最优路径森林分类算法综述*

沈龙凤，宋万干，葛方振，李 想†，杨 忆，刘怀愚，高向军，洪留荣
（淮北师范大学 计算机科学与技术学院，安徽 淮北 235000）

摘 要：针对快速分类算法中最优路径森林（OPF）分类算法进行了研究，进行了 OPF 分类算法研究及应用现状的调查。OPF 算法是近期兴起的一种基于完全图的分类算法，在一些公共数据集上与支持向量机（SVM）、人工神经网络（ANN）等算法的对比中，该算法能取得类似或更好的结果，速度更快。该算法不依赖于任何参数、不需要参数优化、不需要对各类别的形状作任何假设，能够处理多类问题，旨在全面系统地介绍 OPF 算法的研究及应用进展。

关键词：最优路径森林；分类；完全图
中图分类号：TP301.6 **文献标志码**：A **文章编号**：1001-3695(2018)01-0007-06
doi:10.3969/j.issn.1001-3695.2018.01.002

Review on optimum-path forest classification algorithm

Shen Longfeng, Song Wangan, Ge Fangzhen, Li Xiang†, Yang Yi, Liu Huaiyu, Gao Xiangjun, Hong Liurong
(*School of Computer Science & Technology, Huaibei Normal University, Huaibei Anhui* 235000, *China*)

Abstract: This paper did the research on optimal-path forest (OPF) classification algorithm for fast classification algorithm. It investigated the research and application of the OPF classification algorithm. The OPF algorithm is a new classification algorithm based on complete graph. In some public data sets, OPF was compared with support vector machine(SVM) and artificial neural network(ANN), the OPF algorithm could achieve similar or better results, but faster than them. The OPF algorithm does not depend on any parameters, does not need parameter optimization, and also can solve any problems without making any assumptions about the shape of each class. This paper aims to introduce the research status and future research directions of the OPF algorithm to the domestic readers.

Key words: optimum-path forest; classification; complete graph

0 引言

随着数据获取技术及存储技术的快速发展，人类社会的各个行业都积累了大量数据，这不仅促进了数据挖掘技术的飞速发展，也给数据挖掘技术带来了巨大挑战。分类技术是数据挖掘的核心和基础技术之一，目前已广泛应用于保险业、银行信贷、入侵检测、图像识别（掌纹、人脸、遥感卫星等）、医疗诊断、气象预报、案件侦破、工业等行业。目前广泛应用的分类技术主要有神经网络、支持向量机、决策树、贝叶斯等，其中神经网络用得最广泛的是 BP 神经网络。BP（back propagation）网络是 1986 年由 Rumelhart 和 McCelland 为首的科学家小组提出，是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。该算法具有误差小、动态性好、结果客观等优点，但是收敛速度慢，易陷入局部最小化。Jin 等人[1]提出了一种 BP 网络的改进算法，提高了收敛速度，具有算法简单、计算量小等特点，但是仍然存在容易陷入局部最小的问题。2002 年 Zhou 等人[2]提出了“选择性集成”的概念，并证明了从已有的学习基中剔除性能不好、作用不大的学习基，只挑选一些性能好的学习基进行集成，能达到比选择全部学习基更好的预测效果。虽然神经网络算法各方面性能都有了一定改善，避免陷入局部极小和加快收敛速度仍是其需要改进的问题。支持向量机是由 Boser 等人[3]提出的，支持向量机算法对于二分类问题的分类精度和分类速度都很好，但当训练样本很大时，训练速度会很慢，也不擅长处理多类问题。Platt[4]提出了 SMO 算法，并成为最快的二次规划优化算法，特别针对线性 SVM 和数据稀疏时性能更优。2001 年 Keerthi 等人[5]改进了 Platt 提出的 SMO 算法，将一个门限值改成了两个门限参数，训练速度比原 SMO 算法有了很大提升。Dong 等人[6]在 SVM 算法分解框架的基础上，基于有效集成核缓存，整理、收缩策略和停止条件，提出了一种 SVM 的快速训练算法。经实验证明，该算法的训练速度是 Keerthi 等人提出的 SMO 的改进算法的九倍。尽管如此，随着训练集的增大和支持向量的增多，SVM 算法的计算代价也会迅速增加。SVM 最擅长处理二类分类问题，虽然也提出了一些用 SVM 处理多分类问题的方

收稿日期：2016-12-22；**修回日期**：2017-03-01 **基金项目**：安徽省高等学校自然科学基金一般项目（KJ2016B018）；安徽省高等学校自然科学研究重大项目（KJ2017ZD32）；安徽省高校管理大数据研究中心 2017 年招标课题项目经费资助项目（25500119）；淮北师范大学 2017 年校级质量工程项目（12801262，12801240）；安徽省高校管理大数据研究中心 2016 年招标课题项目经费资助项目（12500347）

作者简介：沈龙凤（1981-），男，河北唐山人，讲师，博士，主要研究方向为数据挖掘、大数据技术；宋万干（1963-），男，安徽濉溪人，教授，硕导，主要研究方向为大数据、密码学、信息安全；葛方振（1975-），男，安徽萧县人，教授，博士，主要研究方向为群体智能系统、自适应软件；李想（1983-），男（通信作者），安徽濉溪人，讲师，博士，主要研究方向为物联网工程（lixiang_277@163.com）；杨忆（1980-），男，安徽淮北人，讲师，硕士，主要研究方向为推荐系统；刘怀愚（1979-），男，安徽淮北人，副教授，硕士，主要研究方向为图像处理；高向军，男，山东临沂人，副教授，博士，主要研究方向为医学图像处理；洪留荣（1969-），男，安徽宿松人，教授，博士，研究方向为模式识别、数字图像处理.

