基于 SVM 的化合物分类综述

蒋强荣,马佳佳

(北京工业大学 计算机学院,北京 100022)

摘 要:药物研发是一个难度系数大、耗费时间长的工作。根据结构活性关系规则,具有相似结构的化合物可能具有相似特性。因此,准确地对化合物进行分类具有十分重要的意义。回顾了 SVM 与比较常用的化合物分类方法及各自的优缺点,阐述了对分类方法进行的改进与优化,展望了化合物分类的发展方向。

关键词:SVM;化合物分类;描述符;图核

DOI:10.11907/rjdk.181759

中图分类号:TP301 **文献标识码**:A **文章编号**:1672-7800(2018)012-0004-04

Review of Chemical Compound Classification Based on SVM

JIANG Qiang-rong, MA Jia-jia

(*Department of Computer Science, Beijing University of Technology, Beijing* 100022, *China*)

Abstract:Drug development is a difficult and time-consuming task. According to the rule of structure activity, compounds with similar structures may have similar properties. Therefore, it is very important to classify compounds accurately. Firstly, this paper reviews SVM and the commonly used classification methods of compounds and their respective advantages and disadvantages. Secondly, it introduces the improvement and optimization method of classification methods. Finally, it looks forward to the development direction of compound classification.

Key Words:compound classification;descriptor;graph kernel

0 引言

随着组合化学的快速发展,大大加快了化合物的合成与筛选速度,化合物数量急剧增长。药物发现的目标是从巨大的化学空间里鉴别出对某一特定疾病具有生物活性的分子,然而在数据规模庞大的化学空间上进行详尽的比对搜索十分困难[1]。因此,准确地对化合物进行分类是非常必要的。

化合物分类是一个非线性问题。SVM 可将样本从原始空间映射到一个更高维的特征空间,使样本在该特征空间内线性可分。核函数的优点是可以简化映射空间计算。化学物分类主要有两种方式:基于描述符的分类方法与基于图核的分类方法。

描述符是分子相似性方法中的基本要素[2]。基于描述符分类方法的思想是首先通过一个高维的特征向量描述化合物,该特征向量是由其包含的描述符(如图片段)决定的,然后利用各种基于向量的核函数计算化合物的相似性。描述符分为 1D 描述符、2D 描述符和 3D 描述符。1D 描述符在利用 SMILES[3]表示化合物方面应用较多,其不仅可以表示原子,还可以表示原子间的键;2D 描述符是由 2D 分子图形或结构片段计算得来的,目前在扩展连接性指纹方面应用最多;3D 描述符描述的是分子形状、分子总表面积与电压等。

1 SVM

SVM[4]是建立在统计学习理论[5]基础上的一种数据挖掘方法,可有效处理回归问题(时间序列分析)与模式识别(分类问题、判别分析)问题,并被广泛应用于文本识别[6]、手写字体识别[7]、人脸图像识别[8]与基因分类[9]等。

SVM 的机理是寻找一个满足分类要求的最优分类超平面,使该超平面在保证分类精度的同时,能够使其两侧空白区域最大化。理论上 SVM 能够实现对线性可分数据的最优分类。

以两类数据分类为例,给定训练样本集(x_i, y_i),$i=1$,

收稿日期:2018-05-10

作者简介:蒋强荣(1964-),女,博士,北京工业大学信息学部计算机学院副教授、硕士生导师,研究方向为模式识别与图像处理;马佳佳(1994-),女,北京工业大学信息学部计算机学院硕士研究生,研究方向为模式识别。

$2,\dots,l, x\in \mathbf{R}^n, y\in\{\pm 1\}$,超平面记作$(w\cdot x)+b=0$,为使分类面对所有样本能够正确分类并且具备分类间隔,则要求其满足如下约束:$y_i[(w\cdot x_i)+b]\geqslant 1, i=1,2,\cdots,l$。

