Sensitivity Analysis of a BERT-based scholarly recommendation system

Jie Zhu, Hulin Wu, Ashraf Yaseen

The University of Texas Health Science Center at Houston, Houston, TX, USA
Jie.Zhu@uth.tmc.edu, Hulin.Wu@uth.tmc.edu, Ashraf.Yaseen@uth.tmc.edu

Abstract

With the exponential growth of publicly available datasets, a scholarly recommendation system of datasets would be an essential tool in the field of information filtering. Recommending datasets to users can be formulated as a classification problem where deep learning models can be carefully trained. In such case, when preparing training data for the learning models, one needs to consider different ratios of false and true pairs. Therefore, a sensitivity analysis is necessary. In this work, we conduct a sensitivity analysis using different class ratios on a deep learning model (BERT) for recommending datasets. We found out that our BERT-based recommender model is relatively robust using recommender metrics such as Mean Reciprocal Rank (MRR)@k, Recall@k, etc., except for the extreme class imbalance case (1:5000). Therefore, we conclude that moderate ratio of random negative sampling scheme, (in our case 1:10) is reasonable, sufficient and time efficient in the recommendation system training.

Introduction

With the ever-growing public information online, recommendation systems have proven to be an effective strategy to deal with information overload. In fact, recommenders are thriving in this era of Big Data with wide commercial applications in recommending products, music, movies, books, news articles, and many more.

Applications of recommendation systems are currently expanding beyond the commercial area to include scholarly activities (Bollacker, Lawrence, and Giles 1998; Yoneya and Mamitsuka 2007; Lin and Wilbur 2007; Collins and Beel 2019; Hassan et al. 2019; Achakulvisut et al. 2016). However, the majority of literature belongs to the category of data linking either in the web-service or for academic references (Ellefi et al. 2016; Lopes et al. 2014; Ghavimi et al. 2016; Boland et al. 2012; Piwowar and Chapman 2008). For dataset recommendations, Alghofaily and Ding (2019) used dataset features with meta-learners and factor analysis for the dataset recommendations. Altaf et al. (2019) represented research papers and datasets in the two-layer network using heterogeneous variational graph autoencoder for the recommendation of data. Previous work in scholarly recommendation systems conducted by our team members includes (Patra, Roberts, and Wu 2020; Patra et al. 2020; Zhu, Patra, and Yaseen 2021). Especially, Patra, Roberts, and Wu (2020) experimented with information retrieval paradigms (BM25, TF-IDF, etc.) for Gene Expression Omnibus (GEO) data recommendation to researchers.

There are many public datasets available on the internet which might be useful to researchers for further exploration. A dataset recommendation system for papers is an important and very helpful tool in the field of information filtering. It can enhance the dataset's re-usability and data sharing. Recommending publicly available datasets, for example COVID-19, to make sure public health researchers can promptly access the data is important for the scientific community. This will save time, increase knowledge and help derive actionable measures to tackle population health problems.

When using task-oriented deep learning models such as BERT, we have the option to formulated recommendation as a classification problem (as compared to simple informational retrieval as in our previous work (Patra, Roberts, and Wu 2020). In such case, due to the large number of 'users' (researchers), 'products' (the datasets) and existing associations between them (the datasets citations in the papers), the recommendation problem suffers from data sparsity (Sharma and Gera 2013; Popescul et al. 2013) considering existing user-product associations vs. all available data pairs.

Data imbalance issues were mostly discussed in terms of data sampling techniques and cost-sensitive machine learning algorithms (Sun et al. 2007; Hulse, Khoshgoftarr, and Napolitano 2007; Leevy et al. 2018; Costa et al., 2020; Lin, Chen, and Qi 2019). For experimental evaluations of class imbalance on machine learning model performance, Thabtah et al.(2020) studied the impact of varying class imbalance ratios on the naïve Bayes classifier. (Batista, Silva, and Prati (2012) used artificially modified class distributions using different sampling technique to study the performance loss compared to balanced data (1:1) AUC on a variety of classifiers. Su and Hsiao (2007) compared the classification performances of decision trees, SVMs etc. with different imbalance levels, sample sizes and classification complexities to evaluate the model robustness on ‘g-means’ (Woodall et al. 2003) and the author-proposed measure. For deep learning, however, sensitivity analysis of data imbalance has been concentrated on Convolutional Neural Networks architectures (Buda, Maki, and Mazurowski 2018; Johnson and Khoshgoftaar 2019). Despite a growing demand on big data analytics, there is still limited research that properly evaluates the effect of data imbalance on deep learning models, and therefore the consequences of deep learning on imbalanced data is still largely understudied (Johnson and Khoshgoftaar 2019). To this end, we carried out carefully designed experiments to evaluate how class imbalance will affect BERT-based deep learning recommender.

