基于分形理论的多尺度分类尺度上推算法

李佳星 赵书良 安 磊 李长镜
(河北师范大学数学与信息科学学院 石家庄 050024)
(河北师范大学河北省计算数学与应用重点实验室 石家庄 050024)

摘 要 目前,多尺度数据挖掘的研究多集中于空间图像数据,在一般数据集上的研究已经初见成果,主要包括多尺度聚类以及多尺度关联规则,但还没有研究涉及一般数据下的分类。结合分形理论思想,将多尺度数据挖掘相关理论、知识和方法应用于分类领域,提出基于豪斯多夫距离(HD)的相似性度量方法;相对于以往对权重的经验定义,文中明确通过广义分形维数的相似性定义权重来提高相似性度量方法的精度;提出多尺度分类尺度上推算法(Multi-Scale Classification Scaling-Up Algorithm,MSCSUA);实验采用 4 个 UCI 基准数据集和 1 个真实数据集(H 省部分人口)进行仿真实验,实验结果表明多尺度分类思想可行有效,并且 MSCSUA 算法在不同数据集上的性能均优于 SLAD,KNN,Decision Tree 以及 LIBSVM 算法。

关键词 多尺度数据挖掘,多尺度分类,分形理论,尺度上推
中图法分类号 TP391 **文献标识码** A

Scaling-up Algorithm of Multi-scale Classification Based on Fractal Theory

LI Jia-xing ZHAO Shu-liang AN Lei LI Chang-jing
(College of Mathematic & Information Science, Hebei Normal University, Shijiazhuang 050024, China)
(Hebei Key Laboratory of Computational Mathematics & Applications, Hebei Normal University, Shijiazhuang 050024, China)

Abstract At present, the research of multi-scale data mining mainly focuses on space image data, and recently has produced some results on the general data, including the multi-scale clustering and multi-scale association rules, but it has not been involved in the field of classification mining. Combining with fractal theory, this paper applied the theory, knowledge and methods related to the multi-scale data mining to the areas of the classification mining, and proposed an approach of similarity measure based on Hausdorff. Relative to the definition of weight through experience, this paper clearly defined it by the similarity of generalized fractal dimension to improve the precision of similarity measure. Then, this paper proposed a multi-scale classification scaling-up algorithm named MSCSUA(Multi-Scale Classification Scaling-Up Algorithm). At last, this paper performed experiments on four UCI benchmark data sets and one real data set (H province part of the population). The experimental results show that the thought of multi-scale classification is feasible and effective, the MSCSUA algorithm performs well in terms of classification than SLAD, KNN, Decision Tree and LIBSVM algorithms on different data sets.

Keywords Multi-scale data mining, Multi-scale classification, Fractal theory, Scaling-up

1 引言

尺度的概念是从地学学科中引进的,一般指在研究中所使用的空间或时间单位,亦可指在空间和时间上某一个过程或现象所涉及的范围和发生的频率。研究表明,客观世界中普遍存在尺度现象[1]。

随着地学、物理学、数学、化学以及遥感学等领域的深度研究,多尺度科学已经发展为一门独立的跨学科研究课题。国内外学者在多尺度科学及其在分类数据挖掘领域中的应用方面均取得了一定的成果。在国内,文献[2]提出了一种基于稀疏编码的多尺度空间潜在语义分析的图像分类方法;文献[3]提出一种基于多尺度上下文语义信息的图像场景分类算法;文献[4]根据地类形状指数建立多尺度窗口,并结合主方向权值算法,大大提高了性能,并解决了不确定性问题;文献[5]针对高空间分辨率的遥感影像,提出了一种基于多尺度分割的变化检测算法;文献[6]从多尺度的角度出发,论述了元胞自动机 CA 和多智能体 ABM 模型在土地利用格局和演化中的重要作用,并且阐述了未来它在多类用地模型以及大尺度模型、知识迁移等方面的应用前景。在国外,文献[7]针对光学遥感图像在不同时期可能有着不同几何分辨率的问题,

本文受国家自然科学资金项目(71271067),国家社科基金重大项目(13&ZD091),河北省高等学校科学技术研究项目(QN2014196),河北师范大学硕士基金(xj2015003)资助。

李佳星(1992-),女,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:lijiaxing0322@163.com;**赵书良**(1967-),男,教授,博士生导师,主要研究领域为数据挖掘、智能信息处理,E-mail:zhaoshuliang@sina.com;**安 磊**(1991-),男,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:543080509@qq.com;**李长镜**(1990-),男,硕士生,主要研究领域为数据挖掘、智能信息处理,E-mail:lee_0809hbsd@outlook.com。

提出基于条件随即域(CRF)的方法,大大提高了各个时期的分类的准确率,而且该方法还能识别出不同时期的土地覆盖变化;文献[8]针对图像分类问题中规模较大时所存在的问题,提出了多尺度识别的字典学习方法(ML-DDL)。

但是,近年的相关研究成果表明,多尺度分类多用于空间、图像数据(如遥感图像分类识别),对于具有多尺度特性的一般数据的研究尚未成熟,这大大限制了多尺度科学在分类挖掘领域的应用[9]。

