基于 Markov 性的半监督流行学习算法研究

沈亮^{①②}, 许青松^{①*}, 曹东升^③, 黄新^④

① 中南大学数学与统计学院, 长沙 410083;

② 临沂大学理学院, 临沂 276005;

③ 中南大学药学院, 长沙 410083;

④ 湖南城市学院数学与计算科学学院, 益阳 413099

E-mail: liangshen607@csu.edu.cn, qsxu@csu.edu.cn, dscm66@126.com, xinhuang1516@sina.com

收稿日期: 2015-01-23; 接受日期: 2015-03-14; * 通信作者

国家自然科学基金 (批准号: 11271374) 资助项目

摘要 本文在数据驱动模式的指导下, 以药物动力学中的复杂数据为背景, 利用具有 Markov 性的流行学习来构造半监督分类算法, 并对生物结构与活性关系进行统计建模. 模拟研究表明, 当存在某种流行结构时, 流行学习算法能对具有不同模式的数据进行更好地分类. 通过对不同化合物的生物结构活性关系的三个真实数据集进行分析, 与典型算法相比, 基于不同核函数的流行学习分类算法能充分利用有标签的样本点和无标签的样本点的信息来对样本进行分类, 并产生更稳健、更精确的分类结果.

关键词 转移概率矩阵 Markov 性 模式识别 流行学习

MSC (2010) 主题分类 60J22, 62P10

1 引言

统计模式识别是建立在概率论与数理统计基础上的, 它用特征向量来描述模式. 不同的模式用不同条件的概率来表示, 然后判别未知向量属于哪一类 (参见文献 [1-3]). 分类方法主要有 Bayes 决策方法、线性可分的几何分类方法、非线性可分的几何分类方法、有监督的参数学习分类方法、无监督的参数学习分类方法和聚类分析方法 (参见文献 [3, 4]). 常用的分析方法主要包括线性建模和非线性建模两大类, 其中线性判别分析^[5, 6] (linear discriminant analysis, LDA)、主成分回归^[7-9] (principal component regression, PCR) 和偏最小二乘^[10-13] (partial least squares, PLS) 是普遍采用的线性分类方法. 但是, 由于数据的复杂性, 导致了样本中各因素之间的相关关系难以用简单的线性模型来描述, 因此越来越多的非线性算法被广泛地应用于多领域多行业的研究. 典型的非线性算法有 k -最近邻算法^[14-17] (k -nearest neighbor, k -NN)、支持向量机^[18, 19] (support vector machine, SVM) 和分类回归树^[20] (classification and regression tree, CART) 等.

对于有监督参数学习分类问题, 通过已有的训练样本 (即已知数据以及其对应的标签) 去训练得到一个最优模型, 再利用这个模型对未知样本进行预测从而实现分类的目的, 也就具有了对未知数据进行分类的能力. 典型的有监督学习分类算法包括 LDA、 k -NN 和 SVM 等. 无监督学习则不需要训练样本的标签信息而只需要对数据进行直接建模, 常用的无监督学习方法有主成分分析 (principal component analysis, PCA) 和聚类分析^[1-3] (clustering analysis) 等.

英文引用格式: Shen L, Xu Q S, Cao D S, et al. The research of semi-supervised manifold learning algorithm based on markov property (in Chinese). Sci Sin Math, 2015, 45: 703-712, doi: 10.1360/N012015-00040

在实际分类问题中, 由于对样本进行标签通常需要大量的人力、物力和财力, 所以用于分类的标签点通常是有限的, 这样对于大多数分类算法而言就可能导致预测效果较差. 因此, 分类学习算法 (特别是有监督的) 都面临着如何从大量的无标签训练集中来挖掘数据的信息这个难题. 所以, 如何进行有效的模式识别并改善分类器的预测能力仍然是当今应用概率统计所面临的问题. 在模式识别中核函数的框架下, 本文结合具有 Markov 性的流行学习来构造一类半监督的统计算法, 并对模拟数据和三个真实的药物动力学数据进行分类. 实验结果表明, 一类基于 Markov 性的半监督流行分类算法^[4, 21] (manifold semi-supervised classification, MSSC) 可以挖掘样本中内在的流行结构并且可以有效地处理非线性分类问题.