The work described in this manuscript has the following contributions: 1) addresses the recommendation problem in the context of classification (instead of simple cosine similarity as we addressed before), 2) the data sources encompass wider varieties in the domain of immunology, genetic arrays and sequences, 3) and most importantly, performs sensitivity analysis on how training data class imbalance affects model performances, and therefore provides a practical guide on how we can best train models to efficiently learn. As part of our ongoing effort on the development of Virtual Research Assistant (VRA), a web-based recommender application for population health professionals at http://genestudy.org/recommends/#/dataset, we believe that the work presented here is especially important for further development and improvement of our research. In this paper, we artificially modified the training dataset class imbalance using different proportions of the false association pairs, and we examined how the imbalance in the training data affected the model recommendation performances. In the following sections we describe the data sources, methods and evaluation metrics. After that, we present our results and conclusions.

Data

The experiments were performed with public datasets from two domains: one from immunology: including Immport[1], Immune space[2] and ITN trial share (ITN)[3]; the second from genetic array and sequence: Gene Expression Omnibus (GEO)[4], and Sequence Read Archive (SRA) studies[5]. Associated PubMed articles are crawled through Medline[6]. Data sources are provided in Table 1.

We utilized the citations of PubMed papers within the datasets metadata as our ground truth associations. Then we reversed the citation direction and aggregated all the datasets associated with each publication for recommendation purposes, so that an entry in the final citation is in the format of '20398357' (pmid): ['DRP000002', 'DRP001803'] (data ids). We kept track of all true associations to create different numbers of false pairs using this information. For example, we have two true associations of this example: ('20398357', 'DRP000002'), ('20398357', 'DRP001803'). Additionally, we removed any publications with too many associated datasets (>10 datasets, at 99% of distributions). The rest of the data sources had a maximum of 6 datasets associated. Basic summary statistics of the sources are provided in Table 2.

Data source: Immunology data	
Immport	Immport is a data repository for public data sharing of immunological studies.
Immune space	Immune space allows users to easily explore and analyze datasets from the Human Immunology Project Consortium (HIPC)
ITN	ITN is a clinical trials research portal of the Immune Tolerance Network (ITN) designed to promote transparency, reproducibility and scientific collaboration. It shares information about the ITN’s clinical studies and specimen biorepositories, as well as data and analysis code.
Data source: Genetic array and sequence	
GEO	GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community.
SRA studies	Sequence Read Archive (SRA) studies is the largest publicly available repository of high throughput sequencing data on National Center for Biotechnology Information (NCBI). It has data from all branches of life as well as metagenomic and environmental surveys.

Table 1. Description of data sources

Data sources	Immunology data			Genetic arrays and sequence	
	immport	Immune space	ITN	GEO	SRA studies
Total datasets	354	35	38	96,457	28,710
Total publications associated	259	43	69	73,248	
Total true pairs	354	53	69	103,018	23,558
Max # of associated datasets	6	3	1	10	28,710
Additional notes				Removed >10	5

Table 2. Data sources and basic summary statistics

[1] https://www.immport.org/
[2] https://www.immunespace.org/
[3] https://www.itntrialshare.org/
[4] https://www.ncbi.nlm.nih.gov/geo
[5] https://www.ncbi.nlm.nih.gov/sra
[6] https://www.nlm.nih.gov/medline/index.html

Methods

a. System architecture

We formulated the recommendation as a classification problem, where we trained a BERT model and predicted the probability of a dataset and a PubMed publication being a true match. In our experiments, datasets were represented by their titles and summaries, while publications were represented by their titles and abstracts (extreme long texts were truncated at the end equally from both dataset summaries and publication abstracts). For those predicted as 'matching' pairs, we aggregated the dataset results using predicted probability as the ranking score (descending order). Figure 1 shows our recommender's architecture.

We used the base-BERT (Devlin et al. 2019) for the task. The standardized wordPiece tokenization (Wu et al. 2016) was applied before we fed the pair of PubMed article and dataset in tokens, position ids and segment ids as the inputs to the model. The detailed BERT usage is enlarged in Figure 2.