然而在实际应用中,一般数据的分类与多尺度科学相结合是非常具有现实意义的,主要体现在 3 个方面:1)尺度划分的意义。对于分布复杂的大规模数据集,训练得到的分类模型的复杂度必然也高。将数据集划分为不同的分类识别性高的区域,分布式地为每个区域分别训练一个简单模型,不仅能缩短训练时间,还能提高分类准确度。2)多尺度划分的意义。将数据集划分为具有偏序关系的多尺度数据集,在不同尺度下得到的模型的复杂度和分类准确性不同。尺度越细,模型的复杂度越低,分类的准确性越高。但若尺度过细,则可能导致过拟合的问题;反之,若尺度过粗,则可能达不到尺度划分的目的。3)尺度转换的意义。在分类挖掘中,将训练得到的分类模型作为尺度转化的对象,对基准尺度上的模型进行上推、下推从而得到其他尺度上的模型,不仅避免了繁琐的重复训练,还能减少训练时间。

分形理论作为非线性科学研究中的重要工具和手段,在数据挖掘领域已经取得一定成果,多用于图像的处理。随后推出了广义分形理论,开启了其在一般数据处理方法中的应用前景。分形最显著的特征是自相似性,是指某种结构或者过程的特征,从不同的空间、时间尺度来看它们都是相似的。分形理论强调的这种整体与局部的共性和个性的关系与多尺度数据挖掘中的多尺度数据集划分结构的理念有异曲同工之妙,二者的结合为一般数据的分类研究提供了新的思路和方法。本文将分形理论引入多尺度数据挖掘过程,用于一般数据的分类问题,具有一定的研究意义。

本文第 2 节介绍分形理论的相关知识;第 3 节提出多尺度分类的定义;第 4 节提出多尺度分类尺度上推算法 MSC-SUA;第 5 节通过实验对比验证该算法的有效性和可行性;最后总结全文。

2 分形理论

分形(Fractal)是法国数学家 Mandelbrot 于 20 世纪 80 年代从非规整几何的量测问题的角度出发创立的独立的新型理论,它并没有一个严格的定义,目前最能接受的定义是:一种由许多个与整体有某种相似性的局部所构成的形体。分形理论的主要思想是整体与局部的自相似,而这种自相似可以是严格意义上的结构相似,也可以是广义上的近似[10]。将分形维数应用在一般数据集上,得到广义分形维数。

定义 1 广义分形维数(Generalized Fractal Dimension)[11] 对于 R^n 上的有限数据点集合 A,如果 A 在 $r_{\min}<r<r_{\max}$ 范围内具有自相似性,则 A 的广义分形维数 $D(A)$ 有如下 3 种。分形维数是定量表征自相似性的重要指标。

2.1 计盒维数 D_b[11]

计盒维数又称盒维数(Box-counting,Box Dimension),因数学计算和经验估计都相对较容易,所以其是目前应用得最为广泛的分形维数之一。

定义 2 A 为 R^n 的一个有界子集,用尺度(边长)为 r 的 n 维盒子(n 维立方体)来覆盖 A,该过程称为 X-覆盖,记 $Nr(A)$ 为覆盖 A 的最少盒子数,则集合 A 的计盒维数 $D_b(A)$ 定义为:

$$D_b(A)=\lim_{r\to 0}\frac{\ln N_r(A)}{-\ln r} \tag{1}$$

当 r 充分小且极限存在时,对数比可以近似地看作集合 A 的计盒维数。

计盒维数在一定程度上反映不同数据集的相似度。

2.2 信息维数 D_i[11]

信息维数 D_i 是对计盒维数的一个改进优化。

定义 3 在定义计盒维数的基础上,考虑每个覆盖 U_i 中包含的数据点的个数 N_i,记 $p_i=N_i/N,i=1,2,\cdots,Nr$,$N$ 为数据集的总个数,p_i 表示数据集的元素属于覆盖 U_i 的概率,则集合 A 的信息维数 $D_i(A)$ 定义为:

$$D_i(A)=\lim_{r\to 0}\frac{\sum_{i=1}^{N_r}P_i\ln P_i}{\ln r} \tag{2}$$

当概率相等且 $p_i=1/Nr$ 时,容易得出信息维数等于盒子维数,即 $D_i=D_b$。

信息维数随着尺度的变化而变化,在一定程度上反映了该数据集变化的趋势。

2.3 关联维数 D_c[11]

定义 4 在定义信息维数的基础上,定义关联函数 $C(r)$,集合 A 的关联维数 $D_c(A)$ 定义为:

$$D_c(A)=\lim_{r\to 0}\frac{\ln C(r)}{\ln r} \tag{3}$$

其中,$C(r)=\sum_{i=1}^{N_r}P_i^2$。

理论上,关联维数不会随着尺度的变化而大幅变化,在一定程度上反映了数据集本质特征的属性集个数。

3 多尺度分类

3.1 多尺度分类的定义

多尺度数据挖掘的本质就是经过将原始数据集进行划分以构造为多尺度数据集,从而对不同尺度下的数据集进行全面系统的分析。尺度上推的目的是利用从小尺度数据集中获取的知识和信息来推测大尺度数据集的信息,是一种信息的聚集,如此可避免对数据集的重复学习。对于分类而言,最关键的是提取能刻画数据的模型(即分类模型),用来预测(离散的、无序的)未知数据的类别标签。多尺度分类的定义如下。

定义 5 将数据集划分为多尺度数据集,根据某些条件选择基准尺度数据集,然后在基准尺度数据集上选择现有的或改进的分类方法来训练分类模型,最后根据某种尺度转换机制,由基准尺度上的分类模型推测目标尺度数据集的分类模型。

