1.DataSet 1

The first data set was introduced in 2016 by Ezzat et al[1]. Most new methods of predicting drug-target interactions have evaluated their method and reported results using this dataset[2-4]. This dataset includes three txt files of drug-target interactions, drug features, and target features, and all data in this data set are numeric.

The first data set is described in two parts; In the first part, information about drug-target interactions and in the second part, information about the drug and target used and how they are represented as feature vectors are described.

1-1. Drug-target interaction data

Drug-target interaction file as a two-dimensional matrix $Y \in \mathbb{R}^{n\times m}$ with n drug rows and m target columns. That is, $Y_{ij} = 1$ if drug d_i and target $t_j$interact and $Y_{ij} = 0$ otherwise. In other words, the one indicates the presence of interaction and the zero indicates the lack of interaction between the drug d_i which is in the matrix row and the target t_j which is located in the matrix column. Also, the features of drugs are in the form of $D \in \mathbb{R}^{n\times f}$ matrix and the features of targets are in the form of $T \in \mathbb{R}^{m\times g}$ matrix, where f and g are the number of features of drugs and targets. The interaction data were obtained from the DrugBank [5]. In total, there are 12674 drug-target interactions between 5877 drugs and their 3348 protein interaction partners.

1-2. Drugs and targets information

features for drugs were calculated using the Rcpi package[6]. Examples of drug features include constitutional, topological and geometrical descriptors among other molecular properties. the target features were computed from their genomic sequences with the help of the PROFEAT web server[7]. The features that have been used to represent targets in this work are descriptors related to amino acid composition; dipeptide composition; autocorrelation; composition, transition and distribution; quasi-sequence-order; amphiphilic pseudo-amino acid composition and total amino acid properties.

The file of drug features is in the form of $D \in \mathbb{R}^{n \times f}$ matrix and the target features are in the form of $T \in \mathbb{R}^{m \times g}$ matrix, where f and g are the number of drug features and targets features. The constituents of these matrices are numbers that represent the values of the molecular descriptors of drugs and targets. The properties values of drugs and targets are decimal numbers, all of which are in the range [1,0] to avoid biasing properties in large quantities using the Min-Max normalization method. In this dataset, each drug and target is represented by a 193-dimensional and 1290-dimensional property vectors, respectively; Each drug is displayed as $d = [d_1, d_2, \dots, d_p]$ and each target as $t = [t_1, t_2, \dots, t_q]$. Details of this dataset are given in Table 1.

Table 1. Statistics of first dataset

Drugs	Targets	Interactions
5877	3348	12674

Formation of drug-target vectors for use in machine learning model as $[d_1, d_2, \dots, d_p, t_1, t_2, \dots, t_q, I]$, which label I is a binary number and indicates the presence or absence of interaction Is between the drug and the target in the vector. Label I is identified by the drug-target interaction matrix. Figures 1 and 2 show a view of the drug-target interaction file and the drug features file, respectively.