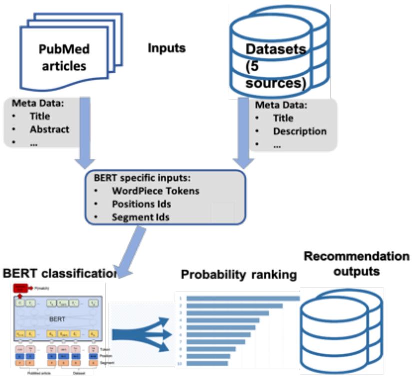

Figure 1. Recommendation system architecture

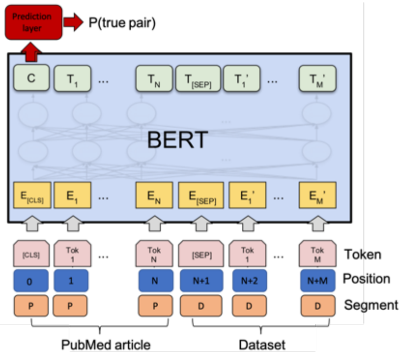

Figure 2. BERT model usage

During the fine-tuning process, the base-BERT architecture with self-attention was kept; its default pretrained parameters were initialized and updated. The pooled output of last hidden layer was used to feed into the classifier. The classifier consists of a dropout layer of $p = 0.1$ and a linear layer as detailed in Table 3. 2-20 epochs were used during preliminary test runs. Since the model converged quickly without considerable performance variations after 4 epochs, 4 epochs were chosen for the final experimentations (results presented in this paper). The best model parameters were saved based on best validation performance. Cross entropy loss together with Adam optimizer with a learning rate of $2 \times 10^{-5}$and an epsilon of 10^{-8} and a linear scheduler were used. Hyperparameters choices of learning rate and epsilon were chosen with Bayesian optimization package Ax[7]. The choices of relatively small number of epochs (fast learning) as well as tuned hyperparameters were also consistent with BERT authors' suggestions (Devlin et al. 2019) on fine-tuning processes.

Architecture	**Sub-architecture**	**Input size**	**Output size**	**Number of parameters**
base-BERT		512[8]	768	110m
Classifier	Dropout	768	768	0
	Linear	768	2	1538

Table 3. Dataset recommender: BERT component architecture

b. Experimental setup

We ran three big sets of sensitivity analysis in consideration of heterogeneity of the datasets. First set of experiments were on immunology datasets only: Immport, Immune space and ITN; the second set on genetic array and sequence only: GEO and SRA studies; the third one on composite data comprising all five sources from the two domains.

The three sets followed the same procedures, namely: 1) set aside the true pairs of association for training, 2) create the false pairs of association for training at different ratios; 3) split for train, validation and test; 4) train and evaluate test performance.

We utilized all available 476 true associations within the immunology datasets, in addition, we randomly sampled 952 true associations from GEO as well as 952 true associations from SRA studies in order not to crowd the composite training data with genetic data sources only. Secondly, false pairs of association were created for each dataset separately for the maximized false vs. true ratio of 50:1. For GEO and SRA studies, we also created a false vs. true ratio of 5000:1 for extreme case experimentations. Then the data were split to 7:1:2 for training, validation and test on unique publication (to preserve all associated datasets for evaluation purposes). For the test data, the false pairs were resampled again at 1:1 ratio and kept aside for use during each experimentation. During each run of experiments, different number of false pairs were randomly sampled to make our desired true vs false ratios for training and validation. The ratios of true vs. false that we experimented with were:

[7] https://ax.dev/

[8] Here 512 refers to the number of (padded) input sequence (sentence pair)

1:0.1(10:1), 1:0.5(2:1), 1:1, 1:2, 1:10, 1:20, 1:50 and one extreme case:1:5000[9].

For each set of the experiments, the metrics was evaluated on the same test data corresponding to that set. Additionally, we were also interested in the computational complexity of the models with the different training ratios, so we recorded the training time for all experiments as well.

The code for experiments is published at https://github.com/ashraf-yaseen/VRA under *dataset_rec/*.

c. Evaluation metrics

For predicted matching pairs, we aggregated the recommendation results at each unique publication level and used the following metrics to evaluate the recommendations. In order to better describe Recall@k and Precision@k, we supplemented the confusion matrix as below.

