在各领域的兴趣爱好权重。

表 1　样本用户兴趣分布表

	军事	体育	财经	计算机	健康	文学	旅游	游戏	影视娱乐	数码电子
用户 1	0.5	0.2	0	0.1	0	0	0	0.1	0	0.1
用户 2	0.1	0.6	0.1	0	0	0	0	0.1	0	0.1
用户 3	0	0	0.4	0.2	0.1	0	0.1	0.1	0.1	0
用户 4	0.1	0.2	0	0.5	0	0	0	0.1	0	0.1
用户 5	0	0.2	0.1	0.1	0.4	0.1	0.1	0	0	0
用户 6	0.1	0	0	0	0	0.6	0	0	0.3	0
用户 7	0	0	0.2	0.1	0.1	0	0.5	0.1	0	0
用户 8	0.1	0	0	0.2	0	0	0	0.6	0.1	0
用户 9	0	0	0	0.1	0	0.2	0	0	0.5	0.2
用户 10	0	0	0	0.1	0	0.1	0	0.2	0.2	0.4

对这 10 组数据分别用模糊角分类(FCC)神经网络和 BP(Back Propagation)神经网络[11]进行训练,模糊角分类神经网络只需训练一次,训练时间 elapsed_time = 0.016s;而 BP 神经网络训练迭代次数为 1 568 次,如图 3,所费时间 elapsed_time = 9.937s,显然采用模糊角分类神经网络在训练时间上的花费更低。

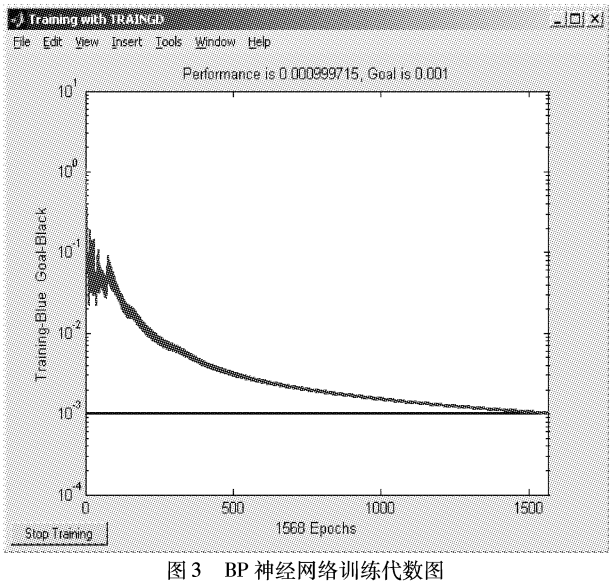


图 3　BP 神经网络训练代数图

模糊角分类神经网络与贝叶斯分类器的分类准确性比较[12]。引用搜索引擎超市(http://searchenginewatch.com/)的 100 组用户数据,分别使用模糊角分类神经网络与贝叶斯分类器对其进行分类,采用模糊角分类神经网络中的 kNN 算法,测试向量隶属于样本向量的 k 最近邻训练样本类;k 值由 1 变化到 10,得到这 100 个用户的隶属度向量:

$\boldsymbol{A}_1^{(1)},\boldsymbol{A}_2^{(1)},\cdots,\boldsymbol{A}_{100}^{(1)},k=1$;

$\boldsymbol{A}_1^{(2)},\boldsymbol{A}_2^{(2)},\cdots,\boldsymbol{A}_{100}^{(2)},k=2$;

…

$\boldsymbol{A}_1^{(10)},\boldsymbol{A}_2^{(10)},\cdots,\boldsymbol{A}_{100}^{(10)},k=10$

$\boldsymbol{A}_i^{(j)}$ 中的 i 表示第 i 个用户,j 表示 $k=j$。

采用贝叶斯分类器分类得到的 100 个用户的隶属度向量为:$\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_{100}$。

