基于RBF神经网络与粗糙集的数据挖掘算法

储 兵,吴 陈,杨习贝
(江苏科技大学 计算机科学与工程学院,江苏 镇江 212003)

摘 要:随着数据挖掘技术的兴起,为了提高数据挖掘的准确性,提出了很多数据挖掘算法。神经网络与粗糙集理论结合的数据挖掘算法一直是基于粗糙集理论数据挖掘研究的热点之一。文中提出利用RBF神经网络收敛速度快、泛化能力强等优势先对数据进行训练,优化数据后传递给粗糙集进行数据挖掘的新思路。并通过对比与未经过RBF神经网络训练的数据挖掘结果,发现RBF神经网络与粗糙集结合算法挖掘的精度有明显的提高,证明了RBF神经网络与粗糙集理论结合的数据挖掘算法是有效的、可行的。

关键词:RBF神经网络;粗糙集;数据挖掘
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2013)07-0087-04
doi:10.3969/j.issn.1673-629X.2013.07.022

Data Mining Algorithm Based on RBF Neural Network and Rough Sets

CHU Bing, WU Chen, YANG Xi-bei
(College of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: With the rise of data mining technology, in order to improve the accuracy of data mining, a lot of data mining algorithms have been put forward. The data mining algorithm which combinesneural networks with rough set theory has been one of the hot spots of data mining research based on rough set theory. Put forward the new idea of training data firstly, then pass to rough sets data mining after eliminating interference data, take the advantages of Radical Basis Function (RBF) neural network: fast convergence rate and strong generalization capability etc. And through the contrast to the data mining results which not using RBF neural network training, the precision of the algorithm which combined RBF neural network with rough set is greatly improved, it shows that the data mining algorithm which combined with neural network and rough sets theory has validity and feasibility.

Key words: RBF neural network; rough sets; data mining

0 引 言

数据挖掘(Data Mining,DM)也称数据库知识发现(Knowledge Discovery in Database,KDD),是一个从数据库的数据中提取隐含的有用信息(或知识)的过程[1]。数据挖掘常采用的算法及理论有粗糙集(Rough Sets)理论、人工神经网络(Artificial Neural Networks)、决策树(Decision Trees)、遗传算法(Genetic Algorithms)等。

粗糙集理论是1982年由Z. Pawlak提出的通过不可分辨关系或者不可分辨类确定没有给定某些特征或者属性情况下的近似区间,从而确定内部属性一些关系的工具。在处理大数据量、消除冗余信息等方面,粗糙集理论有着很好的效果。但是,由于粗糙集理论对错误描述的确定性机制过于简单,所以对对象的噪声比较敏感。人工神经网络由于鲁棒性强,分类精度高,对噪声数据不敏感等优点,在机器学习、模式识别等领域得到了广泛的应用。然而,神经网络面对数据挖掘中的高位和超大规模问题,其学习的速度缓慢,易造成网络训练过度[2],规则生成方面较差等表现出的缺陷更为明显,原有的神经网络算法在效率和可扩展方面都会出现效率低等问题。

由于粗糙集和人工神经网络各自具备很多的优势而独立使用往往无法规避其各自的缺陷。将粗糙集和人工神经网络二者相结合并应用于数据挖掘技术的研究之中,无疑具有十分重要的意义。

针对以上问题,提出了一种融合了RBF神经网络

收稿日期:2012-09-26 **修回日期**:2012-12-31 **网络出版时间**:2013-04-08
基金项目:国家自然科学基金资助项目(61100116/F020512)
作者简介:储 兵(1987-),男,硕士研究生,研究方向为数据挖掘;吴 陈,教授,博士,研究方向为智能信息处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130408.1715.065.html

和粗糙集理论的数据挖掘新方法，应用于大型数据库中挖掘分类的规则。人工神经网络中的 BP 神经网络在数据挖掘中应用广泛，但实际中通过神经网络的数据预处理方法的对比，发现 RBF 神经网络收敛速度更快、精度更高、可靠性更强[3]。

故文中将采用 RBF 神经网络对数据进行训练。其主要的思想是首先利用 RBF 神经网络优点通过网络的训练和学习优化数据，把经过处理的数据传递给粗糙集进行进一步的属性约简和规则抽取，得到最终的挖掘知识。