3.2 多尺度分类任务

3.2.1 确定分类模型

分类的任务就是训练得到良好的分类模型,而对于多尺度分类任务,就是将数据集划分为多尺度数据集,随后得到各层尺度上的分类模型,最终选择分类效果最好的那层尺度上

的分类模型，然而各层上的分类模型并不是直接训练的，而是由基准尺度层上的分类模型经过上推、下推得到的。图 1 给出了一个四层多尺度数据集。

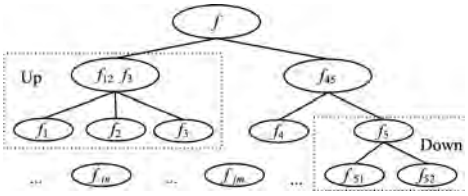

图 1　四层多尺度数据集

选择第三层为基准尺度，一个父层尺度节点的分类模型由它的孩子节点分类模型经过上推得到，如图 1 虚线 Up 部分；相似的孩子节点的分类模型需上推出一个分类模型，极不相似的保持不变。叶子节点分类模型由其双亲节点经过下推得到，如图 1 虚线 Down 部分。

3.2.2　预测方式

得到最优尺度层的分类模型后，对于一个未知样本，先判断其所属于的划分区域（即哪一个节点），再由该节点内的所有分类模型对它进行预测，通过投票的方式确定最终的类别标签。

3.3　多尺度分类转换对象

目前，发展成熟且应用广泛的分类方法众多，如决策树、贝叶斯方法、支持向量机、神经网络等，不同的分类方法训练得到的分类模型不同。

在多尺度数据挖掘中，转换对象是第一要素，如多尺度聚类中的簇心、多尺度关联规则中的频繁项集。

定义 6（多尺度分类转换对象）　它是指在多尺度分类中对分类模型起着关键性作用的对象。例如决策树模型中的属性值与对象值之间的映射关系、支持向量机模型的支持向量、神经网络模型中的神经元等。

不同的转换对象的内部结构大不相同，需要根据不同的相似性度量方法，采明不同的尺度转换机制，由基本尺度上的转换对象推演出其他尺度下的转换对象，当然不可避免地存在着尺度效应，我们的目标是寻求合理的方法来优化尺度转换机制，将尺度效应降到最低。

支持向量机 SVM 处理高维数、非线性的数据集时具有良好的分类效果，在机器学习领域中是继神经网络之后最受关注的研究热点。其模型的转换对象有稳定统一的结构，可以大大减少尺度转换的复杂性。本文主要应用 SVM 的一对一多类别分类方法来研究多尺度分类尺度上推算法。

4　多尺度分类尺度上推算法

4.1　转换对象的存储结构

为了直观地了解 SVM 转换对象的结构并对其进行处理，在此简单地介绍 SVM 一对一多类别分类方法。

顾名思义，一对一就是任意两个类别的数据都要训练一个分类模型，那么对于 n 类的问题，就要训练 $n(n-1)/2$ 个模型。每一个模型都有对应的支持向量、权重以及常数 b，然而在存储时，这些模型的信息并没有分开存储，所有模型的支持向量按类别依次存储，权重也按类别统一存储在一个 $n\times(n-1)$ 的矩阵中，如表 1 所列。

表 1　支持向量的权重存储结构

C_1	C_2	C_3	C_4	…	C_n
C_2	C_1	C_3	C_4	…	C_n
C_3	C_1	C_2	C_4	…	C_n
…	….	…	…	…	C_n
C_n	C_1	C_2	C_3	…	C_{n-1}

第 i 类与第 j 类数据的分类模型的支持向量为统一后的支持向量集的第 i 类与第 j 类数据点，其对应的权重分别为 C_i 行 C_j 列的权重值和 C_j 行 C_i 列的权重值；如果权重为 0，则说明该模型的支持向量不包含对应的数据点。

表 2 列出了一个 3 类问题的权重值。

表 2　3 类问题的权重存储结构

	C2	C3
C1	0	0.848067969653861
	0.275130389542984	0
	C1	C3
C2	−0.978147295959982	0
	0	0.249208838608209
	C1	C2
C3	−0.653560566306045	0
	0	−0.249208838608209

由表 2 可以看出，支持向量总数为 8，对于类别 1 和类别 3 的分类模型，权重如表 2 中的加黑部分，其中第 2 条数据和第 4 条数据的权重为 0，说明这两个数据点不是该分类模型的支持向量。通过这样的统一存储方式避免了由于 n 过大和分类模型的个数过多而导致的存储复杂、繁琐的问题。

定义 7（SVM 转换对象的存储结构模型）　采用一对一多类别分类方法。设一个四元组 $U=(X,C,W,B)$，其中 $X=(X_1,X_2,\cdots,X_n)$ 表示从基准尺度上的 n 个划分区域的数据集上训练得到的支持向量集；$C_i=(C_{i1},C_{i2},\cdots,C_{ic})$，$C_{ij}$ 表示第 i 个支持向量的第 j 类数据点集，c 为类别总个数。$W_i=(W_{i1},W_{i2},\cdots,W_{ic-1})$，$W_{ik}$ 表示第 i 个支持向量的第 k 列权重；$b_i=(b_{i1},b_{i2},\cdots,b_{im})$，$b_{ij}$ 表示第 i 个数据集上训练的第 j 个模型的常数 b，因为采用一对一多类别分类方法，即任意两个类别的数据集之间要训练一个分类模型，所以 $m=c(c-1)/2$，那么 SVM 转换对象的 (X,C,W) 部分的存储结构如图 2 所示。