2 理论及方法

2.1 流行学习的算法原理

作为一种半监督学习算法, 流行学习算法 (MSSC) 是由文献 [4, 21] 针对模式识别问题提出的, 已经被广泛的应用于图像处理、文本分类和信号识别等领域 (参见文献 [21]). 根据有标签的数据集和无标签的数据集所反映出来的内在结构, 流行学习可以更好地进行分类, 特别是对“病态”的分类问题 (标签样本少、维数高) 也能得到有效的、稳定的预测效果.

给定一个数据集

$$\chi = \{x_1, \dots, x_l, x_{l+1}, \dots, x_n\} \subset \mathbb{R}^n,$$

标签类别集合为 $L = \{1, \dots, c\}$, 前 l 个样本 x_i ($i < l$) 是所对应的标签 $y_i \in L$, 其余的点 x_u ($l+1 < u < n$) 是无标签的样本. 令 $\mathcal{F}$ 表示元素非负的 $n \times c$ 的矩阵所组成的集合. 矩阵 $F = [F_1^T, \dots, F_n^T]^T \in \mathcal{F}$ 可以理解成一个向量函数 $F: \chi \rightarrow \mathbb{R}^c$, 每一个向量 F_i 与 x_i 相对应, 每个 x_i 的标签为

$$y_i = \arg \max_{j \leq n} F_{ij}.$$

定义 $n \times c$ 的矩阵 $Y \in \mathcal{F}$, 如果 x_i 是标签样本点且对应的标签 $y_i = j$, 则令 $Y_{ij} = 1$; 否则为 0. 很明显, 矩阵 Y 与初始标签点相一致. 具体算法如下:

第 1 步 首先构建邻接矩阵 W ,

$$W = \{w_{ij} \mid i, j = 1, \dots, n\},$$

其中 w_{ij} 表示 x_i 与 x_j 间的相似性测度. 设 $G = (V, E)$ 表示一个图, 其中 V 代表所有节点 (样本点) 组成的集合, E 表示边. 边的权重用邻接矩阵中的相似性测度 w_{ij} 来表示, 若 $w_{ij} > 0$, 表示节点 x_i 与 x_j 相连且权重为 w_{ij} ; 若 $w_{ij} = 0$, 则称 x_i 与 x_j 是不相连的. 同时, 为了消除节点的自增强性, 令 $w_{ii} = 0$.

第 2 步 对矩阵 W 标准化从而生成转移概率矩阵 S ,

$$S = D^{-1/2} W D^{-1/2},$$

其中 D 为对角矩阵并且 D 为 W 中第 i 行元素之和.

第 3 步 基于转移概率矩阵 S , 利用公式

$$F(t+1) = \alpha S F(t) + (1-\alpha) Y \quad (2.1)$$

进行迭代, 其中 $\alpha \in (0, 1)$. 从而, Y 中标签点的初始得分值 1 就可以通过迭代被不断地传递.

第 4 步 令 F^* 表示序列 $\{F(t)\}$ 的收敛值, 则每个数据 x_i 的标签为 $y_i = \arg \max_{j \leq n} F_{ij}^*$. 收敛性的证明请参见文献 [21].

直观上来说, 如果把算法中的图 G 表示成一个复杂网络的话, 则 w_{ij} 表示从一个节点转移到另一个节点的概率. 此过程相当于一个 Markov 链, 具有 Markov 性. 如果 x_i 没有向外的连接, 则 W 中的第 i 行为 0. 通过第 2 步标准化之后, S 中则不会有行和为 0 的行, S 则为一个随机矩阵. 通过不断地迭代, 最终网络达到一个平稳分布. 平稳分布中各节点的概率值即为节点属于不同类别的得分值.

