基于通用距离测量的机器学习方法用于图像分类和聚类

赵 勇 李怀宇
(长安大学信息工程学院 西安 710064)

摘 要：针对图像分类识别问题，提出了一种用于图像特征提取的新方法。首先定义了基于图像字符串的复杂度和以及通用图像距离(UID)，然后依次提出了测量通用图像距离的 UID 距离测量算法，在维持特征类别之间的固有差异条件下对图像原型进行选择的原型选择算法，利用原型选择算法创建图像的特征向量表示从而生成待分类图像的特征向量的特征向量生成算法，最后基于前述算法提出了对图像的感兴趣区域进行分离的图像分类学习算法。将所提出的方法应用于卫星图像数据的几个监督和非监督学习实验，结果表明文中所提方法效果理想。

关键词：通用图像距离；特征提取；特征向量；图像分类
中图分类号：TN949；TP301.6 **文献标识码**：A **国家标准学科分类代码**：070104

Machine learning methods based on universal distance measurement for image classification and clustering

Zhao Yong Li Huaiyu
(College of Information Engineering, Chang′an University, Xi′an 710064, China)

Abstract: In this paper, a new method for image feature extraction is proposed based on the image classification recognition problem. Firstly defining the complexity of the image string and the universal image distance (UID), and then proposing a UID distance measurement algorithm to measure the general image distance, a prototype selection algorithm for selecting the image prototype under the inherent difference between the maintenance feature categories, a feature vector generation algorithm to generate the eigenvector of the image to be classified by using the prototype selection algorithm to create the feature vector of the image to be classified, and finally an image classification learning algorithm to separate the region of interest of the image based on the aforesaid algorithms. The proposed method is applied to several supervised and unsupervised learning experiments of satellite image data. The results show the feasibility of the proposed method.

Keywords: universal image distance; feature-extraction; feature vector; image classification

0 引 言

图像分类研究的目的是找到一种图像的表示，使得这些图像可以被自动分类成有限的类[1-6]。通常分类图像的算法需要在进行分类之前，对图像进行某种形式的预处理。该过程涉及提取相关特征，且基于一些现有知识将图像分割成子分量[7-8]。为了使分类准确，需要借助于许多方法和算法以及期望分类的图像所具有的任何边缘信息。目前还没有通用的最佳算法能保证对所有类型的问题产生高的分类率[9-10]。

然而，在图像识别领域中，最近已经有了新的字符串距离度量方法，其可以通过比较任何两个字符串来提取图像特征[11-12]，而无需对它们的背景进行任何假设。该方法已经被证明在数据聚类和模式分类的各种模式识别任务上是非常成功的。

在本文引入了一个新的距离函数，称为通用图像距离(UID)，用于测量两个图像之间的距离。UID 首先将两个图像中的每一个转换成来自有限字母表的字符串，然后使用字符串距离来给出图像之间的距离值。两个字符串 x 和 y 之间的距离是字符串的级联 xy 的复杂度与 x 和 y 中的每一个最小复杂度之间的归一化差异，本文的复杂度是指 LZ(Lempel-Ziv)复杂度[13]。本文使用 UID 来创建图像的

收稿日期：2017-02

有限维表示,UID的优点之一是其可比较不同大小的两个图像之间的距离。

1　LZ-复杂度和字符串距离

UID距离函数是基于字符串的LZ复杂度,该复杂度的定义如下[14]:假设S、Q和R是在字母表A上定义的字符串。由$l(S)$表示S的长度,$S(i)$表示S的第i个元素,$S(i,j)$表示S的子串,其由位置i和j(包括其本身)之间的S字符组成。若存在整数$p\leqslant l(S)$,使得对于$k=1,2,\cdots,l(Q)$,$Q(k)=R(p+k-)$,则S的扩展$R=SQ$可以由S再现(表示为$S\rightarrow R$)。例如,对于aacgt → aacgtcgtcg,$p=3$以及对于aacgt → aacgtac,$p=2$。通过首先复制所有的S,然后以顺序方式从S的第p个位置开始复制$l(Q)$个元素,以获得R的Q部分。

如果$S(1,j)\rightarrow S(1,l(S)-1)$,则可由$S(1,j)$再现字符串$S$(表示为$S(1,\mathrm{j})\Rightarrow \mathrm{R}$)。例如,aacgt → aacgtac和aacgt → aacgtacc均具有$p=2$。在复制过程结束时添加额外的"不同"字符,这在复制中是不允许的。

