一种新的基于 SVM 权重向量的云分类器 *

朱 杰[1]，秦亮曦[1]，龙炜哲[1]，苏永秀[2]
(1. 广西大学 计算机与电子信息学院，南宁 530004；2. 广西气象减灾研究所，南宁 530022)

摘 要：提出了一种用支持向量机(SVM)权重向量解决高维对象分类的方法，并结合云理论建立了基于 SVM 权重向量的云分类器。采用云模型建立训练集的各属性模型，分类模型由属性模型集成得到，属性权重根据 SVM 权重向量得到，属性权重越大，其对分类的贡献越大；反之，越小。将新分类器与云模型分类器对积雨云、卷云和卷层云进行分类模拟实验，新分类器的分类准确度比后者总体提升了，经过交叉验证，结果表明新分类器性能稳定。

关键词：支持向量机；云模型；云分类器；交叉验证
中图分类号：TP181 **文献标志码**：A **文章编号**：1001-3695(2009)06-2098-03
doi:10.3969/j.issn.1001-3695.2009.06.029

New cloud classifier based on SVM weight vector

ZHU Jie[1], QIN Liang-xi[1], LONG Wei-zhe[1], SU Yong-xiu[2]
(1. *School of Computer, Electronics & Information, Guangxi University, Nanning* 530004, *China*; 2. *Guangxi Institute of Meteorological Disaster Mitigation, Nanning* 530022, *China*)

Abstract: This paper presented a method of support vector machine weight vector to solve the problem of high-dimensional objects classification, and built cloud classifier based on the cloud theory and the SVM weight vector. Set up the attribute model of training by cloud model. Integrated classification model by every attribute model, and attributed weight came from the SVM weight vector. The larger the weight of an attribute was, the more it would make the contribution to classification. On the contrary, it would reduce the effect of classification. The new classification algorithm and the cloud classifier were applied to classify the radiance profiles as cumulonimbus, cirrus clouds, or cirrostratus clouds. The experimental results show that it must can improve the classification accuracy of spatial data in the overall performance than the latter. This cross validation to prove that the performance of the new classifier is pretty steady.

Key words: support vector machine; cloud model; cloud classifier; cross validation

0 引言

对卫星云图进行定量分析已成为研究的热点，而云分类更是卫星云图分析与应用的难点。早期用于分析卫星云资料的方法是阈值法。阈值法原理简单、计算方便，但阈值法缺乏定量统一的描述。因此，大量学者采用图像处理与统计分析相结合的方法对云图进行自动识别。传统的统计方法主要有判别分析和聚类分析，现代云分类产品很多都是利用聚类分析的方法，如风云-2C 云产品是采用多通道卫星探测数据进行聚类分析，归纳出各种云的类别。近年来智能分类方法也显示出优势，如统计方法、机器学习方法、神经网络方法等，这些智能分类方法在云分类上的应用也取得了一定的成功。但是由于这些分类方法难以处理空间数据模糊边界的不确定性和不精确性，而李德毅等人提出了云理论和基于云模型的分类方法[1~3]，云理论可以建立起定性描述的概念和定量表示的数值之间的转换关系。将训练数据集的类别定义为定性概念，用云模型表达定性概念与样本属于类别中心的隶属度之间的转换关系，从而将待分类样本划分到不同的类别中。云模型划分的边界模糊且不确定，比起其他的硬分类技术更符合实际的数据分布和人的思维方式。目前用云理论进行云分类的研究很少，本文将结合支持向量机和云理论在分类问题上的优势，提出把 SVM 的权重向量与云分类器相结合，目的在于提高空间数据模糊边界的分类精度。

1 云模型算法和支持向量机

1.1 云模型

云模型的表达式为[4]

$$\sigma_i = \mathrm{NORM}(En, He) \tag{1}$$

$$\mu = \exp[-1/2 \times (x_i - Ex)^2 / 2\sigma_i^2] \tag{2}$$

其中：x_i 是论域内的任一数值；Ex 是论域的中心值；En 为论域范围概念的一个熵；NOERM(En, He)表示 En 以为期望值；He 为方差的一个正态随机数。

期望值 Ex 是概念在论域中的中心值，它是最能代表定性概念的值，隶属于定性概念的程度为 100%。熵 En 是定性概

收稿日期：2008-09-18；**修回日期**：2008-11-21 **基金项目**：广西自然科学基金资助项目(桂科自 0728032)；广西研究生教育创新计划资助项目；国家农业科技成果转化资金项目(2007GB2E100249)

作者简介：朱杰(1983-)，男，湖南郴州人，硕士研究生，主要研究方向为数据挖掘、人工智能技术、机器学习(zhujie_831@sina.com)；秦亮曦(1963-)，男，教授，博士，主要研究方向为数据挖掘、进化计算、机器学习；龙炜哲(1983-)，男，硕士研究生，主要研究方向为数据挖掘；苏永秀(1963-)，女，高级工程师，主要研究方向为农业气象预报.

