

معرفی شیوه‌ای نوین و پیشرفته برای پیش‌بینی دامنه تغییرات pH پروتئین‌ها هنگامی که تعداد تکرارها کمتر از تعداد ویژگی‌های مورد بررسی است

باصری سمیه^(۲)، ابراهیمی اسماعیل^{(۱)*}، مینا توحیدی^(۳)

^۱دانشجوی کارشناسی ارشد آمار، دانشکده علوم، دانشگاه شیراز (somayebaseri@gmail.com)

آستادیار بخش زراعت و اصلاح نباتات، دانشکده کشاورزی، دانشگاه شیراز (ebrahimie@shirazu.ac.ir)

آستادیار بخش آمار، دانشکده علوم، دانشگاه شیراز (mtowhidi@susc.ac.ir)

چکیده

برای سال‌ها ارائه مدل پیش‌بینی کننده‌ای کارا و دقیق برای پدیده‌هایی که تحت تاثیر عوامل زیادی رخ می‌دهند اما به دلیل وجود محدودیت‌هایی تعداد تکرار (مشاهده) کمی دارند، امکان‌پذیر نبود. زیرا در این شرایط راهکارهای متداول آماری که بر مبنای زیاد بودن تعداد مشاهدات نسبت به متغیرهای تاثیر گذار شکل گرفته‌اند، قابل استفاده نمی‌باشند. در راستای حل این مشکل، (Srivastava and Kubokawa 2007) برآوردگر Crude را با استفاده از روش‌های بیز تجربی که به عنوان جدیدترین دیدگاه آماری است، معرفی کردند و برتری آن را در مدل بندی دقیق نسبت به دیگر برآوردگرها در این زمینه نشان دادند. در پروتئین بیوانفورماتیک پیش‌بینی دامنه تغییرات pH آنزیم‌های جدید (قبل از تولید آن‌ها در آزمایشگاه) با استفاده از توالی پروتئینی آن‌ها مورد توجه بسیار می‌باشد. این پیش‌بینی و مدل‌بندی به محققین اجازه دستکاری و ایجاد تغییرات در توالی پروتئینی موجود با استفاده از روش جایگزینی آمینو اسید و یا جهش در جهت تولید آنزیم برتر در صنعت را می‌دهد. برای این منظور لازم است خواص و ویژگی‌های زیادی از توالی پروتئین مورد توجه قرار گیرد، در حالیکه تعداد مشاهدات با محدودیت زیادی همراه می‌باشد. با استفاده از برآوردگر Crude مدلی پیش‌بینی کننده برای دامنه تغییرات pH، Xylanase که پروتئین مهمی در صنعت می‌باشد، ارائه شده است. همچنین تعداد عوامل موثر بر تغییرات pH پروتئین Xylanase شده است.

واژه های کلیدی: بیز تجربی، پروتئین بیوانفورماتیک، کاهش بعد

مقدمه

یکی از مهمترین کاربردهای آمار در عرصه‌های مختلف علم از قبیل اقتصاد، ژنتیک، زمین شناسی، فیزیک، بیولوژی و ... مدل بندی و پیش‌بینی می‌باشد. امروزه، مدل‌های رگرسیونی، Data mining، شبکه‌های عصبی مصنوعی و decision trees از مهمترین روش‌های مهم در مدل‌بندی و پیش‌بینی پدیده‌ها می‌باشند (۷)، (۳)، (۸). به طور کلی استفاده از این روش‌ها با در دست داشتن تعداد زیاد مشاهدات امکان‌پذیر است در حالی که در اکثر پدیده‌های بیولوژیکی از قبیل بیماری‌های خاص و بیان ژن‌ها، تعداد متغیرها (یا عامل‌های مؤثر، p) بسیار زیاد و تعداد مشاهدات (یا تکرارها، n) بسیار کم است.

