۱.مجموعه داده اول

مجموعه داده اول در سال ۲۰۱۶ توسط عزت و همکاران [۱] معرفی شده است. اغلب روش‌های جدید پیش‌بینی تعاملات دارو- هدف، با استفاده از این مجموعه داده، روش خود را ارزیابی و نتایج را گزارش کرده‌اند [۴-۱]. این مجموعه داده شامل سه فایل متنی تعاملات دارو- هدف، ویژگی‌های داروها و ویژگی‌های اهداف می‌باشد و تمام داده‌های موجود در این مجموعه داده، از نوع عددی هستند.

مجموعه داده اول در دو قسمت توضیح داده می‌شود؛ در قسمت اول، اطلاعاتی درباره تعاملات دارو-هدف و در قسمت دوم اطلاعات دارو و هدفی که مورد استفاده قرار گرفته شده است و نحوه نمایش آن‌ها به صورت بردارهای ویژگی، تشریح می‌شود.

۱-۱.اطلاعات تعاملات دارو-هدف

عناصر فایل تعاملات دارو- هدف از اعداد صفر و یک تشکیل شده است؛ این فایل به صورت یک ماتریس دوبعدی $Y \in \mathbb{R}^{n\times m}$ دارای n سطر دارو و m ستون پروتئین هدف می‌باشد؛ به طوریکه اگر $Y_{ij} = 1$ باشد، دارو d_i با هدف t_j تعامل دارد و در غیر اینصورت $Y_{ij} = 0$ می‌باشد. به عبارت دیگر عدد یک نشان‌دهنده وجود تعامل و عدد صفر نشان‌دهنده عدم تعامل بین دارو d_i که در سطر ماتریس و هدف t_j که در ستون ماتریس قرار گرفته است، می‌باشد. تعاملات موجود در این مجموعه داده از پایگاه داده DrugBank[5, 6] بدست می‌آید. در این مجموعه داده ۵۸۷۷ دارو، ۳۳۴۸ پروتئین هدف و ۱۲۶۷۴ تعامل شناخته شده وجود دارد.

۲-۱.اطلاعات داروها و اهداف

ویژگی‌های داروهای موجود در این مجموعه داده با استفاده از کتابخانه Rcpi [۷] محاسبه می‌شود. نمونه‌هایی از ویژگی‌های داروها شامل توصیف کننده‌های constitutional، توپولوژیکی[1]، geometrical و دیگر خصوصیات مولکولی می‌باشد [۱]. ویژگی‌هایی از اهداف که بتواند خواص پروتئین‌های مختلف را به طور جامع توصیف کنند از توالی ژنومی پروتئین‌ها با استفاده از وب سرور PROFEAT [۸] تولید می‌شود؛ نمونه‌هایی از ویژگی‌های اهداف شامل توصیف کننده‌های مربوط به ترکیب اسید آمینه[2]، dipeptide composition، همبستگی، amphiphilic pseudo-amino acid و چند ویژگی دیگر است که در وب سرور PROFEAT قابل مشاهده می‌باشد [۱].

---

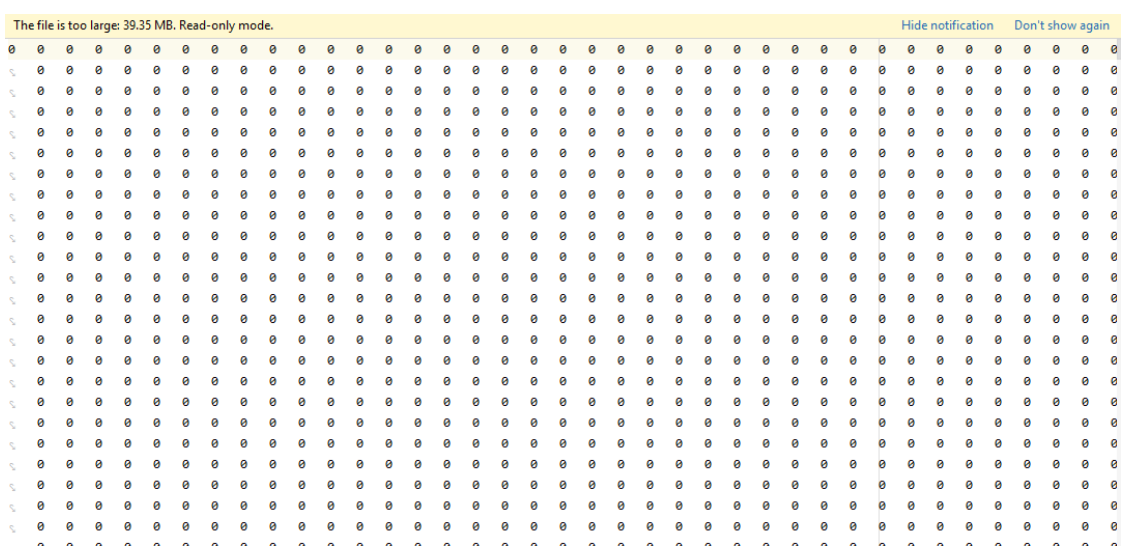
[1] Topological
[2] Amino acid composition