念模糊程度的度量,反映概念在论域中边界波动。超熵 He 是熵 En 的熵,反映论域中数值对定性概念的隶属度的随机性。

云模型的三个数字特征值[5],即期望值 Ex、熵 En 和超熵 He 把模糊性(概念接受的数值范围波动越大, 概念越模糊)和随机性(隶属度的随机性, 反映为云有一定的厚度)完全集成到一起, 构成定性(论域中的概念)和定量(论域中属性值的隶属度)相互间的映射, 作为知识表示的基础。

1.2　支持向量机

支持向量分类器[6]是基于二类分类的分类器。其核心思想就是去寻找一个把二类集合正确地划分并且与二类集合间隔最大的超平面,就是所谓的最优超平面。解决非线性问题,SVM 采用核函数把输入向量映射到高维特征空间,在高维特征空间中,原来的非线性问题变得线性可分了。对于多类分类问题,SVM 把多类问题转换成多个二类分类问题解决。

算法(最大间隔法)[7]:

a)设已知训练集

$$T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in(X\times Y)^l$$

其中:$x_i\in X=R^n$;$y_i\in Y=\{1,-1\}$,$i=1,\cdots,l$。

b)构造并求解对变量 w 和 b 的最优化问题

$$\min_{\omega,b}1/2\,\|w\|^2+C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t. } y_i((w\times x_i)+b)\geqslant 1-\xi_i;i=1,\cdots,l$$
$$\xi_i\geqslant 0;i=1,\cdots,l$$

求得最优解 w^*、b^*。

c)构造划分超平面$(w^*\times x)+b^*=0$,由此求得决策函数 $f(x)=\mathrm{sgn}((w^*\times x)+b^*)$。

2　基于 SVM 权重向量的云分类器

本文提出了一种基于 SVM 和云模型的新分类方法。在云模型分类中,每一维属性被平等计算,但是不同的属性对分类的贡献不同,有些属性的存在甚至会降低分类精度。面对高维的空间数据,必须付出大量的人工分析去验证哪维属性是目标类的特征属性。本文提出的方法就是为了解决这个问题。从 SVM 分类器中提取权重向量,然后把每一维的权重引到云分类器的计算中,使得云分类器对每一维属性对分类的重要性产生偏置,以达到较高的分类精度。

分类问题的训练样本集是从整个数据集中随机抽取出的一部分。建立分类器需要解决的问题:一是类别边界的不确定性; 二是定义各属性在分类中的作用, 确定属性权重。

基于 SVM 权重向量的云分类算法如下:设总样本为 $X=\{X_1,X_2,\cdots,X_p\}$,$|X|$表示总样本所含元素个数,类别 i 的样本集用集合 $X_i=\{x_1,x_2,\cdots,x_n\}$ 表示,它的第 j 个属性集用集合 X_{ij}表示,集合 X_{ij}的期望值为 E_{Xij},熵为 En_{Xij},超熵为 $He_{Xij}(i=1,2,\cdots,p;j=1,2,\cdots,D)$。

$$\sigma=\mathrm{NORM}(En_{Xij},He_{Xij})$$

a)计算每个属性 j 的权 $w_j(j=1,2,\cdots,D)$

SVM 的决策函数 $f(x)=\mathrm{sgn}((w^*\times x)+b^*)$ 中权重向量为 $w^*=\sum_{j=1}^{D}y_ja_j^*x_j=(w_1^*,\cdots,w_n^*)$。它代表了每一维属性与特定的目标类的相关程度,其证明可以参看文献[8]。

b)各类别的分类模型

$$A_i^{\text{weight}}=\sum_{j=1}^{D}w_{(j)}^*\exp(x_{Sj}-E_{Xij})^2/(2\sigma_{ij}^2)$$
$$i=1,2,\cdots,p;0\leqslant S\leqslant|X|$$