2.2 一类流行学习算法的构造

由于大量实际问题是具有多类别的高维复杂模式的识别, 因此研究复杂模式的分析和分类方法是必要而且有意义的. 基于核函数的学习方法 (简称核方法) 本质上是非线性的信息处理工具, 它在处理具有非线性关系的高维复杂模式识别问题时, 有着其他方法无法比拟的优越性. 核方法是解决非线性模式分析问题的一种有效途径, 其核心思想是, 首先, 通过某种非线性映射将原始数据嵌入到合适的高维特征空间; 然后, 利用通用的线性学习器在这个新的空间中分析和处理模式. 核的构建是一个关键的部分, 它体现着数据信息, 可以通过构建不同的核来解决不同的问题, 如线性核、Gauss 径向基核函数、max-min 核、文本核、图核和谱核^[1-3]等. 在实际应用中, 同一种算法可以与不同的核函数相结合, 以解决我们所研究的问题. 这样我们只需要通过选择适当的核函数就能设计许多灵活有效的方法. 在核方法的应用过程中, 核的选择是一个关键的步骤, 我们可以根据数据的特点, 选择不同的核函数来构造流行学习算法.

下面给出构造一类 MSSC 方法的步骤:

第 1 步 结合不同的核函数. 在复杂数据中, 相似性测度在挖掘数据信息的过程中起到了非常重要的作用. Kokare^[22] 验证了 L_1 距离要比常用的 L_2 Euclid 距离对图像信息的提取有显著的改进作用. 因此, 针对我们所研究的药物动力学中高维数据的特征, 本文除了应用最常用的 Gauss 径向基核函数 (RBF-Kernel), 我们还将利用 Laplace 核函数 (LAPL-Kernel) 来构造邻接矩阵 W .

第 2 步 给定一个参数 σ 以及 α , 通过 (2.1) 进行迭代.

第 3 步 根据迭代结果 $y_i = \arg \max_{j \leq n} F_{ij}^*$ 来预测新样本的类别.

为了更直观地表达半监督流行算法的效果, 我们仿照文献 [21] 给出一组模拟数据, 包含两个类别共 160 个样本, 其中两个类别的样本各 80 个. 上半圆形结构的样本属于第一类, 下半圆形结构的样本属于第二类, 见图 1(a). 然后分别用 k -近邻算法 (k -NN)、Gauss 径向基 - 支持向量机 (RBF-SVM) 和 Laplace 核流行算法 (LAPL-MSSC) 来对模拟数据图 1(a) 进行分类. 首先在两类中分别选择一个标签点, 用大 “*” 来表示. 图 1(b) 表示用 k -NN ($k = 1$) 的分类结果, 从图中可以看出, 阴影部分的标号为错分的类别. 图 1(c) 表示用 RBF-SVM 对模拟数据进行分类的结果, 其中选择的最优参数为 $\delta = 0.25$, $C = 1$, 可以看出 RBF-SVM 的效果比 k -NN 要明显好一些, 但还是有一些样本被错分. 图 1(d) 表示用 LAPL-MSSC 的分类结果, 其中选择的最优参数为 $\alpha = 0.9$, $\delta = 0.1$, $\text{iters} = 400$, 可以看出流行学习算法可以完全地根据数据的结构进行正确分类. 此外, 为了说明流行学习算法中参数选择的重要性, 我们演示了随着参数的变化分类结果的差异性, 见图 2. 在图 2 中, (a)–(c) 表示当参数 δ 和 α 固定时, 迭代次数 iters 的变化对模型分类的影响; (d)–(f) 表示当参数 α 和迭代次数 iters 固定时, δ 的变化对于模型分类的影响; (g)–(i) 表示当参数 δ 和迭代次数 iters 固定时, α 的变化对于模型分类的影响.