法,但在处理大数据集时效率低下。

最早的决策树分类算法是 Hunt 等人[7]于 1966 年提出的 CLS(concept learning system)学习算法,该算法首次提出用决策树进行概念学习。Breiman 等人于 1984 年提出 CART(classification and regression trees)分类算法,该算法生成的规则易于理解,计算量不大,但是当类别很多时,错误可能会增加较快。Quinlan[8]于 1986 年提出了 ID3(iterative dichotomizer 3)算法,该算法适合处理大规模数据,因为其选择信息增益作为度量,所以会偏向具有较多取值的属性。Quinlan[9]又于 1993 年提出了 C4.5 算法,该算法使用信息增益率作为度量,克服了 ID3 算法偏向具有较多取值属性的缺点,能够处理不完整数据及连续型属性离散化的问题,但是在处理连续型属性线性搜索阈值时需要付出很大代价。Mehta 等人[10]于 1996 年提出了 SLIQ(supervised learning in quest)算法,该算法可以处理比 C4.5 大得多的样本集,速度更快、精度更高,但是该算法要求类别表常驻内存,从一定程度上限制了数据集的大小。Shafer 等人于 1996 年提出了 SPRINT(scalable parallelizable induction of decision trees)算法,该算法解决了 SLIQ 算法要求类别表常驻内存的问题,速度更快,可伸缩性更好,但该算法很难处理非分裂属性的属性列表,且扩展性不好。Olaru 等人[11]于 2003 年提出了一种完全模糊决策树算法,该算法又称为软决策树,其分类精度明显高于普通决策树。冯少荣[12]于 2007 年提出了一种基于度量的决策树算法(metric based decision tree,MBDT),用该算法构造的决策树能有效地减少决策树的层数,从而提高决策树的分类效率。Kwok 等人[13]于 1990 年提出了一种多决策树方法,基于改进的 ID3 算法来构建多棵性能较好的决策树,用每一棵决策树去预测,将各树的预测值求平均值作为最终的预测值。实验结果表明其预测结果优于单独使用其中的任何一棵决策树,但是该算法不能生成足够多的优良的决策树,不能充分研究增加树的数量所带来的影响。

贝叶斯分类器的原理是通过某对象的先验概率,利用贝叶斯公式计算出其后验概率,选择具有最大后验概率的类作为该对象所属的类。其中朴素贝叶斯分类器(naïve Bayes,NB)的研究非常广泛,其有一个“属性之间必须是相互独立的”先天性假设,假设成立时方法简单、分类精度高、速度快,但在实际应用中这一假设往往不成立,此时分类精度会很低[14~16]。TAN(tree augmented naïve-Bayes)分类器是对朴素贝叶斯分类器的改进,放松了 NB 的独立性假设,每个属性节点最多可以依赖于一个非类节点,类别属性是其他所有属性的父节点,其分类性能明显优于朴素贝叶斯分类器[17,18]。BAN(BN augmented naïve-Bayes)分类器是对 TAN 分类器的进一步扩展,允许各特征变量所对应的节点之间的关系构成一个图,而不只是树[17,18]。GBN(general Bayesian network)分类器是一种无约束的贝叶斯网络分类器[17,18],即把类别节点当做普通节点,不再看成是属性节点的双亲节点。总体上来讲,NB 和 TAN 算法适合小规模数据集,BAN 和 GBN 算法对大数据集效果更佳。

最优路径森林 OPF 算法是由 Papa 等人提出的,是将训练集转换成一个完全图,完全图中的每个节点都是训练集中的一个样本,图中的弧用节点间距离来表示,根据完全图来生成最优路径森林,森林中每棵树上的所有节点都属于同一个类别。在进行分类时,计算待分类样本到哪棵树的距离最近,则其类别就与这棵树的根节点的类别相同[19,20]。

由于 OPF 分类器不依赖于任何参数,训练阶段不需要进行参数优化,所以其训练速度和分类速度都非常快。与其他分类算法相比,OPF 算法的分类精度与 SVM 相近而优于其他方法,训练、分类速度比 SVM 更快,也不需要对类别的形状作任何假设,能处理多类及有一定程度类别重叠的问题。近几年 OPF 算法在国外发展迅速,巴西圣卡洛斯联邦大学的计算机系、坎皮纳斯州立大学的计算所、圣保罗州立大学的电气工程系、圣保罗天主教大学的计算机工程系、福塔莱萨大学技术研究中心等都在进行相关研究,此外葡萄牙波尔图工学院、美国迈阿密大学的电气与计算机工程系、耶鲁大学的医学/精神病学系、宾夕法尼亚大学医学图像处理放射学系、加拿大卡尔加里大学的电气与计算机工程系和计算机科学系也都进行了 OPF 算法的部分研究工作,目前该算法已经应用到了多个领域,但中国对 OPF 算法的研究目前较少。因此,系统地总结整理 OPF 算法最新的理论及应用研究进展有着非常重要的意义。

1 最优路径森林算法原理

OPF 算法是一种基于最优路径树的监督分类方法,该方法分成训练分类器和分类两个阶段。首先根据训练样本之间的距离构建由所有样本构成的完全图的最小生成树,再根据训练样本的类别生成最优路径森林分类器;然后根据构建的最优路径森林分类器对测试样本进行分类。

1.1 训练分类器阶段

将总数据集 Z 分成集合 Z_1 和 Z_2,Z_1 用于训练分类器。假定 Z_1 共有 N 个样本,每个样本 a_i 包含 M 个属性(a_{i1},a_{i2},…,a_{iM})。首先对 Z_1 的 N 个样本构建完全图 A,完全图中的每个节点都是集合 Z_1 中的一个样本,完全图中任意两个节点 a_i 与 a_j 之间都有弧,弧的权值由两个节点之间的欧氏距离来衡量(也可以用其他距离)。

$$d(i,j)=\sqrt{(a_{i1}-a_{j1})^2+(a_{i2}-a_{j2})^2+\cdots+(a_{iM}-a_{jM})^2}\quad 1\leqslant i,j\leqslant N \tag{1}$$

对完全图 A 生成其最小生成树,在得到的最小生成树中,找到连接两个不同类别节点的弧,对应的两个节点就作为最优路径森林中的树的根节点。注:由于连接不同类别节点的弧可能有多个,所以同一类别的树的根节点也可能不止一个。路径是由不同节点所构成的节点序列 $\pi=\langle s_1,s_s,\cdots,s_k\rangle$,其中($s_i$,$s_{i+1}$)$\in A$ 且 $1\leqslant i\leqslant k-1$。当一条路径只包含一个节点时,如 $\pi=\langle s_1\rangle$,称为简单路径。根据路径代价函数 f 可为每条路径 π 指定一个代价 $f(\pi)$。若路径 π 是最优的,当且仅当与该路径具有相同终点 s_k 的任何其他路径 π',都有 $f(\pi)\leqslant f(\pi')$。用 $\pi\cdot\langle s,t\rangle$ 表示以 s 为终点的路径 π 与弧$\langle s,t\rangle$ 所构成的路径。最优路径森林算法的路径代价以路径代价函数 $f_{\max}$ 计量。根据求最优根节点的理论特性,各节点的代价 $f_{\max}$ 初始值及路径代价为

$$f_{\max}(\langle s\rangle)=\begin{cases}0 & \text{当 } s \text{ 是根节点时}\\ +\infty & \text{当 } s \text{ 为其他节点时}\end{cases} \tag{2}$$

$$f_{\max}(\pi\cdot\langle s,t\rangle)=\max\{f_{\max}(\pi),d(s,t)\} \tag{3}$$

当路径 π 为非平凡路径时,$f_{\max}(\pi)$ 的值为路径 π 上任意相邻两点之间距离的最大距离。最优路径就是到相同终点的不同路径中代价最小的路径。最优路径森林算法将找到根节点集合 S 到训练集 Z_1 中每一个节点 t 的最优路径,从而构建最优路径森林,并将得到每个节点的最优路径代价 $C(t)$、最优路径上的前驱节点 $s=\mathrm{Pre}(t)$ 及该节点的类别 $L(R(s))$(其中 $R(s)$ 表示 s 所在最优路径树上的根节点)。节点 t 的代价 $C(t)$ 即是最优路径上 t 的前驱节点 s 的代价 $C(s)$ 与节点 s 和 t

