用于药物活性预报的 Co-Training 方法*)

李 丹[1,2] 李国正[1,2] 陆文聪[3]

(上海大学计算机工程与科学学院 上海 200072)[1]

(南京大学计算机软件新技术国家重点实验室 南京 210093)[2] (上海大学理学院化学系 上海 200444)[3]

摘 要 在药物设计中,可以利用药物分子的构效关系模型进行药物活性的预报,从而降低药物开发的成本、缩短开发的周期。本文尝试结合 Co-Training 方法和嵌入式特征选择方法,提出了一种新的 FESCOT(Feature Selection for Co-Training)算法。算法在药物活性数据集上进行了实验,结果显示结合了特征选择的 Co-Training 方法较之以前泛化能力有所提高。

关键词 药物活性,半监督学习,特征选择

Prediction of Drug Activity by Using Co-Training

LI Dan[1,2] LI Guo-Zheng[1,2] LU Wen-Cong[3]

(Colltge of Computer Engineering and Science, Shanghai University, Shanghai 200072)[1]

(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)[2]

(Department of Chemistry, Shanghai University, Shanghai 200444)[3]

Abstract The activity of drug molecule can be predicted by the QSAR (Quantitative Structure Activity Relationship) model, which overcomes the disadvantages of high cost and long cycle with the traditional experimental method only. With the fact that the number of drug molecule with known activity is less than those of unknown activity, it is important to predict molecular activities with the semi-supervised learning method. However, the numerous features of drug molecule affect the prediction accuracy of the QSAR model. Therefore, a novel algorithm named FESCOT (Feature Selection for Co-Training) is proposed in this paper, which combines Co-Training and an embedded feature selection method. Experiments are carried out on the data set of molecular activities, and the results show that generalization ability of FESCOT is better than that of Co-Training without feature selection.

Keywords Molecular activity, Semi-supervised learning, Feature selection

1 引言

药物的构效关系模型可以指导新药物的设计,通过计算机进行的新药物的活性预报可以节约大量实验费用,同时提高药物设计效率[1]。药物活性取决于药物的分子结构,药物活性预报的首要工作是提取其结构特征参数,利用机器学习方法建立构效关系模型,并利用所得模型指导新药物设计和活性预报。药物的测定工作成本高,药物分子中已测定活性的少,未测定的多,所以如果能够使用未测定的数据提高模型精度则将很有意义。目前,半监督学习方法能利用非标记的数据促进学习性能,正成为机器学习领域中一个热门话题。

药物活性预报中,需要提取分子结构的描述参数,现在提取的方法非常多,一个分子可以得到成千上万的特征参数,这些参数都或多或少地含有分子结构的信息。但是,通常来说,冗余信息会降低学习器的泛化能力。因此如何通过降低冗余的特征信息对于半监督学习的影响,从而提高药物活性预报精度也是一个很重要的课题。许多研究者提出了多种算法,其中嵌入式模型完全依赖于所运用的学习器,精度较高,同时与卷积模型[5]相比,其计算量相对比较低[5]。

本文将研究如何结合半监督学习去除冗余特征,主要是结合 Co-Training 方法和带预报风险准则[6]的嵌入式特征选择方法,提出一种新算法 FESCOT(Feature Selection for Co-Training),该算法预期能够有效去除无关特征和冗余的特征,提高 Co-Training 算法的预报精度。

2 带有特征选择的 Co-Training

2.1 Co-Training

利用不带标记的数据提高监督学习器的泛化能力的半监督学习方法目前已经提出了多种,其中 Co-Training 方法是其中一个比较好的方法[2~4]。该方法最初由 Blum 和 Mitchell 提出[2],它假设数据中存在两个独立并且冗余的特征集或者数据视图,然而,这个条件在现实世界应用中显得过于苛刻。所以,Goldman 和 Zhou 提出了改进的 Co-Training 方法[3],新的方法并不假设特征独立和冗余,相反,它们所采用的是在同一个求解问题上使用两种不同的监督学习方法。基于该思路,Zhou 和 Li 在 CoReg 算法[4]中提出了使用一对具有不同参数的同质监督学习算法,即用不同距离测度的一对 K-近邻算法来进行半监督学习。

本文所运用的 Co-Training 算法继承了 CoReg 算法[4]的思想,Co-Training 算法将具有不同参数设置的一对分类器在同一个数据集上进行训练,比如说 K-近邻方法[4],而这一对 K-近邻方法将通过彼此给非标记的数据进行标记来完成算法。下面将把 CoReg 算法进行修订来运用于分类问题。