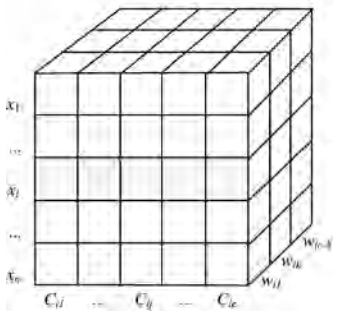

图 2　SVM 转换对象的 (X,C,W) 部分的存储结构

通过对 SVM 转换对象的存储结构的分析，我们得到尺度上推的本质任务在于如何由基准尺度的信息得到上一层尺度上的各分类模型的支持向量集和对应的权重，以及常数 b 的值，即纵向信息的汇总。

为此，本文提出基于豪斯多夫距离的相似性度量方法，根据不同数据集的分类模型的支持向量集的相似度，构造相似矩阵，进而估计上一层尺度上的信息。

4.2 基于豪斯多夫距离的相似性度量

定义 8(豪斯多夫距离[12](HD)) 衡量两个点集之间相似性的度量方法。设 $A=\{a_1,\cdots,a_m\}$ 和 $B=\{b_1,\cdots,b_n\}$ 为两个有限点的集合,则 HD 的定义为:

$$H(A,B)=\max(h(A,B),h(B,A)) \tag{4}$$

其中,单向 HD 的定义为:

$$h(A,B)=\max_{a\in A}\min_{b\in B}\|a-b\| \tag{5}$$

由式(5)可以看出,单向 HD 对距离远的点较敏感,即易受噪声点的影响,因此改进的单向 HD 为:

$$h_{\mathrm{mod}}(A,B)=\frac{1}{|A|}\sum_{a\in A}\min_{b\in B}\|a-b\| \tag{6}$$

即数据集 A 中的每一个点到数据集 B 中的点的最近的距离的和的平均值,这样缓和了远点对 HD 的影响。

定义 9(基于 HD 的相似性度量) 用改进的 HD 衡量支持向量集的相似度,并考虑原本数据集相似度的影响,则基于 HD 的相似性度量定义如下:

$$S(A,B)=W\cdot Sim(A,B) \tag{7}$$

其中,$A=(A_1,A_2,\cdots,A_c)$ 和 $B=(B_1,B_2,\cdots,B_c)$ 为任意两个支持向量集,A_i 与 B_i 表示第 i 类数据点。

$$Sim(A,B)=\frac{1}{1+H(A,B)} \tag{8}$$

本文使用加权的方式改进 HD,$H(A,B)$ 值越小,$Sim(A,B)$ 的值越大,定义如下:

$$H(A,B)=\max(h(A,B),h(B,A)) \tag{9}$$

其中,单向 HD 的定义为:

$$h_w(A,B)=\sum_{i=1}^{c}\frac{1}{n_i}\sum_{a\in A_i}\min_{b\in B_i}\sum_{j=1}^{c-1}\|w_{aj}\cdot a-w_{bj}\cdot b\| \tag{10}$$

其中,c 为类别数,n_i 为第 i 类数据点的总个数,A_i 为 A 中第 i 类数据集,w_j 为数据点对应的第 j 个权重,这是一个至关重要的元素,它是否等于 0 直接影响着两个支持向量的相似度。

权重 W 为支持向量集对应的数据集之间的相似性度量,由分形维数(请参考第 2 节)表示的特征向量求得,定义为:

$$W=Sim(a,b)=\frac{1}{1+\sum_{i=1}^{d}|a_i-b_i|} \tag{11}$$

其中,a 和 b 为数据集的分形维数特征向量,d 为属性个数,若 a 和 b 越相似,则 W 的值越接近为 1,其定义如下:

$$FT=(D_b,D_i,D_c) \tag{12}$$

4.3 构建相似矩阵[13]

支持向量 X_i 的相似度均值为:

$$\overline{S(X_i)}=\frac{1}{N-1}\sum_{j=1,j\neq i}^{N}S(X_i,X_j) \tag{13}$$

其中,N 为相似支持向量的总个数,这里去掉相同元素的 S 值,避免了由于同一个支持向量集相似度过高而导致其他相似度都低于平均值。

支持向量集的相似度矩阵:

$$M(i,j)=\begin{cases}1, & \text{if } S(X_i,X_j)\geqslant\overline{S(X_i)}\\ 0, & \text{otherwise}\end{cases} \tag{14}$$

对于任意两个支持向量集 X_i 和 X_j,将满足 $M(i,k)=M(j,k)=1$ 的 X_k 的个数记作 N_i,若 $N_i\geqslant N-N_i$,则称支持向量集 X_i 和 X_j 存在强不可分辨关系。

4.4 尺度上推机制

通过上述处理,将具有强不可分关系的支持向量集定义为相似的支持向量集,即得到了需要上推出一个分类模型的相似分类模型集。

相似分类模型集记作 $M=(M_1,\cdots,M_n)$,n 为相似的分类模型的个数。

定义 10(基准分类模型) 对基准尺度上的分类模型进行测试,得到分类正确率。在相似的分类模型集中,正确率最高的模型称为基准分类模型。

将基准分类模型作为上推后的分类模型的初始状态,其中权重保持不变,支持向量集以及常数 b 做以下更新。

(1)上推后的分类模型中的支持向量集的第 i 类第 j 个数据点 y_{ij} 为:各相似的分类模型的支持向量集的第 i 类数据中与基准分类模型的第 i 类第 j 个数据点 x_{ij} 的加权距离最近的数据点的均值,如式(15)所示:

$$y_{ij}=\frac{1}{n}\sum_{l=1}^{n}\arg\min_{a_i\in X_{li}}\sum_{k=1}^{c-1}\|w_{ak}\cdot a_l-w_{xk}\cdot x_{ij}\| \tag{15}$$