任何字符串S均可使用再现过程来构建,在其第i步具有再现$S(1,h_{i-1})\Rightarrow S(1,h_i)$,其中$h_i$是字符在第$i$步骤的位置,注意$S(1,0)\Rightarrow S(1,1)$。

S的m步再现过程实现对S的解析,其中$H(S)=S(1,h_1)\cdot S(h_1+1,h_2)\cdot S(h_{m-1}+1,h_m)$被称为$S$的历史,且$H_i(S)=S(h_{i-1}+1,h_i)$被称为$H(S)$的第$i$个分量。例如对于$S=$aacgtacc,有$H(S)=\mathrm{a}\cdot\mathrm{ac}\cdot\mathrm{g}\cdot\mathrm{t}\cdot\mathrm{acc}$作为$S$的历史。

若$S(1,h_i)$不能由$S(1,h_{i-1})$再现,则分量$H_i(S)$被称为穷尽的,意味着复制过程不能继续。每个字符串S均有一个独特的穷举历史[15]。

用$c_H(S)$表示S历史中的分量数目。S的LZ复杂度是$c(S)=\min\{c_H(S)\}$,其中最小值在S的所有历史中取得。可以看出,$c(S)=c_E(S)$,其中$c_E(S)$是S的穷尽历史中分量的数量。

在文献[14]中引入了基于LZ复杂度的字符串的距离,并定义如下:给定两个字符串X和Y,通过XY表示其的级联,然后定义

$$d(X,Y):=\max\{c(XY)-c(X),c(YX)-c(Y)\}。$$

在文献[12]中,发现以下归一化距离在图像的分类和聚类中表现良好:

$$d(X,Y):=\frac{c(XY)-\min\{c(X),c(Y)\}}{\max\{c(X),c(Y)\}}$$

d不是度量,因为其不满足三角不等式,且距离为0意味着两个字符串接近但不一定相同。将d称为通用距离,因为其不依赖于字符串的某个特定表示,也不取决于与其他字符串距离(例如编辑距离)相同的启发法。d只取决于两个字符串中每一个LZ复杂度及其级联,这是一个纯粹的信息量,取决于字符串的内容,而并未其的表示形式。

2　通用图像距离

基于d,现在定义图像之间的距离。该想法是将两个图像I和J中的每一个从有限符号字母表转换成字符串$X^{(I)}$和$X^{(J)}$。一旦形成字符串格式,使用$d(X^{(I)},X^{(J)})$作为I和J之间的距离。这一过程的细节在下面的算法1中进行描述。

算法1:UID距离测量

1)输入:两个彩色图像I、J,jpeg格式(RGB表示);

2)通过根据以下公式形成R、G和B分量的加权和,将RGB矩阵转换为灰度:灰度值$=0.2989\mathrm{R}+0.5870\mathrm{G}+0.1140\mathrm{B}$(使用MATLAB),每个像素现在是一个单一的数值,范围为0～255,并将这组值称为字母表,用A表示;

3)从左上到右下扫描每个灰度图像,并从A形成一串符号,用$X^{(I)}$和$X^{(J)}$表示两个字符串;

4)计算LZ复杂度:$c(X^{(I)})$、$c(X^{(J)})$与其级联的复杂度$c(X^{(I)}X^{(J)})$;

5)输出:$UID(I,J):=d(X^{(I)},X^{(J)})$。

3　原型选择

接下来描述从每个特征类别中选择图像原型的算法,该过程在将图像转换成有限维向量的阶段之前仅运行一次。对于图像I,通过表示I的子图像P,其中P可以是通过在图像I上放置窗口而获得的任何矩形图像,而窗口则完全由I包围。窗口的大小取决于多少单个原型将传达关于相关特征类别的信息。

在下面的算法中,使用聚类作为验证的简单手段,所选择的原型维持特征类别之间的固有差异(聚类算法没有给出特征类别信息,而仅给出原型间距离信息)。

算法2:原型选择

1)输入:M个图像特征类别,以及N个未标记的彩色图像的语料库C_N:$\{I_j\}$,$j=1,\cdots,N$;

2) for (i: = 1 *to* M) do

(1)基于C_N中的任何图像I_j,让用户选择L_i原型图像$\{P_k^{(i)}\}_{k=1}^{L_i}$,并将其设定为特征类别$i$,每个原型被一些图像包含,$P_k^{(i)}\subset I_j$,且$P_k^{(i)}$可以变化,尤其是其可远小于图像$I_j(1\leqslant j\leqslant N)$的大小,(2)结束;