فایل ویژگی‌های داروها به صورت ماتریس $D \in \mathbb{R}^{n\times f}$ و ویژگی‌های اهداف به شکل ماتریس $T \in \mathbb{R}^{m\times g}$ می‌باشد که f و g تعداد ویژگی‌های داروها و اهداف است. عناصر تشکیل دهنده این ماتریس‌ها اعدادی هستند که مقادیر توصیف کننده‌های ملکولی داروها و اهداف را نشان می‌دهند. مقادیر ویژگی‌های داروها و اهداف اعدادی اعشاری هستند که همه آن‌ها برای اجتناب از سوگیری ویژگی‌ها با مقادیر بزرگ با استفاده از روش نرمال سازی Min-Max در بازه [۰,۱] قرار گرفته‌اند. در این مجموعه داده هر دارو و هدف به ترتیب با یک بردار ویژگی ۱۹۳ بعدی و ۱۲۹۰ بعدی نشان داده می‌شود؛ به این صورت که هر دارو به شکل $d = [d_1, d_2, \dots, d_p]$ و هر هدف به صورت $t = [t_1, t_2, \dots, t_q]$ نمایش داده می‌شود. جزئیات این مجموعه داده در جدول ۱ آورده شده است.

جدول ۱.جزئیات مجموعه داده اول

تعداد داروها	تعداد اهداف	تعداد تعاملات
۵۸۷۷	۳۳۴۸	۱۲۶۷۴

تشکیل بردارهای دارو-هدف برای استفاده در مدل یادگیری ماشین به صورت $[d_1, d_2, \dots, d_p, t_1, t_2, \dots, t_q, I]$ می‌باشد که برچسب I یک عدد باینری است و نشان دهنده وجود یا عدم وجود تعامل بین دارو و هدف موجود در بردار می‌باشد. برچسب I از طریق ماتریس تعاملات دارو-هدف مشخص می‌شود. شکل ۱ و ۲ به ترتیب نمایی از فایل تعاملات دارو-هدف و فایل ویژگی‌های داروها می‌باشد.