پیش‌بینی بیماری‌های خاص، بیان ژن‌ها و پروتئین بیوانفورماتیک مثال‌های بیولوژیکی مهمی با تعداد متغیر زیاد و تعداد تکرار کم است که مدل‌بندی دقیق‌تر و مؤثرتر را می‌طلبد. برای مثال، در پروتئین بیوانفورماتیک پیش‌بینی دامنه تغییرات pH، آنزیم‌های جدید (قبل از تولید آن‌ها در آزمایشگاه) با استفاده از توالی پروتئینی آن‌ها مورد توجه بسیار می‌باشد. این پیش‌بینی و مدل‌بندی به محققین اجازه دستکاری و ایجاد تغییرات در توالی پروتئینی موجود با استفاده از روش جایگزینی آمینو اسید و یا جهش در جهت تولید آنزیم برتر در صنعت را می‌دهد. برای این منظور لازم است خواص و ویژگی‌های زیادی از توالی پروتئین مورد توجه قرار گیرد، در حالیکه تعداد مشاهدات با محدودیت زیادی همراه می‌باشد. در تحلیل بیان ژن‌ها، نیز سطوح بیان هر ژن به عنوان یک متغیر در نظر گرفته می‌شود که ممکن است تعداد ژن‌ها تا ۳۰۰۰۰ برسد (۱۱).

رگرسیون خطی مهمترین روش مدل‌بندی می‌باشد. اکثر آماردانان و ریاضیدانان معتقدند که استفاده از مدل‌های رگرسیونی در شرایطی که تعداد متغیرها به مراتب بیشتر از تعداد مشاهدات باشد، امکان‌پذیر نمی‌باشد (۴)، (۶). در سال‌های اخیر تلاش‌هایی برای حل این مشکل به ویژه (۱۲)، (۱۱)، (۱۰) صورت گرفته است. در (۱۲) دیدگاه رگرسیونی بیزی جدیدی برای کلاس‌بندی مطرح می‌شود. اما اصلاحات عمده در این زمینه در (۱۰) انجام شده است. آن‌ها نشان داده‌اند که برآوردگر کمترین توان دوم خطا که مهمترین و متداول‌ترین برآوردگر در مدل‌های رگرسیونی است، خطای پیش‌بینی زیادی نیز ایجاد می‌کند. در (۱۰) با استفاده از دیدگاه بیز تجربی برآوردگرهایی برای β ، بردار ضرایب رگرسیونی، معرفی شده که ریسک به مراتب کمتری نسبت به برآوردگرهای کمترین توان دوم خطا و برآوردگر بیزی در (۱۱) دارند.

از بیز تجربی می‌توان به عنوان سومین و جدیدترین دیدگاه مهم آماری نام برد. بیز تجربی قادر است پاسخ مناسبی به مسائل با تعداد داده محدود دهد و نقطه ضعف دو دیدگاه کلاسیک و بیز را در این زمینه برطرف نماید. اخیراً، بیز تجربی در بررسی و تجزیه و تحلیل بیان ژن‌ها در میکروآرای و پروتئین بیوانفورماتیک جایگاه ویژه‌ای یافته است (۲) و (۹).

در (۱۰) برآورد بیز تجربی با نام برآوردگر Crude را معرفی و برتری آن را نسبت به برآوردگرهایی که تاکنون در این زمینه معرفی شده، نشان داده شده است. با استفاده از برآوردگر Crude مدلی پیش‌بینی کننده برای تعیین دامنه تغییرات pH، Xylanase که پروتئین مهمی در صنعت می‌باشد، ارائه شده است. برآورد Crude می‌تواند گامی بزرگ در مدل‌بندی و پیش‌بینی پدیده‌های بسیاری با تعداد متغیر زیاد و تعداد مشاهده بسیار کم باشد. نکته قابل توجه آن است که این برآوردگر نوعی کاهش بعد در خود دارد که بسیار مفید است و می‌تواند تعداد عوامل تاثیر گذار بر بروز پدیده را مشخص کند.