3)将所有原型枚举为单个未标记集$\{P_k\}_{k=1}^{L}$,其中$L=\sum_{i=1}^{M}L_i$,并计算距离矩阵$\boldsymbol{H}=[UIS(X^{(P_k)},X^{(P_l)})]_{k=1,l=1}^{L}$,其中$\boldsymbol{H}$的$(k,l)$分量是未标记原型$P_k$和$P_l$之间的UID距离;

4)在$\boldsymbol{H}$上运行层次聚类并获得相关的树形图(注意:$\boldsymbol{H}$不包含关于特征类别的任何"标记"信息,因为其基于未标记集);

5)如果存在具有由原型$\{P_k^{(i)}\}_{k=1}^{L_i}$构成的第$i$个簇的$M$个簇,则终止并转到步骤7);

6)否则转到步骤 2);

7)输出:标记原型的集合 $P_L:=\{\{P_k^{(i)}\}_{k=1}^{L_i}\}_{i=1}^{M}$,其中 L 是原型数量。

从学习模式识别的理论可知,特征向量的维数 M 通常被认为比数据大小 N 小。大的 L 将获得图像更好的特征表示精度,但却将增加运行算法的时间。算法 2 的收敛是基于用户选择良好原型图像的能力。因为 UID 允许选择远小于全图像 I_j 大小的原型 $P_k^{(i)}$,因此在本研究中,对于所有特征类别使用 45×70 像素原型大小。这使得用户可轻易的从每个特征类别快速选择典型的代表性原型,即当原型来自不同的特征类别时其是远离的,且当它们来自相同的特征类别时其是接近的。因此,算法 2 通常快速收敛。

作为示例,图 1 显示了由用户从卫星图像的语料库选择的 12 个原型。用户标记原型 1,…, 3 作为特征类别城市的表示;原型 4, …, 6 作为海的表示;原型 7, …, 9 作为道路的表示;原型 10, …, 12 作为旱地的表示。用户便于找到这样的代表性原型,因为其适合于尺寸为 45×70 像素的单个图片(典型图像)。在图 2 中显示了,对于这 12 个原型的集合在算法 2 的步骤 4 中产生的树形图。可以看出,发现了以下 4)个聚类:{10,12,11},{1,2,3},{7,8,9},{4,6,5},这表示在算法 2 中选择的原型是良好的。

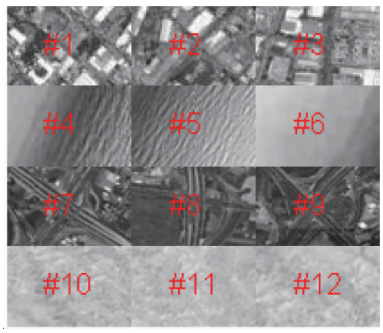

图 1　特征类别城市、海洋、道路和旱地的标签原型(每个特征具有 3 个原型)

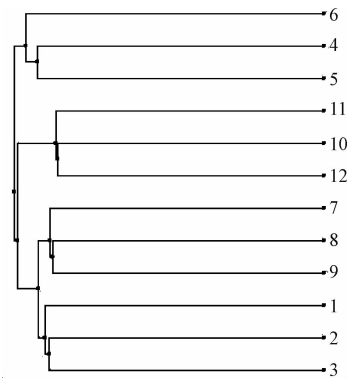

图 2　图 1 中原型的树形图

4　图像特征表示

现在利用算法 2 来创建图像的特征向量表示。其被描述为下面的算法 3。

算法 3:特征向量生成

1)输入:要在特征类别 $1\leqslant i\leqslant M$ 上表示的图像 I,并给出标签原型图像的集合

$P_L:=\{\{P_k^{(i)}\}_{k=1}^{L_i}\}_{i=1}^{M}$(由算法 2 获得);

2)初始化计数变量 $c_i:=0$, $1\leqslant i\leqslant M$;

3)设 W 为尺寸等于最大原型尺寸的矩形;

4)以不重叠的方式从左上到右下跨过 I 扫描窗口 W,并使所获得的 I 的子图像序列表示为 $\{I_J\}_{j=1}^{m}$;

5)for($j:=1$ to m) do

a)for ($i:=1$ to M) do

(1)　i. temp $:=0$

ii. for ($k:=1$ to L_i) do

A.

