۱.مجموعه داده دوم

مجموعه داده دوم، سال ۲۰۱۲ در مقاله [۱] معرفی شده است. 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 دوم 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[2, 3] بدست می‌آید. در این مجموعه داده ۱۸۶۲ دارو، ۱۵۵۴ پروتئین هدف و ۴۸۰۹ تعامل شناخته شده وجود دارد.

۱-۲.اطلاعات داروها و اهداف

داروهای موجود در این مجموعه داده به صورت اثرانگشت‌های PubChem (بردارهای باینری که هر عنصر آن نشان دهنده وجود یا عدم وجود یکی از ۸۸۱ زیرساخت شیمیایی شناخته شده[1] می‌باشد.) هستند. اهداف نیز به صورت اثرانگشت‌هایی نشان داده شده است که نمایانگر وجود یا عدم وجود ۸۷۶ دامنه پروتئین مختلف (بدست آمده از پایگاه داده Pfam [۴]) می‌باشد؛ بنابراین مقادیر ویژگی‌های داروها و اهداف از اعداد صفر و یک تشکیل شده است.

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. در این مجموعه داده هر دارو و هدف به ترتیب با یک بردار ویژگی ۸۸۱ بعدی و ۸۷۶ بعدی نشان داده می‌شود؛ به این صورت که هر دارو به شکل $d =$

[1] Known chemical substructures

$[d_1, d_2, \dots, d_p]$ 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 دوم

تعداد داروها	تعداد اهداف	تعداد تعاملات
۱۸۶۲	۱۵۵۴	۴۸۰۹

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.

The file size (5.82 MB) exceeds the configured limit (2.56 MB). Code insight features are not available.

	3B11_HUMAN	3BHS1_HUMAN	3BHS2_HUMAN	3HIDH_HUMAN	5HT1A_HUMAN	5HT1B_HUMAN	5HT1D_HUMAN	5HT1E_HUMAN	5HT1F_HUMAN	5HT2A_HUMAN	5HT2B_HUMAN	5HT2C_HUM
2	100252	0	0	0	0	0	0	0	0	0	0	0
3	1003	0	0	0	0	0	0	0	0	0	0	0
4	1005	0	0	0	0	0	0	0	0	0	0	0
5	10062751	0	0	0	0	0	0	0	0	0	0	0
6	10132	0	0	0	0	0	0	0	0	0	0	0
7	1014	0	0	0	0	0	0	0	0	0	0	0
8	10180	0	0	0	0	0	0	0	0	0	0	0
9	1021	0	0	0	0	0	0	0	0	0	0	0
10	10214	0	0	0	0	0	0	0	0	0	0	0
11	10267	0	0	0	0	0	0	0	0	0	0	0
12	1030	0	0	0	0	0	0	0	0	0	0	0
13	1045	0	0	0	0	0	0	0	0	0	0	0
14	1046	0	0	0	0	0	0	0	0	0	0	0
15	104741	0	0	0	0	0	0	0	0	0	0	0
16	104758	0	0	0	0	0	0	0	0	0	0	0
17	104799	0	0	0	0	0	0	0	0	0	0	0
18	104850	0	0	0	0	0	0	0	0	0	0	0
19	104865	0	0	0	0	0	0	0	0	0	0	0
20	1050	0	0	0	0	0	0	0	0	0	0	0
21	1051	0	0	0	0	0	0	0	0	0	0	0
22	10517	0	0	0	0	0	0	0	0	0	0	0
23	1053	0	0	0	0	0	0	0	0	0	0	0
24	10531	0	0	0	0	1	1	0	0	1	0	0

شکل ۱ .نمایی از فایل تعاملات دارو-هدف

The file size (3.3 MB) exceeds the configured limit (2.56 MB). Code insight features are not available.