之间距离 $d(s,t)$ 的最大值;t 的类别就是其前驱节点最优路径上的根节点的类别 $L(R(s))$。

输入:具有类别标签的训练集 Z_1,根节点集合 $S\subset Z_1$,任何样本的特征向量 v 与任意两个节点间距离 d。

输出:最优路径森林(即节点的前驱 Pre),节点代价 C,节点类别标签 L。

辅助变量:优先队列 Q,代价临时变量 cst。

1) $\forall s\in Z_1\backslash S$, 设置 $C(s)\leftarrow+\infty$; /* 初始化非根节点代价为正无穷 */

2) $\forall s\in S$,执行 3); /* 初始化根节点代价、前驱、类标签,根节点入 Q */

3) $C(s)\leftarrow 0$, $\mathrm{Pre}(s)\leftarrow\mathrm{nil}$, $L(s)\leftarrow\lambda(s)$,并将 s 插入 Q;

4)若 Q 非空, 执行 5),否则结束;

5)从集合 Q 中移除代价 $C(s)$ 最小的样本 s;

6) $\forall t\in Z_1$,若 $t\neq s$ 且 $C(t)>C(s)$, 执行 7);

7)计算代价临时变量 $\mathrm{cst}\leftarrow\max\{C(s),d(s,t)\}$;

8)若 $cst<C(t)$,则执行 9);

9)若 $C(t)\neq+\infty$, 从集合 Q 中移除 t;

10) $L(t)\leftarrow L(s)$, $C(t)\leftarrow\mathrm{cst}$,并将 t 插入到集合 Q。

第 1) ~3)是对训练集中的非根节点样本初始化代价为无穷大,对根节点初始化其代价、前驱节点及类别。

1.2 分类阶段

对于集合 Z_2 中的任何待分类样本 $t\in Z_2$,计算 t 到 Z_1 中所有样本的距离。考虑 t 通过 Z_1 中任何节点的路径代价,选择所有路径中 t 的代价最小的路径就是最优路径 $P^*(t)$,t 的代价即为最优路径中 t 的代价,t 的类别与该最优路径的根节点的类别相同,即 $L(t)=\lambda(R(t))$。其中 $C(t)=\min\{\max\{C(s),d(s,t)\}\}$, $\forall s\in Z_1$。

2 最优路径森林算法的扩展

2.1 提高 OPF 分类器的分类精度和分类速度的方法

为了提高分类精度,Papa 等人[20]提出学习分类错样本的算法。把数据集分成三个子集合 Z_1、Z_2 和 Z_3,其中 Z_1 用于训练分类器,Z_2 用于评价分类器,Z_3 用于测试。用 Z_1 训练的分类器对 Z_2 进行分类,存在 $L(s)\neq\lambda(s)$ 的情况,即存在分类错误的样本,通常分类错的样本更能代表其类别。将分类错误的样本与 Z_1 中非根节点的相同数量的样本互换,保证训练集大小不变,重新训练分类器,每次训练了新的分类器,都在 Z_3 上测试训练精度。反复迭代 n 次,选择精度最高的分类器。

在用分类器进行分类时每次都要访问训练集中的每个样本,为了提高分类速度,在训练分类器的同时输出训练样本按代价非递减顺序排列的新的训练集 Z_1'。因为 Z_1'是非递减的,在分类时依次计算待分类样本 t 与集合 Z_1'中的样本 s 间的距离 $d(s,t)$ 及路径上的最短代价,即 $\max\{C(s),d(s,t)\}$。当 $\max\{C(s),d(s,t)\}<C(s')$时,停止计算后面所有样本与 t 之间的距离(s'是 s 后继样本),当前已计算的各路径中代价最小的路径就是总的最优路径。假定最优路径上 t 的前驱节点是 m,则 t 的代价就是 $\max\{C(m),d(m,t)\}$,t 的类别是 $\lambda(R(m))$。当训练集很大时,这种方法将对分类速度有很大提升,在最坏情况下,其分类时间与之前的算法相同。为了进一步提高分类速度,Papa 等人[21,22]提出对训练集进行排序及剪枝的算法。用对集合 Z_2 中分类错误样本进行学习得到的分类器对 Z_2 进行分类,记录精度 Acc 及 L_1、P_1、C_1、Z_1':a) $\forall t\in Z_2$,其前驱节点 $\mathrm{Pre}(t)\in Z_1$,从 $\mathrm{Pre}(t)$到根节点的路径上的所有节点并入集合 S,将 Z_1/S 中的所有节点移到 Z_2 中,对新的 Z_1 训练分类器,并在 Z_2 上进行分类,对分类错误进行学习重复 n 次,得到新的分类器(记录精度 Acc1 及 L_1,P_1,C_1,Z_1');b)当 $|\mathrm{Acc}-\mathrm{Acc1}|\leqslant\mathrm{MLoss}$ 且重复次数小于给定重复次数时,重复进行 a)的操作。可以得到缩减了的 Z_1 训练集,并对 Z_1 根据代价非递减排序得到 Z_1',将进一步提高分类速度。

2.2 基于 KNN 的 OPF 算法

基于 KNN 算法的 OPF 算法[23]的 K 值是根据分类精度最大的 K 值来确定的。根据训练集 Z_1 创建 KNN 图 A_k。A_k 定义为:当根据距离 $d(s,t)$ 计算得知 t 是 s 的 K 最近邻时,则 $t\in Z_1$ 是邻近 $s\in Z_1$ 的,A_k 即为训练集 Z_1 的 KNN 图。其中弧 (s,t) 的权值为 $d(s,t)$,节点 $s\in Z_1$ 的权值由密度值来表示:

$$\rho(s)=\frac{1}{\sqrt{2\pi\sigma^2}k}\sum_{\forall t\in A_k(s)}\exp\left(\frac{-d^2(s,t)}{2\sigma^2}\right)\tag{4}$$

