基于 RSKNN 分类改进算法①

兰 天[1,2], 郭躬德[1,2]

[1](福建师范大学 数学与计算机科学学院, 福州 350007)
[2](福建师范大学 网络安全与密码技术福建省重点实验室, 福州 350007)

摘 要: RSKNN 算法是 K 近邻算法的一种改进算法, 该算法基于变精度粗糙集理论, 能在保证一定分类精度的前提下, 有效地降低分类样本的计算量, 并且提高计算效率和分类精度. 由于 RSKNN 算法对属性的依赖度较高, 在分类时容易受到伪近邻的影响, 导致 RSKNN 算法的分类精度受到一定程度的影响. 针对存在问题, 本文提出一种新颖的基于 RSKNN 算法的改进算法 SMwRSKNN, 该算法在 RSKNN 算法的基础上引入类别子空间的思想, 以降低冗余属性和伪近邻对分类的影响. 在 UCI 公共数据集上的实验结果表明, SMwRSKNN 算法比 RSKNN 算法具有更高的分类精度.

关键词: 变精度粗糙集; 类别上、下近似; 互 K 近邻; 类别子空间

Improved RSKNN Algorithm for Classification

LAN Tian[1,2], GUO Gong-De[1,2]

[1](School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)
[2](Network Security and Cryptography key laboratory of Fujian province, Fujian Normal University, Fuzhou 350007, China)

Abstract: RSKNN is an improved algorithm of KNN with better classification performance. The RSKNN algorithm is based on the theory of the variable precision rough set. The algorithm guarantees under the premise of a certain classification accuracy, effectively reduces the computation burden of the classified samples, and improves the computation efficiency and precision of classification. But the degree of dependence on attributes is very high, which can make RSKNN algorithm affected by a certain degree of precision in classification. So the use of the class subspace classification method into RSKNN algorithm can improve the classification accuracy of RSKNN. The experimental results carried out on some UCI public datasets verify the effectiveness of the proposed algorithm.

Key words: variable precision rough set; class upper and lower approximation; mutual K nearest neighbor; class subspace

1 引言

在机器学习中, 分类是一项重要的技术, 并在众多领域取得重大成就, 如今也是重要的研究课题之一. 目前比较常用的分类算法有 K 最近邻、决策树、神经网络、贝叶斯分类器、支撑向量机等.

在大量的分类算法中, K 近邻算法(KNN)是一项较为简单、实用的分类算法. KNN 算法主要依赖于数据集的样本数量 n、距离度量和 k 系数, 从训练集中挑选 k 个距离最近的样本点, 通过这些样本点类别少数服从多数的投票原则决策出待测样本的类别. 近年来, 许多学者在KNN算法的三个依赖关系中进行改善, 并提出了许多有建设性意义的改进算法. 如 Guo G.D 等[1]提出基于模型的 KNN 算法(KNNModel), 该算法通过选择代表点建立分类模型, 并且能够在学习过程中自动确定 K 的取值; Guo G.D 等[7]提出模糊 k 近邻模型在可预测毒物学上的应用, 采用模糊划分的方法, 提高了分类精度; Huang X.M 等[11]提出加权 KNN 模型的数据约简和分类算法,与 KNN 模型的分类算法相比, 进

① 基金项目: 国家自然科学基金(61070062, 61175123);福建高校产学合作科技重大项目(2010H6007)
收稿时间:2013-05-04;收到修改稿时间:2013-06-18

一步降低了异常数据的干扰, 提高了数据分类的精度; Liu H.W[2]在传统的 k 近邻决策分类改进为利用互 k 近邻方法消除伪近邻来提高精度; 张著英等[3]将粗糙集理论应用到传统的 KNN 算法中, 实现了属性约简; 余鹰等[4]基于变精度粗糙集模型改进了 KNN 算法, 该算法基于变精度粗糙集理论, 用上、下近似域来刻画各类的分布特征, 有效地降低分类样本的计算量, 并且提高计算效率和分类精度, 特别是在球状分布的数据中起到了显著的效果; 在距离度量和 k 系数的关系中, Dudani S.A[5]提出了 wKNN 算法, 该算法在取 k 近邻时加入了权重, 降低了错误率. 李荣陆等[8]提出基于密度的 KNN 分类器样本选择方法, 有效地降低训练样本的数量; MKNN[2]算法通过判断待测样本与训练样本是否为互 K 近邻进行选择, 加强了邻居样本与类别之间的关联性, 使 KNN 算法对 K 值的依赖得到了较好的处理, 降低了伪近邻对分类效果的影响. MwKNN 算法[12]又在 MKNN 算法的基础上引入了距离权重, 利用距离加权方式投票, 根据权重最高的类判定归属, 因此能更加有效地降低伪近邻的影响. 张健飞等[9]利用子空间模型簇构造分类模型, 有效分隔了不同样本在全空间中重叠的区域. 李南等[10]提出基于子空间集成的概念漂移数据流分类算法, 有效适应了概念漂移数据流的分类. 卢伟胜等提出了 SMwKNN 算法[12], 在距离计算上对类属性数据引入基尼系数, 根据不同的样本类别构造其特有的子空间, 将待分类样本和训练样本都投射到某一类中, 再根据 MwKNN 算法的方法进行计算样本点在该类子空间下的权值, 最后再根据权值最高的类子空间进行判定归属, 使得在高维空间中的冗余属性和无用属性受到了约束, 提高了精度. RSKNN[4]算法引入了一个精度 β, 构造出上、下邻域空间, 对于落在下领域中的样本点可直接判定归属, 减少计算量, 又不失精度, 在上邻域则根据 KNN 算法快速获得分类结果.

