

分类号 TP3 TP1 Q5 065

密级 _____

UDC _____

编号 _____

中国科学院研究生院 博士学位论文

基于机器学习技术的生物信息检索研究

付 岩

指导教师 _____ 高 文 研究员

中国科学院计算技术研究所

申请学位级别 工学博士 学科专业名称 计算机应用技术

论文提交日期 2007年1月 论文答辩日期 2007年2月

培养单位 _____ 中国科学院计算技术研究所

学位授予单位 _____ 中国科学院研究生院

答辩委员会主席 _____ 陈润生

声 明

我声明本论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，本论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名：

日期：

论文授权使用授权书

本人授权中国科学院计算技术研究所可以保留并向国家有关部门或机构送交本论文的复印件和电子文档，允许本论文被查阅和借阅，可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

（保密论文在解密后适用本授权书。）

作者签名：

导师签名：

日期：

摘要

在知识和数据爆炸的今天，信息检索系统在各个领域都扮演着越来越重要的角色。检索函数(有时也被称为打分函数或排位函数)是所有检索系统的关键组成部分，其任务是对数据库中保存的数据对象与用户查询之间的相关性进行度量。检索函数的设计可以从两个层次上进行，一个是依赖于应用领域的基本相关性度量指标的构造，另一个是与应用领域相对独立的将多种基本相关性度量指标综合起来的检索函数的构造。本文利用机器学习技术，从检索函数设计的以上两个层次，深入研究了生物信息学中蛋白质序列鉴定和蛋白质同源性预测两个重要的检索问题。

在生物信息学领域，串联质谱技术与数据库搜索相结合的肽和蛋白质鉴定是一个重要的生物序列检索问题。肽和蛋白质鉴定软件的核心是对数据库中的候选肽生成实验质谱的可能性进行度量的肽打分函数(即检索函数)。在肽打分函数中，最基本的操作是把实验质谱中的谱峰与从候选肽预测出的理论离子按照质量值进行匹配。由于质量测量的不准确性，随机错误匹配经常发生。为了提高匹配的准确性，本文首先提出了一种更准确的质量匹配误差分布模型，即条件正态分布模型。在该模型中，质量匹配误差分布的均值和标准差不再是恒定不变的，而是分别为离子质量和谱峰强度的函数。其中，质量误差标准差与谱峰强度之间的对数线性关系就作者所知是以前相关文献中没有报道过的。本文并给出了一个迭代学习算法，从训练数据中准确地估计误差模型的参数，刻画串联质谱的质量误差分布。本文接着提出了一种非线性肽打分函数，即核谱向量点积。它是对一大类传统肽打分方法即谱向量点积的非线性扩展。在串联质谱中，碎片离子间的相关性信息对于降低随机匹配是很有帮助的。核谱向量点积利用局部化核函数来强调相关离子的同时匹配。实验表明，核谱向量点积能够显著地提高肽鉴定的精度。基于核谱向量点积肽打分函数的肽和蛋白质鉴定软件 pFind 在多个数据集上的鉴定精度，明显超越了基于谱向量点积的流行商业软件 SEQUEST。在 1%假阳性率下，pFind 比 SEQUEST 多鉴定出了 10%到 30%的肽段数。

由于实际检索问题的复杂性，度量数据对象与查询之间相关性的基本指标往往有多种，构成多维特征向量。如何把多维基本相关性度量指标合并成一个相关性指标，就是检索函数构造问题。从训练数据中学习检索函数是一种常用且有效的检索函数构造方法。一般来讲，检索函数的学习是独立于具体应用的一般性机器学习问题。在这类学习问题中，特征向量是相对于查询计算出来的，因而随所关联的查询不同而分成不同的组(本文称为“块”)。数据的块结构形式是检索函数学习问题独有的特点。本文结合蛋白质同源性预测问题，通过深入挖掘这种块结构包含的丰富信息，提出了一系列旨在提高检索函数学习准确性的方法。这些方法包括用于解决块间数据非独立同分布问题的块内

