International Journal of Engineering and Information Systems (IJEAIS)
ISSN: 2643-640X
Vol. 4 Issue 10, October - 2020, Pages: 17-22

Books' Rating Prediction Using Just Neural Network

Alaa Mazen Maghari, Iman Ali Al-Najjar, Said Jamil Al-laqtah, Samy S. Abu-Naser

Department of Information Technology,
Faculty of Engineering and Information Technology,
Al-Azhar University, Gaza, Palestine

Abstract: *The aim behind analyzing the Goodreads dataset is to get a fair idea about the relationships between the multiple attributes a book might have, such as: the aggregate rating of each book, the trend of the authors over the years and books with numerous languages. With over a hundred thousand ratings, there are books which just tend to become popular as each day seems to pass. We proposed an Artificial Neural Network (ANN) model for predicting the overall rating of books. The prediction is based on these features (bookID, title, authors, isbn, language_code, isbn13, # num_pages, ratings_count, text_reviews_count), which were used as input variables and (average_rating) as output variable for our ANN model. Our model were created, trained, and validated using data set in JNN environment, which its title is "Goodreads-books". Model evaluation showed that the ANN model is able to predict correctly 99.78% of the validation samples.*

Keywords: *Predictive Analysis, Artificial Neural Networks, Books Rates, Goodreads.*

Introduction

In this study, we will analyze the dataset which contains various information of the books on the website of the world's largest book archive and book proposal site GoodReads. This dataset also includes information on the names, writers and spelling languages of books, as well as the rating and total score based on the votes given by various users. Artificial neural networks (ANNs) will be used for the analysis. Artificial neural networks are like biological neural networks and offer a technique, which solves the problem of prediction [3]. Neural networks contain input, hidden and output layers. Hidden layers convert the input into usable thing to the output layer [5]. The ANN Model goes through training and validation on a dataset. Training in which that the network is trained is done on a dataset. Then a configuration is done to the weights of the connections between neurons. Validation in which that the network is validated to determine the prediction of a new dataset [6]. In this study, we used about 33% of the dataset instances for network validation, the remaining 67% for training. ANN Architecture is shown in figure 1.

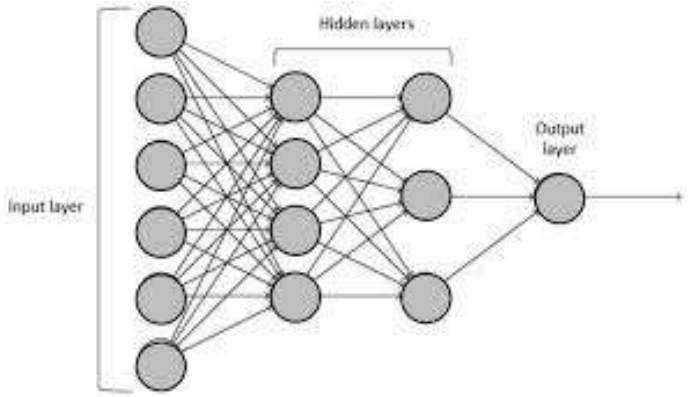

Figure 1: ANN Architecture

1. Literature Review

Artificial Neural Networks have been used many fields. In Education such as: Predicting Student Performance in the Faculty of Engineering and Information Technology using ANN[5], Prediction of the Academic Warning of Students in the Faculty of Engineering and Information Technology in Al-Azhar University-Gaza using ANN[5], Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach[6].

In the field of Health such as: Parkinson's Disease Prediction [7], Classification Prediction of SBRCTs Cancers Using ANN [7], Predicting Medical Expenses Using ANN[8], Predicting Antibiotic Susceptibility Using Artificial Neural Network[8], Predicting Liver Patients using Artificial Neural Network[7], Blood Donation Prediction using Artificial Neural Network[9], Predicting DNA Lung Cancer using Artificial Neural Network[10], Diagnosis of Hepatitis Virus Using Artificial Neural Network[10], COVID-19 Detection using Artificial Intelligence[11].

In the field of Agriculture: Plant Seedlings Classification Using Deep Learning [12], Prediction of Whether Mushroom is Edible or Poisonous Using Back-propagation Neural Network[15], Analyzing Types of Cherry Using Deep Learning[21], Banana Classification Using Deep Learning[13], Mango Classification Using Deep Learning[14], Type of Grapefruit Classification Using Deep Learning[7], Grape Type Classification Using Deep Learning[3], Classifying Nuts Types Using Convolutional Neural Network[2], Potato Classification Using Deep Learning[3], Age and Gender Prediction and Validation Through Single User Images Using CNN[5].