其中:$\sigma=\dfrac{d_f}{3}$(d_f 为 A_k 中所有弧权值中最大的权值)。选择该参数是考虑计算密度时将用到所有节点,因此一个高斯函数覆盖了距离 $d(s,t)\in[0,3\sigma]$内的大部分样本。一系列不同的邻接样本构成一条路径,如 π_t 以某个根节点 $R(t)\in Z_1$ 开始到样本 t 结束。当路径上只包含一个节点时称为平凡路径,如 $\pi_t=\langle t\rangle$。对于路径 $\pi_t=\pi_s\times\langle s,t\rangle$,其中 s 是路径 π_t 上 t 的前驱,使路径代价函数 $f(\pi_t)=f(\pi_s\times\langle s,t\rangle)$最大的路径就是最优路径。所有 $t\in Z_1$ 的样本到根节点的最优路径就构成了最优路径森林。其中路径代价函数定义为

$$f_1(\langle t\rangle)=\begin{cases}\rho(t) & \text{if}\quad t\in R\\ \rho(t)-\delta & \text{其他}\end{cases}\tag{5}$$

$$f_1(\pi_s\cdot\langle s,t\rangle)=\min\{f_1(\pi_s),\rho(t)\}\tag{6}$$

其中:$\delta=\min_{\forall(s,t)\in A_k|\rho(t)\neq\rho(s)}|\rho(t)-\rho(s)|$;$R$ 是根节点的集合,概率密度函数值最大的节点即为根节点,且其前驱为空,根节点的集合是动态生成的。首先对 Z_1 中每个节点 s 初始化其前驱为空,类别为 $\lambda(s)$(函数 $\lambda(s)$用于指定样本 s 属于哪个类别),代价为 $V(s)=\rho(s)-\delta$,并将样本 s 放入优先队列 Q。取出代价最大的样本 s,如果其前驱为空则作为根节点,将其代价更新为 $V(s)=\rho(s)$,并查看其他节点通过 s 能否有更优路径,即节点 t 的路径代价 $V(t)<\min\{V(s),\rho(t)\}$是否成立。若成立,则节点 t 经过 s 到达根节点的路径优于 t 当前的路径 $\pi(t)$,更新 t 的类别为与 s 相同,并且 t 的前驱变为 s,代价 $V(t)=\min\{V(s),\rho(t)\}$。重复进行,直至各节点的代价、前驱不再发生变化,则生成了最优路径森林分类器。

KNN 的 K 值依次取$[1..K_{\max}]$的整数。K 每取一个整数,就按上述方法训练一个分类器,计算其分类精度,精度值最大的分类器所选用的 K 值即是最优的值 K'。其中分类精度 Acc 为

$$E(i)=\frac{FP(i)}{|Z_1|-|NZ_1(i)|}+\frac{FN(i)}{|NZ_1(i)|}\tag{7}$$

$$\mathrm{Acc}=\frac{2c-\sum_{i=1}^{c}E(i)}{2c}=1-\frac{\sum_{i=1}^{c}E(i)}{2c}\tag{8}$$

式(7)中:$FP(i)$ 是属于其他类而错分到第 i 类的样本数;$FN(i)$是属于第 i 类而错分到其他类中的样本数;$NZ_1(i)$是训练集中属于第 i 类的样本数。假定一共有 c 个不同类别,即 $i=1,2,\cdots,c$。选择了最优的 K'之后,应用上述算法重新训练分类器,只是在算法中要将函数 f_1 替换为 f_2。函数 f_2 定义为

$$f_2(\langle t\rangle)=\begin{cases}\rho(t) & t\in R\\ \rho(t)-\delta & \text{其他}\end{cases}\tag{9}$$

$$f_2(\pi_s\cdot\langle s,t\rangle)=\begin{cases}-\infty & \text{如果 }\lambda(t)\neq\lambda(s)\\ \min\{f_2(\pi_s),\rho(t)\} & \text{其他}\end{cases}\tag{10}$$

构建了基于KNN的OPF分类器之后,对测试集 Z_2 进行分类。对于任意 $t \in Z_2$,计算 t 到集合 Z_1 所构建的图 A_k 的K近邻,就像 t 是 A_k 的一部分一样。在从 R 到 t 的所有路径中选择最优路径,选择的依据是所有路径中 $V(t)$ 最大的路径,$V(t)$ 为