数据归一化和块特征向量扩充方法，用于数据去冗余的块选择和支持向量下采样方法，以及用于构造查询适应的检索函数的 K 近块集成排位算法等。使用支持向量机作为基准学习器的实验表明，本文提出的所有这些基于块的方法都明显地比直接应用标准的支持向量机效果要好。其中，块内数据归一化和数据去冗余方法在 2004 年的 ACM KDDCUP 数据挖掘竞赛的蛋白质同源性预测问题上获得了全球并列第一名的总体预测准确度。K 近块集成排位算法在预测精度和训练速度上甚至更胜一筹，在上述蛋白质同源性预测问题上是目前表现最好的算法。

关键词： 生物信息学； 信息检索； 机器学习； 质谱； 肽鉴定； 蛋白质同源性预测

Machine Learning Based Bioinformation Retrieval

FU Yan (Computer Application Technology)

Directed by GAO Wen

In information retrieval systems such as biological sequence search engines, the retrieval functions (also referred to as scoring functions or ranking functions sometimes) that list the search results in the order of their relevance to the query are one of the most important components. The design of retrieval functions can be carried out on two levels, i.e., the domain-dependent construction of basic relevance measures and the relatively domain-independent construction of the final retrieval function that combines multiple basic relevance measures into a single one. In this thesis, two important bioinformation retrieval problems, i.e., the protein sequence identification problem and the protein homology prediction problem, are studied on the above two levels of retrieval function design using machine learning techniques.

Peptide and protein identification via tandem mass spectrometry and database search is an important biological sequence retrieval problem. A key ingredient of peptide and protein identification software is the peptide scoring function (retrieval function) that measures the likelihood of a candidate peptide producing the experimental spectrum. In a peptide scoring function, the most basic operation is to match fragment ions predicted from a candidate peptide to the mass peaks in the experimental spectrum. Due to the imprecision of mass measurement, random mismatches often occur. In this thesis, a more accurate mass match error model, namely conditional normal model, is first proposed to improve the accuracy of matching. This model is based on two important observations on the mass error distribution, i.e. the linearity between the mean of mass error and the ion mass, and the logarithmic linearity between the standard deviation of mass error and the peak intensity. To the best of the author's knowledge, the latter quantitative relationship has never been reported before. An iterative learning algorithm is also proposed to accurately estimate the model parameters from training data to characterize the mass error distribution of tandem mass spectra. The thesis then presents a nonlinear peptide scoring function, namely KSDP, which is a nonlinear extension to the commonly used peptide scoring method, spectral dot product (SDP). The correlation among fragment ions in a tandem mass spectrum is very helpful for reducing random mismatches. In KSDP, localized kernel functions are used to emphasize the co-occurring matches of correlated ions. Experiments show that KSDP can significantly improve the peptide identification accuracy. The KSDP-based peptide and protein identification software tool pFind considerably outperforms the SDP-based popular commercial software

SEQUEST in terms of identification accuracy on several data sets. At the 1% false positive rate, pFind identifies 10% to 30% more peptides than SEQUEST.

Due to the complexity of practical retrieval problems, there are usually more than one basic relevance measures, resulting in multiple-dimensional feature vectors. How to combine the multiple relevance measures into a single one is the problem of retrieval function construction. Learning a retrieval function from training data is a common and effective strategy. In general, retrieval function learning is independent of specific domains. In this class of machine learning problem, the feature vectors of database items are computed based on queries and thus they are grouped into blocks by queries. The block structure of data is a unique feature of retrieval function learning problems. This thesis describes a series of approaches for more accurate learning of retrieval functions based on the block structure. These approaches range from the intra-block data normalization and block feature expansion methods for solving the non-i.i.d. (independent and identically distributed) problem, the block selection and support vector under-sampling methods for reducing redundant data, and the K-nearest-block ensemble method for designing query-adaptive retrieval functions. Experimental results with the support vector machine (SVM) used as the benchmark learner demonstrate that all of these block-based approaches significantly outperform the straightforward application of SVMs. The intra-block data normalization and data reduction methods have contributed to our original winning of the protein homology prediction task in the ACM KDDCUP-2004 competition. The K-nearest-block ensemble approach is even more attractive in both prediction accuracy and training speed. It obtains the best prediction result at present on the protein homology prediction task mentioned above.

Keywords: bioinformatics; information retrieval; machine learning; mass spectrometry; peptide identification; protein homology prediction