In other fields such as : Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods [14], Predicting Overall Car Performance Using Artificial Neural Network [8], Glass Classification Using Artificial Neural Network [9], Tic-Tac-Toe Learning Using Artificial Neural Networks[14], Energy Efficiency Predicting using Artificial Neural Network[15], Predicting Titanic Survivors using Artificial Neural Network[14], Classification of Software Risks with Discriminant Analysis Techniques in Software planning Development Process[13], Handwritten Signature Verification using Deep Learning[12], Email Classification Using Artificial Neural Network[14], Predicting Temperature and Humidity in the Surrounding Environment Using Artificial Neural Network[12], English Alphabet Prediction Using Artificial Neural Networks[9].

2. Methodology

We downloaded a data set from *kaggle* that contains books information from *goodreads* application/website. This dataset created by the user *Soumik* [19]. We did some preprocessing on the data, and then we trained our ANN model and validated it.

3. Original Dataset Description

Table 1: Original Dataset Description

#	Attribute	Description	Type
1.	bookID	A unique Identification number for each book.	Integer
2.	title	The name under which the book was published.	String
3.	authors	Names of the authors of the book. Multiple authors are delimited with -.	String
4.	average_rating	The average rating of the book received in total.	Real
5.	isbn	Another unique number to identify the book, the International Standard Book Number.	Long
6.	language_code	Helps understand what is the primary language of the book. For instance, eng is standard for English.	String
7.	isbn13	A 13-digit ISBN to identify the book, instead of the standard 11-digit ISBN.	Long
8.	# num_pages	Number of pages the book contains.	Integer
9.	ratings_count	Total number of ratings the book received.	Integer
10.	text_reviews_count	Total number of written text reviews the book received.	Integer

3.1 Dataset Preprocessing

We wanted to use this dataset to build an ANN model to predict the overall rating of the books (attribute number 4).
The first thing we had to do, is choose a suitable factors for this prediction, and delete the unnecessary ones, we chose these factors to be our input to the predictive model: #num_pages, rating_count, text_reviews_count, language_code.
Moreover, the dataset contain 11128 instances. After preprocessing it becomes 11122 which is a large a number to a neural network to deal with, so, we divided these samples to 7451 training instances, and 3670 validation instances.
In addition, because of the integer numbers of the inputs are too large comparing with the real rate values, we did a normalization to them so all the data are real.

Normalization formula was:

$$\text{Normalized value (xi)} = \frac{\text{Xi}-\text{Xmin}}{(\text{Xmax}-\text{Xmin})} .$$

While checking the instances, it has been noticed that there are a conflict between some instances; which means, there at least two books with the same input values but different rates, we excluded for the secondary ones. Moreover, there were validation instances that are out of range, we converted them to training. Now, the dataset is ready for training and validation.

3.2 Our ANN Model

The resulted predictive ANN model is shown in Figure 2 and Figure 6.