$$V(t) = \max\{\min\{V(s), \rho(t)\}\} \quad \forall s \in Z_1 \tag{11}$$

假定 $s' \in Z_1$ 满足式(11),则 s' 的类别与最优路径上根节点的类别是相同的,即 t 被分类到 s' 的类别。

3 最优路径森林算法的应用

OPF算法一经提出,便引起了各方的广泛关注,目前已经应用到了很多领域。表1给出了OPF算法的主要应用领域,并给出了在相应领域具体的应用方法和效果。

表1 OPF算法应用领域

应用领域	具体应用	方法及效果	参考文献
模式识别领域	红外人脸识别	使用OPF算法进行人脸红外识别,使用PCA进行特征提取,实验结果表明OPF算法的识别率比ANN算法高很多,很接近SVM算法,但其计算时间比SVM和ANN算法有了大幅度下降	[24]
	基于特征脸的人脸识别	基于OPF分类器提出一种新的、快速而精确的整体分析方法,分类精度与SVM相当,远高于MLP神经网络,但计算代价比SVM和MLP神经网络大幅度缩减	[25]
	人脸、人类行为识别	将OPF分类算法应用于面向张量的应用环境,评估了该算法使用多线性主成分分析在人脸和人类行为识别中的空间转换任务的鲁棒性,证明了OPF分类算法在应用于基于张量的特征空间时能够获得更高的识别率	[26]
	指纹分类	提出一种新的满足离散及连续域的方法对指纹图片数据集进行查询与检索。该方法的主要特征为:a)基于多分辨率分解方法的纹理图像描述符对全局和局部指纹信息的有效编码,用于指纹匹配的目的相似措施;b)一种基于OPF分类器的新的多类目标识别方法,该方法具有较高的分类率,证明了其用于表征指纹图片的高可行性和高有效性	[27]
	说话情感识别	提出一种基于声门气流信号特征的说话情感识别方法,将OPF算法的有效性与GMM、SVM、ANN-MLP、KNN、BC、C4.5等算法进行了比较测试,OPF和SVM算法的分类精度高于GMM、ANN-MLP、KNN、BC、C4.5等算法,运算速度上KNN最快,OPF第二,综合性能OPF算法最优	[28]
	鲁棒而快速的元音识别	自动元音识别的艺术状态系统基于传统的机器学习算法,如ANN和SVM等,但这些算法很难同时保证高效性和有效性。基于OPF提出一种用于自动元音识别的算法,该算法在训练时间和分类精度上均超越了SVM和ANN及其他算法	[29]
	手写体数字识别	使用人物签名的方法进行特征提取,使用OPF进行手写体数字分类识别,结果表明该方法使用曼哈顿距离进行分类时的结果是令人满意的,平均精度可达99.53%,且其训练时间和测试时间比其他方法要低,这也是OPF方法的主要特点	[30]
	快速的机器人语音接口	提出并设计了一种鲁棒的、可扩展的用语音控制移动机器的监督分类体系结构,其主要贡献是构建了可用的语音指令数据集及将OPF分类器引入该领域。在该领域进行了OPF与Bayes、SVM-RBF、SVM-NoKernel、ANN-MLP等算法的对比实验,OPF与SVM-RBF、Bayes获得了相似的结果,但OPF的总耗时最少	[31]
遥感图像	土地使用分类	首次将OPF分类器应用于农业科学领域、识别土地使用情况以达到环境管理的目的。该研究比较了OPF与ANN-MLP、BC、SVM分类器的优劣,结果表明OPF与SVM的分类精度类似而优于其他算法,但OPF的速度比SVM要快约65倍	[32]
	土地使用遥感图片分类	近年来土地使用分类已被广泛研究,因为可以识别非法土地使用及监视滥伐区域,提出一种OPF聚类的方法并首次应用于土地使用的非监督识别。实验在OPF、均值漂移、K-均值等算法间开展,结果表明OPF聚类方法的预测结果比另外两种方法好很多	[33]
	卫星图片	主要贡献有两点:a)提出非离散的马尔可夫模型的优化框架;b)提出一个后处理的方法,以避免在高频区矫枉过正,从而获得关于标准OPF-MRF的改进。实验证明所改进的方法可以提供比原OPF-MRF更好的结果	[34]
	土地覆盖范围	提出一种连同马尔可夫随机场的局部适应OPF分类器,该分类器优于其传统版本。基于四卫星图像的实验结果表明,所提出的方法性能比之前的方法更优,并且学习马尔可夫随机场参数的速度比之前版本快很多	[35]
石油勘探	快速石油钻井监测	对OPF进行了快速而健壮的修订,提出了一种高效的OPF,称为EOPF。进行了两轮应用监督型分类器的实验,第一轮是OPF、ANN-MLP、SVM及BC的性能比较,实验结果表明OPF在精度和效率上都优于其他几种方法;第二轮是比较OPF与EOPF的效率与有效性,结果表明EOPF的精度率优于OPF,但其速度是OPF的1.41倍	[36]
音乐领域	音乐流派分类	首次将OPF分类器应用到音乐流派分类领域,在两个公开数据集上比较了OPF、SVM、BC分类器,所有的分类器都能达到相近且很好的分类精度,但OPF的训练速度和分类速度更快。使用PSO-OPF、HS-OPF和GSA-OPF等方法进行特征选择,可以减少特征数量却不影响识别率	[37]
医学领域	喉癌病理学检测	基于OPF分类器提出一种喉癌病理学检测方法,在三个公共数据集上对OPF和SVM算法进行了比较,OPF算法在两个数据集上的精度都优于SVM算法,在所有的数据集上都比SVM算法的运算速度快	[38]
	人脑磁共振图像分类	提出了一种适合大数据集的OPF聚类算法的扩展方法,并用于人脑磁共振T1图像上灰质和白质的自动分类。其分类正确率很高,与近期报道的方法不相上下,但其速度却要快30倍,不依赖于其他模板,且在实像中的GM分类精度更高	[39]
	人脑组织的磁共振图片分割	提出了一种基于OPF聚类的高效、精确、通用、稳定的方法进行人脑组织分割。基于DOC曲线的目标评价结果表明,该分割方法在大多数情况下比当前最流行的脑组织分割方法FAST和PVC更精确、更稳定,分析结果也表明,OPF方法的参数比其他方法更容易调节到最优值	[40]
	帕金森疾病的识别	提出一种基于OPF的帕金森疾病自动识别方法,该方法不需要对形状、类别的可分性、特征空间作任何假设。实验结果表明,OPF比SVM、ANN等常用的帕金森疾病识别方法的效果更好 基于进化技术实现了帕金森疾病自动识别的特征选择任务,特征选择是用OPF在评价集上实现的。实验结果表明,所用的每种特征选取技术都比直接在原始(未进行特征选择)数据集上的结果好,其中HS-OPF和GSA-OPF的精度最高,HS-OPF的速度最快	[41] [42]
	核磁共振脑组织图片分割	结合OPF聚类算法与概率图集方法提出一种改进人脑组织皮质下层磁共振图片分割的自动兹共振图片分割方法,12个脑MR图像中丘脑区域组织分割的实验结果证明,该方法的性能优于广泛应用的SPM和FSL方法	[43]