1)CoReg 算法中,Δ_u 定义在回归问题中用于估计标记

*)本工作受到国家自然科学基金(20503015)、上海市教委自然科学一般项目(05AZ67)和上海市教委 E 研究院-上海高校网格项目(20030301)的资助。**李国正** 副研究员,博士,主要研究方向:机器学习,数据挖掘。

的置信度。本文所研究的是分类问题，所以Δ_u定义如下：

$$\Delta_u=\sum_{\Omega}acc'-acc$$

Ω代表某一未标记样例x_u的K个近邻的带标记样例集合，acc是在没有加入某个新的未标记样例的训练集所训练的模型的分类正确率，acc'是在加入了某个新的原先未标记样例$(x_i,\hat{y}_u)$后的训练集所训练的模型的分类正确率，$\hat{y}_u$是由原分类器得到的分类结果。

2)与CoReg算法的第二个不同之处在于其算法的输出。CoReg算法解决的是回归问题，所以其输出是两个回归器预报值的平均值。然而，对于分类问题而言，测试样例x的标记是由两个通过训练而得的K-近邻学习器预测所得的在样例x附近的K-近邻标记样例的标记决定，K-近邻的标记数目最多的类的标记就是该测试样例x的标记。如果存在类的标记数目相同的情况，那么，测试样例x的标记将随机地由所有标记样例中数目最多的标记而决定。所以，其输出就变成：

$$h^*(x)=\arg\max\{N_a(x),\cdots,N_a(x),N_m(x)\}$$

此处$N_a(x)$表示属于第Ci^{th}类的样例的数目，这些样例就是由Co-Training方法训练所得分类器中样例x的K-近邻。

3)与CoRge算法的另一个不同之处在于添加置信度最大的未标记样例到标记样例集中时未标记样例集的范围。CoReg算法中随机地在未标记样例集中选取100个未标记样例，本文中采用了多组不同比例分割的数据做实验，所以在FESCOT算法中未标记样例集的范围是整个未标记样例集，而不是部分样例。

2.2 FESCOT

FESCOT（Feature Selection for Co-Training)算法在Co-Training方法中结合了嵌入式特征选择方法，并且采用预报风险准则来评价各个特征。预报风险准则由Moody等提出[6]，该准则通过计算当所有样例在某个特征的数值被其平均值代替时的训练正确率的变化来评价各个特征

$$S_i=acc-acc(\bar{x}_i) \tag{1}$$

对应S_i值最小的特征将被删除，因为该特征值的变化对结果的影响最小，也就是说该特征的重要性不大。

在FESCOT算法中，使用后向序列搜索方法来生成特征子集，然后用一个验证样本来决定最优特征子集。FESCOT算法的详细步骤如下：

FESCOT算法

假设特征子集$u=[1,2,\cdots,M]$，删除特征的序列为$r=[]$，正确率的序列为$a=[]$。将训练样本$x_{r0}=[x_r^1,\cdots,x_r^i,\cdots,x_r^M]^T$和$y_r$(注意：因为$x_{r0}$中的某些样例是没有标签的，所以它们的$y_r$为空)，以及验证样本$x_v$一起作为函数FESCOT的输入。

步骤1：将训练样本的特征序列限制在一个合适的序列中$x_r=x_{r0}(:,u)$，第一次迭代时，$x_r=x_{r0}$。

步骤2：训练半监督分类学习器，以获得模型$COT(x_r,y_r)$。

步骤3：用验证样本对所得的模型进行测试并计算分类正确率，使$a_t=COT(x_v(:,u),y_v)$，然后更新a序列，使$a=[a_t,a]$。

步骤4：在训练集上利用等式(1)计算每个特征的预报风险S_i。

步骤5：找出预报风险最小m个特征集合h，当前特征个数大于100时，m取80，当当前特征个数大于10且小于等于100时，m取8，否则m取当前特征个数减1(因为原始的实验数据集中有562个特征，如果每个特征依次判断，则时间代价过高)。

步骤6：更新删除特征序列号$r=[u(h),r]$，删除u中出现在集合r中的特征序列号，更新u。如果$length(u)>1$，则继续步骤1。

步骤7：找出分类准确率最佳的特征子集，$h=\arg\max(a)$，并得到最优特征子集$u_h=[u,r(1:h)]$。

步骤8：输出最优特征子集u_h。

3 在药物活性预报上的应用

3.1 药物活性数据集

本文使用Mutagenicity数据集进行实验，(可在http://www.niss.org上进行下载)。有1863个样例，结构参数包括47个CONS描述符，260个拓扑指数，64个BCUT描述符，247个FRAG描述符，共618个，去除不可用特征，共有562个用于我们的实验。