شکل ۱ .نمایی از فایل تعاملات دارو-هدف

The file size (6.48 MB) exceeds the configured limit (2.56 MB). Code insight features are not available.

```
0.74017 0.01967 0.24839 0.23409 0.23077 0   0   0   0   0   0.23077 0   0   0.66667 0.2 0.11111 0   0.33333 0   0   0.2175  0.10824 0.822
0.77503 0.0099534   0.20324 0.18898 0.17949 0.16667 0.5 0   0.16667 0.14286 0   0.20513 0   0   0.33333 0   0   0   0   0   0   0.17125 0.088254
0.77162 0.010758    0.29128 0.26601 0.25641 0.16667 0   0   0   0   0   0.25641 0   0   0.33333 0.2 0   0   0.16667 0   0   0.24125 0.10624 0.82
0.84958 0.00017337  0.25819 0.25141 0.15385 0   0   0   0   0   0   0.17949 0   0   0   0.4 0   0   0   0   0.4 0.24375 0.089449    0.82397 0.82
0.76321 0.01288 0.20109 0.18744 0.20513 0   0.5 0   0.33333 0.14286 0   0.23077 0   0   0.33333 0   0.11111 0   0   0   0   0.17    0.084837
0.79337 0.0061565   0.21689 0.18973 0.17949 0   0   0   0.33333 0   0   0.17949 0   0   0   0   0.11111 0   0.16667 0.25    0   0.17    0.069826
0.83887 0.00065599  0.043929    0.035165    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0.03125 0.029281    0.92733 0.85
0.83527 0.00088729  0.26454 0.25528 0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0.22625 0.020701    0.96091 0.83373 0.17
0.82023 0.0022332   0.086367    0.073563    0.025641    0   0   0   0   0   0.14286 0.051282    0   0   0   0   0   0   0   0   0   0.0675  0.04
0.82983 0.0013039   0.030515    0.024233    0.025641    0   0   0   0   0   0   0.025641    0   0   0   0   0   0   0   0   0   0.02375 0.014919
0.8071  0.003908    0.070662    0.066079    0.025641    0   0   0   0   0   0   0.025641    0   0   0   0   0   1   0   0   0   0.06125 0.040021
0.85476 5.1187e-005 0.014376    0.010469    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0.01125 0.014746    0.97967 0.82
0.86246 3.198e-006  0.039448    0.029701    0.025641    0   0   0   0   0   0   0.025641    0   0   0   0   0   0   0   0   0   0.0275  0.010025
0.84538 0.00032589  0.042468    0.04196 0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0.03875 0.01839 0.95523 0.84023 0.17
0.84128 0.00052035  0.023237    0.022464    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0.025   0.0084866   0.97708 0.83
0.82662 0.0015871   0.025917    0.027001    0.025641    0   0   0   0   0   0   0.025641    0   0   0   0   0.11111 0   0   0   0   0.02875 0.01
0.83371 0.00099868  0.029167    0.029726    0.025641    0.16667 0   0   0   0   0   0.025641    0   0   0   0   0   0   0   0   0   0.03125 0.01
0.83897 0.00064999  0.029135    0.023293    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0.02375 0.026036    0.97091 0.81
0.81821 0.0024606   0.043922    0.046372    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0.04875 0.0078048   0.98199 0.83
0.84405 0.00038376  0.021203    0.017483    0.025641    0   0   0.5 0   0   0   0.025641    0   0   0   0   0   0   0   0   0   0.01875 0.018201
0.83021 0.001272    0.023591    0.023169    0.025641    0   0   0   0   0   0   0.025641    0   0   0   0   0   0   0   0   0   0.025   0.011076
0.83553 0.0008695   0.023706    0.02251 0.025641    0   0   0   0   0.14286 0   0.025641    0   0   0   0   0   0   0   0   0   0.02375 0.015912
0.81864 0.0024119   0.058498    0.049273    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0.045   0.046482    0.8929  0.84
0.8597  2.0128e-006 0.064148    0.063653    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0.06125 0.0083392   0.9806  0.83
```

شکل ۲ .نمایی از فایل ویژگی‌های دارو

فهرست منابع

[۱] A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, "Drug-target interaction prediction via class imbalance-aware ensemble learning," *BMC bioinformatics,* vol. 17, no. 19, p. 509, 2016.

[۲] A. Sharma and R. Rani, "BE-DTI': ensemble framework for drug target interaction prediction using dimensionality reduction and active learning," *Computer methods and programs in biomedicine,* vol. 165, pp. 151-162, 2018.

[۳] A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, "Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey," *Brief Bioinform,* vol. 8, 2018.

[٤] A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, "Drug-target interaction prediction using ensemble learning and dimensionality reduction," *Methods,* vol. 129, pp. 81-88, 2017.

[٥] C. Knox *et al.*, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic acids research,* vol. 39, no. suppl_1, pp. D1035-D1041, 2010.

[٦] D. S. Wishart *et al.*, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research,* vol. 34, no. suppl_1, pp. D668-D672, 2006.

[۷] D.-S. Cao, N. Xiao, Q.-S. Xu, and A. F. Chen, "Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions," *Bioinformatics* ,vol. 31, no. 2, pp. 279-281, 2014.

[۸] Z.-R. Li, H. H. Lin, L. Han, L. Jiang, X. Chen, and Y. Z. Chen, "PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence," *Nucleic Acids Research* ,vol. 34, no. suppl_2, pp. W32-W37, 2006.