续表 1

材料领域	合成材料孔隙度分割	提出了一种 OPF 分类器完成合成材料孔隙率分割的新的应用和评价,并对光学显微镜获得的图像进行量化。合成材料样本图像分析结果表明,OPF 分类器的结果与 SVM-RBF 相似,都优于 SVM-LINEAR 分类器,但 OPF 的速度是 SVM-RBF 的 4.01 倍,因此,OPF 分类器更适合应用于该领域	[44]
	铁合金样本快速自动微结构分割	提出了一种用 OPF 分类器快速自动分割铁合金的方法,用 1% 的输入图像作训练,99% 的图像作测试。实验结果表明,OPF 在效率(训练时间 + 测试时间)、有效性上都优于 SVM-RBF 和 SVM-LINEAR,只是在可锻铸铁上几个分类器才会达到相似的结果	[45]
天气预报方面	降雨量估计	用 OPF 分类器进行了卫星降雨量的估算,将 OPF、SVM 和 ANN-MLP 等监督算法均用于降雨量估算。实验结果表明,所有分类器所获得的结果类似,但 OPF 分类器速度比其他分类器快很多。使用不同大小的训练集进行预测发现,OPF 在大数据集上优势更明显。对于 OPF 算法,用堪培拉距离比欧氏距离的效果更好	[46]
		提出一种 OPF 分类器新的学习算法,通过剪掉不相关的样本来减少训练集的大小。实验结果表明,该方法具有较好的分类精度、可观的存储空间收益和较少的分类计算时间。当数据集更大时,这些优势将更加明显。在该领域的任何情况下,该分类器在效率和精度上都比 SVM 具有明显优势	[47]
生物繁殖方面	鱼生殖细胞自动分类	生精周期的监测是保留物种的繁殖非常重要的信息,对开发方法来处理可能的问题也是非常重要的。将物种最先进的监督模式识别技术应用到鱼生殖细胞自动分类,OPF、SVM-RBF 和 SVM-LINEAR 获得了相似的结果,但 OPF 在训练和分类上都获得了更快的速度。这也是首次有人将 OPF 分类器应用于生物环境	[48]
网络安全方面	计算机网络入侵检测	ANN 和 SVM 等算法已广泛应用于入侵检测系统,但是这些算法在学习新攻击时的时间代价非常高,导致其在实时性再训练中难以生存。首次提出用 OPF 分类器进行入侵检测,并证明在公开数据集上 OPF 分类器适合应用于计算机网络的入侵检测,并且其学习新攻击的速度更快	[49]
		提出了一种基于 OPF 分类器的入侵检测方法。对该方法的实验主要包括三个步骤,得到的结果分别是:a)OPF、SVM-RBF、贝叶斯分类器具有相似的结果,OPF 的训练时间与测试时间总和最小;b)将 OPF 的剪枝算法应用于入侵检测领域,减小训练集的大小,使其能够保证实时检测;c)使用了 PSO-OPF、HS-OPF、GSA-OPF 进行特征选择,提高了预测精度,其中 PSO-OPF 的速度最快	[50]
		为了找到 K 近邻 OPF 聚类的合适值,提出一种基于和声搜索的方法,并在两个公开的数据集上表明所提出的方法可以找到合适的 K 值,且基于和声的搜索比传统用穷举法的搜索能更快速地找到合适的 K 值。将其应用于入侵检测中,取得了较好的效果	[51]
电力公司	非技术性损失检测	提出一种基于 OPF 分类器的新的框架以缩减大数据集上的计算代价,以及元启发式算法求解组合优化的代价。在两个公共数据集上的实验结果表明,该框架确实可以提高 OPF 的有效性,大大降低了数据存储成本	[52]
		将 OPF 分类器应用于电力公司快速非技术性损失的识别,并对 OPF、SVM-RBF、SVM-LINEAR 和 ANN-MLP 进行了比较,OPF 和 SVM-RBF 性能相近且优于其他算法,但 OPF 算法要快得多	[53]
		提出用 OPF 分类器进行非技术性损失检测,比较了 OPF、SVM-RBF、SVM-LINEAR、ANN-MLP 和 SOM 神经网络在工业和商业数据集上的效果,OPF 在效率和有效性上优于大多数常用的技术。通过使用剪枝算法,可以最高将训练集减小到 50% 而不影响测试集的分类精度,有时还可能提高其分类精度,从而加速测试分类	[54]
		将 OPF 聚类算法应用于电力公司消费用户是否欺诈的预测,实验在巴西某电力公司的数据集上的行为进行评估,实验结果表明,当 OPF 的参数 kmax 确定在[100 ~ 140]时,OPF 聚类算法的训练及分类速度是最快的	[55]
		提出了一个从数据预处理一直到分类输出的整体框架,分类中使用了 CS-SVM、One-class SVM、OPF、C4.5 分类器相结合的方法,分类精度高于任何单个分类器。结果证明该框架非常有效,节约了时间和费用	[56]
		将 OPF 聚类算法用于识别来自巴西电力公司的商业和工业消费者的不定期和定期的概况,基于 OPF 算法实现了非技术性损失的非监督检测,结果表明,OPF 在无监督的非技术性检测及异常检测上都具有健壮性	[57]

4　结束语

综上所述,OPF 算法是基于图的方法,根据不同的距离函数可以生成不同的最优路径森林。其优点在于:不依赖于任何参数,不需要参数优化;不需要对各类别的形状作任何假设,能够处理多类问题;分类速度快;分类精度高;算法简单易理解。近年来,OPF 算法在理论和应用方面都得到了快速发展。研究结果表明,OPF 算法在分类速度、分类精度、运算复杂性上比其他算法有较大的优势。目前对于最优路径森林算法的理论及应用研究较少,本文的目的正是为了引起更多学者对最优路径森林算法的理论、方法和应用研究的关注。未来 OPF 算法的研究重点主要是与其他算法的结合,设计实现更加高效的、适用于不同领域的分类算法。

参考文献:

[1] Jin Wen, Li Zhaojia, Wei Luosi, *et al*. The improvements of BP neural network learning algorithm[C]//Proc of the 5th International Conference on Signal Processing Proceedings. [S.l.]:IEEE Press,2000:1647-1649.

[2] Zhou Zhihua, Wu Jianxin, Tang Wei. Ensembling neural networks: many could be better than all[J]. Artificial Intelligence,2002,137(1-2):239-263.

[3] Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers[C]//Proc of the 5th Annual Workshop on Computational Learning Theory. New York:ACM Press,1992:144-152.

[4] Platt J C. Fast training of SVMs using sequential minimal optimization [C]//Advances in Kernel Methods. Cambridge: MIT Press,1999:185-208.

[5] Keerthi S S, Shevade S K, Bhattacharyya C, *et al*. Improvements to Platt's SMO algorithm for SVM classifier design[J]. Neural Computation,2001,13(3):637-649.

[6] Dong Jianxiong, Krzyzak A, Suen C Y. A fast SVM training algorithm [J]. International Journal of Pattern Recognition and Artificial Intelligence,2003,17(3):367-384.

[7] Hunt E B, Marin J, Stone P T. Experiments in induction[M]. New York:Academic Press,1966.

[8] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986,1(1):81-106.

[9] Quinlan J R. C4.5:programming for machine learning[M]. San Francisco:Morgan Kauffmann Publisher,1993:27-48.

[10] Mehta M, Agrawal R, Rissanen J. SLIQ: a fast scalable classifier for data mining[C]//Proc of International Conference on Extending Database Technology. Berlin: Springer,1996:18-32.

[11] Olaru C, Wehenkel L. A complete fuzzy decision tree technique[J]. Fuzzy Sets and Systems,2003,138(2):221-254.

[12] 冯少荣.决策树算法的研究与改进[J].厦门大学学报:自然科学版,2007,46(4):496-500.

[13] Kwok S W, Carter C. Multiple decision trees[C]//Proc of the 4th Annual Conference on Uncertainty in Artificial Intelligence. [S. l.]: North-Holland Publishing Co,1990:327-338.

[14] Murphy K P. Naive Bayes classifiers[D]. Vancouver: University of British Columbia,2006.