如果样本 x_S 在模型 i 中获得最大输出,则样本属于类别 i。

3　实验及其讨论

《风云二号》C 星(简称 FY-2C)主要有效负载为红外和可见光自旋扫描辐射器 VISSR。FY-2C 卫星云分类产品是利用卫星遥感技术,采用多通道卫星探测数据进行聚类分析,归纳出各种云的类别,分别代表地面、中低云、高层云、卷层云、密卷云、积雨云等。由于目前探测能力的限制,对于中低云很难作出准备分类。对于高云可以准确地分出积雨云、密卷云、卷层云、高层云或雨层云等类别。本章选用较为准确可信的积雨云、卷云、卷层云样本作为实验数据,用实验证明本文的新分类方法也是一个有效的云分类方法,并且更加智能。数据说明:从 NSMC 的官方网下载到 2008 年 5 月 26 日的云分类产品,从云产品中提取像元的云类标签(15 表示积雨云、14 表示卷云、13 表示卷层云),然后提取对应于云标签的每个像元的两个红外通道、一个水气通道、一个中波红外和一个可见光通道数据作为此类云的目标特征。

3.1　实验设计一

从遥感数据中选取各云类型 300 个点作为训练样本,总训练样本 900 个,随机抽取测试样本 2 515 个。本文选用 C-SVM 与 RBF(径向基)核函数,$C=2$,云模型的超熵 $He=0$。把权重映射到 0 ~ 100,最不重要的属性权重为 0,最重要的属性权重为 100。图 1 显示了实验的整个流程。

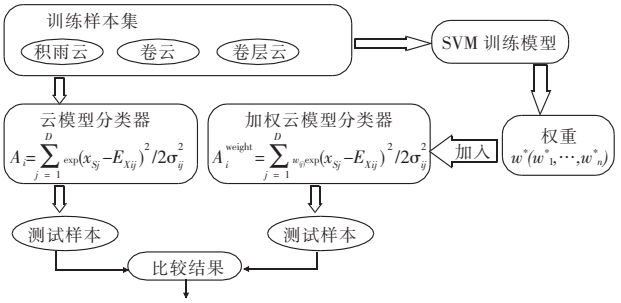

图 1　实验流程图

对比表 1 和 2 可以看出,加权云模型分类器的分类精度比不加权的云模型分类器有所提高,特别是对卷层云的分类精度提高最快。卷层云与积雨云基本上能完全分离,卷层云错分成卷云也减少 66 个。但是卷云被误分卷层云的数目增加了 9 个,被误分为积雨云的增加了 11 个,积雨云被误分为卷云也多了 4 个,这是因为加权(表 3)后,权重因子改变了实际波谱特征而造成的,但是对于整体精度,特别是积雨云与卷层云的识别基本上能非常准确,误分情况极少。由实验一可以证明本文提出来的加权云模型分类器有效地改进了不加权的云模型分类器的分类精度,实验二将证明本文提出的分类器是稳定的。

表 1　云模型分类器的分类结果

类别	预测类别					
	积雨云	卷云	卷层云	总数	正确率/%	总正确率/%
积雨云	362	23	0	385	94.03	
卷云	31	512	81	630	81.27	88.50
卷层云	7	147	1 346	1 500	89.73	

表 2　基于权重的云模型分类器的分类结果

类别	预测类别					
	积雨云	卷云	卷层云	总数	正确率/%	总正确率/%
积雨云	358	27	0	385	92.99	
卷云	42	488	100	630	77.46	90.06
卷层云	0	81	1 419	1 500	94.60	

表 3　各属性权重表

属性	红外一通道	红外二通道	水气通道	中波红外通道	可见光通道
权重	42.752 9	56.408 2	58.808 0	100	0

3.2　实验设计二

实验采用了交叉验证的方法,每种云类型各选取 1 000 个点,共 3 000 个数据样本点。每次将 3 000 个样本随机平均分成 10 份,用其中 9 份做训练样本,剩余的一份做测试样本。实验结果如表 4 所示。权重云模型分类器分类精度大部分高于云模型分类器,而且正确率稳定在 79% 以上,且波动振幅小于云模型分类器,如图 2 所示,分类正确率最高达到 91.5%,最低为 79%,十次分类有三次的正确率达到 90% 以上。

表 4　交叉验证及对比实验表　　%

权重云模型分类器分类精度	91	85.5	86	79	89	90	83	91.5	80.5	89
云模型分类器分类精度	86	76	88	84	79.5	82	85.5	81.5	71	80