图 1 模拟数据的分类图. (a) 两类各含 80 个点的原始数据; (b) k -NN 的分类结果; (c) SVM 的分类结果; (d) Laplace 核的流行分类结果

图 2 参数选择对于 MSSC 模型的重要性. (a) $\delta = 0.01$, $\alpha = 0.9$, iters = 5; (b) $\delta = 0.01$, $\alpha = 0.9$, iters = 10; (c) $\delta = 0.01$, $\alpha = 0.9$, iters = 300; (d) $\delta = 1$, $\alpha = 0.9$, iters = 400; (e) $\delta = 0.2$, $\alpha = 0.9$, iters = 400; (f) $\delta = 0.05$, $\alpha = 0.9$, iters = 400; (g) $\delta = 0.01$, $\alpha = 0.1$, iters = 400; (h) $\delta = 0.01$, $\alpha = 0.6$, iters = 400; (i) $\delta = 0.01$, $\alpha = 0.9$, iters = 400

3 数据与软件

结构与活性关系^[23] (structure activity relationship, SAR) 研究是现今分析化学较为活跃的领域之一, 特别是在药物设计中得到了广泛应用. 其目的是, 通过现代统计方法建立起化合物的生物活性或某种性质 (如药效学性质和药物的毒性等) 同理化性质参数或者结构参数 (包括二维、三维分子结构参数等) 之间的定量关系. 我们选择三个与药物药效属性有关的数据集, 从 SAR 的两分类上来评估 MSSC 算法的性能. 下面给出这三个数据集的简要说明, 其中包括样本的数量、类别和描述子 (变量) 等. 详见表 1.

3.1 HIA 数据集

人类肠细胞膜对药物化合物的吸收性对选择一个候选药物来说是一个重要的属性. 在肠细胞膜的吸收 (HIA) 数据集中有 131 个可吸收的 (HIA+) 和 65 个不可吸收 (HIA-) 的化合物, 按照 70% 以上的药物被吸收则认为可吸收, 否则认为不可吸收这个标准进行分类. HIA 数据集及下面的 P 糖蛋白培养基 (P-gP) 数据集和药物引起的尖端扭转型室 (TdP) 数据集均来自文献 [24, 25]. 我们使用 Xue 等人给出的 159 个描述子来讨论这三个数据集, 其中包括 13 个量子化学属性、16 几何属性、18 简单分子属性、28 分子连接和形状的描述子及 84 电拓扑描述子.

3.2 P-gP 数据集

P-gP 是一种跨膜蛋白质, 它可以把大量的不同种类的抗癌药物运送出细胞, 从而阻碍化学疗法. P-gP 的增加与多向性抗药有关 (multidrug resistance, MDR). 许多研究人员正努力研究 MDR 的逆转化合物是否在临床中有显著的效果. P-gP 培养基 (P-gP+) 表示能够被 P-gP 或者 P-gP 的 MDR 反物质转换, P-gP 非培养基 (P-gP-) 表示不能被 P-gP 转换. 数据集 P-gP 包含两类, 116 个培养基和 85 个非培养基.

3.3 TdP 数据集

TdP 是一种潜在致命的多形室性心动过速疾病 (potentially fatal polymorphic ventricular tachycardia). 这种疾病可能是由于那些引起 QT 延长的药物的不良反应所产生. 临床中, 这种不良的反应存在于很多不同的治疗药剂中, 如抗组织胺类 (antihistamines) 和抑郁药 (antidepressants) 等. TdP 数据集含有两类诱导剂, 85 个 TdP 诱导剂 (TdP+) 和 276 个非诱导剂 (TdP-).

3.4 软件

我们使用统计分析中的 R 软件 (3.1.1 版本) 并结合软件包 (“kernlab package”) 来对算法编程, 计算机运行环境为 Intel Pentium processors4/2.6GHZ 4.00 GB RAM.

表 1 使用的三个数据集

数据集	所预测的属性	化合物数目	类别 1	类别 2	描述子
HIA	Absorbable/Nonabsorbable	196	131	65	159
P-gP	P-gP Substrates/Nonsubstrates	201	116	85	159
TdP	TdP-including agents/Nonincluding	361	85	276	159

4 结果与讨论

为了评估 MSSC 方法在 SAR 研究中的性能, 作为参考标准, 我们选择 k 近邻 (k -NN)、线性判别分析 (LDA)、线性函数与 Gauss 径向基核函数的支持向量机 (L-SVM 和 RBF-SVM) 来作为对照的比较方法. 模型主要是通过两种检验方式来进行, 分别为三折内部交互检验^[26, 27] (three-fold cross validation) 和外部独立检验 (independent validation set) 来评估.