$temp:=temp+(UID(I_j,P_k^{(i)},P_k^{(i)}))^2$

B. 结束

iii. $r_i:=\sqrt{temp}$

iv. 结束

(2)令 $i*(j):=argmin_{1\leqslant i\leqslant M}r_i$,这是所确定的子图像的特征类别 I_j;

(3)递增计数 $c_j*(j):=c_i*(j)+1$

(4)结束;

6)归一化计数,$v_i:=\dfrac{c_i}{\sum_{l=1}^{M}c_i}$, $1\leqslant i\leqslant M$

7)输出:作为图像 I 的特征向量表示的归一化向量 $v(I)=[v_1,\cdots,v_M]$。

5　监督和无监督学习的图像

给定图像的语料库 C 和一组标记的原型,使用算法 3 来生成对应于 C 中的每个图像 I 的特征向量 $v(I)$。在这一点上,具有大小等于 C 的数据库 D 由 C 中的所有图像的特征向量组成。该数据库可用于无监督学习,例如发现感兴趣的图像簇,也可用于监督学习。让用 T 表示类目标变量,数据库 D_T 由 D 的特征向量与相应目标类值组成。以下算法描述了图像的学习分类的过程。

算法 4:图像分类学习

1)输入:(1)目标类别变量 T,(2)数据库 D_T,(3)任何监督学习算法 A;

2)使用 n 重交叉验证将 D_T 分成训练和测试集;

3)训练和测试算法 A 并产生分类器 C,其将特征空间 $[0,1]^M$ 映射到 T 中;

4)定义图像分类器如下:对于任何图像,分类为 $F(I):=C(v(I))$,其中 $v(I)$ 是 I 的 M 维特征向量;

5)输出:分类器 F。

6　实验设置和结果

从各种类型区域的 GoogleEarth 创建了大小为 670×1 364像素的 60 个图像语料库 C。图 3 显示了这种图像的一些按比例缩小的示例。从这些图像中,让用户定义 4 个特征类别:海、城市、旱地和道路,并从每个特征类别中选择尺寸为 45×70 像素的 3 个相对较小的图像原型,即在 $M=4$ 和 $L_i=3$ 条件下运行算法 2。然后运行算法 3 为语料库中的每个图像生成特征向量,并获得数据库 D。再让用户通过目标变量标记图像湿度,可能值为 0 或 1。如果区域为低湿度,则图像标记为 0;若湿度较高,则标记为 1。注意到,低湿度区域的图像可以在旱地(干燥)区域中,或者也可在并非是旱地的高海拔地区。由于在用户已经选择的特征类别中高程信息不可用,所以分类问题是困难的,因为学习算法需要发现潮湿区域和仅由上述 4 个特征类别表征的区域之间的依赖性。有了标签信息,产生了标签数据库 $D_{潮湿}$,使用算法 4 来学习具有目标几何的图像分类器。作为学习算法 A,使用以下标准的监督算法:J48、CART、NaiveBayes 和多层感知器(反向传播),所有这些均在 WEKA ©工具包中提供。进行 10 重交叉验证,并将其准确性与基线分类器(表示为 ZeroR)进行比较,其具有与最高先验经验概率的类别值相应的单个决定。J48、CART、NaiveBayes 和反向传播分别以 86.5%、81.5%、89.25%和 87.25%的精确度进行,而基线 ZeroR 分类器实现 50%的精确度。因此,基于显著性水平为 0.05 的 T 检验得出,这几种学习算法均显著优于基线分类器。