分别计算 $\boldsymbol{A}_i^{(j)}$ 与 $\boldsymbol{B}_i$ 的欧氏距离 $d_i^{(j)}$,$i=1,2,\cdots,100$;$j=1,2,\cdots,10$。

$$d^{(j)}=\sqrt{\sum_{i=1}^{100}d_i^{(j)}},\ j=1,2,\cdots,10$$

结果如图 4 所示。

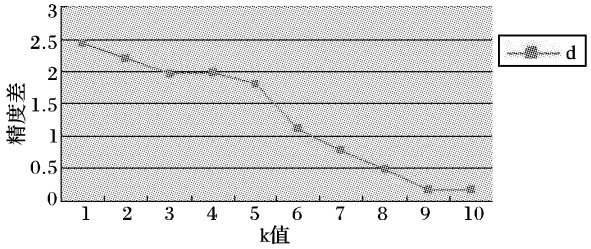


图 4　模糊角分类与贝叶斯分类精度差图

随着 k 值的增大,其分类效果趋近于贝叶斯分类算法。由此可见,模糊角分类神经网络是高效的。

3　结语

经典的数据挖掘中分类技术的目标是判定待判定数据的所属类别,而一个数据只能隶属一个数据类。但在现实生活中,大量的存在着一个数据隶属于多个类别的情形。信息推荐是搜索引擎的一个重要方面。对信息检索中的个性化推荐问题,一个最基本的要求是个性化推荐实现的时间应尽可能的少,用户的行为对应着用户兴趣的变更,用户的所属类别也相应的动态变化,对用户的多隶属判断问题,模糊角分类神经网络满足线性的时间需求,是可行的。

参考文献:

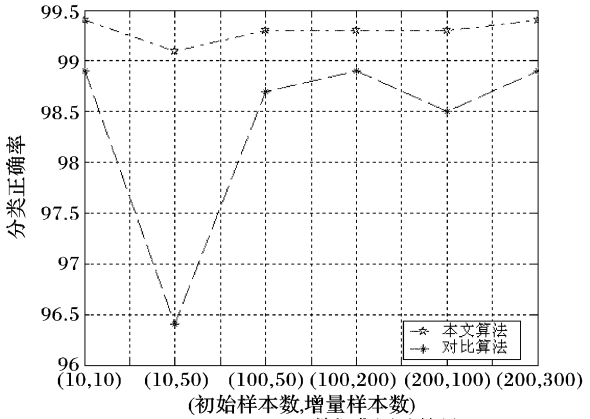
[1]　徐宝文,张卫丰. 搜索引擎与信息获取技术[M]. 北京:清华大学出版社,2003.

(下转第 2443 页)

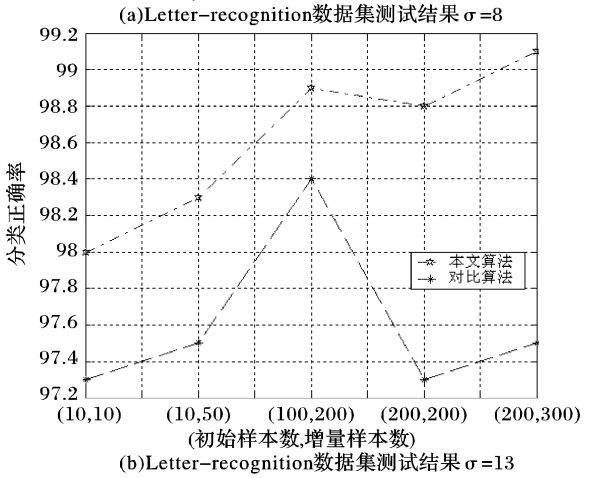
进行训练,选取其中的 600 个样本作为训练样本,1 000 个样本作为测试样本。

实验表明,在其他各项条件相同的条件下,采用径向基函数比采用其他核函数效果要好。图 4、图 5 给出了以径向基函数作为 SVM 核函数的实验结果,惩罚参数 $C = 1000$,阈值 $\xi = \eta = 1.0001$。图中,横坐标值标明了增量学习初始样本个数和增量样本个数,纵坐标为对应的分类正确率。

从图 4、图 5 可以看出:本文的对等增量 SVM 算法与已有算法相比精度明显提高。



(a)Letter-recognition数据集测试结果 σ=8



(b)Letter-recognition数据集测试结果 σ=13

图 5　对 Letter-recognition 数据集的实验结果

在对比算法中,使用 SV∪ASV∪ESV 作为新增样本加入后的训练样本集进行训练,其中 SV 为原训练样本集中的 SV 集,ASV 为增量样本集中的错分向量,ESV 为被正确分类的,与最优分类面邻近,处在最优分类面和间隔平面之间的样本,实际上,ASV∪ESV 就是违背原 SVM 的 KKT 条件的新增样本集,SV∪ASV∪ESV 是违背原 SVM 的广义 KKT 条件的新增样本集。