首先对数据预处理, 使每一个预测变量的均值为零, 长度等于单位长度. 在分类问题中, 我们采用一些常见的分类指标来评估 k -NN、LDA、SVM 和 MSSC 方法的性能: 正确的正分类 (true positives, TP)、正确的负分类 (true negatives, TN)、错误的正分类 (false positives, FP) 和错误的负分类 (false negatives, FN). 综合这些指标可以得到几个常用的评价模型性能的标准: 敏感度 (sensitivity, SE)、特异性 (specificity, SP)、综合预测精度 (the overall prediction accuracy, Acc) 和 Matthews 相关系数 (Mcc), 它们对应的公式如下:

$$SE = \frac{TP}{TP + FN}, \quad (4.1)$$

$$SP = \frac{TN}{TN + FP}, \quad (4.2)$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \quad (4.3)$$

$$Mcc = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}. \quad (4.4)$$

4.1 三个 SAR 数据的内部验证结果

对于三折交互检验, 首先把数据集大致均分为三部分, 然后用其中的两部分拟合模型, 另一部分计算误差. 这个过程重复 3 次以使得每一部分都能作为验证集被预测. 表 2 列出三个数据集的三折交互检验结果. 根据经验, 当迭代次数大约为 200 时, 公式 (2.1) 达到收敛. 为此我们设定迭代次数 $iters = 400$, 同时对 k -近邻的最优近邻参数 k 通过网格搜寻进行优化. 对 L-SVM 和 RBF-SVM, 正则化参数 C 和 Gauss 核函数的宽度 δ 通过交互检验结合网格搜寻来确定. 对 RBF-MSSC 和 LAPL-MSSC, 宽度 δ 和参数 α 也通过交互检验结合网格搜寻来确定模型的最优参数值.

从表 2 来看, LDA 和 L-SVM 算法的预测精确度最低. 这可能暗示在三个 SAR 数据集中存在非线性结构. 同时, 它也表明 MSSC 能够被用来处理非线性问题. 对 HIA 数据, LAPL-MSSC 取得了 82.14% 的最好综合预测精度, 分别获得了 90.08% 的敏感度、66.15% 的特异性和 58.6% 的 Matthews 相关系数. 同样, RBF-MSSC 也取得了不错的分类效果. 对 P-gP 数据, RBF-MSSC 分别赢得了 78.11% 的预测精度、85.34% 的敏感度、68.24% 的特异性和 54.73% 的 Matthews 相关系数. 这些结果在所考虑的方法中是最好的. 比 RBF-SVM 93.97% 的敏感度, RBF-MSSC 方法的敏感度有点低. 然而, MSSC 的精度和特异性明显地优于 RBF-SVM (78.11% 与 77.61%, 68.24% 与 55.29%). 此外, LAPL-MSSC 也达到了较好的分类效果. 对 TdP 数据, LAPL-MSSC 和 RBF-MSSC 的预测结果优于 k -NN 和 LDA, 且 LAPL-MSSC 的预测精度与 L-SVM 一致, 但是略低于 RBF-SVM. 这可能是由于数据的不平衡性所引起的. 总体上, 比较 k -NN、LDA、L-SVM 和 RBF-SVM, MSSC 算法的预测性能是优良的或可与其他方法相比较的.

通过比较其他 k -NN、LDA 和 SVM 模型, 可以清楚地看到在构建一类基于流行学习算法的过程中, 我们利用了有标签的样本点而且还利用无标签的样本点的信息, 从而能够有效地发现样本之间的

相似性. 而在构造 k -NN、LDA 和 SVM 监督的模型中, 我们仅仅利用标签样本点的信息. 故当样本标签点较少时, 流行学习算法就能发挥较大的作用. 所以, MSSC 算法所构建的模型属于半监督的, 而 k -NN、LDA 和 SVM 所构建的模型属于有监督的. 而且, 由于结合了模式识别中的核方法, MSSC 可以有效地处理非线性问题. 通过与 LDA 和 L-SVM 的比较, 这一点已经被证实.