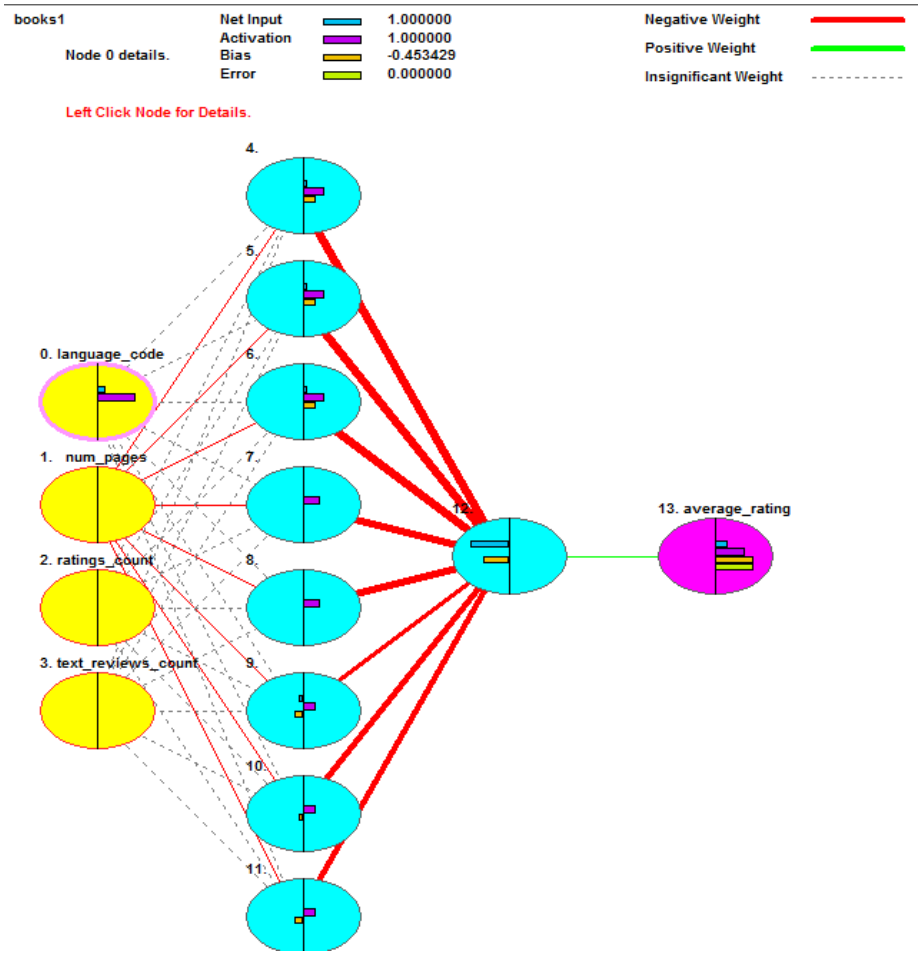

Figure 2: Our ANN Model

3.3 Validation

Our ANN model was able to predict the books' overall rate with 99.78% accuracy, with about 0.005 errors as seen in figure (3). Furthermore, The Model showed that the most effective factor in a book's rate is the rating_count. More details are shown in figure (4).

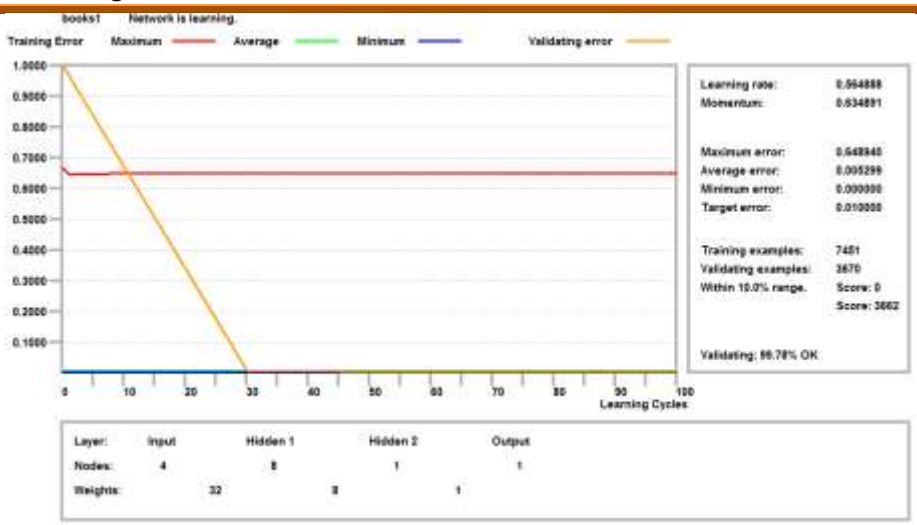

Figure 3: Validation and Errors

books1 101 cycles. Target error 0.0100 Average training error 0.005299
The first 4 of 4 Inputs in descending order.