在当样本的统计性质比较差时,历史样本和增量样本分布不相似,甚至分布差异十分显著时,若用 SV∪ASV∪ESV 作为新增样本加入后的训练样本集进行训练,结果将会严重偏离真正的最优分类面,因此,SV∪ASV∪ESV 还不能完全表示新增样本加入后训练样本集所含的信息。

本文提出的对等增量 SVM 算法考虑了满足广义 KKT 条件的样本中与分类间隔距离较近的样本,而这些样本也可能成为增量学习后的 SV,从而在及时淘汰对后继分类影响不大的样本的同时保留了含有重要分类信息的样本。

4　结语

本文提出了一种新的 SVM 增量学习算法——对等增量 SVM 算法。算法在增量学习中考虑了可能成为增量学习后的新 SV 的训练样本,即违背广义 KKT 条件的样本、以及满足广义 KKT 条件的样本中与原分类间隔距离较近的样本,从而在及时淘汰对后继分类影响不大的样本的同时保留了原样本集和新增样本集中含有重要分类信息的样本。使增量学习的结果能够准确反映训练样本集的变化。对标准数据集的实验结果表明,本算法可以在新增样本加入后有效淘汰无用样本,同时保留含有重要分类信息的样本,从而获得了较好的分类性能。

参考文献:

[1] BURGES CJC. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121 - 167.

[2] VAPNIK V. 统计学习理论本质[M]. 北京: 清华大学出版社, 2000.

[3] CAUWENBERGHS G, POGGIO T. Incremental and decremental support vector machine learning[J]. Machine Learning, 2001, 44(13): 4098 - 4151.

[4] 萧嵘, 王继成, 孙正兴, 等. 一种 SVM 增量学习算法——ISVM[J]. 软件学报, 2001, 12(12): 1818 - 1824.

[5] SYED N, LIU H, SUNG KK. Incremental learning with support vector machines[A]. Proc. Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence (IJCAI-99)[C]. Stockholm, Sweden, 1999.

[6] 滕月阳, 唐焕文, 张海霞. 一种新的支持向量机增量学习算法[J]. 计算机工程与应用, 2004, 36: 77 - 80.

[7] 周伟达, 张莉, 焦李成. 支撑向量机推广能力分析[J]. 电子学报, 2001, 29(5): 590 - 594.

(上接第 2439 页)

[2] 曾春, 邢春晓, 周立柱. 个性化服务技术综述[J]. 软件学报, 2002, 13(10): 1952 - 1961.

[3] SAKAGAMI H, KAMBA T, SUGIURA A. Effective personalization of push-type systems: visualizing information freshness[J]. Computer Networks and ISDN Systems, 1998, 30(1 - 7): 53 - 63.

[4] CHEN EH, ZHEN YZ, XU FW. An extended corner classification neural Network based document classification approach [J]. Journal of Software, 2002, 13(5): 871 - 878.

[5] 朱大铭, 马绍汉. 二进制神经网络分类问题的几何学习算法[J]. 软件学报, 1997, 8(8): 622 - 629.

[6] 张振亚. 基于文本信息检索的知识发现技术研究[D]. 安徽: 中国科学技术大学计算机系, 2004.

[7] TANG KW, KAK SC. A new corner classification approach to neural network training[J]. Circuits Systems Signal Processing, 1998, 17(4): 459 - 469.

[8] KAK S. On generalization by neural networks[J]. Information Sciences, 1998, 111: 293 - 302.

[9] TANG KW, KAK S. Fast Classification Networks for Signal Processing[J]. Circuits Systems Signal. Processing, 2002, 21(2): 207 - 224.

[10] 韩小云, 周建平, 刘瑞岩. 广义聚类神经网络 GC[J]. 数据采集与处理, 1999, 14(1): 1 - 4.

[11] RAINA P. Comparison of learning and generalization capabilities of the Kak and the back propagation algorithms[J]. Information Sciences 1994, 81: 261 - 274.

[12] 欧洁, 林守勋. 基于贝叶斯网络模型的信息检索[J]. 微电子学与计算机, 2005, 5(1): 83 - 87.