本文主要针对 RSKNN 算法中存在的缺点进行改进, 在变精度粗糙集的基础上引入了类别子空间对属性加权的方法(SMwRSKNN), 该方法保留了 RSKNN 在一定精度下降低样本分类时计算量的特点, 并将待测样本投影到不同类别的子空间进行分类, 降低了高维空间中冗余或无用属性对分类的影响, 使该算法相比于 KNN 算法和 RSKNN 算法的精度在一定程度上有了提高, 并能保留 RSKNN 算法中利用集合刻画样本空间分布并进行相似性比较的优点.

2 背景知识

KNN 算法简单、实用、结合性强, 应用广泛, 其基本思想是先给定一个参数 $k(k\geq 1)$, 搜索训练集 $S=(X_1, X_2, \cdots, X_n)$(其中 n 表示训练集样本数量, X_i 表示样本点), 并按照距离度量(如: 欧几里得距离, 简称“欧式距离”)找出距离待测样本 v 最近的 k 个“最近邻”. 接着待测样本 v 的 k 个“最近邻”按照自身类别的投票方式决定待测样本 v 的类别. 本文提出的 SMwRSKNN 算法以 KNN 算法为基础, 结合了 RSKNN 算法、MKNN 算法, MwKNN 算法、SMwKNN 算法的优点. RSKNN 算法在刻画样本空间分布的基础上, 提高了同类样本相关性的对比; MKNN 则是提高了邻居样本与类别的关联性, 降低伪近邻的影响; MwKNN 则是在距离的度量上加以改进, 使得越紧密的邻居样本选择度更高; SMwKNN 算法使得在高维的数据空间中, 冗余或无用的属性得到约束, 与类别关联性大的属性占有更高权重, 使分类精度得到较大提高.

2.1 RSKNN 算法的基本思想

RSKNN 算法是基于变精度粗糙集的 KNN 算法改进而来, 在对模型训练中做了较大工作. 该算法先根据类别对训练集样本进行分类 $S=(X_1, X_2, \cdots, X_n)$(其中 n 表示训练集类别数量), 并引入一个精度 β ($0\leq\beta\leq 0.5$). 精度 β 也可以理解为错误容忍度, 在精度 β 的基础上计算出训练集 S 中每个类别的上、下近似域半径, 然后样本空间中划分出每个类的下邻域、上邻域和外界, 精度 β 代表下邻域中的异类样本分布率. 在分类样本时, 每个待测样本 v 先计算出与每个类中心点的距离, 如果样本 v 处于某类的下近似域中(又称下邻域), 那么就直接判断该待测样本的归属; 如果样本 v 处于某些类的上近似域中(又称上邻域), 那么只需在这些类的上近似域中寻找 k 近邻来判定归属; 如果样本 v 处于所有类的上近似域之外(该区域称为外界), 则该样本 v 属于距离上近似边界最近的那个类.

如图 1, 两个类的中心点分别为 $O(X_1)$和 $O(X_2)$, 其中 $O(X_1)$的下邻域半径为 $\underline{r}$, 上邻域半径为 $\bar{r}$. 此时为下邻域与另一个类的上邻域相交的情况.

从经典粗糙集理论对样本点分布进行描述:

设 v 表示待测样本点, X_i 表示训练集中第 i 类, $\overline{R}(X_i)$表示第 i 类的上邻域, $\underline{R}(X_i)$表示第 i 类的下邻

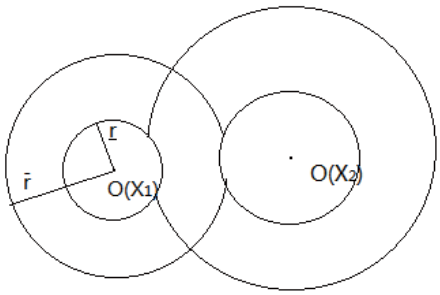

图 1 一个类的下邻域与另一个类的上邻域相交

域, 则

1) 若 $\exists v \in \underline{R}(X_i)$, 则 $v \in \overline{R}(X_i)$

2) 若 $\exists v \in \underline{R}(X_i)$, 则 $v \notin \overline{R}(X_j)(j \neq i)$

计算类 X_i 的 $\bar{r}$ (上近似域半径)和 $\underline{r}$ (下近似域半径)的算法:

Step1: 将训练集中的样本根据类别进行分组.