其中,X_{li} 为第 l 个相似的分类模型中支持向量集的第 i 类数据集。

(2)同样地,根据均值的思想,上推后的分类模型的常数 b 为:

$$b=\frac{1}{n}\sum_{i=1}^{n}b_i \tag{16}$$

其中,b_i 为第 i 个相似分类模型常数 b 的值。

尺度上推是宏观上的信息聚集的过程,需模糊掉细节,即特例的信息,因此本文采用均值的思想。

4.5 算法描述

多尺度分类上推算法如算法 1 所示。

算法 1 多尺度分类上推算法

输入:原始数据集

输出:输出目标尺度上的分类模型

1. 数据预处理,将原始数据集经过尺度划分,构造多尺度数据集,记作 DS。
2. 选择基准尺度数据集,记作 BS,BS=(d_1,d_2,…,d_n),n 为目标层尺度数据集划分区域的个数。
   $d_i=(d_i^1,d_i^2,\cdots,d_i^m)$
3. Foreach d_i do begin
4. $M_i=\emptyset$;
5. For d_i^j do begin
6. Libsvm(d_i^j);//调用 Libsvm 挖掘算法,为每个区域训练分类模型;
7. $FT(d_i^j)=(D_b,D_i,D_c)$;//计算分形维数;
8. End For
9. For 任意两个分类模型 M_p,M_q
10. $S(M_p,M_q)$;//根据式(7)计算相似性度量 S;
11. End For
12. Get M;//根据式(14)构造相似度矩阵;
13. For 每组具有强不可分关系的支持向量集 do begin
14. Get M_X;//根据分类正确率,确定基准分类模型;
15. $M_X\leftarrow$Get Y,Get b;//根据式(15)、式(16)得到父层尺度数据集上分类模型的信息;
16. $M_i\leftarrow(M_i\cup M_x)$;
17. End For
18. $M_i\leftarrow(M_i\cup M)$;//对于极不相似的支持向量集对应的分类模型集 M
19. Return M_i
20. End Foreach

5　实验分析

本文以 SVM 一对一多类别分类方法为基本分类算法，采用 RBF 核函数，基于 MATLAB 实现多尺度分类思想以及 MSCSUA 算法。同时采用 4 个 UCI 标准数据集以及 1 个真实数据集进行实验，并与 KNN，Decision Tree 以及 LIBSVM 算法进行实验对比，以验证本文算法的有效性与可行性，其中 LIBSVM 方法同样采用一对一多类别分类方法和 RBF 核函数，因此在本文实验中 MSCSUA 算法的实现是建立在 LIBSVM 算法之上的。

5.1　数据集

本文采用的 UCI 标准数据集有 Ionosphere，Pima Indians Diabetes(PID)，Spambase 以及 wine，真实数据集采用 H 省部分人口数据，一共 5 个数据集，各个数据集的样本数量、特征维数和类别数都不尽相同。数据集的相关信息如表 3 所列。

表 3　数据集的相关信息

数据集	样本数	特征数	类别数
Ionosphere	351	34	2
PID	768	8	2
Spambase	4601	57	2
wine	178	13	3
H 省部分人口数据	6311	7	3

5.2　性能指标

为了衡量所提方法的分类性能，采用常用的度量标准：正确率(Accuracy，Acc)、F1-Measure、标准化互信息(NMI)以及运行时间(Run Time)。

5.2.1　Acc

分类的正确率(Acc)表示两部分之间的一对一关系，即正确划分的样本个数占全部样本的比例，计算公式如下：

$$Acc=\frac{1}{n}\sum_{i=1}^{n}\delta(C_i,map(P_i)) \tag{17}$$

其中，n 为全部样本数量，c_i 为第 i 个样本的真实类别标号，$map(c_i)$ 表示实验结果 p_i 到真实类别标签的最优映射，$\delta(x,y)$是一个函数，当 $x=y$ 时，$\delta(x,y)=1$，否则 $\delta(x,y)=0$。Acc 的值越高，表示分类效果越好。

5.2.2　F1-Measure

F1-Measure 是数据挖掘领域中一个重要评价指标，其值越大，说明分类效果越好。对于多类分类问题，F1-Measure 值的计算公式如下：

$$F1\text{-}Measure=\frac{1}{c}\sum_{i=1}^{c}F1\text{-}Measure_i \tag{18}$$

其中，$F1\text{-}Measure_i$ 采用一对多的形式，将第 i 类作为正类，其余类作为负类，这样就产生了 c 个 F1-Measure 值，再求和取均值。

5.2.3　NMI

标准化互信息(NMI)的计算可以借助混淆矩阵，计算公式如下：

$$NMI=\frac{\sum_{ij}\frac{n_{ij}}{n}\cdot\log\frac{n\cdot n_{ij}}{n_i\cdot n_j}}{\sqrt{(\sum_i n_i\cdot\log\frac{n_i}{n})(\sum_j n_j\cdot\log\frac{n_j}{n})}} \tag{19}$$