[15] Duda R O, Hart P E. Pattern classification and scene analysis[M]. New York: Wiley,1973.

[16] Langley P, Iba W, Thompson K. An analysis of Bayesian classifiers [C]//Proc of the 10th National Conference on Artificial Intelligence. 1992:223-228.

[17] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers [J]. Machine Learning,1997,29(2-3):131-163.

[18] Cheng Jie, Greiner R. Comparing Bayesian network classifiers[C]// Proc of the 15th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc,1999:101-108.

[19] Papa J P, Falcão A X, Suzuki C T N, *et al*. A discrete approach for supervised pattern recognition[C]//Proc of International Workshop on Combinatorial Image Analysis. Berlin: Springer,2008:136-147.

[20] Papa J P, Falcão A X, Miranda P A V, *et al*. Design of robust pattern classifiers based on optimum-path forests[C]//Proc of International Symposium on Mathematical Morphology and Its Applications to Image and Signal Proces-sing. 2007:337-348.

[21] Papa J P, Falcão A X. A learning algorithm for the optimum-path forest classifier[C]//Proc of International Workshop on Graph-Based Representations in Pattern Recognition. Berlin: Springer,2009:195-204.

[22] Papa J O P, Falcã A X, De Albuquerque V H C, *et al*. Efficient supervised optimum-path forest classification for large datasets[J]. Pattern Recognition,2012,45(1):512-520.

[23] Papa J P, Falcão A X. A new variant of the optimum-path forest classifier[C]//Proc of International Symposium on Visual Computing. Berlin: Springer,2008:935-944.

[24] Chiachia G, Marana A N, Papa J P, *et al*. Infrared face recognition by optimum-path forest[C]//Proc of the 16th International Conference on Systems, Signals and Image Processing. [S. l.]: IEEE Press,2009:1-4.

[25] Papa J P, Falcao A X, Levada A L M, *et al*. Fast and accurate holistic face recognition using optimum-path forest[C]//Proc of the 16th International Conference on Digital Signal Processing. [S. l.]: IEEE Press,2009:1-6.

[26] Lopes R, Costa K, Papa J. On the evaluation of tensor-based representations for optimum-path forest classification[C]//Proc of IAPR Workshop on Artificial Neural Networks in Pattern Recognition. [S. l.]: Springer International Publishing,2016:117-125.

[27] Montoya-Zegarra J A, Papa J P, Leite N J, *et al*. Novel approaches for exclusive and continuous fingerprint classification[C]//Proc of Pacific-Rim Symposium on Image and Video Technology. Berlin: Springer,2009:386-397.

[28] Iliev A I, Scordilis M S, Papa J P, *et al*. Spoken emotion recognition through optimum-path forest classification using glottal features[J]. Computer Speech & Language,2010,24(3):445-460.

[29] Papa J P, Marana A N, Spadotto A A, *et al*. Robust and fast vowel recognition using optimum-path forest[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. [S. l.]: IEEE Press,2010:2190-2193.

[30] Lopes G S, De Silva D C, Rodrigues A W O, *et al*. Recognition of handwritten digits using the signature features and optimum-path forest classifier[J]. IEEE Latin America Transactions,2016,14(5): 2455-2460.

[31] Nakamura R, Pereira L, Silva D, *et al*. Fast robot voice interface through optimum-path forest[C]//Proc of the 16th International Conference on Intelligent Engineering Systems. [S. l.]: IEEE Press, 2012:67-71.

[32] Pisani R J, Papa J P, Zimback C R L, *et al*. Land use classification using optimum-path forest[C]//Proc of the 14th Brazilian Symposium on Remote Sensing. 2009:7063-7070.

[33] Pisani R, Riedel P, Ferreira M, *et al*. Land use image classification through optimum-path forest clustering[C]//Proc of IEEE International Geoscience and Remote Sensing Symposium. 2011:826-829.

[34] Osaku D, Pereira D R, Levada A L M, *et al*. Fine-tuning contextual-based optimum-path forest for land-cover classification[J]. IEEE Geoscience and Remote Sensing Letters,2016,13(5):735-739.

[35] Osaku D, Levada A L M, Papa J P. A block-based Markov random field model estimation for contextual classification using optimum-path forest[C]//Proc of IEEE International Symposium on Circuits and Systems. 2016:994-997.

[36] Guilherme I R, Marana A N, Papa J P, *et al*. Fast petroleum well drilling monitoring through optimum-path forest[J]. Journal of Next Generation Information Technology,2010,1(1):77-85.

[37] Marques C M, Guilherme I R, Nakamura R Y M, *et al*. New trends in musical genre classification using optimum-path forest[C]//Proc of the 12th International Society for Music Information Retrieval Conference. 2011:699-704.

[38] Papa J P, Spadotto A A, Falcao A X, *et al*. Optimum path forest classifier applied to laryngeal pathology detection[C]//Proc of the 15th International Conference on Systems, Signals and Image Processing. [S. l.]: IEEE Press,2008:249-252.

[39] Cappabianco F A M, Falcao A X, Rocha L M. Clustering by optimum path forest and its application to automatic GM/WM classification in MR-T1 images of the brain[C]//Proc of the 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 2008:428-431.

[40] Cappabianco F A M, Falcão A X, Yasuda C L, *et al*. Brain tissue MR-image segmentation via optimum-path forest clustering[J]. Computer Vision and Image Understanding,2012,116(10):1047-1059.

[41] Spadoto A A, Guido R C, Papa J P, *et al*. Parkinson's disease identification through optimum-path forest[C]//Proc of Annual International Conference of the IEEE Engineering in Medicine and Biology. 2010:6087-6090.

[42] Spadoto A A, Guido R C, Carnevali F L, *et al*. Improving Parkinson's disease identification through evolutionary-based feature selection[C]//Proc of Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2011:7857-7860.

[43] Cappabianco F, Ide J S, Falcão A, *et al*. Automatic subcortical tissue segmentation of MR images using optimum-path forest clustering [C]//Proc of the 18th IEEE International Conference on Image Processing. 2011:2653-2656.

[44] Albuquerque V H C, Papa J P, Falcão A X, *et al*. Application of optimum-path forest classifier for synthetic material porosity segmentation [C]//Proc of the 17th International Conference on Systems, Signals and Image Processing. 2010:1-4. (下转第23页)

[46] Natarajan S, Sehar S. A novel algorithm for distributed data mining in HDFS[C]//Proc of the 5th International Conference on Advanced Computing. 2013.