实验前首先要对数据进行分割。数据集中25%的数据作为测试数据集，其余的75%的数据作为训练数据集，该训练数据集又以不同的比例分割为标记部分和未标记部分。验证数据集来自于训练数据集中的标记数据部分的10%。整个药物数据集用此法随机地分割10次，实验最终结果是由这10次实验的结果取平均而得到的。

3.2 FESCOT的计算结果

实验研究两个问题：1)Co-Training是否优于在标记样本上的完全监督学习Full-Training；2)FESCOT是否优于不带特征选择的Co-Training。所涉及的算法分别在药物活性的数据集上进行了对比实验。表1是Full-Training，Co-Training和FESCOT三种算法在Mutagenicity数据集上当标记样例个数与未标记样例的比例分别为2∶8，3∶7，5∶5和9∶1时的结果，R表示训练数据集中，标记样例数与未标记样例个数的比例，F-T表示Full-Training方法，C-T表示Co-Training方法，ratio表示删除特征数目占总特征的比例。

表1 药物活性数据上的三种算法的对比实验结果(%)

R	F-T	C-T	FESCOT	ratio
2∶8	56.93	63.27	64.56	81.50
3∶7	57.20	66.19	66.84	76.87
5∶5	59.31	69.87	72.04	71.35
9∶1	65.29	74.49	76.26	46.62
平均	59.68	68.46	69.92	69.09

从表1可以看出，各次实验中Co-Training的预报正确率均比Full-Training的预报正确率有接近10个百分点的幅度的提高，半监督学习优于单单在已标记样本上的完全监督学习；再者，各次实验中，FESCOT的预报正确率均比Co-Training的预报正确率高，带特征选择的半监督学习优于不带特征选择的半监督学习。4组实验结果也显示了实验数据中存在着大量的冗余特征，而冗余特征的删除明显提高了药物活性的预报正确率。

以上4组实验表明，随着实验中标记样例数比例的增加，半监督学习Co-Training和单单在标记样本上进行完全监督学习Full-Training预报正确率都不断提高，而且Co-Training比Full-Training总是有所提高，同时带特征选择的半监督学习FESCOT的泛化能力也逐渐高于半监督学习Co-training

的泛化能力。

结论 本文将半监督学习技术运用于药物活性预报中，并且用特征选择的方法克服了结构特征参数中无关和冗余特征对预报性能的影响。实验证明，半监督学习比单单在标记样本上进行完全监督学习具有更好的性能，同时特征选择能够提高半监督学习在药物活性预报中的精度。进一步的工作包括如何提高半监督学习的效率、如何选择合适的基学习方法以提高半监督学习的泛化能力，从而提高化合物活性预报的效率和精度。

参 考 文 献

1 Xu L, Wu Y, Hu C, Li H. A QSAR of the toxicity of aminobenzenes and their structures. Science in China (Series B), 2000, 43(2): 130～136

2 Blum A, Mitchell T. Combining labeled and unlabeled data with Co-Training. In: Proceedings of the 1998 COLT, Morgan Kaufmann Publishers, 1998. 92～100

3 Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data. In: Proceedings of the 17th ICML, San Francisco, CA, Morgan Kaufmann, 2000. 327～334

4 Zhou Z H, Li M. Semi-supervised regression with Co-Training. In: Proceedings of the 19th IJCAI'05, Edinburgh, Scotland 2005. 908～913

5 Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research, 2003(3): 1157～1182

6 Moody J, Utans J. Principled architecture selection for neural networks: Application to corporate bond rating prediction. In: NIPS 4, Morgan KaufmannPublishers, Inc., 1992. 638～690

(上接第 144 页)

程——要素检验和结构检验，也可以用类似的方法描述子过程的自明度。

在分析广义对象语义块的内部构成阶段，需要确定组成广义对象语义块的各个词语之间的概念关联。假设组成广义对象语义块的词语总数为 N，判定出的词和词之间的关系总数为 n，则广义对象语义块构成的自明度可以表示为 $M_{43}=n/N$。对于词与词之间的组合层次关系，以及出现语句嵌套的广义对象语义块等子过程的自明度，也可以根据子过程的分析方法和分析步骤计算。

