文章编号:1007-1423(2018)25-0027-05 DOI:10.3969/j.issn.1007-1423.2018.25.006

基于 MIMLNN 的玉米蛋白质功能预测

陈彦明

(上海海事大学信息工程学院,上海 201306)

摘要:

玉米作为世界上最重要的谷物之一,对其蛋白质进行预测具有重要的现实意义。为实现对玉米蛋白质的自动化功能预测,从蛋白质序列数据库提取数据并进行处理,使用优秀的多示例多标记学习算法进行玉米的蛋白质功能预测,应用主流的评价指标对预测结果进行评价,并与同类预测结果进行对比。结果显示,预测效果良好,具有一定的实用价值。

关键词:

玉米;蛋白质功能预测;多示例多标记学习;机器学习

0 引言

谷物及其制品,提供了人类 40%-70%的食品[1],玉米是世界上最主要的谷物,大约 1 万年墨西哥南部的土著人首先种植了玉米[2],现今玉米已成为世界许多地区的主食,总产量超过小麦、大米。然而,并不是所有的玉米都直接被人类消费,一些玉米用于生产乙醇、动物饲料和其他玉米产品,如玉米淀粉和玉米糖浆。谷类对人体健康有非常重要的积极影响,玉米中的纤维素和植物化学素等成分对人体而言具有良好的营养保健作用。

对玉米的蛋白质功能进行注释以便对它功能蛋白的生理意义进行理解,对于玉米蛋白质组学的研究显然非常重要。在世界上较为主流的蛋白质序列数据库中,已有一定量的经人工注释复核的玉米蛋白质数据可供使用,但同时仍有大量未经注释且功能未知的玉米蛋白质序列。面对这些没有经过注释且功能未知的玉米蛋白质,显然手工注释的方法已经跟不上数据的脚步,非常需要一种自动化的方法来对玉米的蛋白质进行功能预测。

在这样的时代背景下,不管是从玉米蛋白质研究的角度来说,还是从玉米对于我国经济社会发展的重要性来说,研究使用计算机技术实现对玉米的蛋白质自动化地进行功能预测具有不言而喻的现实意义。而机器学习技术的兴起发展为解决此类问题提供了一种优秀的解决方案,其中一部分技术则非常适合解决此类预测问题。

1 算法概述

多示例多标记学习(Multi-Instance Multi-label Learning,MIML)由 Zhou 提出[4],提出后产生了很大的影响,作为一种新颖的机器学习框架得到了很好的发展,如今整个多示例多标记学习的生态已经日益繁荣[4-7]。

传统的监督学习使用一个示例(instance)来描述一个对象(object),这里的示例亦即一个特征向量,同时使用一个类别标记(label)与此对象对应。令 X 表示示例空间(或特征空间),Y 表示类别标记的集合,传统监督学习的任务是从给定数据集 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ 中学习函数 $f:X\rightarrow Y$,其中 $x_i\in X$ 是一个实例,$y_i\in Y$ 是 x_i 的已知标记。

这种传统的监督学习框架适用于一些问题,但有很多现实世界的问题不适合这个框架。它的缺点在于每个对象只属于一个概念,相应的示例只对应于单个的类别标记。然而大多数现实世界的对象并非这样简单,可能同时对应于多个的类别标记。于是,多实例多

标签学习框架应运而生，在此框架中，一个对象由多个示例描述，与多个类别标记相关联。对比上述传统的监督学习，MIML 框架对于表示复杂的现实世界对象更方便自然。文献[4]中提出，多示例多标记学习使用多个特征向量来描述一个对象，得到多个示例，同时，使用多个类别标记来与此对象对应。形式上设 X 表示示例空间，Y 表示类别标记的集合。在形式上，多示例多标记学习任务被定义为[4]：从给定数据集 $\{(X_1,Y_1),(X_2,Y_2),\cdots,(X_m,Y_m)\}$ 中学习函数 $f:2^X\rightarrow 2^Y$，其中 $X_i\subseteq X$ 是一组示例 $\{x_{i1},x_{i2},\cdots,x_{i,n_i}\}$ 的集合，$x_{ij}\in X(j=1,2,\cdots,n_i)$，$Y_i\subseteq Y$ 是一组标记 $\{y_{i1},y_{i2},\cdots,y_{i,l_i}\}$ 的集合，$y_{ik}\in Y(k=1,2,\cdots,l_i)$。这里 n_i 表示 X_i 中的示例数量，l_i 表示 Y_i 中的标签数量。

文献[4]基于 MIML 框架提出了多种 MIML 算法，MIMLNN（Multi-Instance Multi-Label Neural Network）是其中一种较优秀的算法。下面简要概述 MIMLNN 算法的主要思想和过程，并使用伪代码进行描述。