可以计算出分类间隔为$2/\|w\|$,因此构造最优超平面问题则转化为在约束式下求解:

$$\min\Phi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}(w'\cdot w) \tag{1}$$

为了解决该约束最优化问题,引入 Lagrange 函数:

$$L(w,b,a)=\frac{1}{2}\|w\|-a(y((w\cdot x)+b)-1) \tag{2}$$

式(2)中,$a_i>0$ 为 Lagrange 乘数。约束最优化问题的解由 Lagrange 函数的鞍点决定,并且最优化问题的解在鞍点处满足对 w 和 b 的偏导为 0。将该二次型规划问题转化为相应的对偶问题,即:

$$\max Q(a)=\sum_{j=1}^{l}a_j-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_i a_j y_i y_j(x_i\cdot x_j)$$

$$\text{s.t.}\sum_{j=1}^{l}a_j y_j=0\ \ j=1,2,\cdots,l, a_j\geqslant 0, j=1,2,\cdots,l \tag{3}$$

因此,求得最优解。

计算最优权值向量 w^* 与最优偏置 b^*,分别为:

$$w^*=\sum_{j=1}^{l}a_j^* y_j x_j \tag{4}$$

$$b^*=y_i-\sum_{j=1}^{l}y_j a_j^*(x_j\cdot x_i) \tag{5}$$

式(4)和式(5)中,下标 $j\in\{j\mid a_j^*>0\}$。因此,得到最优分类超平面$(w^*\cdot x)+b^*=0$,而最优分类函数为:

$$f(x)=\text{sgn}\{(w^*\cdot x)+b^*\}=sgn\left\{\left(\sum_{j=1}^{l}a_j^* y_j(x_j\cdot x_i)\right)+b^*\right\}, x\in R^n \tag{6}$$

对于线性不可分情况,SVM 的主要思想是将输入向量映射到一个高维特征向量空间,并在该特征空间中构造最优分类面。

将 x 从输入空间 Rn 到特征空间 H 进行 Φ 变换,得到:

$$x\rightarrow\Phi(x)=(\Phi(x),\cdots,\Phi(x))^T \tag{7}$$

以特征向量 $\Phi(x)$ 代替输入向量 x,则可得到最优分类函数为:

$$f(x)=\text{sgn}(w\cdot\Phi(x)+b)=\text{sgn}\left(\sum_{i=1}^{l}a_i y_i\Phi(x_i)\cdot\Phi(x)+b\right) \tag{8}$$

在以上对偶问题中,无论是目标函数还是决策函数,都只涉及到训练样本之间的内积运算,从而在高维空间中避免了复杂的高维运算。

2　描述符研究现状

在化学中,图可以用来直接模拟化合物结构的主要拓扑与几何特征。图中顶点表示原子,边表示原子间的连接关系。将化合物表示的分子图中除去 H,分子中的重原子(C,N,O)对应图中顶点,原子间的键(单键、双键、三键、芳香键)对应图中的边,如图 1 所示。

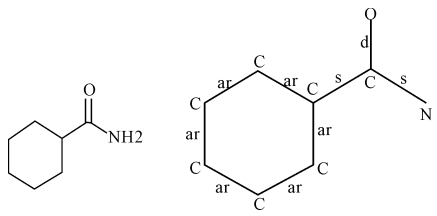

图 1　化合物分子图转换

本节将介绍当前流行的从分子图提取片段的描述符与描述符常用核函数。

2.1　描述符

2.1.1　指纹

指纹[10]是指将化合物的结构特征编码成固定位的向量,指纹中具体位字符串的生成依赖于键的数量、设置位数量、哈希函数与位字符串长度。指纹描述符的优点是能将化合物包含的大量子结构紧凑地表示出来。

2.1.2　Maccs Keys

Maccs Keys[11]是指基于给定化合物结构与预先由该领域专家定义结构片段的模式匹配。每一个结构片段就是一个键,在描述空间中占据一个固定位置。因此,该方法依赖于预先定义的规则封装分子描述符,而没有从数据集中学习。