模型最优参数选择如下: HIA 数据集: 对 k -NN, $k = 5$; 对于 L-SVM, $C = 1$; 对于 RBF-SVM,

$$\delta = 0, \quad \delta = 0.0002, \quad C = 100;$$

对于 LAPL-MSSC, $\alpha = 0.4$, $\delta = 1$; 对于 RBF-MSSC,

$$\alpha = 0.3, \quad \delta = 0.125.$$

P-gP 数据集: 对于 k -NN, $k = 5$; 对于 L-SVM, $C = 0.1$; 对于 RBF-SVM,

$$\delta = 0.002, \quad C = 1;$$

对于 LAPL-MSSC, $\alpha = 0.1$, $\delta = 1$; 对于 RBF-MSSC,

$$\alpha = 0.5, \quad \delta = 0.125.$$

TdP 数据集: 对于 k -NN, $k = 3$; 对于 L-SVM, $C = 0.1$; 对于 RBF-SVM, $\delta = 0.004$, $C = 10$; 对于 LAPL-MSSC, $\alpha = 0.6$, $\delta = 1$; 对于 RBF-MSSC, $\alpha = 0.7$, $\delta = 0.125$.

表 2 三个数据集的 3 折交叉验证结果

数据集	方法	TP	FN	TN	FP	SE	SP	Acc	Mcc
HIA	k -NN	122	9	29	36	0.9313	0.4462	0.7704	0.4495
	LDA	84	47	36	29	0.6412	0.5538	0.6122	0.1859
	L-SVM	104	27	39	26	0.7939	0.6000	0.7296	0.3924
	RBF-SVM	116	15	41	24	0.8855	0.6308	0.8010	0.5380
	LAPL-MSSC	118	13	43	22	0.9008	0.6615	0.8214	0.5860
	RBF-MSSC	121	10	35	30	0.9237	0.5385	0.7959	0.5173
P-gP	k -NN	96	20	51	34	0.8276	0.6000	0.7313	0.4419
	LDA	68	48	48	37	0.5862	0.5647	0.5771	0.1493
	L-SVM	94	22	58	27	0.8103	0.6824	0.7562	0.4973
	RBF-SVM	109	7	47	38	0.9397	0.5529	0.7761	0.5490
	LAPL-MSSC	96	20	60	25	0.8276	0.7059	0.7761	0.5384
	RBF-MSSC	99	17	58	27	0.8534	0.6824	0.7811	0.5473
TdP	k -NN	246	30	42	43	0.8913	0.4941	0.7978	0.4092
	LDA	226	50	56	29	0.8188	0.6588	0.7812	0.4450
	L-SVM	245	31	44	41	0.8877	0.5176	0.8006	0.4239
	RBF-SVM	250	26	56	29	0.9058	0.6588	0.8476	0.5718
	LAPL-MSSC	251	25	38	47	0.9094	0.4471	0.8006	0.3985
	RBF-MSSC	263	13	29	56	0.9529	0.3412	0.8089	0.3891

4.2 三个 SAR 数据的外部验证结果

为进一步表明 MSSC 模型的预测能力, 我们还采用独立的测试集来评估, 即外部验证. 首先数据集被随机地分成训练集和测试集两部分. 为使两类样本能够比较均衡地分成两部分, 数据集按类别随机划分, 训练集和测试集分别占总数据样本的 75% 和 25%. 测试集的作用是用来评估模型的性能. 在一定范围内通过网格搜寻策略, 应用训练集来优化参数值, 最优参数值见表 3. 表 3 列出了六种方法在测试集上的预测结果, 从表中可清楚地看到 MSSC 算法取得了满意的预测性能.