	SUB1	SUB2	SUB3	SUB4	SUB5	SUB6	SUB7	SUB8	SUB9	SUB10	SUB11	SUB12	SUB13	SUB14	SUB15	SUB16	SUB17	SUB18
2	100252	1 1	0 0	0 0	0 0	0 1	1 1	0 0	1 1	1 0	1 1	1 1	0 0	0 0	0 0	0 0	1 0	0 0
3	1003	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	1 1	1 0	0 0	0 0	0 0	0 0	1 0	0 0
4	1005	1 0	0 0	0 0	0 0	0 1	0 0	0 0	0 0	0 0	1 1	1 0	0 0	0 0	0 0	0 0	1 0	0 0
5	10062751	1 1	0 0	0 0	0 0	0 0	1 1	1 0	0 1	0 0	0 1	1 1	0 0	0 0	0 0	0 0	0 0	0 0
6	10132	0 0	0 0	0 0	0 0	0 1	0 0	0 0	0 0	0 0	1 1	1 0	0 0	0 0	0 0	0 0	0 0	0 0
7	1014	1 1	0 0	0 0	0 0	0 1	1 0	0 0	1 0	0 0	1 1	1 0	0 0	0 0	0 0	0 0	1 0	0 0
8	10180	1 1	1 0	0 0	0 0	0 1	1 1	1 0	1 1	0 0	1 0	0 0	0 0	0 0	0 0	0 0	0 0	1 0
9	1021	1 1	0 0	0 0	0 0	0 1	1 1	0 0	1 1	0 0	1 1	1 0	0 0	0 0	0 0	0 0	0 0	0 0
10	10214	1 1	0 0	0 0	0 0	0 1	1 1	0 0	1 1	1 0	1 1	1 1	0 0	0 0	0 0	0 0	1 1	0 1
11	10267	1 1	0 0	0 0	0 0	0 1	1 0	0 0	0 0	0 0	1 1	1 1	0 0	0 0	0 0	0 0	1 1	0 0
12	1030	1 1	0 0	0 0	0 0	0 0	0 0	0 1	0 0	0 0	1 1	0 0	0 0	0 0	0 0	0 0	0 0	0 0
13	1045	1 1	0 0	0 0	0 0	0 1	1 0	0 0	1 1	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
14	1046	1 0	0 0	0 0	0 0	0 1	1 0	0 0	1 1	0 0	1 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
15	104741	1 1	1 1	0 0	0 0	0 1	1 1	1 1	0 0	0 0	1 1	0 0	0 1	1 1	0 0	0 0	0 0	0 1
16	104758	1 1	1 0	0 0	0 0	0 1	1 1	1 0	1 1	1 0	1 1	1 0	0 0	0 0	0 0	0 0	0 0	0 1
17	104799	1 1	1 0	0 0	0 0	0 1	1 1	0 0	1 1	0 0	1 1	1 0	0 0	0 0	0 0	0 0	1 0	0 0
18	104850	1 1	1 0	0 0	0 0	0 1	1 1	1 0	1 1	1 0	1 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
19	104865	1 1	1 0	0 0	0 0	0 1	1 1	1 0	1 1	1 0	1 1	1 0	0 0	0 0	0 0	0 0	0 0	0 1
20	1050	1 1	0 0	0 0	0 0	0 1	1 1	0 0	1 0	0 0	1 1	0 0	0 0	0 0	0 0	0 0	0 0	0 0
21	1051	1 1	0 0	0 0	0 0	0 1	1 1	0 0	1 0	0 0	1 1	1 0	0 0	0 0	0 0	0 0	1 0	0 0
22	10517	1 1	1 0	0 0	0 0	0 1	1 1	1 0	1 0	0 0	1 1	0 0	0 0	0 0	0 0	0 0	0 0	0 0
23	1053	1 1	0 0	0 0	0 0	0 1	1 1	0 0	1 1	0 0	1 1	1 0	0 0	0 0	0 0	0 0	1 0	0 0
24	10531	1 1	1 1	0 0	0 0	0 1	1 1	1 1	1 1	1 0	1 1	1 0	0 0	0 0	0 0	0 0	0 0	0 0
25	1054	1 1	0 0	0 0	0 0	0 1	1 1	0 0	1 0	0 0	1 1	0 0	0 0	0 0	0 0	0 0	0 0	0 0

شکل ۲ .نمایی از فایل ویژگی‌های دارو

فهرست منابع

[۱] Y. Tabei, E. Pauwels, V. Stoven, K. Takemoto, and Y. Yamanishi, "Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers," *Bioinformatics,* vol. 28, no. 18, pp. i487-i494, 2012.

[۲] 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.

[۳] 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.

[٤] R. D. Finn *et al.*, "The Pfam protein families database: towards a more sustainable future," *Nucleic acids research,* vol. 44, no. D1, pp. D279-D285, 2016.