图 1 为数据挖掘流程图

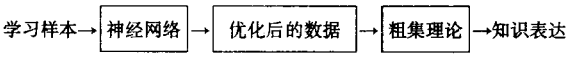

图 1 数据挖掘流程图

文中将融合了 RBF 神经网络与粗糙集的数据挖掘方法应用于开发区高新技术企业参数数据挖掘中，同时将对比未用神经网络处理的粗糙集数据挖掘方法，验证了该方法的优越性、有效性。

1 RBF 神经网络的基本原理

径向基函数(Radial-Basis Function)由 Powell 在 1985 年提出。径向基函数神经网络是由 J. Moody 和 C. Darken 于 20 世纪 80 年代末提出的一种神经网络结构，它是具有单隐层的三层前向网络。RBF 神经网络根据隐单元个数分为正规化网络和广义网络两种。广义网络因为在处理大量样本时更优，得到了更广泛的应用。

其原理如下：

假设有训练样本 N 组，输入层有 M 个神经元，隐含层有 I 个神经元，第 i 个隐单元输出称做“基函数” $\varphi(X,t_i)$，其中 $t_i=[t_{i1},t_{i2},\cdots,t_{im},\cdots,t_{iM}]$，$i=(1,2,\cdots,I)$ 是基函数的中心，输出层含 J 个神经元。输出层与输入层间的权值以 $\omega 1_{mi}(m=1,2,\cdots,M;i=1,2,\cdots,I)$ 表示；隐含层和输出层间的权值表示为 $\omega 2_{ij}(i=1,2,\cdots,I;j=1,2,\cdots,J)$。另外在隐含层设置阈值单元 φ_0，其输出恒为 1，与输出单元之间的权值为 $\omega 2_{0j}$。

设 $X=[X_1,X_2,\cdots,X_n,\cdots,X_N]^{\mathrm{T}}$ 为一个训练样本集。

其中任意一列 $X_n=[X_{n1},X_{n2},\cdots,X_{nm},\cdots,X_{nM}]^{\mathrm{T}}(n=1,2,\cdots,N)$ 为一个训练样本，对应的实际输出为 $Y_n=[y_{n1},y_{n2},\cdots,y_{nj},\cdots,y_{nJ}](n=1,2,\cdots,N)$，目标矢量集为 $\boldsymbol{D}$，其中任一目标矢量为 $\boldsymbol{D}_k$。

设输入一个训练样本 X_n，输出层的一个神经元的输出为：

$$y_{nj}(X_n)=\omega_{0j}+\sum_{i=1}^{I}\omega_{ij}(X_n,t_i),j=1,2,\cdots,J \quad (1)$$

基函数是高斯函数时，可表示为：

$$\varphi(X_n,t_i)=G(\|X_n-t_i\|)$$

$$=\exp\left(-\frac{1}{2\sigma_i^2}\|X_n-t_i\|^2\right)$$

$$=\exp\left(-\frac{1}{2\sigma_i^2}\sum_{m=1}^{M}(x_{km}-t_{im})^2\right) \quad (2)$$

上式中 σ_i 为高斯函数方差。常用的是另外一个参数 C_i，称为扩展常数，两者满足 $\sigma_i=0.8491C_i$ 的关系式[4]。

神经网络的训练要分为两个阶段：第一，无监督(无教师)的学习，仅根据网络的输入调整权值和阈值。第二，有监督(有教师)的学习：学习规则由一组描述网络行为的训练样本集(已知的输出/输入数据)给出，然后学习规则调整网络的权值与阈值。在训练开始之前，需要提供训练样本集 X 和对应的目标矢量集 $\boldsymbol{D}$ 以及径向基函数的扩展常数 C。训练的目的是求两个层之间的阈值和最终的权值 ω_1,ω_2。

2 粗糙集理论

粗糙集理论的思想是在保持信息系统的分类能力不变的前提下，通过知识约简[5,6]，导出问题的决策或分类原则。

2.1 粗糙集基本概念

(1)信息表。在粗糙集中使用信息表描述域中各种数据集合。信息表与关系数据库中的数据模型类似，相当于一张二维表，信息表的每一行称为对象、记录或者实例，每一列称为属性，每一个对象具有多个属性，通过这些属性取值去描述对象。一组由多个属性描述的对象的集合，称之为信息表。