模型最优参数选择如下: HIA 数据集: 对 k -NN, $k = 4$; 对于 L-SVM, $C = 0.1$; 对于 RBF-SVM,

$$\delta = 0.008, \quad C = 10;$$

对于 LAPL-MSSC, $\alpha = 0.4$, $\delta = 1$; 对于 RBF-MSSC, $\alpha = 0.3$, $\delta = 0.125$. P-gP 数据集: 对于 k -NN, $k = 3$; 对于 L-SVM, $C = 0.1$; 对于 RBF-SVM,

$$\delta = 0.002, \quad C = 100;$$

对于 LAPL-MSSC, $\alpha = 0.6$, $\delta = 1$; 对于 RBF-MSSC, $\alpha = 0.6$, $\delta = 0.25$. TdP 数据集: 对于 k -NN, $k = 3$; 对于 L-SVM, $C = 1$; 对于 RBF-SVM,

$$\delta = 0.008, \quad C = 10;$$

对于 LAPL-MSSC, $\alpha = 0.5$, $\delta = 1$; 对于 RBF-MSSC, $\alpha = 0.2$, $\delta = 0.125$.

表 3 六种方法在独立测试集上的预测结果

数据集	方法	TP	FN	TN	FP	SE	SP	Acc	Mcc
HIA	k -NN	30	0	8	11	1	0.4211	0.7755	0.5551
	LDA	23	7	37	39	0.7667	0.4868	0.5660	0.2318
	L-SVM	27	3	13	6	0.9000	0.6842	0.8163	0.6070
	RBF-SVM	29	1	12	7	0.9667	0.6316	0.8367	0.6602
	LAPL-MSSC	30	0	11	8	1	0.5789	0.8367	0.6761
	RBF-MSSC	30	0	11	8	1	0.5789	0.8367	0.6761
P-gP	k -NN	19	8	14	9	0.7037	0.6087	0.6600	0.3137
	LDA	18	9	12	11	0.6667	0.5217	0.5378	0.1400
	L-SVM	18	9	18	5	0.6667	0.7826	0.7200	0.4493
	RBF-SVM	18	9	17	6	0.6667	0.7319	0.7000	0.4048
	LAPL-MSSC	23	4	13	10	0.8519	0.5652	0.7200	0.4388
	RBF-MSSC	23	4	14	9	0.8519	0.6087	0.7400	0.4782
TdP	k -NN	13	11	54	12	0.5417	0.8182	0.7444	0.3553
	LDA	14	10	57	9	0.5833	0.8636	0.7889	0.4532
	L-SVM	11	13	56	10	0.4583	0.8485	0.7444	0.3208
	RBF-SVM	15	9	59	7	0.6250	0.8939	0.8222	0.5340
	LAPL-MSSC	15	9	58	8	0.6250	0.8788	0.8111	0.5108
	RBF-MSSC	9	15	64	2	0.3750	0.9697	0.8111	0.4654

5 总结

基于核方法的现代统计分析已经成为解决复杂数据的一个强有力的新技术. 本文针对药物动力学中复杂数据的结构特点, 结合不同的核方法构造了一类具有 Markov 性的半监督流行学习分类算法. 在核方法的框架下, 流行学习算法可以利用有标签和无标签样本点的信息, 从而能够挖掘出数据的结构特征. 实验表明, 这类流行学习算法能够有效的发现隐藏在数据中的模式, 其分类性能有明显改进.

致谢 非常荣幸能在侯振挺教授 80 华诞之际, 与我的博士生导师许青松合作这篇文章来表达祝福. 侯振挺教授和阎国军教授是我的研究生导师, 在中南大学概率论与数理统计研究所读书期间, 侯振挺教授不但教授我做学问的方法, 还教导我做人的道理, 使我终生受益匪浅.