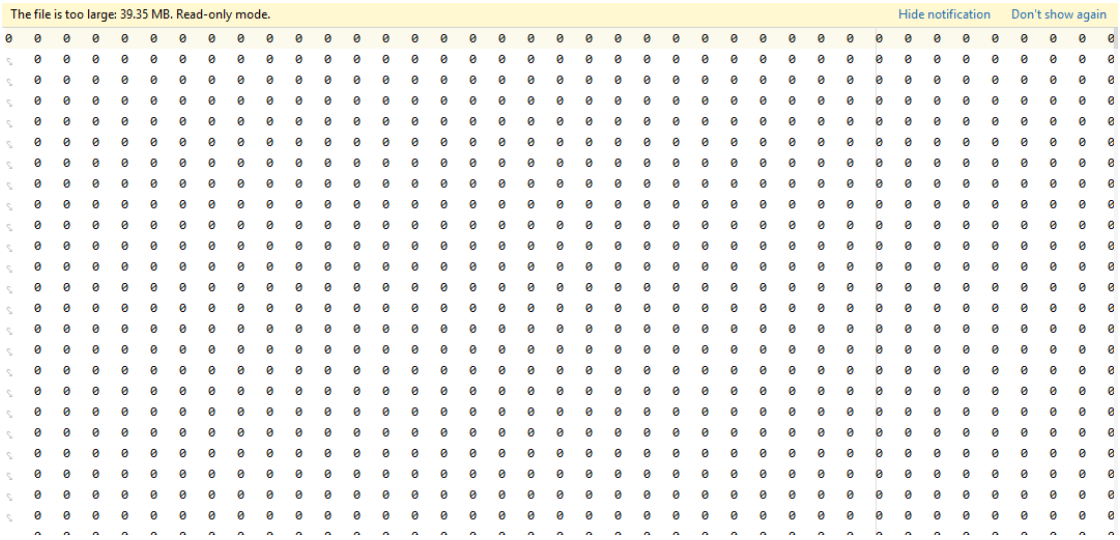

Figure 1. view of the drug-target interaction file

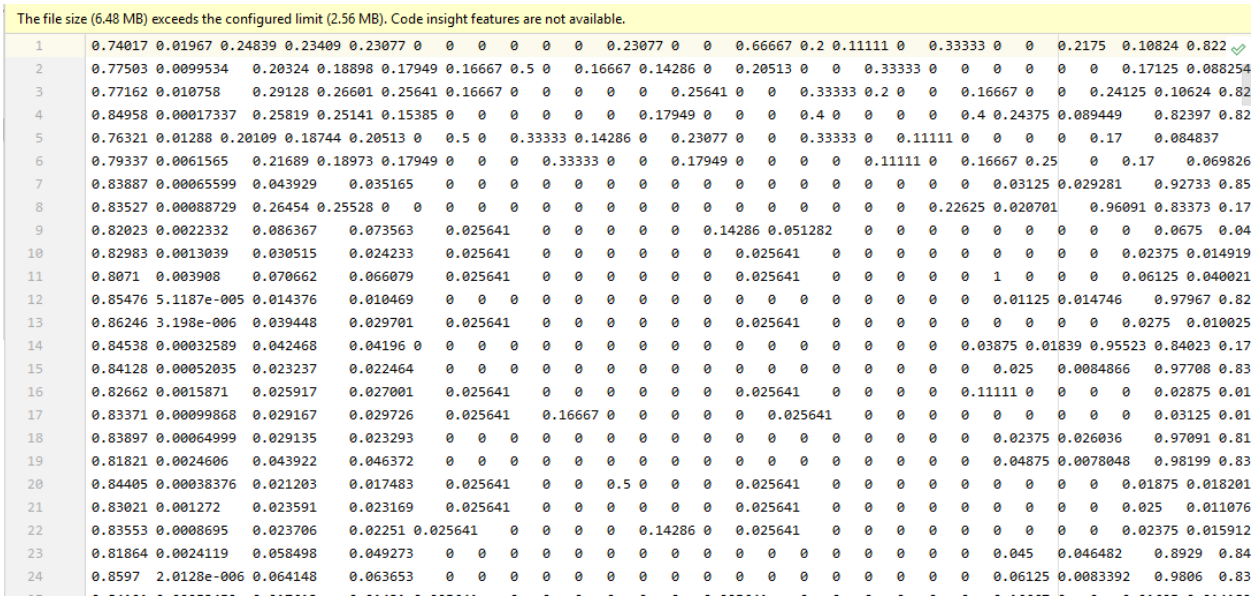

Figure 2. view of the drug features file

References:

[١] 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.

[٢] 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.

[٣] 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, pp. 267-276, 2016.

[٤] A. Ezzat, M. Wu, X. Li, and C.-K. Kwoh, "Computational prediction of drug-target interactions via ensemble learning," in *Computational methods for drug repurposing*: Springer ,٢٠١٩ ,pp. 239-254.

[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. 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.

[٧] 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.