Step2: 计算类 X_i 的中心点 $O(X_i)$, 并将该类所有其他点按(1)式计算欧拉距离排序存放于 D_i(D_i 表示第 i 类的样本集合)中.

Step3: 将 D_i 中最大值作为上近似域半径, 即 $\bar{r}$ =max(D_i).

Step4: 搜索其他类集合中位于类 X_i 上近似域中的点, 并计算与 $O(X_i)$的距离, 按距离顺序插入到 D_i 中.

Step5: 从 D_i 中第一个点开始, 依次取前 k 个, 并且遵循下列公式 1[4]

$$\frac{NUM(k)}{k} \leq \beta, \quad \underline{r} = dist(v_k, O(X_i))$$

否则停止下半近似域半径的计算.

在模型训练完成之后, 便将样本空间根据类别进行了划分. 在此基础上根据待测样本的空间分布进行分类, 下面对分类算法进行描述:

RSKNN 分类算法:

Step1: 计算待测样本 v 到训练集中各类中心点的距离, 并找出距离最近的类中心点.

Step2: 比较待测样本 v 到该类中心点的距离 d 与该类上、下邻域半径的关系:

若 $d < \underline{r}$, 表明待测样本处于该类的下邻域中, 则直接判定待测样本属于该类;

若 $r < d \leq \bar{r}$, 表明待测样本处于该类的上邻域中, 则根据 KNN 算法判定归属;

若 $d > \bar{r}$, 表明待测样本处于该类的外界, 则将样本空间中上邻域最靠近待测样本 v 的类判定为待测样本的类.

2.2 MKNN 算法基本思想

MKnn 算法是基于 KNN 算法的改进算法, 对于一个集合 $S = \{s_1, s_2, \cdots, s_n\}$($n$ 表示样本数量), 给定一个系数 $k \in \{k | 0 < k < n, k \in Z\}$, 对于其中的每一个样本 s_i 都有对应的一个 k 最近邻集合 $N_i = \{t_1, t_2, \cdots, t_k\}$ ($N_i \subseteq S$). 若 s_i 与 s_j 两个样本为互 k 最近邻关系, 则有 $s_i \in N_j$ 并且 $s_j \in N_i$. 反之则称 s_i 与 s_j 不成互 k 最近邻关系. 拥有互 K 最近邻关系的 s_i 和 s_j, 我们称 s_i 为 s_j 的互 K 最近邻, s_j 为 s_i 的互 K 最近邻. 如果一个集合都是由样本 s 的互 k 最近邻组成的, 则称这个集合为样本 s 的互 K 最近邻集合. MKnn 算法能有效地降低伪近邻的影响.

2.3 MwKNN 算法基本思想

MwKNN 算法是基于权重的互 K 近邻算法, 互 K 近邻算法主要作用是消去了 KNN 算法中可能存在的“伪邻居”, 并且 MwKNN 在 KNN 算法的基础中引入了距离权重, 利用距离加权方式投票, 算出每个类别在互 K 最近邻集合中占有的距离权重比, 从中选中比例最大的类别作为待分类样本的类标签, 因此在距离度量上又提高了一定精度. MwKNN 算法的距离权重采用: w_i=1/d_i, 其中 d_i 为到待测样本点的距离.

2.4 SMwKNN 算法基本思想

SMwKNN 是基于类别子空间距离加权的互 K 近邻算法, SMwKNN 算法依据类别求得各个子空间, 增加不同类别属性上的区分度, 相比于传统的互 K 最近邻算法, 进一步增强了邻居之间的关系.

给定一个具有 n 个样本数量的数据集 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ (x_i 表示数据集中第 i 个样本, y_i 表示对应样本的类别)以及具有 $m(m > 1)$ 个类别标签的集合 $Y = \{1,2,...,m\}$. x_i 是一个 D 维的空间向量, $x_i = \{x_{i1}, x_{i2}, ..., x_{iD}\}(D > 1)$.

权重算法为(FWkM)[6]:

$$w_{ld} = \left(\sum_{t=1}^{D} \left(\frac{\sum_{i=1}^{n} \left((x_{id} - v_{id})^2 + \delta \right)}{\sum_{i=1}^{n} \left((x_{it} - v_{it})^2 + \delta \right)} \right)^{\frac{1}{\beta - 1}} \right)^{-1}$$

其中, v_{id} 为对应维度 d 上的均值, δ 为很小的一个值, 主要是为了避免分母为 0, β 为加权参数. 在后面的实验中 δ, β 分别取 $\delta = 10^{-4}$, $\beta = 1.5$.