[47] Liao Jinggui, Zhao Yuelong, Long Saiqin. MRPrePost:a parallel algorithm adapted for mining big data[C]//Proc of IEEE Workshop on Electronics, Computer and Applications. 2014:564-568.

[48] Deng Zhihong, Wang Zhonghui, Jiang Jiajian. A new algorithm for fast mining frequent itemsets using N-lists[J]. Sciece China Information Sciences,2012,55(9):2008-2030.

[49] Pei Jian, Han Jiawei, Lu Hongjun, *et al*. H-mine: fast and space-preserving frequent pattern mining in large databases[J]. IIE Transactions,2007,39(6):593-605.

[50] Pasquier N, Bastide Y, Taouil R, *et al*. Discovering frequent closed itemsets for association rules[C]// Proc of International Conference Database Theory. 1999:398-416.

[51] Han Jiawei, Pei Jian, Yin Yiwen. *et al*. Mining frequent pattern without candidate generation[C]//Proc of ACM SIGMOD International Conference on Management of Data. 2003.

[52] Wang Jianyong, Han Jiawei, Pei Jian. CLOSET +: searching for the best strategies for mining frequent closed itemsets[C]//Proc of Internationall Conference Knowledge Discovery and Data Mining. 2003: 236-245.

[53] Grahne G, Zhu Jianfei. Efficiently using prefix-trees in mining frequent itemsets[C]//Proc of IEEE ICDM Workshop on Frequent Itemset Mining Implementations. 2003.

[54] Zaki M J, Hsiao C. Charm: an efficient algorithm for closed itemset mining[C]//Proc of SIAM International Conference Data Mining. 2002:57-473.

[55] Lucchese C, Orlando S. Perego R. Fast and memory efficient mining of frequent closed itemsets[J]. IEEE Trans on Knowledge and Data Engineering,2006,18(1):21-35.

[56] Cheung D W, Han Jia, Ng V T, *et al*. Maintenance of discovered association rules in large databases: an incremental updating technique [C]//Proc of the 12th IEEE International Conference on Data Engineering. 1996:106-114.

[57] Cheung D W, Lee S D, Kao B. A general incremental technique for maintaining discovered association rules[C]//Proc of the 5th International Conference on Database for Advanced Applications. 1997:185-194.

[58] Ayan N F, Tansel A U, Arkun E. An efficient algorithm to update large itemsets with early pruning[C]//Proc of SIGKDD. 1999:287-291.

[59] Srikant R, Agrawal R. Mining sequential patterns: generalizations and performance improvements[C]//Proc of the 5th Conference on Extending Database Technology. 1996:3-17.

[60] Hong T P, Wang C Y, Tao Yuhui. A new incremental data mining algorithm using pre-large itemsets[J]. Intelligent Data Analysis, 2001,5(2):111-129.

[61] Koh J L, Shieh S F. An efficient approach for maintaining association rules based on adjusting FP-Tree structures[C]//Proc of Database Systems for Advanced Applications. 2004:417-424.

[62] Cheung W, Zaïane O R. Incremental mining of frequent-patterns without candidate gneration or support constraint[C]//Proc of International Database Engineering and Applications Symposium. 2003: 111-116.

[63] Leung C K S, Khan Q I, Hoque T, *et al*. CanTree:a tree structure for efficient incremental mining of frequent patterns[C]//Proc of the 5th IEEE InternationalConference on Data Mining. 2005.

[64] Totad S G, Geeta R B, Reddy P P. Batch processing for incremental FP-tree construction[J]. International Journal of Computer Applications, 2011,5(5):28-32.

[65] Han Jiawei, Kamber M, Pei Jian, *et al*. 数据挖掘:概念与技术[M]. 范明,孟小峰译. 3 版. 北京:机械工业出版社,2012:186-188.

(上接第 12 页)

[45] Papa J P, De Albuquerque V H C, Falcão A X, *et al*. Fast automatic microstructural segmentation of ferrous alloy samples using optimum-path forest[C]//Proc of International Symposium Computational Modeling of Objects Represented in Images. Berlin: Springer,2010:210-220.

[46] Freitas G M, Avila A M H, Papa J P, *et al*. Optimum-path forest-based rainfall estimation[C]//Proc of the 16th International Conference on Systems, Signals and Image Processing. 2009:1-4.

[47] Papa J P, Falcao A X, De Freitas G M, *et al*. Robust pruning of training patterns for optimum-path forest classification applied to satellite-based rainfall occurrence estimation[J]. IEEE Geoscience and Remote Sensing Letters,2010,7(2):396-400.

[48] Papa J P, Gutierrez M E M, Nakamura R Y M, *et al*. Automatic classification of fish germ cells through optimum-path forest[C]// Proc of Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2011:5084-5087.

[49] Pereira C, Nakamura R, Papa J P, *et al*. Intrusion detection system using optimum-path forest[C]//Proc of the 36th Conference on Local Computer Networks. 2011:183-186.

[50] Pereira C R, Nakamura R Y M, Costa K A P, *et al*. An optimum-path forest framework for intrusion detection in computer networks [J]. Engineering Applications of Artificial Intelligence,2012,25 (6):1226-1234.

[51] Costa K, Pereira C, Nakamura R, *et al*. Boosting optimum-path forest clustering through harmony search and its applications for intrusion detection in computer networks[C]//Proc of the 4th International Conference on Computational Aspects of Social Networks. 2012: 181-185.

[52] Costa K A P, Pereira L A M, Nakamura R Y M, *et al*. A nature-inspired approach to speed up optimum-path forest clustering and its application to intrusion detection in computer networks[J]. Information Sciences,2015,294(C):95-108.

[53] Ramos C C O, Souza A N, Papa J P, *et al*. Fast non-technical losses identification through optimum-path forest[C]//Proc of the 15th International Conference on Intelligent System Applications to Power Systems. 2009:1-5.

[54] Ramos C C O, De Sousa A N, Papa J P, *et al*. A new approach for nontechnical losses detection based on optimum-path forest[J]. IEEE Trans on Power Systems,2011,26(1):181-189.

[55] Ramos C C O, Souza A N, Nakamura R Y M, *et al*. Electrical consumers data clustering through optimum-path forest[C]//Proc of the 16th International Conference on Intelligent System Application to Power Systems. 2011:1-4.

[56] Júnior L A P, Ramos C C O, Rodrigues D, *et al*. Unsupervised non-technical losses identification through optimum-path forest[J]. Electric Power Systems Research,2016,140:413-423.

[57] Di Martino M, Decia F, Molinelli J, *et al*. A novel framework for nontechnical losses detection in electricity companies[M]//Pattern Recognition-Applications and Methods. Berlin: Springer, 2013: 109-120.