参考文献

- 1 Vapnik V N, Vapnik V. Statistical Learning Theory. New York: Wiley, 1998
- 2 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 2000
- 3 Hastie T, Tibshirani R, Friedman J, et al. The elements of statistical learning: Data mining, inference and prediction. Math Intelligencer, 2005, 27: 83–85
- 4 Zhou D Y, Burges Christopher J C. Spectral clustering and transductive learning with multiple views. In: Proceedings of the 24th international conference on Machine learning (ICML). New York: Association for Computing Machinery, 2007, 1159–1166
- 5 Abdi H. Discriminant correspondence analysis. Encyclopedia Meas Statist, 2007, 1: 270–275
- 6 Yu H, Yang J. A direct LDA algorithm for high-dimensional data—with application to face recognition. Pattern Recogn, 2001, 34: 2067–2070
- 7 Hand D J, Mannila H, Smyth P. Principles of Data Mining. Cambridge: MIT press, 2001
- 8 Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometr Intell Lab Syst, 1987, 2: 37–52
- 9 Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. J Comput Graph Stat, 2006, 15: 265–286
- 10 Wold S, Sjöström M, Eriksson L. PLS-regression: A basic tool of chemometrics. Chemometr Intell Lab Syst, 2001, 58: 109–130
- 11 Geladi P, Kowalski B R. Partial least-squares regression: A tutorial. Anal Chim Acta, 1986, 185: 1–17
- 12 Barker M, Rayens W. Partial least squares for discrimination. J Chemometr, 2003, 17: 166–173
- 13 Stone M, Brooks R J. Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. J R Stat Soc Ser B Stat Methodol, 1990, 52: 237–269
- 14 Wilson D L. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans Syst Man Cybern, 1972, 3: 408–421
- 15 Yu K, Ji L, Zhang X. Kernel nearest-neighbor algorithm. Neural Process Lett, 2002, 15: 147–156
- 16 Hart P E. The condensed nearest neighbor rule. IEEE Trans Inf Theory, 1968, 16: 515–516
- 17 Wilson D L. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans Syst Man Cybern, 1972, 3: 408–421
- 18 Burges C J C. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov, 1998, 2: 121–167
- 19 Li H, Liang Y, Xu Q. Support vector machines and its applications in chemistry. Chemometr Intell Lab Syst, 2009, 95: 188–198
- 20 Kingsford C, Salzberg S L. What are decision trees? Nat Biotechnol, 2008, 26: 1011–1013
- 21 Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency. Adv Neural Inf Process Syst, 2004, 16: 321–328
- 22 Okare M, Chatterji B N, Biswas P K. Comparison of similarity metrics for texture image retrieval. In: Conference on Convergent Technologies for the Asia-Pacific Region. Bangalore: IEEE, 2003, 571–575
- 23 Izrailev S, Agraftis D. A novel method for building regression tree models for QSAR based on artificial ant colony systems. J Chem Inform Comput Sci, 2001, 41: 176–180
- 24 Xue Y, Li Z R, Yap C W, et al. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. J Chem Inform Comput Sci, 2004, 44: 1630–1638

- 25 Xue Y, Yap C W, Sun L Z, et al. Prediction of P-glycoprotein substrates by a support vector machine approach. *J Chem Inform Comput Sci*, 2004, 44: 1497–1505
- 26 Xu Q S, Liang Y Z. Monte Carlo cross validation. *Chemometr Intell Lab Syst*, 2001, 56: 1–11
- 27 Stone M. Cross-validated choice, assessment of statistical predictions. *J R Stat Soc Ser B Stat Methodol*, 1974, 36: 111–147

The research of semi-supervised manifold learning algorithm based on markov property

SHEN Liang, XU QingSong, CAO DongSheng & HUANG Xin

Abstract On the background of complex data in pharmacokinetics, we use manifold classification algorithm which has Markov property to model the QSAR. Simulation study show that when there exists some manifold structures, the manifold classification algorithm can give better results of classification. Three real datasets related to different categorical bioactivities of compounds are used to test the performance of MSSC. Statistical evaluation shows that the proposed algorithm is a promising and alternative modeling approach compared to the baseline models.

Keywords transition probability matrix, Markov property, pattern recognition, manifold learning

MSC(2010) 60J22, 62P10

doi: 10.1360/N012015-00040