特征权重矩阵为:

$$Weight_l = \begin{pmatrix} w_{l1} & & & \\ & w_{l2} & & \\ & & \cdots & \\ & & & w_{lD} \end{pmatrix}$$

其中 $\sum_{d=1}^{D} w_{ld} = 1; \forall d = 1,2,\cdots,D : w_{ld} \geq 0$

矩阵中的每一个元素代表子空间里某个维度的权重. 维度的权重越大, 表示该维度与类别的相关性越大. 反之, 权重越小则说明该维度与对应类别的相关性越小.

X_i 与 X_j 之间的距离权重公式为:

$$Dist_{w_l}(x_i, x_j) = \sqrt{\sum_{d=1}^{D} w_{ld}(x_{id} - x_{jd})^2}$$

SMwKNN 算法基本步骤:

输入 s: 待分类样本

K: 初始选择的最近邻个数

T={t_1, t_2, …, t_n}: 带有 n 个具有类标签样本的训练集合

输出: 待分类样本 s 的类别标签

Begin

Step1: 将 n 个训练样本依据对应的类别标签分为 m 个集合 $\{T_1, T_2, ..., T_m\}$;

Step2: 计算出每个属性的均值 $v_i (1 < i \leq D)$.

Step3: 依据公式(2-4)计算出每个类别对应的特征权重矩阵 $Weight_i$, 其中 i=1, 2, …, m;

Step4: 依据每个特征权重矩阵, 计算待测样本 s 的 K 个“最近邻”(其中公式(2-6)作为距离计算公式); 分别计算待测样本在每个子空间下类别距离权重比值;

Step5: 累加在各个子空间下类别的距离权重比值, 然后选择其中权重比值最大的类别作为待分类样本 s 的类别标签;

Step6: 返回待分类样本 s 的类别标签;

End

3 基于RSKNN模型的类别子空间加权距离的互K近邻算法思想(SMwRSKNN)

3.1 SMwRSKNN 算法概念

SMwRSKNN 算法是基于 RSKNN 算法基础上的一种改进算法. SMwRSKNN 算法的工作主要是降低 RSKNN 算法在距离度量上易受到伪近邻影响、对参数 K 值的依赖性较高、对于高维数据空间中冗余属性没有相应较好的处理方法这些问题进行改进. SMwRSKNN 算法保留了 RSKNN 算法利用变精度粗糙集理论的上、下近似描述各个类的分布, 提高同类样本的相似度比较, 并引入了类别子空间距离加权算法, 使高维数据空间中样本的冗余或无用属性得到了约束, 并在加权距离的互 K 近邻算法的应用下, 使邻居样本较好的和类别相联系, 降低了伪近邻对于分类效果的影响. SMwRSKNN 优势主要体现在属性球状分布的样本数据集, 及高维属性数据集上.

3.2 SMwRSKNN 算法步骤

SMwRSKNN 算法基本步骤

输入 s: 待分类样本

T={t_1, t_2, …, t_n}: 带有 n 个具有类标签样本的训练集合

K: 初始选择的最近邻个数

β: 精度参数

输出: 待分类样本 s 的类别标签

Begin

Step1: 将 n 个训练样本依据对应的类别标签分为 m 个集合 $\{T_1, T_2, ..., T_m\}$;

Step2: 计算出同类样本集合中, 每个属性的均值, 构造类 T_i 中心参考点 $v'_i = \{v_{i1}, v_{i2}, ..., v_{iD}\} (1 \leq i \leq m)$, 找出每个类中距离类 T_i 中心参考点最近的点为中心点 $O(T_i)$.

Step3: 依据公式 (3) 计算出每个类别对应的特征权重矩阵 $Weight_i$, 其中 i=1, 2, …, m;

Step4: 计算类 T_i 中所有点到中心点 $O(T_i)$的距离, 并按照欧拉距离排序存放于 D_i 中.

Step5: 将 D_i 中最大值作为上近似域半径, 即 $\bar{r} = \max(D_i)$.

Step6: 查找其他类中位于类 T_i 上近似域中的点, 并计算与 $O(T_i)$的距离, 插入到 D_i 中.

Step7: 从 D_i 中第一个点开始, 依次取前 K 个, 并且遵循下列公式:

$$\frac{NUM(k)}{k} \leq \beta, \quad \underline{r} = dist(v_k, O(T_i))$$

否则停止下半近似域半径的计算.

Step8: 计算待分类样本 s 与各类中心点的距离, 判断待分类样本 s 所处的位置. Step8 如果分类样本 s 处于某类的下近似域中, 则直接判定 s 归属此类; 如果分类样本 s 处于某些类的下近似域中, 则根据 SMwKnn 算法 Step 5 判定 s 的归属; 如果分类样本 s 处于所有类的边界域之外, 则 s 属于距离上近似边界最近的那个类.