与指纹描述符相比,Maccs Keys 没有哈希函数作用在子结构上。其优点在于子结构的任意拓扑可形成描述符空间的一部分,缺点是不能适应特殊数据集与分类问题。

2.1.3　环树表示法(CT)

CT[12]是指将化合物表示成环和特定树的集合,主要思想是首先识别分子图中的互连组件(也称为块),一旦这些块被识别,通过从块中枚举具有确定数量的简单环,第一个特征集合随之产生。所有环被识别之后,分子图中的块则被删除,此时的图是剩余树组成森林的集合,每一个树作为一个描述符。最终的描述符空间是环与剩余树的集合。CT 表示法用到树模型的具体拓扑与大小取决于分子图中块的位置。

2.1.4　频繁子结构(FS)

FS 是指在给定 σ 的前提下,在数据集中寻找出现次数大于 σ 的子结构。因此,与 Maccs Keys 不同的是,当 σ 改变时,FS 的描述符空间也会改变;与指纹描述符不同的是,其不考虑子图大小(键的数量),所有子图构成描述符空间。FS 的缺点是 σ 值选取过大或过小都可能导致分类效果不理想。

2.1.5　扩展连接性指纹(ECFPs)

ECFPs 由摩根算法[13]的变体派生而来,生成过程分为 3 步:①初始分配阶段,为每个原子分配整数标识符;②迭代更新阶段,更新每个原子的标识符,以对每个原子邻居的标识符作出反应;③重复标识符移除阶段[14],如果两

个特征是经过不同次数迭代生成的,则经过更大迭代次数生成的特征将被移除,如果两个特征是经过相同次数迭代生成的,则哈希标识符值更大的特征将被拒绝。

2.2 核函数

描述符空间常用的核函数有 Tanimoto coefficient 核与 Min-Max 核,满足 Mercer 条件[15],两者实际上都是统计两个被比较对象的共有特征占两个对象所有特征之和的比例,值在[0,1]之间。

2.2.1 Tanimoto coefficient 核

Tanimoto coefficient 核适用于二进制向量,计算的核定义如下:

$$K(X,Y)=\frac{\sum_{i=1}^{M}X_iY_i}{\sum_{i=1}^{M}(X_i^2+Y_i^2-X_iY_i)} \tag{9}$$

其中,M 表示 X、Y 均由 M 维二进制向量表示,X_i、Y_i 分别表示 X 和 Y 的第 i 维向量。

2.2.2 Min-Max 核

在二进制向量的情况下,Min-Max 核退化为 Tanimoto 系数。Min-Max 核定义如下:

$$K(X,Y)=\frac{\sum_{p\in P}\min(\varphi_p(X),\varphi_p(Y))}{\sum_{p\in P}\max(\varphi_p(X),\varphi_p(Y))} \tag{10}$$

其中,P表示 X 和 Y 的所有特征集合,$\varphi_p(\cdot)$统计 p 出现的次数。

2.3 描述符领域创新

随着研究的深入,为了提高化合物分类的准确率,研究者将重心放在提出或改进新的描述符,以及改进或组合描述符空间的核函数等方面。Gong-Hua Li[16] 提出新的分子指纹描述方法,将每个化合物的对应模式在活性与非活性化合物中所占比例作为权重系数,并将单个核函数进行两两相乘组合,最终取得 85% 左右的成功率;翟璨等[17] 采用 ECFPs 描述符表示分子图,根据不同长度描述符应具有不同权重改进了 Min-Max 核,并在 PTC 和 HIV 数据集中进行测试,使分类准确率都得到了提高;王山等[18] 采用计数型布隆过滤器对指纹描述符分子相似性进行改进,并采用 DUD LIBVS 1.0 数据集对改进方法进行了比较验证,与其它原始分子相似性方法相比,其在相似性判断的准确性与骨架跃迁潜能上均有所提高。此外,还有很多学者提出多核组合方式,以更好地对化合物进行分类。