首先，收集每个 MIML 示例 $(X_u,Y_u)(u=1,2,\cdots,m)$ 中的 X_u 并将其放入数据集 Γ 中。然后，对 Γ 使用 k-Medoids 算法[8]聚类。由于 Γ 中的每个数据项，即 X_u，是一个未标记的多示例包而不是单个示例，因此基于最大豪斯道夫距离[4]对含有每个标记的训练样本进行 k-Medoids 聚类，并保留每个聚类簇的中心点。

在数学中，豪斯道夫距离（Hausdorff Distance），也称为 Pompeiu-Hausdorff 距离。常被用于计算机视觉等领域。这个距离最早是由豪斯多夫在他 1919 年首次出版的书《人民报》中提出的。简单来说，如果一个集合中的每个点都接近另一个集合的某个点，那么两个集合在 Hausdorff 距离上是接近的。

对于两个示例的包（bag），$A=\{a_1,a_2,\cdots,a_{n_i}\}$，$B=\{b_1,b_2,\cdots,b_{n_j}\}$，两者间的最大豪斯道夫距离为：

$$H^{\max}(A,B)=\max\{\max_{a\in A}\min_{b\in B}\|a-b\|,\ \max_{b\in B a\in A}\min_{a\in A}\|a-b\|\} \quad (1)$$

式中 $|.|$ 为集合的元素数目，$\|.\|$ 为示例之间的欧氏距离（Euclidean Distance）。

在聚类过程之后，数据集 Γ 被划分为 k 个分区，其中中心点（Medoids）为 $M_t(t=1,2,\cdots,k)$。根据这些中心点，原始的多实例 X_u 被转换成 k 维数值向量 z_u，其中 z_u 的第 $i(i=1,2,\cdots,k)$ 分量是 X_u 和 M_i 之间的最大豪斯道夫距离，即 $H^{\max}(X_u,M_i)$。因此，最初的 MIML 例子 $(X_u,Y_u)(u=1,2,\cdots,m)$ 已经被转化为多标签的例子 $(z_u,Y_u)(u=1,2,\cdots,m)$。然后，从数据集中学习多标签学习函数 f_{MLL}，因为 ，故可以得到所需的 MIML 函数。在 MIMLNN 算法中，使用反向传播（Back Propagation，BP）神经网络来实现 f_{MLL}。

2 数据提取和处理

蛋白质数据从世界上主流的蛋白质序列数据库 UniProtKB 取得。UniProtKB 分为 Swiss-Prot 和 TrEMBL 两个子数据库，Swiss-Prot 的注释经过人工手动完成并复核，本文选用 Swiss-Prot 中的数据进行训练和验证，这样得到的结果更有说服力。

使用关键词检索，从 Swiss-Prot 获得了 153 条玉米蛋白质数据，每条玉米蛋白质数据，均包含两个部分：蛋白质结构域（Domain）序列数据和基因本体（Gene Ontology，GO）编号表示的分子功能（Molecular Function）数据。

蛋白质结构域是给定蛋白质序列和蛋白质（三级）结构的保留部分，它可以独立于蛋白质链的其余部分进化、作用和存在。每个结构域形成一个紧凑的三维结构，往往可以独立稳定和折叠。大多数蛋白质由不止一个结构域组成，同样的一个结构域可能出现在各种不同的蛋白质中。分子进化使用结构域作为基本的结构单元，这些结构可以以不同的排列进行重组，以创建具有不同功能的蛋白质。结构域长度从约 25 个氨基酸到 500 个氨基酸长度不等。此概念最早由 Wetlaufer 在 1973 年提出[9]。Wetlaufer 将结构域定义为蛋白质结构的稳定单位，可以自动折叠。大自然通常将几个结构域结合在一起形成具有多种可能性的多域和多功能蛋白质。在多域蛋白质中，每个结构域都可以独立地完成自己的功能，或者以与其邻居一致的方式完成它自己的功能。

基因本体论（GO）是一项重要的生物信息学计划。在生物学领域没有通用的标准术语，术语用法可能特定于物种、研究领域甚至特定的研究小组而异，而此计划旨在解决这些混乱的表示方法。简单来说，GO 提供了一种统一的编号方法来表示所有物种中基因和基因产物的属性，它涵盖三个领域：细胞成分、分子功能、生物过程，本文中我们使用 GO 分子功能的编号来

表示蛋白质的功能。

GO 本体文件可以从 GO 网站以各种格式免费获得。表 1 展示了一个编号为 GO:0000005 的用来描述某种分子功能的 GO 条目。

表 1 GO 本体示例

[Term]
id: GO:0000005
name: obsolete ribosomal chaperone activity
namespace: molecular_function
def: "OBSOLETE. Assists in the correct assembly of ribosomes or ribosomal subunits in vivo, but is not a component of the assembled ribosome when performing its normal biological function." [GOC:jl, PMID:12150913]
comment: This term was made obsolete because it refers to a class of gene products and a biological process rather than a molecular function.
synonym: "ribosomal chaperone activity" EXACT []
is_obsolete: true
consider: GO:0042254
consider: GO:0044183
consider: GO:0051082

