A Comparative Analysis of Machine Learning Algorithms

by Indrashis Mitra

Submission date: 04-Apr-2022 09:07AM (UTC+0530)
Submission ID: 1800930318
File name: Major_II_-_Abstract.docx (186.23K)
Word count: 1251
Character count: 6952

A Comparative Analysis of Machine Learning Algorithms

Indrashis Mitra, Souvik Karmakar, Sudeshna Dutta, Kinjal Sarkar and Pratyay Basu

Guided by: Prof. Kananbala Ray

Abstract - India is a major player in the global pharmaceutical sector. There is a large hub of budding engineers and scientists in the nation that can take it to the next level. Antiretroviral drugs manufactured in India currently account for more than 80% of the world's AIDS medicine supply. However, people in India are taking desperate steps to keep loved ones alive as a disastrous spike of new coronavirus infections overwhelms the country's health-care infrastructure. They are turning to dubious medical therapies in some circumstances, and to the underground market for life-saving pharmaceuticals in others. Hence our project aims to solve this by developing a medicine recommendation system so that pharmacists can know in advance what medicines are being bought together the most, and keep stocks accordingly. We do not recommend which medicine is to be taken, rather the focus is on finding the drug combinations bought frequently together so that an idea can be had of the most-selling medicines; thus, keeping their stock replenished would eliminate the black market, help to earn profits and help the customers. Furthermore, the focus is also on an exploratory analysis of different classifiers for DNA classification to understand how they can be modified to suit a particular requirement.

Index Terms - medicine, recommendation, DNA, classification, Apriori, ECLAT, supervised learning

I. INTRODUCTION

Healthcare has become one of the world's fastest-growing industries, undergoing a total transformation and revolution on a worldwide scale.It has an essential role in enhancing people's health and well-being all over the world. Our main objective is to employ machine learning to help in the supply of medicines. Using the Apriori algorithm's support metrics, we plan to develop a recommendation system for the medicine that a particular customer is most likely to buy, resulting in a win - win for both the customer and the shop owner: the customer gets the required medicine they want at all times and does not have to deal with the annoyances of out of stock medicines; and the pharmacist learns the particular combination of medicines that is made available easily, wiping out the need for out-of-stock medicines. We also intend to compare the Apriori and Eclat algorithms to see what the differences are between the two recommendation systems.

II. LITERATURE REVIEW

All across the world, healthcare plays an important part in improving people's health status and well-being. In healthcare data classification, ambiguity and high-dimensionality are two factors that add to the difficulty [3]. Patients, physicians, and medical treatments are all recorded in the healthcare big data set, which grows in volume so quickly that typical data analytics tools are not able to keep up with it and evaluate it effectively. Some machine learning technologies are used in conjunction with the big data analytics framework as a means of addressing these issues. Data mining has emerged as a critical study issue in the advancement of computing applications in health care and biology [5]. Several large tech companies, including IBM and Google, have developed machine learning tools that can help doctors uncover novel treatment options for patients. Precision medicine is a significant concept in this discussion since it involves developing new ways to treat complex disorders and uncovering the underlying causes. However, even though several semi-supervised approaches have been presented to give additional training data, automatically produced labels are frequently too noisy to adequately retrain models [2]. In this project we have tried to make things easier in whatever way we can. Yoo et al. investigated the benefits and drawbacks of using data mining techniques in the biomedical field[6].To increase the performance of recurrent neural networks (RNNs) for anomaly detection, Yadav et al. employed ordinary differential equations (ODEs) to generate time series[7].

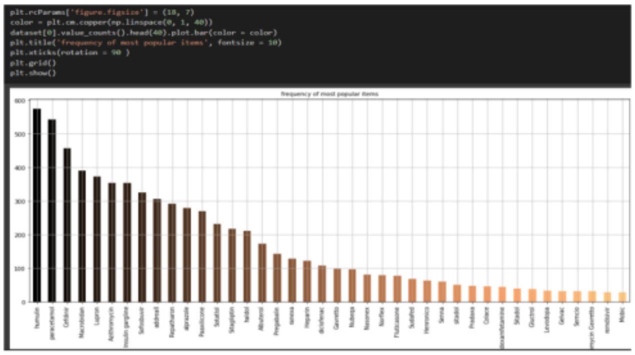
III. BASIC CONCEPTS

R. Agrawal and R. Srikant presented the Apriori algorithm in 1994 for locating frequent itemsets in a dataset for boolean association rules. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. An iterative approach horizontal level-wise search is done to find k+1 itemsets by using k-frequent itemsets mimicking the Breadth-First Search.Apriori property is used to improve the efficiency of level-wise generation of frequent itemsets by reducing the search space.

Equivalence Class Clustering and Bottom-up Lattice Traversal is the acronym for the ECLAT algorithm. It is one of the most widely used Association Rule mining techniques. It is a faster and more scalable version of the Apriori algorithm since works vertically, mimicking the Depth-First Search of a graph.

Classification is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In classification, a program learns from the given dataset or observations and then classifies new observations into a number of classes or groups.