Column	Input Name	Importance	Relative Importance
1	num_pages	0.3347	
3	text_reviews_count	0.0478	
2	ratings_count	0.0432	
0	language_code	0.0099	

Figure 4: Attributes Importance

books1

	language_c+	num_pages	ratings_co+	text_revie+	average_ra+
#0	0.0000	0.0374	0.0011	0.0054	0.8420
#1	0.0000	0.0414	0.0014	0.0049	0.7960
#2	0.0000	0.0243	0.0009	0.0037	0.8100
#3	0.0000	0.0389	0.0005	0.0019	0.7880
#4	0.0000	0.0318	0.0002	0.0011	0.8340
#5	0.0000	0.0590	0.0010	0.0055	0.7320
#6	0.0000	0.0354	0.0011	0.0016	0.8340
#7	0.0476	0.0991	0.4556	0.2927	0.9140
#8	0.0476	0.1323	0.4683	0.3100	0.8980
#9	0.0476	0.0535	0.0014	0.0026	0.8840
#10	0.0476	0.0661	0.5089	0.3853	0.9120
#11	0.0476	0.4091	0.0090	0.0017	0.9560
#12	0.0476	0.5082	0.0061	0.0086	0.9460
#13	0.0476	0.1239	0.0008	0.0027	0.8760
#14	0.0476	0.1239	0.0543	0.0433	0.8760
#15	0.0476	0.0327	0.0011	0.0049	0.8440
#16	0.0476	0.0009	0.0003	0.0027	0.8440
#17	0.0476	0.1239	0.0006	0.0021	0.8760
#18	0.0476	0.0827	0.0541	0.0997	0.8420
#19	0.0476	0.0084	0.0016	0.0053	0.6880
#20	0.0476	0.0389	0.0005	0.0014	0.7740
#21	0.0476	0.0509	0.0158	0.0450	0.8140
#22	0.0476	0.0462	0.0107	0.0235	0.7800
#23	0.0476	0.0455	0.0099	0.0239	0.7660
#24	0.0476	0.0386	0.0106	0.0237	0.7720
#25	0.0476	0.0493	0.0175	0.0350	0.7820
#26	0.0476	0.0411	0.0062	0.0221	0.7860
#27	0.0476	0.2628	0.0220	0.0164	0.9180
#28	0.0476	0.1800	0.0004	0.0010	0.9000
#29	0.0476	0.0605	0.4630	0.1450	0.8720
#30	0.0476	0.0332	0.0043	0.0005	0.9060