3 图核研究现状

化合物分子图采用邻接矩阵表示,两个顶点如果有边相连,则值为 1。邻接矩阵定义如下:

$$[A]_{ij}=\begin{cases}1 & if(v_i,v_j)\in E\\ 0 & otherwise\end{cases} \tag{11}$$

其中,A 为化合物图 G 的邻接矩阵,v_i 和 v_j 为 G 的顶点,假设图 G 有 n 个顶点,则 $0\leqslant i<n$,$0\leqslant j<n$。

图核函数主要分为 3 类:①基于游走的图核函数;②基于路径的图核函数;③基于子树的图核函数。

3.1 基于游走的图核函数

基于游走的图核函数主要是随机通路核[19]。随机通路核通过计算两个图的公共通路数目度量两个图的相似性,两个图 $g(V_1,E_1)$和 $g'(V_2,E_2)$的匹配通路数可以通过计算其直积图 $g\times g$得到。设 $A_\times$ 为直积图 $g\times g'$的邻接矩阵,则随机通路核函数表示如下:

$$K(g\times g')=\sum_{i,j=1}^{|V_\times|}\left[\sum_{n=0}^{\infty}\lambda^nA_\times^n\right]_{ij} \tag{12}$$

其中 λ是使和收敛的衰减因子,$\lambda<1$,确保对于足够大的 n可以忽略不计。$V_\times$ 为直积图 $g\times g'$的任一顶点,时间复杂度为 $O(n^6)$。

3.2 基于路径的图核函数

基于路径的图核函数主要是最短路径核[20]。最短路径核通过比较两个图 G_1 和 G_2 的所有最短路径度量两个图的相似性。$G_1'=(V_1',E_1')$和 $G_2'=(V_2',E_2')$分别是 G_1 和 G_2 的最短路径图,所有最短路径对核的值构成最短通路核,最短路径可通过弗洛伊德算法求出。最短路径核的优点是可以完全避免路径回溯问题。最短路径核表示如下:

$$K(G_1,G_2)=\sum_{s_1\in SP(G_1)}\sum_{s_2\in SP(G_2)}k(s_1,s_2) \tag{13}$$

其中 $SP(\cdot)$表示图的所有最短路径集合,$k(s_1,s_2)$定义为狄拉克核函数,当 s_1 与 s_2 的长度一样时值为 1,否则为 0。

3.3 基于子树的图核函数

基于子树的图核函数主要是 Weisfeiler-Lehman 子树核[21]。它的思想是基于一维 Weisfeiler-Lehman 同构判定算法,寻找一对图结构中同构的子树结构。基于 Weisfeiler-Lehman 图核的一般表示形式为:

$$k_{WL}^{(h)}(G,G)=k(G_0,G_0)+k(G_1,G_1)+\cdots+k(G_n,G_n) \tag{14}$$

其中 h表示迭代次数,$G_\times$、$G'_\times$分别表示图 G 和 G'对应的 WL 序列。假设 $\sum_i\in\sum$ 表示 WL 算法在第 i 次迭代后,在图 G 和 G'中出现至少一次的顶点标签集中所构成的字母集合,定义一个映射 $C_i:\{G,G'\}\times\sum_i\rightarrow N$,$C_i(G,\sigma_{ij})$表示图 G 中字母 σ_{ij} 出现的次数,则 Weisfeiler-Lehman 子树核的表示形式为:

$$k_{WLST}^{h}(G,G')=<\varphi_{it}^{h}(G),\varphi_{it}^{h}(G')> \tag{15}$$

其中 h表示迭代次数或层数,$\varphi_{it}^{h}(G)$和 $\varphi_{it}^{h}(G)$分别为 G 和 G'对应的映射特征,$\varphi_{it}^{h}(G)=(c_0(G,\delta_{h1}),\cdots,c_h(G,\delta_{h1}),\cdots,c_h(G,\delta_{h|\sum|}))$,$\varphi_{it}^{h}(G')=(c_0(G',\delta_{h1}),\cdots,c_h(G',\delta_{h1}),\cdots,c_h(G',\delta_{h|\sum|}))$。