End

3.3 SMwRSKNN 算法流程及分析

SMwRSKNN 算法流程见图 2. 首先在算法开始先给定一个精度 β, 精度 β 的合理取值范围在 0 到 0.5 之前, 在文献 4 中有进行说明. 给定一个参数 k, 在互近邻判断中作为参考值. 然后根据以上提到算法对各个类进行空间划分. 在上诉步骤完成之后, 便可以根据待测样本在样本空间所处的位置进行相应的分类算法.

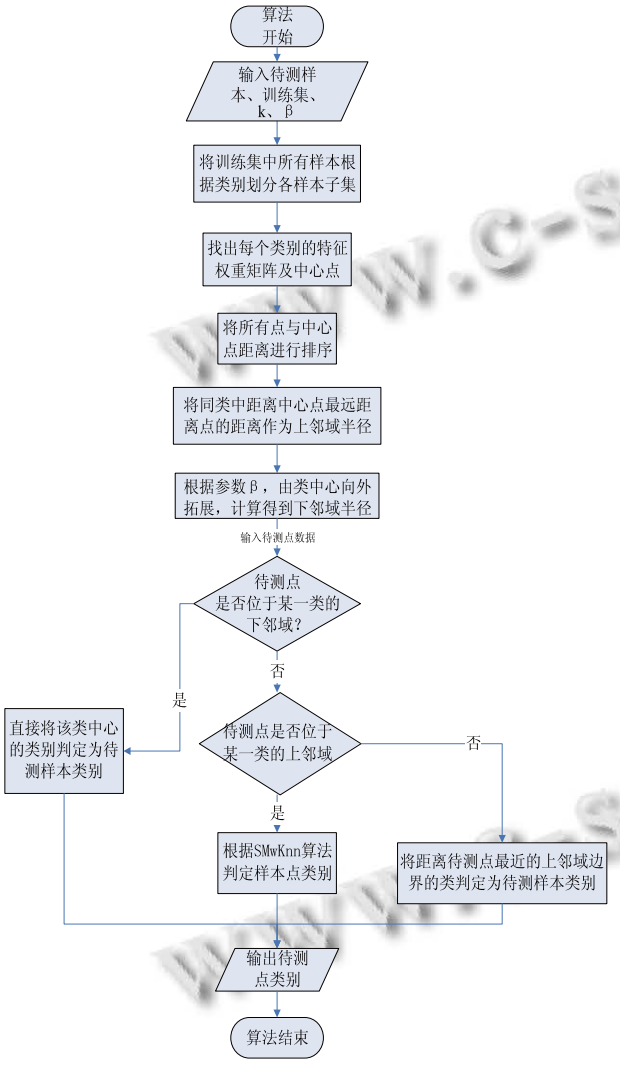

图 2 SMwRSKNN 程序流程图

可知 SMwRSKNN 算法相比于 RSKNN 算法在训练集的处理上需要更多的成本投入. 引入变精度粗糙集刻画样本分布空间的关键在于求出上、下邻域的范围, 并对训练集样本的空间分布进行记录. 从算法理论上, 同类样本的与中心点越紧密, 下邻域半径就越大, 同类相似性比较的优点就更突出. 待测样本点位于下邻域半径和上邻域边界之外的点可简化分类过程, 直接判断类别; 位于上邻域内的待测样本点便进行类别子空间投影, 然后根据互 K 近邻的思想进行分类.

4 实验环境与实验结果

4.1 实验环境和实验数据集

实验使用 15 个经典数据集进行实验, 数据集来自 UCI 机器学习公共数据集, 实验数据集相关信息见表 1. 为了验证 SMwRSKNN 算法的有效性, 实验还加入 RSKNN 算法及 KNN 算法进行同等条件实验来对比参照. 本实验对选取的 15 个数据集根据不同的精度 β 和 k 值进行实验比较, 并选取代表性的实验结果进行分析.

表 1 实验数据集相关信息

序号	数据集	实例个数	属性数目	类别数目	类分布
1	contact-lenses	24	4	3	15:5:4
2	column_2C	310	6	2	210:100
3	haberman	306	3	2	225:81
4	Hayes-Roth	132	5	3	51:51:30
5	breast-w	683	9	2	444:239
6	iris	150	4	3	50:50:50
7	wine	178	13	3	59:71:48
8	ecoli	336	7	8	143:77:2:2:35:20:5
9	ionosphere	351	34	2	225:126
10	balance-scale	625	4	3	49:288:288
11	Liver-disorders	345	6	2	145:200
12	tae	151	5	3	52:50:49
13	diabetes	768	8	2	268:500
14	glass	214	9	6	70:17:9:76:29:13
15	heart-statlog	270	13	2	120:150

4.2 实验结果

本实验采用 10 折交叉验证方法, 将进行实验的某一数据集随机均分为 10 个子集, 轮流选取一个子集作为待测样本集, 剩余 9 组作为训练样本集. 在 10 次验证中, 分别将待测样本集作为算法的输入, 累计记录每一次验证中的分类准确率, 最终求出 10 次验证的分类准确率平均值作为实验结果. 即分类准确率定义为 10 次验证中的平均分类精度.