其中，n_i 为真实标签为 i 的样本个数，n_j 为实验预测标签为 j 的样本个数，n_{ij} 为真实标签为 i 但实验预测的标签为 j 的样本个数。NMI 的值越高，表示分类效果越好，但是当 Acc 的值降低时，NMI 的值大大减小，因此，当两个 Acc 值接近时，其 NMI 值会更接近，反之，差异会更大。

5.2.4　Run Time

上文提出的 MSCSUA 算法的步骤如算法 1 所示，但是一旦步骤 1－步骤 12 完成，即确定了相似的分类模型，根据多尺度数据挖掘的性质，步骤 13－步骤 19 的尺度上推机制是可以重复利用的，当尺度层次规模扩大时，步骤 1－步骤 12 的运行时间可以忽略不计。因此，MSCSUA 算法的 Run Time 值仅限于 4.4 节中的尺度上推机制。

本文实验将数据集划分为 3 层，MSCSUA 算法应用在第二层，因此其他算法的 Run Time 值也是指在第二层数据集上的运行时间。

5.3　实验结果分析

本文首先根据等级理论和概念划分，按照某些特征值将数据集划分为 3 层尺度的多尺度数据集，第一层为原始数据集，第二层划分为 2 个部分，第三层划分为 2～5 个部分，如图 1 所示。

本文首先将不同的分类算法(KNN，Decision Tree 和 LIBSVM)在不同尺度数据集上进行对比实验，以体现多尺度划分的优势；其次，将其与本文提出的尺度上推算法 MSCSUA 进行对比分析；最后，与文献[15]提出的新方法 SLAD 做比较。

表 4 列出了不同算法在不同尺度层的数据集上的 Acc 值。从表中可以明显看出，经过多尺度划分后，不同算法在不同尺度层的数据集上的 Acc 值呈现上升趋势，并且第三层较第一层的 Acc 平均提升大约 3%。

表 4　不同算法的 Acc 值

(单位：%)

数据集	KNN			Decision Tree			LIBSVM			MSCSUA
	第一层	第二层	第三层	第一层	第二层	第三层	第一层	第二层	第三层	
Ionosphere	77.1429	78.8571	81.1429	76.0000	77.1429	78.8571	81.1429	82.2857	84.0000	**86.2857**
PID	72.2008	72.9730	74.5174	71.8147	72.2008	74.9035	72.2008	72.9730	74.1334	**76.8333**
Spambase	78.8679	80.1258	81.0063	78.2390	81.0063	81.6352	78.8679	80.2516	81.2579	**82.1384**
wine	88.7640	93.2584	95.5056	83.1461	86.5169	87.6404	92.1348	93.2584	95.5056	**97.7528**
H 省部分人口数据	92.8775	92.9421	93.4297	92.8063	96.2251	96.7949	94.3732	96.7949	97.2925	**97.5071**

整个数据集分布无规律，学习的过程复杂多样，即使训练的正确率很高，也有可能导致过拟合的问题，但是经过多尺度的划分，可以降低分布的复杂性及学习到的分类模型的复杂性，加上多尺度的划分是建立在经验的概念分层、分形以及等级理论之上的，因此目标性较强。文献[14]提到的 CBB 方法的主要思想是先聚类再分类，但是仅局限于球形分布的数据集，本文提出的多尺度划分不限制数据集类型。

本文提出的 MSCSUA 方法是建立在 LIBSVM 算法之上

的，选择 LIBSVM 算法下的第三层为基准尺度数据集，过尺度上推算法得到上推后的第二层的分类模型。从表 4 中看到，MSCSUA 算法的 Acc 值较 LIBSVM 的第一层有明显的提升，平均大约提升了 4%，MSCSUA 算法虽然是建立在 LIBSVM 算法的第三层之上的，但是较第三层平均提升了 1.5%左右；并且经过 MSCSUA 算法得到的第二层的 Acc 较直接划分得到的第二层的 Acc 值平均提升了 2.8%左右。

图 3 给出了不同算法在第一层数据集上的 Acc 值的对比结果。从图中也可以明显看出，MSCSUA 算法较其他单一算法有明显的优势。

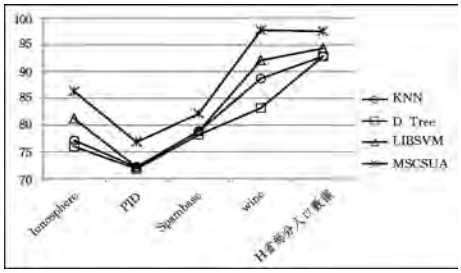

图 3 不同算法在第一层数据集上的 Acc 值

F1-Measure 是数据挖掘领域的一个重要评价指标。如表 5 所列，经过多尺度划分后，不同算法在不同尺度层的数据集上的 F1-Measure 值呈现上升趋势，尤其是在 Ionosphere 数据上 LIBSVM 算法提升了约 0.08；而本文提出的 MSCSUA 算法在多数数据集上的 F1-Measure 也是最高的，虽然在第三层 PID 和 Spambase 数据集上，仅次于 Decision Tree 算法，但是相对于 LIBSVM 第一层数据集上的 F1-Measure 值已经提高了平均约 0.09。图 4 给出了不同算法在第一层数据集上的 F1-Measure 值，本文提出的 MSCSUA 算法的 F1-Measure 值均高于其他算法。