	Recommended	**Not recommended**	**Total**
Relevant	True positive (TP)	False negative (FN)	Total relevant
Not Relevant	False positive (FP)	True negative (TN)	Total Not relevant
Total	Total recommended	Total Not recommended	Overall Total

Table 4. Confusion metrics for recommender systems

Mean reciprocal rank (MRR)@k: The Reciprocal Rank (RR) measures the reciprocal of the rank at which the first relevant document was retrieved. RR is 1 if the relevant document was retrieved at rank 1, RR is 0.5 if document is retrieved at rank 2, and so on. When we average the top k retrieved items across the queries Q, the measure is called the MRR@k. In our case, we chose k=10.

$$MRR@k = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

Recall@k: At the k-th retrieved item, this metric measures the proportion of relevant items that are retrieved. We evaluated both recall@1 and recall@10.

$$recall@k = \frac{TP@k}{TP@k + FN@k}$$

Precision@k: At the k-th retrieved item, this metric measures the proportion of the retrieved items that are relevant. In our case, we are interested in precision@1.

$$precision@k = \frac{TP@k}{TP@k + FP@k}$$

Results and Discussion

Using the setups explained in Methods, we presented the results below. We additionally provided performance percentage loss (Batista, Silva, and Prati 2012) in the bracket, where the percentage loss is calculated as:

$$percentage\ loss = \frac{performance - best\ performance}{best\ performance} \times 100\%$$

Experiment group (True: False ratio)	**MRR@10 (percentage loss)**	**Recall@1 (percentage loss)**	**Recall@10 (percentage loss)**	**Precision@1 (percentage loss)**
1:0.1 (10:1)	0.703 (-24.8%)	0.579 (-30.8%)	0.677 (-19.9%)	0.639 (-31.1%)
1:0.5 (2:1)	0.806 (-13.8%)	0.695 (-17.0%)	0.756 (-10.5%)	0.753 (-18.8%)
1:1	0.777 (-16.9%)	0.658 (-21.4%)	0.746 (-11.7%)	0.703 (-24.2%)
1:2	0.785 (-16.0%)	0.657 (-21.5%)	0.749 (-11.3%)	0.718 (-22.5%)
1:10	0.856 (-8.4%)	0.731 (-12.7%)	0.783 (-7.3%)	0.817 (-11.9%)
1:20	0.867 (-7.2%)	0.736 (-12.0%)	0.836 (-1.1%)	0.782 (-15.6%)
1:50	**0.935 (0%)**	**0.837 (0%)**	**0.845 (0%)**	**0.927 (0%)**

Table 5. Sensitivity analysis with different class ratios on immunology datasets, with the best performance highlighted in bold and corresponding percentage loss in parenthesis

Experiment group (True: False ratio)	**MRR@10 (percentage loss)**	**Recall@1 (percentage loss)**	**Recall@10 (percentage loss)**	**Precision@1 (percentage loss)**
1:0.1 (10:1)	0.784 (-9.8%)	0.616 (-10.9%)	0.617 (-11.4%)	0.783 (-9.4%)
1:0.5 (2:1)	0.849 (-2.3%)	0.667 (-3.5%)	0.670 (-3.8%)	0.847 (-2.0%)
1:1	0.856 (-1.5%)	0.677 (-2.0%)	0.678 (-2.6%)	0.855 (-1.0%)
1:2	0.864 (-0.6%)	0.683 (-1.2%)	0.684 (-1.7%)	**0.864 (0%)**
1:10	0.865 (-0.5%)	0.689 (-0.3%)	0.692 (-0.6%)	0.862 (-0.2%)
1:20	0.865 (-0.5%)	0.688 (-0.4%)	0.692 (-0.6%)	0.861 (-0.3%)
1:50	**0.869 (0%)**	**0.691 (0%)**	**0.696 (0%)**	**0.864 (0%)**

Table 6. Sensitivity analysis with different class ratios on genetic array and sequence datasets, with the best performance highlighted in bold and corresponding percentage loss in parenthesis

[9] 1:5000 created for GEO and SRA, while the immunology data kept at 1:50 due to too limited records to create extreme false ratios. Thus we only reported the 1:5000 results on the composite datasets.