在实验中采用 SMwRSKNN、RSKNN、KNN 三种算法对这 15 个数据集进行对比实验, K 的取值每次递增 10 个单位, 精度 β 的取值递减 0.05 个单位. 表 2 记录各算法在实验中能达到的最高准确率. 表 3 记录当 k=20、β=0.1 时, 各算法的分类准确率.

表 2　各算法在实验中得到的最高分类准确率(%)

数据集	SMwRSKNN	RSKNN	KNN
contact-lenses	86.67	58.3	58.33
column_2C	74.52	73.23	73.23
haberman	71.94	76.83	76.83
Hayes-Roth	76.46	41.69	42.46
breast-w	96.47	96.62	96.47
iris	96.00	97.33	97.33
wine	74.12	66.47	61.76
ecoli	74.55	75.45	75.15
ionosphere	94.00	82.57	82.57
balance-scale	87.90	88.39	88.39
liver-disorders	52.65	66.47	66.47
tae	70.00	32.67	32.67
diabetes	64.34	75.00	74.87
glass	62.86	61.90	63.81
heart-statlog	78.89	67.78	67.78

表 3　当 k=20、β=0.1 时各算法分类准确率(%)

数据集	SMwRSKNN	RSKNN	KNN
contact-lenses	86.67	58.33	58.33
column_2C	72.58	72.58	73.23
haberman	71.61	76.83	76.50
Hayes-Roth	76.46	41.02	41.49
breast-w	95.74	95.74	96.18
iris	94.00	94.00	97.33
wine	60.59	60.00	60.00
ecoli	74.24	73.03	71.82
ionosphere	92.29	82.57	81.43
balance-scale	87.90	88.39	88.39
Liver-disorders	52.35	64.71	64.12
tae	70.00	32.67	32.67
diabetes	63.55	74.74	74.87
glass	61.90	61.90	63.81
heart-statlog	78.89	67.78	67.78
平均值	75.92	69.62	69.86
平均值	77.42	70.72	70.54

从表 2 的数据中可以看出, 在 15 个数据集的实验中, SMwRSKNN 算法与 RSKNN 算法和 KNN 算法相比, 有 7 个数据集的实验结果准确率最高, 并且在 15 个数据集上的平均分类结果最好. 并且 SMwRSKNN 算法与 RSKNN 算法相比, 有 8 个数据集的实验结果准确率更高. 因此可以认为基于变精度粗糙集算法引入投影子空间概念的算法改进对提高分类精度是有效的. SMwRSKNN 能有效地从样本空间分布入手, 对密集的同类样本较准确的分类, 并且在高维的数据空间中能有效地降低伪近邻的影响, 减少样本的冗余属性对距离度量的影响. SMwRSKNN 算法是 RSKNN 算法的一种有效改进.

在实验结果中, 尤其对于数据集 ionosphere、tae、heart-statlog, SMwRSKNN 算法能有较好的分类效果. 这 3 个数据集的属性维度较高, 因此通过将样本投影到子空间的方法能有效地将与类别最为紧密的属性赋予更大的权重, 在距离度量的时候更加准确, 并且在互 K 近邻的运用上能降低伪近邻对于分类效果的影响.

4.3 实验分析

4.3.1 参数 β 和参数 k 的取值对分类精度的影响

[实验一] β 固定, K 取 5, 10, 20, 30 值时, SMwRSKNN, RSKNN 和 KNN 算法在 15 个数据集上的平均分类精度见图 3 所示.

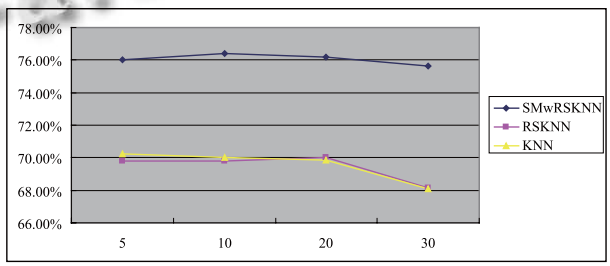

图 3　β 固定, K 不同取值时, 三种不同算法在 15 个数据集上的平均分类精度

由图 3 可以发现, SMwRSKNN 算法在 15 个数据集上的平均分类精度显著大于 RSKNN 算法和 KNN 算法, 并且 SMwRSKNN 算法的精度曲线随着 k 的增大有上升趋势, 随后在下降一部分后趋于稳定. 可见 k 值对于 SMwRSKNN 算法的精度影响很大, 同样也看出随着 k 值的增大, SMwRSKNN 算法与 RSKNN 算法和 KNN 算法的精度的区分度也逐渐拉大.