使用文献[10]中提出的基于 Conjoint Triad 法[11]的氨基酸序列特征向量提取方法，对上述每个条目中的结构域进行特征向量的提取，每个结构域得到对应的一个特征向量，即为一个“示例”。同时，每个 GO 编号则对应的作为一个“标记”。以这种逻辑关系得到一个完整的玉米多实例多标记样本库，导入 MIMLNN 算法中进行训练，并进行功能预测。

3 结果与对比

使用 3 种主流的多标记学习评价指标对结果进行评价。

Hamming Loss 指标[12-13]用来评价所得结果与实际情况之间的差异大小，也就是样本实际上拥有标记 Y_i，却没有被成功预测，或者，实际上没有拥有标记 Y_i，但是被误认为拥有的可能性，其值越小则预测效果越好。定义如下：

$$HL(x_i,y_i)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{x_i\oplus y_i}{L} \tag{2}$$

式中 $|D|$ 为样本数量，$|L|$ 为标记数量，x_i 为预测值，y_i 为真实值。

maF1、miF1 指标[14-15]分别对 F1 值（F1 Measure）应用宏平均（macro average）和微平均（micro average）。

maF1 先对单一分类标记的个体计算 F1 分数，再对所有标记进行平均。其计算方法如下：

$$maF1(x_i,y_i)=\frac{2}{|L|}\sum_{l=1}^{|L|}\frac{\sum_{i=1}^{|D|}y_{i,l}x_i}{\sum_{i=1}^{|D|}y_{i,l}+\sum_{i=1}^{|D|}x_i} \tag{3}$$

式中 $|D|$ 为样本数量，$|L|$ 为标记数量，x_i 为预测值，y_i 为真实值，$y_{i,l}$ 为 y_i 的第 l 个元素。

miF1 先对所有示例和标记直接进行平均。其计算方法如下：

$$miF1(x_i,y_i)=\frac{2\sum_{i=1}^{|D|}<x_i,y_i>}{\sum_{i=1}^{|D|}y_i+\sum_{i=1}^{|D|}x_i} \tag{4}$$

式中 $|D|$ 为样本数量，$|L|$ 为标记数量，x_i 为预测值，y_i 为真实值，$<\cdot>$ 为数量积。

使用第 2 节处理得到的玉米蛋白质数据，使用 MIMLNN 算法在最优参数条件下进行蛋白质功能预测，使用上述三种主流的评价标准进行评价，结果如表 2 所示，一共进行 10 次预测实验，采用 10 折交叉验证（保留 3 位小数）得到，在表的末尾列出了 10 次实验结果的平均值以及方差。如上文所述，Hamming Loss 的值越小越好，其余两者反之。

表 2 三种指标下玉米蛋白质功能预测结果

试验序号	Hamming Loss	miF1	maF1
1	0.015	0.350	0.046
2	0.015	0.352	0.045
3	0.015	0.344	0.043
4	0.015	0.350	0.044
5	0.014	0.368	0.049
6	0.015	0.364	0.050
7	0.014	0.379	0.048
8	0.014	0.376	0.050
9	0.014	0.362	0.048
10	0.014	0.378	0.052
平均值	0.015	0.362	0.048
标准差	0.000	0.013	0.003

表 3 中展示了本文得出的结果和文献[16]中对于两种微生物的蛋白质功能预测的结果对比，表中数据均以平均值±标准差的形式给出。

表 3 与同类预测结果的对比

物种	Hamming Loss	miF1	maF1
玉米	0.015±0.000	0.362±0.013	0.048±0.003
Haloarcula marismortui	0.012±0.003	0.069±0.037	0.005±0.003
Azotobacter vinelandii	0.013±0.003	0.126±0.044	0.007±0.003

可见，在 Hamming Loss 指标下，本文中的预测结果取得了近似同等的表现，而在其余两种指标下，本文预测结果皆显著更好。

4 结语

玉米作为重要的谷物之一，对其蛋白质进行预测具有显而易见的现实意义。本文应用了一种优秀的多示例多标记学习算法 MIMLNN 进行玉米的蛋白质功能预测，通过对比，证明取得了良好的结果，因此具有一定的实用价值。同时，在机器学习技术日新月异的今天，这类方法仍有较大的改进空间以提高预测效果。

参考文献：

[1]吴伟,李彤,蔡勇建,等. 三种稻米在贮藏过程中蒸煮特性变化的比较[J]. 食品与机械,2014(3):122-126.