Experiment group (True: False ratio)	MRR@10 (percentage loss)	Recall@1 (percentage loss)	Recall@10 (percentage loss)	Precision@1 (percentage loss)
1:0.1 (10:1)	0.776 (-10.5%)	0.628 (-11.9%)	0.632 (-12.6%)	0.773 (-10.1%)
1:0.5 (2:1)	0.816 (-5.9%)	0.660 (-7.4%)	0.664 (-8.2%)	0.813 (-5.5%)
1:1	0.834 (-3.8%)	0.678 (-4.9%)	0.686 (-5.1%)	0.826 (-4.0%)
1:2	0.839 (-3.2%)	0.681 (-4.5%)	0.690 (-4.6%)	0.830 (-3.5%)
1:10	0.864 (-0.3%)	0.702 (-1.5%)	0.709 (-1.9%)	**0.860 (0%)**
1:20	0.864 (-0.3%)	0.709 (-0.6%)	0.714 (-1.2%)	0.859 (-0.1%)
1:50	**0.867 (0%)**	**0.713 (0%)**	**0.723 (0%)**	0.857 (-0.3%)
1:5000*	0.(-100%)	0. (-100%)	0. (-100%)	0. (-100%)

Table 7. Sensitivity analysis with different class ratios on composite datasets, with the best performance highlighted in bold and corresponding percentage loss in parenthesis

For three sets of experiments, the general trend was that with the inclusion of more false pairs, the recommendation results improved in terms of nearly all metrics, except for precision@1 where it reached the maximum at ratio 1:2 for genetic arrays and sequence and at ratio 1:10 for composite datasets. The extreme training ratio 1:5000, not surprisingly, produced the worst results.

Specifically, for the immunology data, however, where the data is relatively scarce compared to more abundant genetic arrays and sequence as well as composite sources, the increase of performance was more prominent than the latter two cases. Once again, this confirms the need of sufficient data for training deep learning based models. But overall, we observed that variations of metrics were not significant within a range of moderate class ratios (1:0.5-1:50).

All experiments were performed on an HPE server with 36 cores-72 threads, 768GB memory, and NVIDIA V100 16GB GPU. We configured the models to utilize the GPU in all of our experiments. The total training time for 4 epochs for different ratios for the three sets are shown in Table 8. Training time increased almost exponentially with the increasing negative class ratios, especially after 1:10.

Experiment group (True: False ratio)	Immunology (min)	Genetic arrays & sequence (h)	Composite (h)
1:0.1 (10:1)	1.10	0.08	0.12
1:0.5 (2:1)	1.67	0.10	0.15
1:1	2.10	0.13	0.20
1:2	3.03	0.25	0.27
1:10	11.27	0.73	1.02
1:20	21.38	1.50	1.90
1:50	55.15	3.50	4.50
1:5000*	NA	NA	211.87

Table 8. Training time with different class ratios on all experiments

Overall, the training time increased exponentially with the increasing negative class ratios, especially after 1:10. Taking into consideration of recommender metrics in Table 5-7, the model was relatively robust in a range of ratios. Between ratio 1:10 to 1:50, the metrics stayed relatively optimal, especially for genetic array and composite data where the maximum percentage loss is only -1.9%. Considering the time vs. performance trade-off, we concluded that a relative balanced ratio was sufficient for both model performance and training efficiency (and in our particular case, 1:10 was optimal).

Conclusion

In this work, we performed a sensitivity analysis on training class imbalance ratios in the dataset recommender system. We carefully designed the analysis to provide practical guidance on how training data imbalance affected the deep-learning based recommender performances. We found out that even though the recommender results improved moderately with the incorporation of more false pairs in our experimental setups (with the exception of the extreme ratio), the performance gain was not cost-efficient considering the time increase in the training. Therefore, relatively balanced training (in our case 1:10), was optimal in terms of both model performance and training efficiency.

References

Achakulvisut, T.; Acuna, D. E.; Ruangrong, T.; and Kording, K. 2016. Science concierge: A fast content-based recommendation system for scientific publications. *PloS one* 11(7):e0158423.

Alghofaily, B. I., and Ding, C. 2019. Data mining service recommendation based on dataset features. *Service Oriented Computing and Applications* 13(3):261–277.

Alharthi, H.; Inkpen, D.; and Szpakowicz, S. 2018. A survey of book recommender systems. *Journal of Intelligent Information Systems* 51(1):139–160.

Altaf, B.; Akujuobi, U.; Yu, L.; and Zhang, X. 2019. Dataset recommendation via variational graph autoencoder. In *2019 IEEE International Conference on Data Mining (ICDM)*, 11–20. IEEE.

Alves, G. E. D. A. P.; Silva, D. F.; Prati, R. C.; et al. 2012. An experimental design to evaluate class imbalance treatment methods. In 2*012 11th International Conference on Machine Learning and Applications*, volume 2, 95–101. IEEE.

Boland, K.; Ritze, D.; Eckert, K.; and Mathiak, B. 2012. Identifying references to datasets in publications. In *International Conference on Theory and Practice of Digital Libraries*, 150–161. Springer.

