●余肖生 周 宁 张芳芳

基于 KNN 的图像自动分类模型研究*

摘 要 所谓图像自动分类是指利用图像自动分类器把待分类的图像分配到预定义的图像类的过程。用于图像自动分类的方法有多种。其中 K 近邻算法是一种基于实例学习的方法,是一种较理想的自动分类器。本文在它的基础上提出了图像自动分类模型,整个图像自动分类过程包括图像预处理、特征表示、机器学习和图像分类 4 个步骤。表 1。图 1。参考文献 13。

关键词 图像自动分类 K 近邻算法 支持向量机 贝叶斯分类法
分类号 TP391.41

ABSTRACT Automatic image categorization is a process of categorizing images into defined categories by using automatic image categorizers. There are many methods for the automatic categorization of images, among which the K-Nearest Neighbor algorithm is a case-based learning method and is a comparatively ideal automatic categorizer. On the basis of the K-Nearest Neighbor algorithm, the authors proposes a model for the automatic categorization of images, including the preprocessing, characteristic presentation, machine learning and categorization of images. 1 tab. 1 fig. 13 refs.

KEY WORDS Automatic image categorization. K-nearest neighbor. Support vector machine. Bayes categorization.
CLASS NUMBER TP391.41

互联网上各种各样的图像与日俱增,人们需要一些新的方法存取它们。传统的数据库仅允许基于元数据的文本查询,将图像作为一个文件存储在文件系统中,而数据库模式仅保持这些图像资源的引用。图像可能包含一些不能通过文本描述来传递的语义信息。按传统方法检索,结果往往不尽人意。Niblack 等人提出 QBE(Query By Example)模式,用图像的像替代文本来描述图像,用这种方式检索的结果是数据库中与给定查询图像相似的那些图像[1]。而这种模式成功的关键在于图像的分类。MeSH[2]、UMLS[3]对图像分类进行了有益探索,并取得了一定成绩,但这些分类都是建立在手工基础上的,效率比较低。人们于是开始图像自动分类方面的研究,提出了一些实用模型,如人工神经网络与决策树相结合模型[4],小波变换和马尔可夫随机场模型等[5],但这些模型的自动分类效果不是很理想。本文从图像自动分类的方法着手,通过图像自动分类方法的比较实验,可知 KNN 是一种较理想的自动分类器。在此基础上,笔者提出了基于 KNN 的图像自动分类模型。

1 图像自动分类方法

所谓图像自动分类是指利用图像自动分类器将待分类的图像分配到预定义的图像类的过程。用于图像自动分类的方法主要有 K 近邻算法、支持向量机、贝叶斯算法等[6-9]。

1.1 K 近邻算法(KNN:k-Nearest Neighbor)

K 近邻算法是一种基于实例学习的方法。它的主要思想是训练样本用 n 维数值属性描述,每个样本代表 n 维空间的一个点,所有样本存放于 n 维模式空间中,给定一个未知样本。该算法搜索模式空间,找出最接近的未知样本的 k 个训练样本作为未知样本的近邻,未知样本被分配到 k 个最邻近者中的最公共的类。一个实例的最近邻是根据距离来定义的,即把任意的实例 x 表示为下面的特征向量:

$$\langle v_1(x), v_2(x), \cdots, v_n(x) \rangle$$

其中 $v_i(x)$ 表示实例 x 的第 i 个属性值,则两个实例 x_i, x_j 之间的距离定义为 $d(x_i, x_j)$,其中最常用的计算距离的方法是欧氏距离:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (v_k(x_i) - v_k(x_j))^2}$$

1.2 支持向量机(SVM:Support Vector Machine)

支持向量机是由 Boser 等人于 1992 年提出的一种基于统计学习理论的模式识别方法。它的基本思想是找到一个超平面,能够尽可能多地将两类数据点

* 本文系国家自然科学基金项目(70473068)和教育部社会科学研究重大课题攻关项目(05JZD0024)的研究成果之一。

正确分开,并使分开的两类数据点距离分类面最远。可以自动寻找那些对分类有较好区分能力的支持向量,由此构造出的分类器可以将类与类的间隔最大化,因而有较高的分类准确率。经过10多年的发展,现今它在文本分类、图像分类、手写识别、生物信息学等领域获得较好应用。