(2)信息系统。信息系统是信息表的理论化表示形式，形式上，四元组 $S=(U,A,V,f)$ 是一个知识表达系统，其中：

U：对象的非空有限集合，称为论域；

A：属性的非空有限集合；通常分为条件属性 C 与决策属性 D；

V：$V=\bigcup_{a\in A}V_a$

$f:U\times A\rightarrow V$ 是一个信息函数，它为每个对象的每个属性赋予一个信息值，即 $\forall a\in A,x\in U,f(x,a)\in V_a$。

(3)属性约简、依赖、核。

属性约简是在不损失原有信息的前提下去掉信息表中的冗余部分，揭示各属性间的依赖关系。

设属性集合 $R,P\subseteq Q,R\subset P$，$\pounds$ 为粗糙集的划分，如果 $r_R(\pounds)=r_p(\pounds)$ 且不存在属性集合 $R'\subset R$，使 $r_R(\pounds)=r_p(\pounds)$，则称 R 为 P 的 $\pounds$ 约简，记为 $\mathrm{RED}_{\pounds}(P)$。

由定义可知,约简前后 ε 划分的趋近精度没有发生变化,即约简前与约简后系统所含信息量不改变。信息系统约简一般有多个,约简的交集称为 P 的 ε 核,记为:

$$\mathrm{CORE}_{\varepsilon}(P)=\cap\ \mathrm{RED}_{\varepsilon}(P)$$

由此可知,信息系统最重要的属性包含于核中[7,8]。

设 $R,P\subseteq Q$,如果 P 产生的任意的等价类都包含在 R 产生的某个等价类中,那么就称为 R 依赖于 P,记为:$P\rightarrow R$。

显然,如果集合 R 中任意属性取值都能由 P 中的属性取值唯一确定,那么集合 R 依赖于集合 P。

2.2 常见的属性约简算法介绍

属性约简算法众多,以下主要介绍一些经典算法。

1)基于系统熵的属性约简方法。

输出:属性集 $REDU$,约简后的属性集合。

(1) $REDU$ =核;

(2)样本属性集 $AR=C-REDU$;

(3)找出 AR 中具备最大属性重要性性质的 $SGF(\alpha,R,D)$ 的属性 a;

(4)如果有多个属性 $a_i(i=1,2,\cdots,m)$ 最大重要性是相同的,那么选择与 $REDU$ 拥有属性取值组合数量最小的属性 a_j;

(5) $REDU=REDU\cup\{a_j\}$;

(6)如果 $K(REDU,D)=1$,那么算法结束,否则回到(3)再执行。

其中:$SGF(\alpha,R,D)=K(R\cup\{a\},D)-K(R,D)$

$K(R,D)=\mathrm{card}(POS_R(R,D))/\mathrm{card}(POS_C(D))$

2)MIBARK 算法[9,10]——基于互信息的属性约简方法。

(1)首先对条件属性集 C 与决策属性集 D 的互信息 $I(C,D)$ 进行计算;

(2)对条件属性集 C 相对于决策属性集 D 的核 C_0 进行计算;

(3)令 $B=C_0$,

①如果 $|B|!=0$,则计算互信息 $I(B,D)$;转到④;

②对于每个属性 $p\in C-B$,计算 $I(p,D\mid B)$,称为条件互信息;

③遴选出使互信息 $I(p,D\mid B)$ 达到最大时属性值,记为 p,并且 $B\Leftarrow B\cup\{p\}$,若有多个属性在同一时刻达到最大值,则从中选择一个与 B 的属性值组合数最少的属性;

④若 $I(B,D)=I(C,D)$ 则终止,B 即为约简;否则转②。

3 基于RBF神经网络与粗糙集理论的基本思想与算法设计

基于对RBF神经网络与粗糙集理论的分析与研究,提出如下一种算法设计思想:首先利用RBF神经网络对规则的预测机制与对原始的属性预测,对比于真实决策属性分析出干扰数据将其删除。再通过属性的离散化将数据处理成粗糙集挖掘要求的数据传递给粗糙集进行属性的约简与规则的提取。

根据以上的算法设计思想,算法实现的步骤如下:

Step 1 首先对属性对象在概念层进行属性的泛化,分析表中的条件属性与决策属性,形成一个具有不可分辨关系的属性表。

Step 2 将第一步中属性表对应的数据传递给RBF神经网络进行训练与预测,多次调节 spread 值,直到达到一个最佳的预测效果为止,确定 spread。

Step 3 通过对神经网络的预测曲线及其误差曲线的分析,得出其误差较大的值,将其从数据中删除。

Step 4 将由RBF神经网络优化的数据进行数据的离散化。

Step 5 将离散化的数据在粗糙集理论基础上进行属性的约简,从不可分辨关系中求出分辨矩阵,并求出其核。

Step 6 通过上下近似集的计算得出上下近似集,边界域,负域。

Step 7 规则生成与通过精度设置提取规则。

4 应用实例

4.1 数据的预处理

全国开发区众多,每个开发区的信息都有许多属性值,数据量是相当庞大的。在提取出具有代表性地区后,首先将通过属性删除和属性泛化等操作,得到如表1所示的属性表。

表1　属性表

	条件属性1	条件属性2	条件属性3	条件属性4	决策属性
开发区	企业数(个)	从业人员(人)	总产值(万元)	出口总额(万美元)	总收入(万元)
天津	3463	285560	18450694	484532	30164172
石家庄	453	64003	9626958	66484	12522658
…	…	…	…	…	…
青岛	145	79254	11059996	182945	13078584

4.2 RBF神经网络对数据进行处理

设计一个RBF神经网络,将决策表的条件属性作为网络的训练样本,决策属性作为网络的训练目标。选取十六组样本为训练样本,另外的四十组数据为测试样本,运用 newgrnn 函数建立RBF神经网络模型。由于选取的数据较大,所以先要进行归一化处理,然后

经过多次 spread 数值的调节，确定 spread = 0.38。

由 Matlab 运行后得到了如图 2 与图 3 所示的真实值与测试值曲线，可以发现这两组曲线形状相似。说明建立的网络能够正确地对数据进行估计，也说明网络正确地对各个参数之间的函数关系进行了拟合，可以用来对奇异点等异常数据进行数据处理。