5.5 小结

理解自明度的计算是以从形式结构到语义结构的句类分析过程为基础，其基本标准是句类检验能否通过。通过计算机程序的句类检验的，就作为正确的结果计入自明度，不能通过句类检验的只能作为候选集合总数的一部分来影响自明度的大小。自明度计算公式实际上反映了从候选集中找到通过句类检验的唯一元素的可靠性，即自明度越趋向于 1，找到的可靠性越大；当自明度为 0 时，表示无法找到。自明度描述了构成理解过程的子过程结果的可靠性，可以逐级深入。整个句类分析过程的理解自明度可以用下面的式子表示：

$$M=(M_1+M_2+M_3+M_{41}+M_{42}+M_{43})/6 \qquad (1\text{-}9)$$

结束语 在计算机理解汉语语句的过程中，既需要从外部制定符合要求和预期的指标，也需要从系统内部给出过程数据和结果的分析评价。语句理解自明度就是计算机内部的一个自我评价指标，它不仅考核过程结果的正确性，而且考核过程结果的数据来源依据和可靠性。自明度从数据的使用和结果等方面说明了理解过程的解模糊程度，从语句形式结构到语义结构的过程中经常要从多个候选中选出正确的，在这个选择过程中自明度越高说明系统解模糊的能力程度越强。

本文把从语句形式结构到语义结构的理解过程划分为五个子过程，对每一个子过程中重要的可能出现模糊的数据给出了自明度的计算公式。这些公式只是粗略地反映了理解过程中计算机进行“多义选一”操作的效果，还没有反映出详细的支撑知识和规则。对于如何记录理解过程中使用的规则和知识，并评价它们对理解过程的影响，以及如何计算它们的自明度，将是本文进一步深入研究的内容和方向。

参 考 文 献

1 房玉清. 实用汉语语法[M]. 北京：北京语言学院出版社，1992

2 黄曾阳. HNC(概念层次网络)理论—计算机理解自然语言的新思路[M]. 北京：清华大学出版社，1998

3 黄曾阳. 语言概念空间的基本定理和数学物理表示式[M]. 北京：海洋出版社，2004

4 晋耀红. HNC 句类分析的“自知之明”. 见：第一届 HNC 与语言学研究学术研讨会，武汉，2001

5 池毓焕. 汉语动词形态困扰的分析与处理[D]：[博士论文]. 北京：中国科学院声学研究所，2005

6 苗传江. HNC(概念层次网络)理论导论[M]. 北京：清华大学出版社，2005

7 [英]玛格丽特. 博登，编著. 人工智能哲学[M]. 刘西瑞，王汉琦，译. 上海：上海译文出版社，2001

8 徐波，孙茂松，靳光谨，主编. 人中文信息处理若干重要问题[M]. 北京：科学出版社，2003

(上接第 147 页)

成了统计数据存在一定的不完美，对于这一类的准确划分仍需要深入的配合句群的研究工作展开。

其次，辅块省略标记符号还原为辅块时问题仍然存在，主辅两可块的问题有待深入解决，也造成了统计数据存在一定的不完美。例如：这条新闻电视台播过。可还原为“电视台||播过||这条新闻”，也可还原为“这条新闻||～在电视台||播过”。国外纺织企业已普及无梭织机。可还原为“国外的纺织企业||已普及||无梭织机”，也可还原为“在国外～||纺织企业||已普及||无梭织机”。

再次，对于规则的判断，需要用到具体的语义块构成以及语义块之间的信息，其中涉及到的深层次问题仍然存在，规则的具体形式化工作仍需进一步结合计算机形式语言的特征深入展开。

最后，对于“主谓谓语句”的统计数据是建立在人为搜索语料后得出的，这其中必然含有人为的遗漏、误添，实际的统计数据在此基础上一定存在一定的偏差，更精确的数据还需要在更大规模的真实语料的基础上展开。

总结 本文在现有的 HNC 研究所取得的成果的基础上，用 HNC 理论中的句类理论和语句格式知识来分析“主谓谓语句”这一现代汉语语言现象，并归纳得出了一套解决现代汉语主谓谓语句的语句格式判断规则，该规则采用 HNC 理论句类知识和语句格式知识为核心，利用语言概念空间中的概念及其关联知识来解决主谓谓语句给计算机带来的模糊不清问题，并试图通过规则得出判断结果。初步实验表明，这些规则是高效可行的。今后的工作将集中在规则进一步验证和将规则的学习机制引入知识库的构造过程的研究上。能否很好地解决这些问题将是整个计算机能否解决主谓谓语句带来的语义模糊的关键所在。

参 考 文 献

1 朱德熙. 语法答问. 北京：商务印书馆，1985

2 徐青. 现代汉语. 修订版. 上海：华东师范大学出版社，1977. 295～302

3 黄曾阳. HNC(概念层次网络)理论. 北京：清华大学出版社，1998. 6～9

4 张全，萧国政. HNC 与语言学研究. 武汉：武汉理工大学出版社，2001

5 黄曾阳. 语言概念空间的基本定理和数学物理表达式. 北京：海洋出版社，2004