1.3　贝叶斯分类法

贝叶斯分类法是统计模式识别中的一个经典方法,它是通过运用Bayes公式计算后验概率来实现模式分类。

假设有n个模式分类 $\omega_1,\omega_2,\cdots\cdots,\omega_n$;X是某一待识别模式。令

$P(\omega_i)$:模式属于 ω_i 的先验概率;

$P(\omega_i/X)$:当给定输入模式属于 ω_i 类时,模式X出现的条件概率;

$P(\omega_i/X)$:当给定输入模式X时,该模式属于 ω_i 类的后验条件概率。

由Bayes公式,可得:

$$P(\omega_i/X)=\frac{P(X/\omega_i)\times P(\omega_i)}{\sum_{i=1}^{n}P(X/\omega_i)\times P(\omega_i)}$$

后验概率是一种客观概率,它表明随机试验中事件发生的相对频率,值越大,表示的相对频率越高。因此,若存在 $i\in\{1,2,\cdots\cdots,n\}$,使得对所有的 $j(j\neq i)$ 均有

$$P(\omega_i/X)>P(\omega_j/X)$$

则 $X\in\omega_i$。

在此基础上,现今已经衍生出多种相近的方法,如朴素贝叶斯分类方法、多项朴素贝叶斯分类方法等。

此外还有决策树、Rocchio等算法。为了知道上述哪个分类器更好一些,Yang和Liu对一些分类器进行对比实验[10],结果如下:

$$\{SVM,KNN\}>LLSF>>MNB$$

其中,SVM表示支持向量机,KNN表示K近邻算法,LLSF(Linear Least Square Fit)表示线性最小二乘方拟合,MNB(Multinomial Naive Bayes)表示多项朴素贝叶斯分类。

在Joachims进行的另一项比较实验中,得出了如下结果[11]:

SVM(0.864) > KNN(0.823) > C4.5(0.794) > naive Bayes(0.72)

其中,SVM,KNN含义同上,C4.5是决策树算法之一,naive Bayes表示朴素贝叶斯分类。

KNN和SVM是这些分类器中最好的,而KNN算法有易于实现和高效等特点。本文构建图像自动分类模型时采用KNN算法。

2　基于KNN的图像自动分类模型

本文在KNN的基础上提出了图像自动分类模型(如图1所示)。整个图像自动分类过程大致包括以下几个步骤。

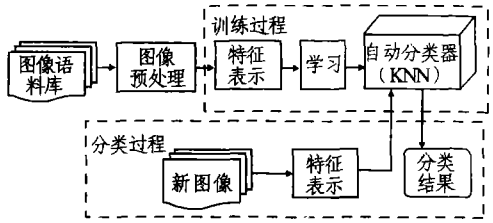

图1　基于KNN的图像自动分类模型

(1)图像预处理。因为图像输入设备的限制,输入的图像可能会出现一些"失真"现象。为了达到更好效果,有必要进行一些预处理,主要方法有:傅立叶变换、直方图变换和灰度变换等。

(2)特征表示。从图像的像素中提取图像的各个特征向量(颜色、纹理、形状),其中颜色特征向量可从色度直方图中得到,纹理特征向量可从共生矩阵中获得,而形状特征向量可从边界方向直方图和区域内部的不变矩中得到。图像特征的表示在自动分类中占有相当重要地位,直接关系到分类的正确率。

(3)机器学习。通过相关规则的学习,有效提高自动分类器的准确性。

(4)图像分类。对于一个新图像,计算出它们的特征向量 Do 与图像语料库中各个图像的特征向量 $d_i(i=1,2,\cdots\cdots,s)$ 的相似度,其中图像语料库中含有n类s个图像:设定阈值,得到 $m(m<s)$ 个与该图像距离最近的图像,这m个图像属于 $k(k<n)$ 类;依次计算这k类中每类的权重;计算新图像的特征向量 Do 与比较向量 $D_j(j=1,2,\cdots\cdots,k)$ 的权重;最后得到结果,新图像属于权重最大的那个类别。