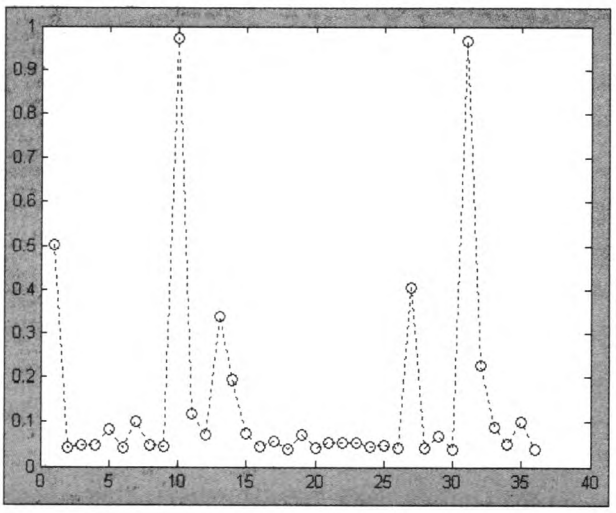

图 2 预测值曲线

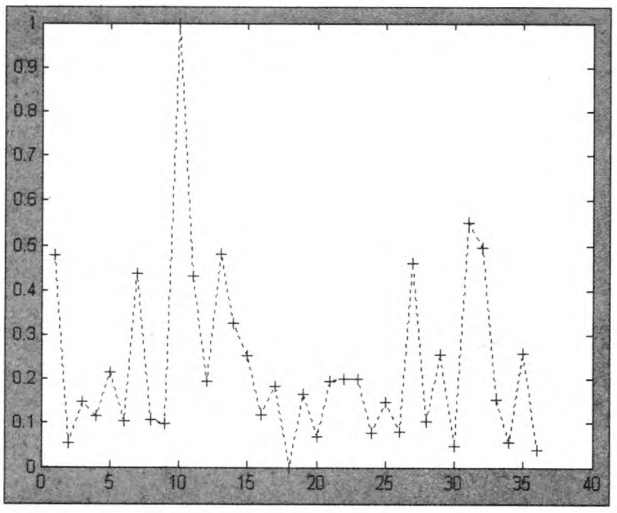

图 3 真实值曲线

对照其反应出决策属性的曲线，删除了奇异点及间断点等异常数据。对于删除前后数据的精度的变化情况，如表 2 所示：

表 2 精度对比表

	训练精度(SSE 值)	测试精度(SSE 值)
删除前	0.0268	0.8204
删除后	0.0268	0.762

通过表 2 删除前后的 SSE 值对比，显示出删除前后训练精度虽没有什么变化，但是测试精度却高于删除前的网络。所以用 RBF 神经网络对于数据的预处理，提高了系统的泛化能力。

4.3 基于粗糙集的数据挖掘

4.3.1 属性的离散化

为了要取得良好的数据挖掘效果，用粗糙集进行数据挖掘中，首先必须要将数据离散化，离散化的方法众多，有等距离离散化方法、等频离散化方法、Naive scaler 方法[11]等。

通过等距离离散化算法(Equal Interval Width)进行属性离散化，得到如表 3 所示的属性表：

表 3 离散化后属性表

	企业数	从业人员	总产值	出口总额	总收入
天津	4	4	2	0	2
保定	0	0	0	0	0
…	…	…	…	…	…
南宁	0	0	0	0	0

4.3.2 属性的约简及规则的抽取

通过使用 Rosetta 数据挖掘软件，进行对数据的约简、规则生成和抽取。一共得到 15 条规则，形式如下：

规则 1：从业人员(4) AND 总产值(2) AND 出口总额(0) = > 总收入(2)

规则 2：从业人员(0) AND 总产值(0) AND 出口总额(0) = > 总收入(0)

规则 3：从业人员(2) AND 总产值(1) AND 出口总额(1) = > 总收入(1)

规则 4：从业人员(1) AND 总产值(3) AND 出口总额(0) = > 总收入(2)

规则 5：从业人员(4) AND 总产值(4) AND 出口总额(4) = > 总收入(4)

规则 6：从业人员(1) AND 总产值(3) AND 出口总额(1) = > 总收入(2)

规则 7：从业人员(1) AND 总产值(1) AND 出口总额(0) = > 总收入(0)

规则 8：从业人员(4) AND 总产值(3) AND 出口总额(3) = > 总收入(2)

规则 9：从业人员(3) AND 总产值(2) AND 出口总额(4) = > 总收入(1)

规则 10：从业人员(3) AND 总产值(0) AND 出口总额(0) = > 总收入(1)

规则 11：从业人员(0) AND 总产值(0) AND 出口总额(1) = > 总收入(0)

规则 12：从业人员(0) AND 总产值(1) AND 出口总额(0) = > 总收入(1)

规则 13：从业人员(4) AND 总产值(3) AND 出口总额(0) = > 总收入(2)

规则 14：从业人员(2) AND 总产值(1) AND 出口总额(0) = > 总收入(1)

规则 15：从业人员(2) AND 总产值(2) AND 出口总额(1) = > 总收入(1)

将上述规则应用于训练样本中，发现测试精度

(87.5%)高于没有进行RBF数据预处理的粗糙集挖掘预测精度(81.25%)。分析所得结果,由于RBF神经网络的预处理,使得传递给粗糙集进行挖掘的数据更加精确。通过这种方法可以大大降低数据中一些不可靠的数据对于数据挖掘的影响,使得挖掘的效果更加显著。尤其是在大型数据挖掘项目上,优化后的数据在降低错误率、提高精度方面有着积极的作用。

当然,随着对RBF神经网络与粗糙集的研究的深入,可以通过优化RBF神经网络的算法及粗糙集数据挖掘的算法提高RBF神经网络与粗糙集在数据挖掘中的准确性。

5　结束语

文中融合RBF神经网络与粗糙集理论等数据挖掘技术,利用RBF训练速度快,泛化能力强的优点,提出一种新的先由RBF神经网络优化数据,再传递给粗糙集进行数据挖掘的新技术。通过对比未经过RBF神经网络处理的数据挖掘结果,反应出RBF神经网络与粗糙集理论结合的算法的良好效果。

参考文献:

[1] Chen M S, Han J, Yu P S. Date mining: an overview from a database perspective[J]. IEEE Trans on Knowledge and Data Engineering, 1996, 8(6): 866-883.