Figure 5: imported pre-processed Dataset

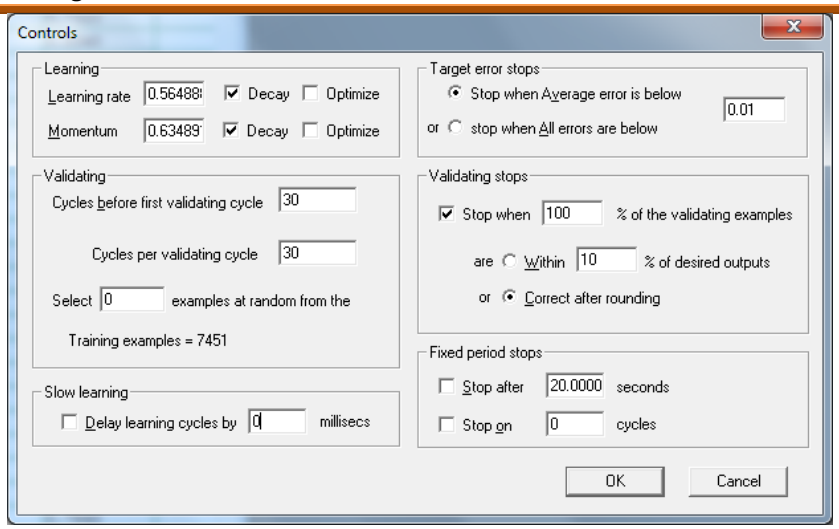

Figure 6: Parameter values of the ANN Model

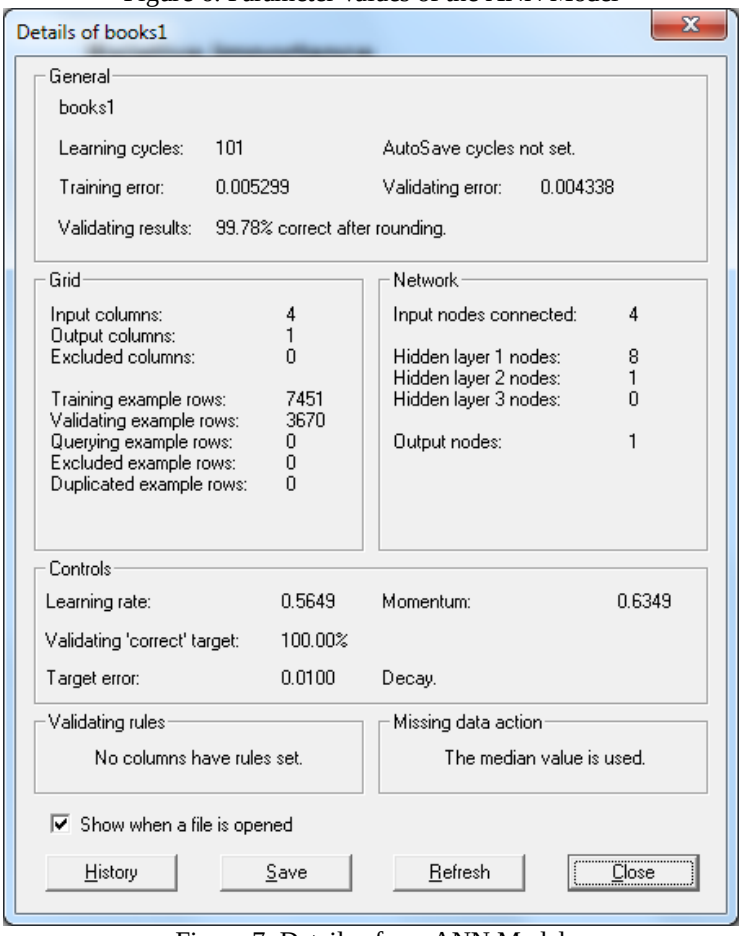

Figure 7: Details of our ANN Model

4. Conclusion

A predictive Artificial Neural Network Model for predicting books' rating was developed. The Model trained and validated using a dataset from the goodreads application/ website. We did some preprocessing on the dataset to make it suitable as input to our ANN model. Validation showed that the model is 99.78% accurate.

References

1. Dormehl, L. (2019). "Digital Trends." https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/.

2. Livieris, K. et al. (2012). "Predicting students' performance using artificial neural networks," in 8th PanHellenic Conference with International Participation Information and Communication Technologies in Education, Volos, Greece, 2012.
3. https://www.kaggle.com/jealousleopard/goodreadsbooks
4. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications. 2009 Mar 1;36(2):3240-7.
5. Yeh WC, Chang WW, Chung YY. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. Expert Systems with Applications. 2009 May 1;36(4):8204-11.
6. Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications. 2011 Aug 1;38(8):9573-9.
7. Chaurasia V., Pal., S, Tiwari., BB.: Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, Vol. 12(2), pp. 119–126. DOI: http://dx.doi.org/10.1177/1748301818756225. (2018).
8. Verma, D., Mishra., N.:Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques. In Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS), pp 533-538, (2017).
9. Quinlan, "Improved use of continuous attributes in C4. 5," Journal of artificial intelligence research, vol. 4, pp. 77-90, 1996.
10. Pena-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis," Artificial intelligence in medicine, vol. 17, pp. 131-155, 1999.
11. Nauck and R. Kruse, "Obtaining interpretable fuzzy classification rules from medical data," Artificial intelligence in medicine, vol. 16, pp. 149-169, 1999.
12. Setiono, "Generating concise and accurate classification rules for breast cancer diagnosis," Artificial Intelligence in medicine, vol. 18, pp. 205-219, 2000.
13. Albrecht, G. Lappas, S. A. Vinterbo, C. Wong, and L. Ohno-Machado, "Two applications of the LSA machine," in Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on, 2002, pp. 184-189.
14. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers," Pattern Recognition Letters, vol. 24, pp. 2195-2207, 2003.
15. Kiyan and T. Yildirim, "Breast cancer diagnosis using statistical neural networks," IU-Journal of Electrical & Electronics Engineering, vol. 4, pp. 1149-1153, 2004.
16. Übeyli, "Implementing automated diagnostic systems for breast cancer detection," Expert systems with Applications, vol. 33, pp. 1054-1062, 2007.
17. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification," Journal of Biomedical Informatics, vol. 43, pp. 15-23, 2010.
18. Salama, M. Abdelhalim, and M. A.-e. Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers," Breast Cancer (WDBC), vol. 32, p. 2, 2012.
19. UCI Machine Learning repository (https://archive.ics.uci.edu/ml/datasets.html)
20. EasyNN Tool