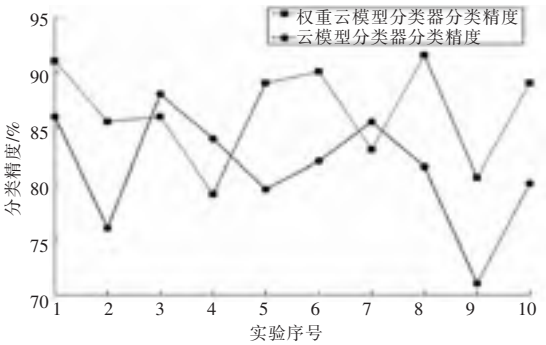

图 2　十次交叉验证及对比分类精度比较图

4　结束语

本文提出了结合 SVM 权重与云理论为一体的分类器,并把它应用于空间遥感数据的云分类中。第三章用两个实验证明了:a)基于 SVM 权重的云模型分类器在分类精度上比云模型分类器有所提高,总体精度提高了 1.56%,卷层云与积雨云基本上能完全分离,卷层云与卷云的识别也大大改善;b)本文提出的新分类器较稳定。由于 SVM 的权重向量的引入,使得云模型分类器对各维属性有了偏置,对类别边界变得非常敏感,分类精度提高。从实验可以看出本文提出的基于 SVM 权重的云模型分类器是一个有应用前景的分类方法。

参考文献:

[1] 宋远骏,李德毅,杨孝宗.电子产品可靠性云模型评价方法[J].电子学报,2000,28(12):74-76.

[2] 杨朝辉,李德毅.二维云模型及其在预测中的应用[J].计算机学报,1998,21(11):962-968.

[3] 邸凯昌,李德毅,李德仁.云理论及其在空间数据发掘和知识发现中的应用[J].中国图象图形学报,1999,4(11):930-935.

[4] 李德仁,王树良,李德毅.空间数据挖掘理论与应用[M].北京:科学出版社,2006.

[5] 刘桂花,宋承祥,刘弘.云发生器的软件实现[J].计算机应用研究,2007,24(1):46-48.

[6] VAPNIK V. Statistical learning theory[M]. New York:Wiley,1998.

[7] 邓乃扬,天英杰.数据挖掘中的新方法——支持向量机[M].北京:科学出版社,2005.

[8] CHUNG Y Y,CHOI L E H C,IU Li-wei,*et al*. A new hybrid audio classification algorithm based on SVM weight factor and Euclidean distance[C]//Proc of Annual Conference on International Computer Engineering and Applications. USA:WSEAS,2007:152-157.

(上接第 2085 页)

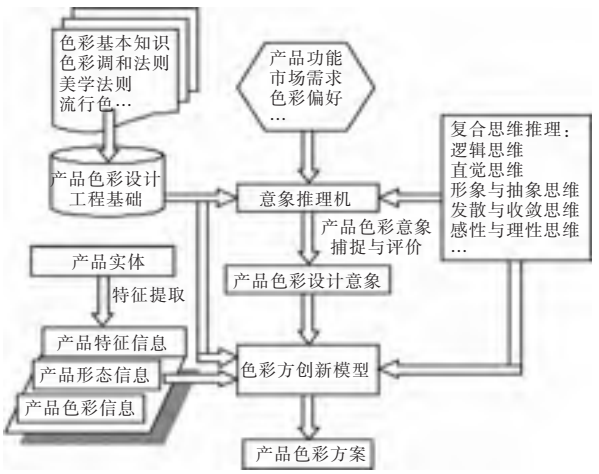

图 3　PCID 的复合思维体系示意图

4　结束语

本文从思维和设计的角度论述了复合思维理论与技术研究的基础,主要阐述了思维的生理、心理、哲学以及人工智能方面的研究。根据思维对象和思维操作划分思维的类型,构建了相应的单一思维模型和复合思维概念模型。根据可拓学理论,构建了 PCID 的思维对象模型和思维操作模型。根据产品色彩设计思维对象转换过程和操作机制,建立了色彩设计的复合思维模型和色彩设计的复合思维体系,为 PCID 研究奠定了智能理论基础,阐明了解决 PCID 问题的途径。

参考文献:

[1] 张全,陆长德,许占民.复合思维——设计智能化研究的关键技术[C]//2006 年国际工业设计研讨会.2006.

[2] JACOBSON N,BENDER W. Color as a determined communication[J]. IBM Systems Journal,1996,35(3&4):526-538.

[3] WYSZECKI G,STILES W S. Color science[M]. New York:Wiley,2000.

[4] 刘芳,余隋怀,陆长德.工业设计中油漆配色专家系统的构建[J].机械科学与技术,1997,16(3):557-560.

[5] 张全,于明玖,许占民,等.面向三维数字化产品的油漆配方预测系统[J].计算机工程与应用,2004,40(36):4-6.

[6] 苑寅秋,王珉,左敦稳.面向机电产品的色彩设计专家系统(MC-SES)框架研究[J].小型微型计算机系统,2001,22(3):376-378.

[7] 王可,陆长德,余隋怀.面向计算机辅助工业设计的色彩设计系统[J].计算机辅助设计与图形学学报,2004,16(10):1425-1429.

[8] 徐雪松,诸静.多模态函数优化的免疫算法[J].浙江大学学报,2004,38(5):530-533.