برآوردهای کمترین توان دوم خطا (LSE)

در نظر می‌گیریم که n متغیر وابسته y_1, \dots, y_n با p متغیر رگرسیونی x_1, \dots, x_p ارتباط خطی به فرم $y = \beta_0 \mathbf{1}_n + X\beta + \varepsilon$ داشته باشند. در این مدل $y = (y_1, \dots, y_n)^t$ ، X ماتریسی $n \times p$ از مشاهدات بر اساس p متغیر مستقل x_1, \dots, x_p که میانگین نمونه‌ای متناظرشان از آن‌ها کسر شده است و $\mathbf{1}_n = (1, \dots, 1)^t$ برداری n -تایی با مولفه‌های یک می‌باشند. همچنین β_0 ثابتی نامعلوم، β برداری p -بعدی از پارامترهای رگرسیونی β_1, \dots, β_p به فرم $\beta = (\beta_1, \dots, \beta_p)$ و $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$ برداری از خطاهاست. همچنین فرض می‌کنیم $y | \beta \sim N_n(\beta_0 \mathbf{1}_n + X\beta, \sigma^2 I_n)$ که I_n ماتریس همانی از مرتبه n است. تجزیه مقدار تکین (SVD) از ماتریس X به فرم $X^t = ADB$ ، $A^t A = I_r$ ، $BB^t = I_r$ در نظر می‌گیریم. در این تجزیه، A ماتریسی $p \times r$ از r بردار ویژه متناظر با r مقدار ویژه غیر صفر ماتریس $X^t X$ و $D = \text{diag}(d_1, \dots, d_r)$



می‌خواهیم مدل ۹۹/۵ درصد تغییر پذیری را توجیه کند. این مساله ذاتا نوعی کاهش بعد در خود دارد و می‌توان نتیجه گرفت که از ۷۳ متغیر تنها ۲۱ متغیر ۹۹/۵ درصد تغییر پذیری را در بر دارند. نتایج حاصل از مدل‌بندی در جدول زیر آمده است که نشان می‌دهد برآوردگر Crue به نحو بارزی دقیق‌تر و کاراتر از برآوردگر LS عمل می‌کند.

No. (observation, #)	Y (Actual response, pH range)	Y_Is (prediction Least Square method)	Y_C (prediction by Crude estimator)
1	3.0	3.0414576	3.0000118
2	4.0	5.4452857	3.9999950
3	0.5	7.0718161	0.5000462
4	4.0	6.7851466	4.0000149
5	1.0	7.5894497	1.0000230
6	4.7	3.1961093	4.6999910
7	3.1	0.6023857	3.1000359
8	3.0	4.7507216	2.9999620
9	0.6	4.6730216	0.6000243
10	0.2	4.4371186	0.1999862
11	6.0	4.1928801	6.0000090
12	7.0	4.7645279	6.9999768
13	6.0	0.6443653	5.9999477
14	4.0	2.0228184	4.0000533
15	3.0	4.8661313	3.0000425
16	5.0	3.6931024	5.0000041
17	7.0	2.9227417	6.9999368
18	6.5	3.0719017	6.4999648
19	5.0	7.2498874	4.9999892
20	7.0	2.8627370	6.9999901
21	2	-0.1150655	1.9999929
22	2.5	1.3314599	2.5000025
Coefficient of Determination, R^2		0.4300244	0.9995

نتیجه و بحث

در این مقاله روی راهکاری نوین در مدل بندی پدیده‌هایی با تعداد متغیرهای مستقل بسیار زیاد و تعداد تکرار کم بحث کردیم. براساس تحقیق انجام شده در (۱۰) در بین برآوردگرهای بیز تجربی و برآوردگر معرفی شده در (۱۱)، برآوردگر Crude دارای کمترین میزان خطای پیش‌بینی و مدل‌بندی در زمینه مدل‌بندی پدیده‌ها با تعداد تکرار کم و تعداد متغیر بسیار می‌باشد. با استفاده از این برآوردگر مدل پیش‌بینی کننده کارا و دقیقی برای تعیین دامنه تغییرات pH، Xylanase که پروتئین مهمی در صنعت می‌باشد، ارائه شده است. این پیش‌بینی و مدل‌بندی به محققین اجازه دستکاری و ایجاد تغییرات در توالی پروتئینی موجود با استفاده از روش جایگزینی آمینو اسید و یا جهش در جهت تولید آنزیم برتر در صنعت را می‌دهد. همچنین مشخص شده که از ۷۳ عامل تنها ۲۱ عامل تاثیر به سزایی در دامنه تغییرات pH پروتئین Xylanase دارند.