Bollacker, K. D.; Lawrence, S.; and Giles, C. L. 1998. Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the second international conference on Autonomous agents*, 116–123.

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106:249–259.

Collins, A., and Beel, J. 2019. Document embeddings vs. keyphrases vs. terms for recommender systems: a large-scale online evaluation. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 130–133. IEEE.

Costa, A. J.; Santos, M. S.; Soares, C.; and Abreu, P. H. 2020. Analysis of imbalance strategies recommendation using a meta-learning approach.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv*:1810.04805.

Ellefi, M. B.; Bellahsene, Z.; Dietze, S.; and Todorov, K. 2016. Dataset recommendation for data linking: An intensional approach. In *European Semantic Web Conference*, 36–51. Springer.

Ghavimi, B.; Mayr, P.; Vahdati, S.; and Lange, C. 2016. Identifying and improving dataset references in social sciences full texts. *arXiv preprint arXiv*:1603.01774.

Hassan, H. A. M.; Sansonetti, G.; Gasparetti, F.; Micarelli, A.; and Beel, J. 2019. Bert, elmo, use and infersent sentence encoders: The panacea for research-paper recommendation? In *RecSys (Late-Breaking Results)*, 6–10.

Johnson, J. M., and Khoshgoftaar, T. M. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6(1):1–54.

Leevy, J. L.; Khoshgoftaar, T. M.; Bauder, R. A.; and Seliya, N. 2018. A survey on addressing high-class imbalance in big data. *Journal of Big Data* 5(1):1–30.

Lin, J., and Wilbur, W. J. 2007. Pubmed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics* 8(1):1–14.

Lin, E.; Chen, Q.; and Qi, X. 2020. Deep reinforcement learning for imbalanced classification. *Applied Intelligence* 50(8):2488–2502.

Lopes, G. R.; Leme, L. A. P. P.; Nunes, B. P.; Casanova, M. A.; and Dietze, S. 2014. Two approaches to the dataset interlinking recommendation problem. In *International Conference on Web Information Systems Engineering*, 324–339. Springer.

Patra, B. G.; Maroufy, V.; Soltanalizadeh, B.; Deng, N.; Zheng, W. J.; Roberts, K.; and Wu, H. 2020a. A content- based literature recommendation system for datasets to improve data reusability–a case study on gene expression omnibus (geo) datasets. *Journal of Biomedical Informatics* 104:103399.

Patra, B. G.; Soltanalizadeh, B.; Deng, N.; Wu, L.; Maroufy, V.; Wu, C.; Zheng, W. J.; Roberts, K.; Wu, H.; and Yaseen, A. 2020b. An informatics research platform to make public gene expression time-course datasets reusable for more scientific discoveries. *Database* 2020.

Piwowar, H., and Chapman, W. 2008. Identifying data sharing in biomedical literature. *Nature Precedings* 1–1.

Popescul, A.; Ungar, L. H.; Pennock, D. M.; and Lawrence, S. 2013. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. *arXiv preprint arXiv*:1301.2303.

Sharma, L., and Gera, A. 2013. A survey of recommendation system: Research challenges. *International Journal of Engineering Trends and Technology (IJETT)* 4(5):1989– 1992.

Su, C.-T., and Hsiao, Y.-H. 2007. An evaluation of the robustness of mts for imbalanced data. *IEEE Transactions on knowledge and data engineering* 19(10):1321–1332.

Sun, Y.; Kamel, M. S.; Wong, A. K.; and Wang, Y. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern recognition* 40(12):3358–3378.

Thabtah, F.; Hammoud, S.; Kamalov, F.; and Gonsalves, A. 2020. Data imbalance in classification: Experimental evaluation. *Information Sciences* 513:429–441.

Van Hulse, J.; Khoshgoftaar, T. M.; and Napolitano, A. 2007. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, 935–942.

Woodall, W. H.; Koudelik, R.; Tsui, K.-L.; Kim, S. B.; Stoumbos, Z. G.; and Carvounis, C. P. 2003. A review and analysis of the mahalanobis—taguchi system. *Technometrics* 45(1):1–15.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv*:1609.08144.

Yoneya, T., and Mamitsuka, H. 2007. Pure: a pubmed article recommendation system based on content-based filtering. *Genome informatics* 18:267–276.

Zhu, J.; Patra, B. G.; and Yaseen, A. 2021. Recommender system of scholarly papers using public datasets. In *AMIA Annual Symposium Proceedings*, volume 2021, 672. American Medical Informatics Association.