3.4 图核领域创新

Xu 等[22]考虑到已有 Weisfeiler-Lehman 图核忽视的结构信息,提出一个 Weisfeiler-Lehman 图核混合框架,并将其运用于 Weisfeiler-Lehman 图序列上,取得了很好的分类结果;Bai 等[23]提出一个新的用 Jensen-Shannon 方法表示顶点的图核,可识别两个图顶点之间的对应关系;

Kondor & Pan 提出多尺度拉普拉斯图核,可捕获个别顶点及子图之间的拓扑关系。此外,研究者们也不断致力于提出新的图核或改进已有图核,以提高化合物分类的准确率。

4 结语

本文从两方面对基于 SVM 的化合物分类进行了详细介绍,分别介绍了 SVM 理论、描述符与图核的研究现状及发展,并对目前常用的化合物分类方法进行了简要叙述。目前的化合物分类方法很难达到 95% 以上的成功率,因此还需要作进一步深入研究,捕捉化合物结构间的特征,以提出更好的比较化合物相似性的方法,进一步提高化合物分类的准确率。

参考文献:

[1] RANU S. Querying and mining chemical databases for drug discovery[M]. University of California at Santa Barbara, 2012.

[2] MALDONADO A G, DOUCET J P, PETITJEAN M, et al. Molecular similarity and diversity in chemoinformatics: from theory to applications[J]. Molecular diversity, 2006, 10(1):39-79.

[3] WEININGER D. SMILES I: introduction and encoding rules[J]. Journal of Chemical Information and Computer Sciences, 1988.

[4] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.

[5] VLADIMIR N VAPNIK, 张学工.统计学习理论的本质[M].北京:清华大学出版社,2000.

[6] 陈佳希.基于支持向量机的文本分类[J].电子世界,2017(7):64.

[7] 董婉君.基于 SVM 的手写字体识别[J].工程技术:全文版,2016(2):00288.

[8] 郭慧敏,丁军航.基于支持向量机的人脸特征分类技术[J].青岛大学学报:工程技术版,2016,31(4):56-61.

[9] 王晶,周旷.基于支持向量机的肿瘤基因识别[J].计算机与数字工程,2011,39(9):3-6.

[10] DAYLIGHT INC. Mission Viejo CA USA [EB/OL] http://www.daylight.com.

[11] DURANT J L, LELAND B A, HENRY D R, et al. Reoptimization of MDL keys for use in drug discovery[J]. Journal of Chemical Information and Modeling, 2002, 42(6):1273 - 1280.

[12] WROBEL S. Cyclic pattern kernels for predictive graph mining[C]. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2004:158-167.

[13] MORGAN H L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service[J]. Journal of Chemical Documentation, 1965, 5(2):107-113.

[14] ROGERS D, HAHN M. Extended-connectivity fingerprints[J]. Journal of Chemical Information & Modeling, 2010, 50(5):742-54.

[15] SWAMIDASS S J, CHEN J, BRUAND J, et al. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity[J]. Bioinformatics, 2005, 21(1):359 - 368.

[16] LI G H, HUANG J F. CDRUG: a web server for predicting anticancer activity of chemical compounds[J]. Bioinformatics, 2012, 28(24):3334-3335.

[17] JIANG Q, ZHAI C, XIONG Z. Chemical compound classification based on improved Max-Min kernel[J]. Journal of Chemical & Pharmaceutical Research, 2014.

[18] 王山,孙莉,吴杰,等.一种基于计数型布隆过滤器的分子相似性算法研究[J].计算机科学,2017,44(b11):552-556.