[2]Benz B F. Archaeological Evidence of Teosinte Domestication from Guila Naquitz,Oaxaca[J]. Proceedings of the National Academy of Sciences,2001,98(4):2104-2106.

[3]Zhou Z H,Zhang M L,Huang S J,et al. Multi-Instance Multi-Label Learning[J]. Artificial Intelligence,2012,176(1):2291-2320.

[4]Zhou Z H,Zhang M L. Multi-Instance Multi-Label Learning with Application to Scene Classification[C]. Advances in Neural Information Processing Systems,2007:1609-1616.

[5]Briggs F,Fern X Z,Raich R. Rank-Loss Support Instance Machines for MIML Instance Annotation[C]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2012:534-542.

[6]Wu Q,Ng M K,Ye Y. Markov-MIML: A Markov Chain-Based Multi-Instance Multi-Label Learning Algorithm[J]. Knowledge and Information Systems,2013,37(1):83-104.

[7]Shen C,Jing L,Ng M K. Sparse-MIML: a Sparsity-Based Multi-Instance Multi-Learning Algorithm[C]. International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. Springer,Berlin,Heidelberg,2013:294-306.

[8]Kaufman,L. and Rousseeuw,P.J. Clustering by Means of Medoids, in Statistical Data Analysis Based on the L1-Norm and Related Methods[J],1987.

[9]Wetlaufer D B. Nucleation,Rapid Folding,and Globular Intrachain Regions in Proteins[J]. Proceedings of the National Academy of Sciences,1973,70(3):697-701.

[10]WU J S,HU D, XU X,et al. A Novel Method for Quantitatively Predicting Non-Covalent Interactions from Protein and Nucleic Acid Sequence[J]. Journal of Molecular Graphics and Modelling,2011,31:28-34.

[11]Shen J W,Zhang J,Luo X M,et al. Predicting Protein-Protein Interactions Based only on Sequences Information[J]. Proceedings of the National Academy of Sciences,2007,104(11):4337-4341.

[12]Schapire R E,Singer Y. BoosTexter: A Boosting-Based System for Text Categorization[J]. Machine Learning,2000,39(2-3):135-168.

[13]Ghamrawi N,Mccallum A. Collective Multi-Label Classification[C]. Proceedings of the 14th ACM International Conference on Information and Knowledge Management. ACM,2005:195-200.

[14]Zhang M L,Zhou Z H. A k-Nearest Neighbor Based Algorithm for Multi-Label Classification[C].Granular Computing,2005 IEEE International Conference on IEEE,2005, 2:718-721.

[15]Rogati M Yang Y. High-Performing Feature Selection for Text Classification[C]. Proceedings of the Eleventh International Conference on Information and Knowledge Management. ACM,2002:659-661.

[16]Wu J S,Huang S J,Zhou Z H. Genome-wide Protein Function Prediction through Multi-Instance Multi-Label Learning[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2014,11(5):891-902.

作者简介：

陈彦明(1995-)，男，安徽淮南人，硕士，研究方向为机器学习和生物信息学

收稿日期：2018-05-09　　修稿日期：2018-07-12

（下转第 36 页）

Item Collaborative Filtering Algorithm Based on Improved Similarity

GAO Xing-qian, WANG Xiao-feng

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306)

Abstract:

In personalized recommendation algorithm, similarity computation is one of the key factors to decide the accuracy of recommendation algorithm. In the MovieLens data set, the traditional calculation does not consider the common user similarity score scores of time and user items, the average score, that cannot be a good measure of similarity between users in the case of sparse data, leading to the recommendation result is not accurate, because of this, the above problems, adds three improved correction factor the traditional similarity calculation method. In the MovieLens data set, the experimental results show that the average absolute error of the improved collaborative filtering algorithm (MAE) than the improved similarity measure project collaborative filtering recommendation algorithm based on (ICF_IPSS) 7.4% lower than the weighted fusion preference and structural similarity in collaborative filtering algorithm (MCF) is 6% lower than the similarity of cooperation filtering fusion algorithm (ICF_SI) low 1%, visible the improved algorithm has a significant improvement in the accuracy of recommendation.

Keywords:

Collaborative Filtering; Recommendation Algorithm; Person Similarity; Common Scoring Project

(上接第 30 页)

Prediction of Maize Protein Function Based on MIMLNN

CHEN Yan-ming

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306)

Abstract:

As one of the world's most important cereals, maize has important practical significance in predicting its protein. In order to realize the automatic function prediction of maize protein, extracts the data from the protein sequence database and processes, and performs the protein function prediction of maize using an excellent Multi-Instance Multi-Label learning algorithm. Uses mainstream evaluation indicators to evaluate the prediction results, and compares the results with similar experimental results. The results show that the prediction effect is good and has a certain practical value.

Keywords:

Maize; Protein Function Prediction; Multi-Instance Multi-Label Learning; Machine Learning