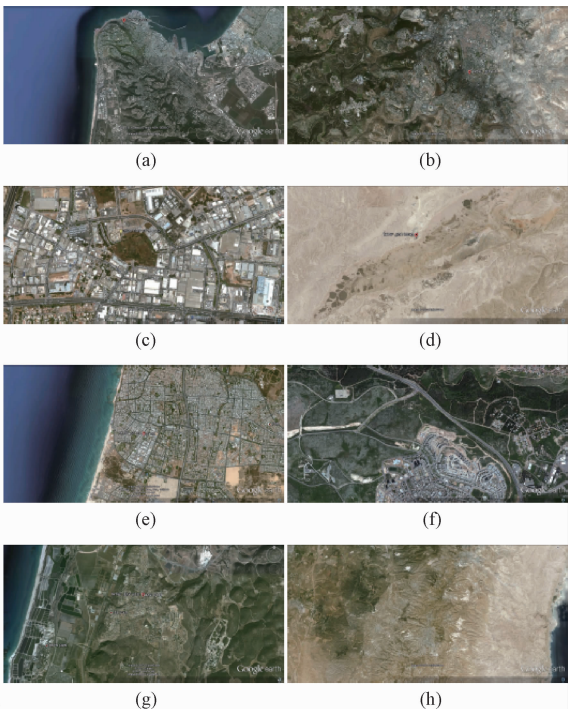

图 3　语料库中的图像示例

7　结　　论

本文介绍了一种用于自动定义和测量彩色图像的特征方法。该方法基于通过计算两个图像的字符串表示及其级联的复杂度而测量的通用图像距离。图像由特征向量表示,该特征向量表示从图像到由用户定义的小图像原型的固定集合的距离。无需要任何复杂的基于数学的图像分析或预处理,因为通用图像距离将图像视为包含图像的所有相关信息的符号串。本方法的简单性使得其对于快速和可缩放的实现具有一定的吸引力。通过将本文的方法应用于卫星图像上的监督和非监督机器学习,结果表明基于图像的特征向量表示,标准机器学习算法执行良好。

参考文献

[1] YU J, RUI Y, TANG Y Y, et al. High-order distance-based multiview stochastic learning in image classification[J]. IEEE Transactions on Cybernetics, 2014, 44(12):2431.

[2] 鲁萌萌,赵凤军,李宁. 基于词包模型的高分辨率 SAR 图像特征提取[J]. 国外电子测量技术,2015,34(6):62-69.

[3] 高炜欣,胡玉衡,武晓朦,等. 埋弧焊 X 射线焊缝缺陷图像分类算法研究[J]. 仪器仪表学报,2016,37(3):518-524.

[4] 安健,张扬. 基于 Otsu 和模糊核聚类算法的极化 SAR 图像分类[J]. 电子科技,2014,27(2):42-45.

[5] VU T H, MOUSAVI H S, MONGA V, et al. Histopathological image classification using discriminative feature-oriented dictionary learning[J]. IEEE Transactions on Medical Imaging, 2016, 35(3):738-751.

[6] SAMAT A, LI J, LIU S, et al. Improved hyperspectral image classification by active learning using pre-designed mixed pixels [J]. Pattern Recognition, 2016, 51(C):43-58.

[7] 张靖. 面向高维小样本数据的分类特征选择算法研究[D]. 合肥:合肥工业大学,2014.

[8] SPANHOL F A, OLIVEIRA L S, PETITJEAN C, et al. A dataset for breast cancer histopathological image classification [J]. IEEE Transactions on Biomedical Engineering, 2016, 63(7):1455-1462.

[9] SLAVKOVIKJ V, VERSTOCKT S, NEVE W D, et al. Unsupervised spectral sub-feature learning for hyperspectral image classification [J]. International Journal of Remote Sensing, 2016, 37(2):309-326.

[10] KAMAROL S K A, JAWARD M H, PARKKINEN J, et al. Spatiotemporal feature extraction for facial expression recognition [J]. IET Image Processing,

2016, 10(7):534-541.

[11] YUAN F, SHI J, XIA X, et al. High-order local ternary patterns with locality preserving projection for smoke detection and image classification [J]. Information Sciences, 2016, 372(C):225-240.

[12] 詹曙,姚尧,高贺. 基于随机森林的脑磁共振图像分类[J]. 电子测量与仪器学报,2013, 27(11):1067-1072.

[13] HUDETZ D A G, LIU X, PILLAY S, et al. Propofol anesthesia reduces Lempel-Ziv complexity of spontaneous brain activity in rats[J]. Neuroscience Letters, 2016, 628(8):132-135.

[14] OUT H H, SAYOOD K. A new sequence distance measure for phylogenetic tree construction [J]. Bioinformatics, 2003, 19(16):2122-2130.

[15] LIMNIOTIS K, KOLOKOTRONIS N, KALOUPTSIDIS N. On the nonlinear complexity and lempel-ziv complexity of finite length sequences[J]. IEEE Transactions on Information Theory, 2007, 53(11):4293-4302.

作者简介

赵勇 1978 年出生,硕士,工程师,研究方向为计算机视觉、软件工程。

李怀宇,1973 年出生,博士研究生,工程师,研究方向为计算机视觉、智能交通。