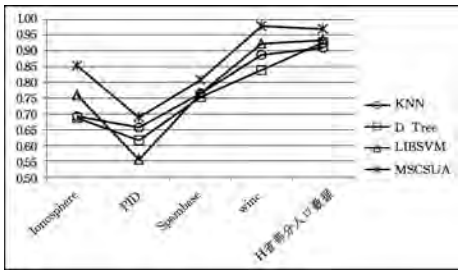

图 4 不同算法在第一层数据集上的 F1-Measure 值

标准互信息 NMI 值主要衡量真实的类别标签集与实验所得的类别标签集之间的差异，差异越小，NMI 值越高。值得一提的是，当差异增大时，NMI 值会迅速降低，这样就放大了差异，效果更直观。从表 6 中可以看出，经过多尺度划分后，不同算法在不同尺度层的数据集上的 NMI 值都呈现上升趋势，并且第三层较第一层的 NMI 平均提升大约 19%，而 Decision Tree 算法在 PID 的第三层数据集上较第一层数据集上的 NMI 值大约提升了 86%。MSCSUA 算法的 NMI 值较 LIBSVM 算法在第二层数据集上的 NMI 均有显著提升，尤其是在 PID 数据集上提高了大约 59%。图 5 给出了不同算法在第一层数据集上的 NMI 值。从中可以看出，MSCSUA 算法的 NMI 值均高于其他算法。

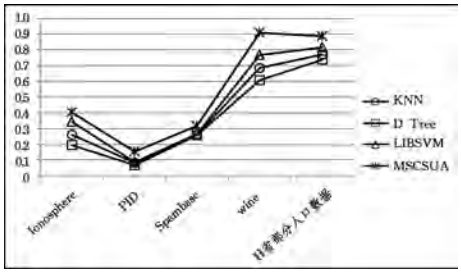

图 5 不同算法在第一层数据集上的 NMI 值

表 5 不同算法的 F1-Measure 值结果

数据集	KNN			Decision Tree			LIBSVM			MSCSUA
	第一层	第二层	第三层	第一层	第二层	第三层	第一层	第二层	第三层	
Ionosphere	0.6917	0.7213	0.7676	0.6859	0.7166	0.7604	0.7584	0.7730	0.8126	**0.8522**
PID	0.6567	0.6574	0.6678	0.6151	0.6698	**0.6962**	0.5539	0.5789	0.6345	0.6887
Spambase	0.7644	0.7894	0.7986	0.7545	0.8048	**0.8089**	0.7640	0.7785	0.7938	0.8076
wine	0.8882	0.9338	0.9583	0.8388	0.8672	0.8757	0.9225	0.9342	0.9552	**0.9780**
H 省部分人口数据	0.9106	0.9565	0.9635	0.9258	0.9537	0.9627	0.9334	0.9597	0.9671	**0.9699**

表 6 不同算法的 NMI 值结果

数据集	KNN			Decision Tree			LIBSVM			MSCSUA
	第一层	第二层	第三层	第一层	第二层	第三层	第一层	第二层	第三层	
Ionosphere	0.2617	0.2958	0.2974	0.1953	0.1973	0.2181	0.3429	0.4019	0.3458	**0.4026**
PID	0.0835	0.0899	0.1107	0.0672	0.0915	0.1254	0.0853	0.0967	0.1052	**0.1545**
Spambase	0.2628	0.2669	0.2858	0.2573	0.2910	0.2989	0.2642	0.3032	0.3110	**0.3178**
wine	0.6856	0.7766	0.8731	0.6062	0.6354	0.6473	0.7677	0.7732	0.8326	**0.9088**
H 省部分人口数据	0.7704	0.8610	0.8687	0.7395	0.8573	0.8624	0.8147	0.8752	0.8827	**0.8879**

运行时间是衡量一个算法是否有效可行的一个重要指标。表 7 与图 6 展示的是不同算法的运行时间，单位都为 s。从中可以看出，KNN，Decision Tree 以及 LIBSVM 算法的 Run Time 值随着数据集的样本数量以及特征维数的增加呈现上升趋势，特别是 LIBSVM 算法，在 Spambase 和 H 省全员数据数据集上的运行时间发生了较大的波动。而本文提出的 MSCSUA 算法由于没有训练过程，仅需要从基准尺度层经过上推机制得到，因此 Run Time 值始终处于较低的水平，波动不大。

表 7 不同算法在第二层数据集上的运行时间

（单位：s）

数据集	KNN	D Tree	LIBSVM	MSCSUA
wine	0.018	0.031	0.007	**0.002**
Ionosphere	0.012	0.04	0.009	**0.002**
PID	0.029	0.061	0.012	**0.002**
Spambase	0.108	0.224	0.398	**0.004**
H 省部分人口数据	0.07	0.047	0.355	**0.002**

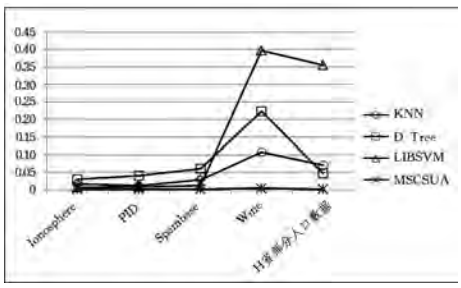

图 6　不同算法在第二层数据集上的运行时间

表 8 列出了本文提出的 MSCSUA 算法与文献[15]中所提出的方法 SLAD 的 Acc 和 NMI 的对比结果。SLAD 算法基于 LAD 框架并通过数据的统计信息来提高性能。从表 8 可以看出，本文在第一层数据集即原始数据集上实现的 LIBSVM 的各指标低于文献[15]中实现的 LIBSVM 的指标，但差异不大，原因可能在于训练集与测试集的随机抽样，因此差异不可避免。