[实验二] k 值固定, β 取 0.2, 0.15, 0.1, 0.05, 0 值时, SMwRSKNN, RSKNN 和 KNN 算法在 15 个数据集上的平均分类精度见图 4 所示.

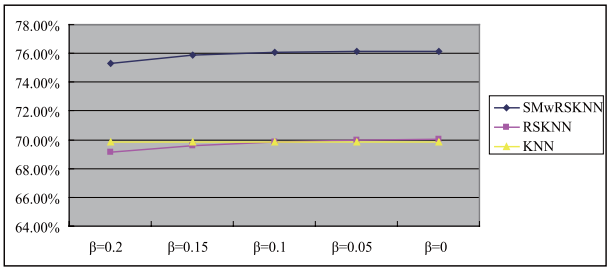

图 4　k 固定, β 不同取值时, 三种不同算法在 15 个数据集上的平均分类精度

由图 4 可以看出, k 固定时, β 的不同取值对分类精度有所影响, 随着 β 的取值由大变小, SMwRSKNN 算法和 RSKNN 算法的平均分类精度都有一定程度上的提高.

由以上两个实验可以看出, SMwRSKNN 算法相比于原算法能在一定程度上可以提高分类精度.

4.3.2 下邻域与上邻域外界的分类精度与简化效率分析

由于经典粗糙集的理论运用, 使待测样本点在处于某一类的下邻域或外界区域时可以直接进行简单的分类, 实现了对分类过程的简化. 因此, 简化效率可以定义为处于处于以上两个区域的待测样本点占所有待测样本点的比值.

[实验三] β 取 0.2, 0.15, 0.1, 0.05, 0 值时, 下邻域中的分类精度和简化效率见表 4. 在表 4 中, LC 表示下邻域分类精度, LR 表示下邻域简化效率. (由于实验中 k 值对样本在下邻域的划分上没有作用, 因此不做比较.)

表 4 下邻域中的分类精度和简化效率

序号	数据集	β =0.2		β =0.15		β =0.1		β =0.05		β =0	
		LC	LR	LC	LR	LC	LR	LC	LR	LC	LR
1	contact-lenses	70	11.67	100	0	100	0	100	0	100	0
2	column_2C	62.25	38.06	61.78	27.1	75	7.42	90	5.81	90	5.81
3	haberman	98.26	8	97.27	4	97.5	3	97.5	3	97.5	3
4	Hayes-Roth	46.67	6.82	46.67	6.82	46.67	6.82	46.67	6.82	46.67	6.82
5	breast-w	95.69	98.98	95.65	98.54	95.61	97.66	96	69.05	99.43	62.45
6	iris	94.57	86.67	94.36	84.67	97.13	74	98.26	61.33	98.26	60.67
7	wine	92.05	28.68	91.55	27.69	91.55	27.69	93.5	18.26	92..5	15.91
8	ecoli	58.38	44.59	58.75	43.38	58.75	41.56	76.67	12.47	77.12	11.91
9	ionosphere	90.94	44.67	93.09	41.24	95.3	37.52	96.4	25.29	98.57	7.12
10	balance-scale	99.38	22.1	99.58	16.94	99.57	14.52	99.17	7.9	100	5.65
11	liver-disorders	95	2.61	90	1.47	100	1.47	100	1.47	100	1.47
12	tae	90	2.67	90	2.67	90	2.67	90	2.67	90	2.67
13	diabetes	90.04	19.63	85.03	9.46	85.33	4.2	85	1.17	85	1.17
14	glass	84.17	13.1	84.17	13.1	79.17	14.06	87.5	12.15	87.5	12.15
15	heart-statlog	85	5.19	96.67	2.96	96.67	2.96	96.67	2.96	96.67	2.96

由表 4 可以看出随着 β 值的降低, 15 个数据集的分类精度总体上得到了提高, 但是简化效率有所下降. 简化效率同时也反映出同类样本空间分布的密集程度, 尤其 breast-w 和 iris 的密集程度最高, 同时在下邻域中的分类精度也都在 90%以上. 由此说明了同类样本在空间分布较为密集时, 既能达到较好的简化效率又能得到较高的分类精度.

[实验四] β 取 0.2, 0.15, 0.1, 0.05, 0 值时, 上邻域外界的分类精度和简化效率见表 5. (由于实验中 k 值和 β 值对外界区域的划分没有作用, 因此不做比较.)