[2] 何炎祥,陈莘萌. Agent和多Agent系统的设计与应用[M]. 武汉:武汉大学出版社,2001.

[3] 李映颖,朱立贵,张德全,等. 基于BP和RBF神经网络对试飞数据预处理比较研究[J]. 计量与测试技术,2009,36(2):1-2.

[4] 唐昌盛,曲建岭. 基于RBF神经网络的飞参数据预处理[J]. 计测技术,2007,27(5):11-16.

[5] 汪小燕,杨思春. 一种基于分辨矩阵的新的属性约简算法[J]. 计算机技术与发展,2008,18(2):77-79.

[6] 陈　贞. 基于属性权重的区分举证启发式约简算法[J]. 莆田学院学报,2007,14(5):15-18.

[7] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341-356.

[8] Pawlak Z. Rough sets: theoretical aspects of reasoning about data[M]. Boston: Kluwer Academic Publishers, 1991: 72-80.

[9] 王国胤,于　洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报,2002,25(7):759-766.

[10] 丁宝桢,桑　琳,朱全英,等. 基于信息熵的粗糙集属性约简及其应用[J]. 计算机工程与应用,2007,43(3):245-248.

[11] 王国胤. Rough集理论与知识获取[M]. 西安:西安交通大学出版社,2001:93-116.

(上接第86页)

```
    UGL_STATUS (* info) (struct ugl_ugi_driver * pDriver,
    UGL_INFO_REQ infoRequest, void * info);
    UGL_STATUS (* destroy) (struct ugl_ugi_driver * pDriver);
    ......
} UGL_UGI_DRIVER;
```

4　结束语

文中介绍了自主研制的基于FPGA的GPU芯片原型的驱动软件的设计实现,其主要的难点在于WindML的移植和使用,需要对WindML SDK层和DDK层接口进行分析结合已有驱动开发,GLUT和GLU的封装实现需透彻理解其接口的定义,结合对mesa3D源码的分析实现。目前3D处理仅支持OpenGL1.3接口,使用WindML提供的简单窗口系统,下一步可在已有的接口之上封装实现更高版本的OpenGL接口以及移植功能更加强大的窗口系统,以满足更多的应用需求。

参考文献:

[1] 饶志恒. 图形处理器管线的研究与实现[D]. 长沙:湖南大学,2010.

[2] 阙　恒. 嵌入式图形处理器设计[D]. 南京:南京航空航天大学,2007.

[3] 周启平,张　扬,吴　琼. Vxworks开发指南与Tornado使用手册[M]. 北京:中国电力出版社,2004.

[4] 董英英,王启峰. 基于S3C2440的WindML图形驱动设计[J]. 现代电子技术,2010(16):69-71.

[5] WindML DDK Programmer's Guide, 3.0[M]. America: Wind River Systems Inc., 2002.

[6] 李海亮,石鹏程. Vxworks的WindML图形界面程序的框架分析[J]. 工业控制计算机,2007,20(1):46-49.

[7] WindML SDK Programmer's Guide, 3.0[M]. America: Wind River Systems Inc., 2002.

[8] 赵　俊,张克环,李仁发. 嵌入式通用图形加速芯片的研究与设计[J]. 计算机工程与应用,2008,44(26):74-76.

[9] Gamma E, Helm R, Johnson R, et al. Design Patterns: Elements of Reusable Object-oriented Software[M]. USA: Pearson Education, 1995.

[10] 杨国东. 嵌入式图形处理器设计与实现[D]. 济南:山东大学,2010.

[11] 张继伟. 基于WindML环境下的显卡驱动设计[J]. 现代电子技术,2010(14):78-80.

[12] 姚宇峰,邓志杰,陈光武,等. Vxworks图形驱动研究[J]. 单片机与嵌入式系统应用,2007(7):26-28.

基于RBF神经网络与粗糙集的数据挖掘算法

作者： 储兵， 吴陈， 杨习贝， CHU Bing， WU Chen， YANG Xi-bei
作者单位： 江苏科技大学计算机科学与工程学院,江苏镇江,212003
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2013,23(7)