表 8　不同算法的 Acc 和 NMI 对比结果

数据集		LIBSVM（第一层）	LIBSVM（文献[15]）	LAD	SLAD	MSCSUA
Ionosphere	Acc	81.1429	82.66	76.58	**85.14**	**86.2857**
	NMI	0.3429	0.36	0.21	**0.40**	**0.4026**
PID	Acc	72.2008	72.74	72.80	**75.54**	**76.8333**
	NMI	0.0853	0.15	0.10	**0.15**	**0.1545**

SLAD 算法通过改进 LAD 算法而得到，其效果得到显著提高，本文提出的 MSCSUA 方法是建立在 LIBSVM 算法之上的，但是在 LIBSVM 的 Acc 值远小于 SLAD 的情况下，其 Acc 值却更甚之；文献[15]中的 NMI 值由于都仅保留了小数点后两位，因此不能和本文的 NMI 做精确比较，但是从表 8 中可以看出，SLAD 算法和 MSCSUA 算法的 NMI 值的差异不大。由此可见，本文提出的 MSCSUA 算法具有一定的优势。

通过以上实验的对比结果可知，本文提出的多尺度分类思想以及 MSCSUA 算法相对于其他单一的算法具有显著的优势，同时其具有有效性与可行性。

结束语　本文结合分形理论，将多尺度数据挖掘理论和方法应用在分类领域，分析了多尺度分类的任务，明确定义了转换对象，提出了基于 HD 的相似性度量方法，并将广义分形维数的相似性应用于权重的定义，提高了相似性度量方法的精度；在此基础上提出了多尺度分类尺度上推算法 MSCSUA。首先，根据概念分层、等级理论等将一般数据集划分为多尺度数据集；其次，选择基准尺度，并获取基准尺度上的分类模型；然后，根据基于 HD 的相似性度量方法，明确相似的分类模型；最后，根据尺度上推机制，得到目标尺度上的分类模型。

在未来的研究工作中，我们将致力于研究多尺度分类尺度下推算法，以及其他分类方法（决策树、贝叶斯等）的转换对象、尺度转换机制以及更适宜的相似性度量方法，并且寻找基准尺度选择的评价指标，以完善多尺度分类算法的理论和方法。

参 考 文 献

[1] 韩玉辉，赵书良，柳萌萌，等. 多尺度聚类挖掘算法[J]. 计算机科学，2016，43(8)：244-248.

[2] 赵仲秋，季海峰，高隽，等. 基于稀疏编码多尺度空间潜在语义分析的图像分类[J]. 计算机学报，2014，37(6)：1251-1260.

[3] 张瑞杰，李弼程，魏福山. 基于多尺度上下文语义信息的图像场景分类算法[J]. 电子学报，2014，4：646-652.

[4] 兰泽英，刘洋. 领域知识辅助下基于多尺度与主方向纹理的遥感影像土地利用分类[J]. 测绘学报，2016，45(8)：973-982.

[5] 佃袁勇，方圣辉，姚崇怀. 多尺度分割的高分辨率遥感影像变化检测[J]. 遥感学报，2016，20(1)：129-137.

[6] 李少英，刘小平，黎夏，等. 土地利用变化模拟模型及应用研究进展[J]. 遥感学报，21(3)：329-340.

[7] HOBERG T，ROTTENSTEINER F，FEITOSA R Q，et al. Conditional Random Fields for Multitemporal and Multiscale Classification of Optical Satellite Imagery[J]. IEEE Transactions on Geoscience and Remote Sensing，2015，53(2)：659-673.

[8] SHEN L，SUN G，HUANG Q M，et al. Multi-Level Discriminative Dictionary Learning With Application to Large Scale Image Classification[J]. IEEE Transactions on Image Processing，2015，24(10)：3109-3123.

[9] 柳萌萌，赵书良，陈敏，等. 多尺度关联规则挖掘的尺度上推算法[J]. 计算机应用研究，2015，32(10)：2924-2929.

[10] 栾海军，田庆久，余涛，等. 根据分形理论与五指标评价体系构建 NDVI 连续空间尺度转换模型[J]. 遥感学报，2015，19(1)：116-125.

[11] BELUSSI A，FALOUTSOS C. Estimating the selectivity of spatial queries using the Correlation Fractal dimension[C]// Proceedings of the 21st International Conference on Very Large Data Bases（VLDB'95）. San Francisco，CA，USA：Morgan Kaufndnn，1995：1-26.

[12] 孙力帆，张森，冀保峰，等. 基于改进豪斯多夫距离的扩展目标形态估计评估[J]. 光学学报，2017，37(7)：0728003.

[13] 高新波. 模糊聚类分析及其应用[M]. 西安：西安电子科技大学出版社，2004：42-46.

[14] MILLER D，SOH L K. Cluster-Based Boosting[J]. IEEE Transactions on Knowledge and Data Engineering，2015，27(6)：1491-1504.

[15] BRUNI R，BIANCHI G. Effective Classification Using a Small Training Set Based on Discretization and StatisticalAnalysis[J]. IEEE Transactions on Knowledge and Data Engineering，2015，27(9)：2349-2361.