[19] GÄRTNER T, FLACH P, WROBEL S. On graph kernels: hardness results and efficient alternatives[J]. Lecture Notes in Computer Science, 2003, 2777:129-143.

[20] BORGWARDT K M, KRIEGEL H P. Shortest-path kernels on graphs[C]. IEEE International Conference on Data Mining. IEEE, 2006:74-81.

[21] SHERVASHIDZE N, SCHWEITZER P, JAN VAN LEEUWEN E, et al. Weisfeiler-Lehman Graph Kernels[J]. The Journal of Machine Learning Research, 2011, 12(3):2539-2561.

[22] XU L, XIE J, WANG X, et al. A mixed Weisfeiler-Lehman graph kernel[J]. Lecture Notes in Computer Science, 2015, 9069:242-251.

[23] BAI L, ZHANG Z, WANG C, et al. A graph kernel based on the Jensen-Shannon representation alignment[C]. International Conference on Artificial Intelligence. AAAI Press, 2015:3322-3328.

(责任编辑:黄　健)

(上接第 3 页)

[4] 王帆.基于卡尔曼滤波和粒子滤波的移动机器人同时定位与地图创建研究[D].西安:西安工程大学,2012.

[5] 刘书池.面向工业互联网的井下无人机单目视觉 SLAM 定位方法[D].北京:北京交通大学,2017.

[6] 杨晓晓.室内机器人单目视觉同时定位与地图构建技术研究与实现[D].成都:成都信息工程学院,2014.

[7] 冯少江,徐泽宇,石明全,等.基于改进扩展卡尔曼滤波的姿态解算算法研究[J].计算机科学,2017,44(9):227-229.

[8] 李志雄,王妲.基于无迹卡尔曼滤波的运动机器人定位研究[J].轻工科技,2015(11):73-74,76.

[9] 韩同辉,沈超,沈静,等.基于 PF 算法的移动机器人定位研究[J].机电一体化,2012,18(3):13-16.

[10] 王聪伟.基于扩展卡尔曼滤波的足式机器人运动速度估计研究[D].哈尔滨:哈尔滨工业大学,2014.

[11] 金峰,蔡鹤皋.机器人 IMU 与激光扫描测距传感器数据融合[J].机器人,2000(6):470-473.

[12] 骆云祥.非线性滤波在移动机器人 SLAM 中的应用[D].南京:南京理工大学,2009.

[13] MURPHY K. Bayesian map learning in dynamic environments[C]. Proc of the Conf on Neural Information Processing Systems, NIPS, Denver, USA, 1999:1015-1021.

[14] DOUCET A, FREITAS J DE, MURPHY K, et al. Rao-Blackwellized particle filtering for dynamic Bayesian networks[C]. Proc of the Conf on Uncertainty in Artifical Intelligence, UAL, Stanford, USA, 2000:176-185.

[15] MONTEMERLO M, THRUN S, ROLLER D, et al. FastSLAM 2.0: an improved particle filtering algorithm for simultaneous localization and mapping that provably converges[J]. Proc. int. conf. on Artificial Intelligence, 2003, 133(1):1151-1156.

[16] MONTEMERLO M, THUN S, KOLLER D, et al. FastSLAM: a factored solution to simultaneous mapping and localization[C]. Proceedings of the National Conference on Artificial Intelligence, 2002:590-605.

[17] KIM C, SAKTHIVEL R, WAN K C. Unscented FastSLAM: a robust algorithm for the simultaneous localization and mapping problem[C]. IEEE International Conference on Robotics and Automation. IEEE, 2007:2439-2445.

[18] SHOJAIE K, AHMADI K, SHAHRI A M. Effects of iteration in Kalman filters family for improvement of estimation accuracy in SLAM[C]. IEEE International Conference on Advanced Intelligent Mechatronics, 2007:1-6.

(责任编辑:黄　健)