IV. IMPLEMENTATION AND RESULTS



```
plt.rcParams['figure.figsize'] = (18, 7)
color = plt.cm.copper(np.linspace(0, 1, 40))
dataset[0].value_counts().head(40).plot.bar(color = color)
plt.title('frequency of most popular items', fontsize = 18)
plt.xticks(rotation = 90 )
plt.grid()
plt.show()
```

Fig 1 - Frequency of most popular medicines

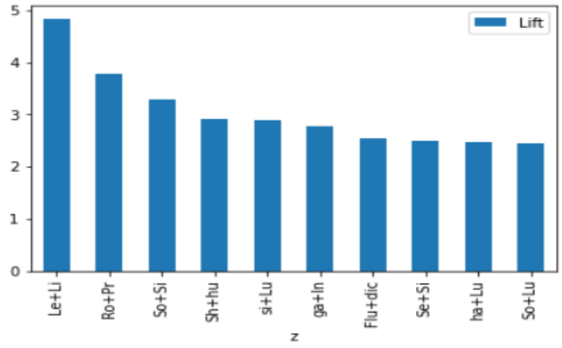


Fig 2- Visual depiction of top 10 medicines bought together - Apriori model result

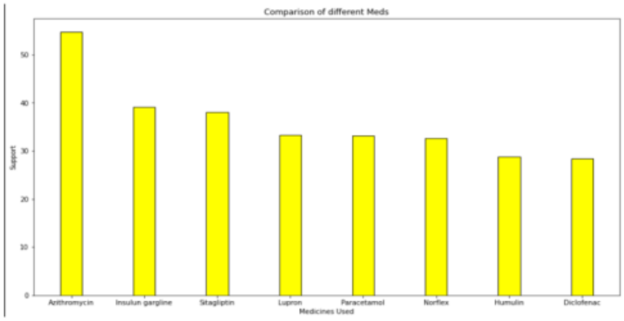


Fig 3 - Most commonly bought medicines - ECLAT model result

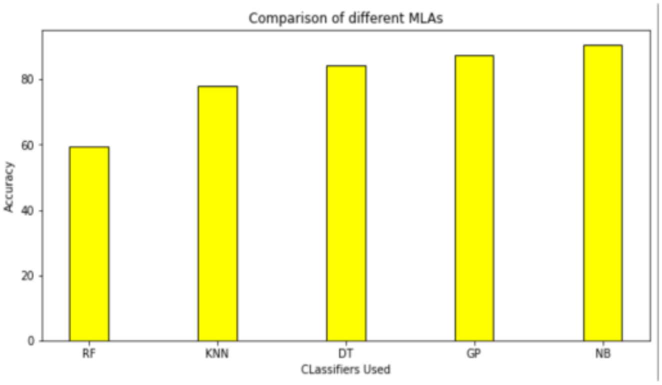


Fig 4 - Comparison of classifiers

The best model for this particular classification is found to be the Naive-Bayesian algorithm with an accuracy of 90.625%.

V. CONCLUSION

We have mainly focused on two objectives. The first one was a medicine recommendation system that will be helpful for the healthcare sector. People won't have to face the problem of unavailable medicines, since the stores will be stocked well in advance since they can know which medicines are most likely to be bought. Moreover, the economy will be helped since the medical black market will be eliminated as medicines are readily available so there is no shortage, thus no scope of dishonest people to dupe others by profiteering from selling medicines at exorbitant rates to needy people. Secondly, our focus is on the study and comparison of various classifiers. We have taken 5 classifiers and processed the same dataset containing DNA sequences through each of them, so that we can understand which classifier works best in our case.

VI. REFERENCES

[1].J. Bouwens, "Embracing Change: The healthcare industry focuses on new growth drivers and leadership requirements."

[2]. "Intro to Machine Learning | Udacity." Intro to Machine Learning | Udacity. Accessed April 27, 2016.

[3]. Thanh Nguyen, Abbas Khosravi, Douglas Creighton, Saeid Nahavandi, Classification of healthcare data using genetic fuzzy logic system and wavelets, Expert Systems with Applications, Volume 42, Issue 4, 2015, Pages 2184-2197

[4]. Schölkopf, Bernhard, Christopher J. C. Burges, and Alexander J. Smola,Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT Press, 1999.

[5]. Witten, I. H., and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufman, 2005

[6]. Zhiyong Ma, Juncheng Yang, Taixia Zhang, Fan Liu. (2016). An Improved Eclat Algorithm for Mining Association Rules Based on Increased Search Strategy. International Journal of Database Theory and Application, 9(5), 251-266

[7]. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, et al. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst 2012; 36:2431–48.

[8]. Yadav M, Malhotra P, Vig L, Sriram K, Shroff G. Ode-augmented training improves anomaly detection in sensor data from machines. arXiv preprint arXiv:1605.01534,2016:1–5.

[9]. Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-color Illustrations. New York: Springer, 2001

[10]. O.Stephen et.al.,"An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare",Journal of Healthcare Engineering ,Volume 2019

A Comparative Analysis of Machine Learning Algorithms

ORIGINALITY REPORT

15%	10%	6%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	4%
2	Submitted to University of Cincinnati Student Paper	4%
3	Submitted to Universidade Nova De Lisboa Student Paper	3%
4	Submitted to Queen Mary and Westfield College Student Paper	2%
5	Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li. "A Dictionary-based Oversampling Approach to Clinical Document Classification on Small and Imbalanced Dataset", 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2020 Publication	1%
6	Submitted to Victoria University of Wellington Student Paper	1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		