对于上邻域之外的分类精度和简化效率分析, 由表 5 可以看出上邻域之外的空间只有极少数的样本分布, 并且与 β 取值没有明显的关系. 但是在这少数的分布情况下, 15 个数据集的平均分类精度达到了 91.11%. 由此说明, 在上邻域之外的待测样本只需根

表 5 上邻域外界的分类精度和简化效率

数据集	分类精度（%）	简化效率（%）
contact-lenses	90	8.33
column_2C	100	0.65
haberman	90	0.33
Hayes-Roth	100	0
breast-w	100	0.15
iris	100	1.33
wine	90	1.76
ecoli	73.33	5.45
ionosphere	100	0.57
balance-scale	100	0
liver-disorders	100	0.29
tae	90	0.67
diabetes	90	0.13
glass	53.33	6.19
heart-statlog	90	1.11

据上邻域边界距离它最近的类就可以决策出待测样本的类别, 并且能得到很高的准确率.

由以上两个实验可以看出, 在下邻域和上邻域外界中分布的待测样本可以简化分类过程的计算, 并且能保持较高的精度. 因此不需要在这两块区域内进行类别投影, 降低分类过程一定的成本.

4.4 SMwRSKNN 算法缺点分析

从 SMwRSKNN 算法在不同数据集上的实验结果中可以得知, SMwRSKNN 算法具有如下优点: 1.在高维数据中, 分类精度能起到比较明显的提高; 2.在球状分布的样本空间中, 同类的相似性更加显著, 分类精度有所较高; 3.算法对属性和 k 值的依赖度较低, 能较好地处理伪近邻的影响; 4.算法思路简单、清晰, 可结合性较强, 能和其他 KNN 改进方法相结合.

同样, SMwRSKNN 算法在下列情况下存在一定的局限性: 1.如果数据集的空间分布没有呈现球状分布(如: 带状分布), 变精度粗糙集算法就失去了它的优势, 并在一定程度上可能造成由算法本身存在的 β 参数导致算法结果的精度下降; 2.如果两种不同类别的样本对同一或某几个属性具有相同或相近的依赖程度, 就有可能因为类别子空间属性的加权导致两种类别的待测样本在距离度量上的区分度降低, 导致分类精度下降.

5 总结

本文基于变精度粗糙集分类算法, 引入了类别子空间距离加权的概念. SMwRSKNN 算法以 RSKNN 算法为基础, 刻画好各类的分布范围, 再根据待测样本点处于不同的空间位置采用不同的处理方法, 使下邻域中和上邻域外界分布的待测样本可以在保证精度的前提下, 并且简化分类计算. 上邻域中, 通过样本投影到不同类子空间进行加权运算, 保证了在高维数据空间中能消减伪近邻的影响. 在类的判定中, 通过互 k 最近邻关系的筛选, 增强了分类样本与邻居样本类别之间的相关性, 同时降低了传统算法中对 k 值的依赖, 在总体上起到了提高分类精度的效果. 下一步研究的重点在于如何更加准确的刻画类的分布, 在保证邻域半径的前提下, 降低下邻域中的异类样本. 同时降低训练集中刻画各类分布的成本.

致谢

本项目受国家自然科学基金(No.61070062, No.61175123)和福建高校产学合作科技重大项目(No.2010H6007)的资助, 特此表示感谢.

参考文献

1 Guo GD, Wang H, Bell D. KNN model-based approach in classification. Proc. of the OTM Confederated International Conference on CoopIS, DOA, and OD BASE. Catania, Italy. 2003. 986–996.

2 Liu HW, Zhang SC. Noisy data elimination using mutual k-nearest neighbor for classification mining. The Journal of Systems and Software, 2012, (85): 1067–1074.

3 张著英,黄玉龙,王翰虎.一个高效的 KNN 分类算法.计算机科学,2008,35(3):170–172.

4 余鹰,苗夺谦,刘财辉,王磊.基于变精度粗糙集的 KNN 分类改进算法.模式识别与人工智能,2012,25(4):618–623.

5 Dudani SA. The distance-weighted kNearest neighbor rule. IEEE Trans. on System, Man and Cybernetics, 1976, SMC-6(4): 325–327.

6 Huang JZ, Ng MK, Rong H. Automated variable weighting in k-means type clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657– 668.

7 Guo GD, Neagu D. Fuzzy kNNModel applied to predictive toxicology data mining. International Journal of Computational Intelligence and Applications, 2005, 5(3): 321–333.

8 李荣陆,胡运发.基于密度的 KNN 文本分类器训练样本裁剪方法.计算机研究与发展,2004,41(4):539–545.

9 张健飞,陈黎飞,郭躬德,李南.多代表点子空间分类算法.计算机科学与探索,2011,(11):1037–1048.

10 李南,郭躬德.基于子空间集成的概念漂移数据流分类算法.计算机系统应用,2011,20(12):241–248.

11 Huang XM, Guo GD, Neagu D. Weighted kNN model-based data reduction and classification. Fuzzy Systems and Knowledge Discovery, 2007.

12 卢伟胜,郭躬德,严宣辉,陈黎飞.SMwKNN:基于类别子空间距离加权的互 K 近邻算法.计算机科学,2013.

13 UCI Repository of Machine Learning Databases http://archive.ics.uci.edu/ml/.