3　实例分析[12-13]

3.1　实例原理

为了满足医学基于内容的图像检索应用和数据

挖掘要求，提高图像检索效率，作为这一过程重要一步的图像自动分类自然成为人们关注的重点。以德国阿亨工业大学(Aachen University of Technology)医学系的Thomas M. l.为首的研究小组以医学图像为对象，对图像自动分类进行了深入研究。实验中以KNN作为图像自动分类器，为了便于交互式地表示分类结果，k值取1或5。不同的图像特征抽取方法对自动分类结果影响非常大，因此他们对Tamura等人提出的粗度、对比度、方向性来描述一个图像的纹理特征，Castelli等人提出的使用多个纹理特征来描述图像特征，Zhou和Huang提出的在一个图像内获取边缘特征等图像特征抽取方法进行了比较研究。结果显示：就单一方法而言，Tamura方法比其他方法效果好一些；通过两种或两种以上方法同时提取效果更好。在实验中，他们采用了Tamura & IDM来提取图像特征，其中Tamura纹理特征是全局特征，而IDM则保持局部像素邻居之间的信息，这样既保持了全局特征又保持了局部特征信息，取得了较好效果。采用KNN作为自动分类器，使用Tamura & IDM来提取图像特征，他们分别对四组不同的图像数据进行分类，结果如表1所示。

表1 1-NN和5-NN分类器给定的正确率 (%)

语料库	1-NN	5-NN
所有图像(6231张图像，81个类)，每类最少5张图像	85.48	85.36
所有图像(6155张图像，70个类)，每类最少10张图像	85.69	85.65
X光片(5776张图像、57个类)、每类最少5张图像	85.01	85.01
X光片(5756张图像，54个类)，每类最少10张图像	85.15	85.18

3.2 评价

图像自动分类成功与否的关键在于自动分类器的选择与图像特征的抽取方法。以Thomas M. L.为首的研究小组采用KNN作为医学图像自动分类器并取得大了良好效果。无论是1-NN还是5-NN，其分类正确率均在85%以上，基本上能够满足医学基于内容的图像检索应用要求。表明KNN作为图像自动分类的分类器是成功的，基于KNN的图像自动分类模型是可行的。

参考文献

1,12 Lehmann T. m. ,et al. Automatic categorization of medical images for content-based retrieval and data mining. Computerized Medical Imaging and Graphics,2005(29)

2 the Medical Subject Heading. [2006-03-12]. http://nlm.nih.gov/mesh

3 the Unified Medical Language System. [2006-03-12]. http://nlm.nih.gov/research/umls

4 李飞雪等. 基于人工神经网络与决策树相结合模型的遥感图像自动分类研究. 遥感信息,2003(3)

5 刘国庆等. 基于小波变换和马尔可夫随机场的极化SAR图像自动分类. 电子与信息学报,2003(3)

6 段宏等. 一种基于Web挖掘的信息自动分类系统. 华中科技大学学报(自然科学版),2003(7)

7 王伟等. 自动分类模型及算法研究. 微电子学与计算机，2004(5)

8 方兰等. 文本自动分类技术及其应用. 计算机与现代化，2004(7)

9 王永庆. 人工智能原理与方法. 西安：西安交通大学出版社,1998

10 Yiming Yang, Xin Liu. A Re-Examination of Text Categorization Methods. Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1999

11 T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference of Machine Learning, Berlin,1998

13 Mark O. G. ,et al. Comparison of Global Features for Categorization of Medical Images. [2006-03-22]. http://libra.imib.rwth-aachen.de/irma/ps-pdf/SPIE_2004-5371-35.pdf

余肖生 武汉大学信息管理学院04级博士生。通信地址：武汉。邮编430072。

周 宁 武汉大学信息管理学院教授。通信地址同上。

张芳芳 武汉大学信息资源研究中心工作。通信地址同上。

(来稿时间：2006-04-19)