منابع:

- (۱) باصری، س. بررسی مدل‌های خطی و توابع پیش‌بینی کننده خطی با استفاده از برآوردهای بیز تجربی. ۱۳۸۷. پایان‌نامه.
- 2) Efron B, Tibshirani R, Empirical Bayes methods and false discovery rates for microarrays, *Genetic Epidemiology*, 2002, **23**, 70-86.
- 3) Elson A, Tailor S, Salim RB, Hillaby K, Jurkovic D, Expectant management of tubal ectopic pregnancy: prediction of successful outcome using decision tree analysis, *Ultrasound in Obstetrics and Gynecology*, 2004, 6: 552 – 556.
- 4) Montgomery DC, Peck EA, Introduction linear regression analysis, John Wiley, New York, 1982.
- 5) Natesh R, Bhanumoorthy P, Vithayathil PJ, Sekar K, Ramakumar S, M. A. Viswamitra MA, Crystal structure at 1.8 Å resolution and proposed amino acid sequence of a thermostable xylanase from *Thermoascus aurantiacus*, *Journal of Molecular Biology*, 1999, 999-1012.
- 6) Rencher AC, Schaalje GB, *Linear Models in Statistics*, 2nd ed., John Wiley, Hoboken, New Jersey, 2007.
- 7) Roddick JF, Hornsby K, Spiliopoulou M, An Updated Bibliography of Temporal, Spatial, and Spatio-temporal Data Mining Research, *Lecture Notes in Computer Science Springer Berlin*, 2001.
- 8) Schuize FH, Wolf H, Jansen H, Vander VP, Applications of artificial neural networks in integrated water management : fiction or future?, *Water science and technology*, 2005, 52: 21-31.
- 9) Smyth GK, *Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments*, *Statistical Application in Genetics and Molecular Biology*, 2004, 3, Article3.
- 10) Srivastava MS, Kubokawa T, Empirical Bayes regression analysis with many regressors but fewer observations, *Journal of Statistical Planning and Inference*, 2007, 137, 3778-3792.
- 11) West M, Bayesian factor regression models in the 'large p , small n ' paradigm, *Bayesian Statist*, 2003, 7, 723-732
- 12) West M, Nevins JR, Marks JR, Spang R, Zuzan H, (2000) DNA microarray data analysis and regression modeling for genetic expression profiling, ISDS Discussion Paper, 2000, www.isds.duke.edu

Introduction of an expert system for prediction of protein pH resistance when the number of replications are less than the number of features

Somaye Baseri¹, Esmaeil Ebrahimie², Mina Towhidi³

¹Department of Statistics, College of Science, Shiraz University, Shiraz, Iran (somayebaseri@gmail.com)

²Department of Crop Production & Plant Breeding, College of Agriculture, Shiraz University, Shiraz, Iran (brahimiet@gmail.com)

³Department of Statistics, College of Science, Shiraz University, Shiraz, Iran (mtowhidi@susc.ac.ir)

Abstract

For years, presenting an efficient and precise predicting model for those phenomena, occurring under diverse circumstances, with few repetition numbers as a result of the limiting factors, was impossible. As under such conditions, the statistical methods which were based on the abundance of the observation occurrences per the effective variables, are not applicable. To solve this problem Srivastava and Kubokawa (2007) introduced the Crude function, using the empirical Bayes which is the most recent statistical viewpoint. They demonstrated its prominence in a precise and efficient modeling, compared to the other functions. There is a great interest in protein bioinformatics to predict the pH tolerance of new enzyme (before their generation in laboratory) using protein sequence. This prediction and modeling allows scientists to smartly manipulate the sequence of the available protein by amino acid substitution method or mutation to produce super enzyme for industry. For this purpose it is necessary to calculate a large number of traits for a protein sequence though the number of observations have lots of limitations. Using the crude estimator, a predictive modeling for the pH resistance of Xylanase (which is an important protein in industry) and the affecting factors on the resistance of its pH is presented.

Keyword: empirical Bayes, protein bioinformatics, dimension reduction