International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 2533–2539

Health Mantra: Machine Learning based Solution for Usage of Drugs from Social Network

Shridevi Soma
Dept., CSE, PDACE, Kalaburgi
Karnataka, India
shridevisoma@gmail.com

Supriya B D
Dept., CSE, PDACE, Kalaburgi Karnataka, India
supriya.b.deodurg@gmail.com

Abstract

Medication through Social Network is a need of hour in the situation like COVID19 and for the people placed away from the health centers. In a social media when somebody finds some good medicines by accident, such types of drugs are considered as serendipitous drugs. Finding such medicines is based on the health condition, mental status of the patient, cost and reviews of the medicines. This requires a kind of intelligence to fulfill the need of the patient. Hence, Health Mantra: A Machine Learning based solution proposed in this work provides an optimal solution for serendipitous drugs. This work contains various modules such as add patient details, view patient details, search drug and get drug details, view all types of drug reviews, view drug score result, drug review count. The total of 161297 drug data are considered from a standard webMD data set for experimentation, which includes the features like drugID, drug name, condition, review, ratings, useful count, date of manufacture the drug. Apriori algorithm is used to perform association analysis on the characteristics of drugs and balancing the data set. A Naives Bayes classifier is used in the proposed system and Optimal results are obtained from Naives Bayes classifier. Over 53766 drug data are tested against whole data set from standard dataset i.e. 161297 that resulted with an Accuracy of 100% for Naives Bayes classifier against existing system[1] where the authors experimented with 15714 dataset with Deep neural network classifiers and obtained an accuracy of 93%. The future scope of this work is to design a system for differentiating Drugtraffickers and Consumers

.

Keywords

Naives Bayes Algorithm, Apriori Algorithm, Support vector machine, Random forest, Serendipity.

I. INTRODUCTION

Serendipity is when accidently somebody finds some good medicines, such types of accidentely found drugs are serendipitous drugs. Usage of Serendipitous drugs refers to the unexpected relief of comorbid diseases or symptoms when taking a medication for a different known indication. Researchers in previous work[1] built an entire computational pipeline to investigate the feasibility of mining a new data source. They have used filtered drug reviews taken from WebMD standard dataset for experimentation. A comparative analysis of Deep neural network such as Convolutional neural network, long short memory network, Convolutional long short memory network have demonstratred performace of Area Under Curve is 0.815-0.919, precision is 0.156-0.783 and recall is 0.286-0.683 to non neural network algorithm such as SVM, Adaboost, Random forest classifiers has been carried out and evaluated the classification performance of Area Under Curve is 0.532-0.670, pecision is 0.232-0.981, recall is 0.064-0.365 and with n-grams AUC is 0.937, precision is 0.857, recall is 0.476 on the independent test dataset of size 15,714, but it could not maintain a large storage for data. This could lead to the problem for adding drugs to the database. In this proposed work, serendipity is implemented by using machine learning techniques. Machine learning focuses on computer programs being created that can access data and use it to learn for themselves. The proposed work helps to

ISSN: 2233-7857 IJFGCN

access the data from standard dataset from webMD and also scale-up drug data to be stored in the database. Experimentation is carried on total of 161297 drug data. Datasets contains information about the drugs, user details, ratings and manufactured date etc. Classification and Regression modeling are designed by dividing the dataset into train and test data. Apriori algorithm is used to perform association rule among the data inside the dataset, analysis on the characteristics of drugs and balancing the data set and also reduces search space. A Naives Bayes classifier is used in the proposed system and Optimal results are obtained from Naives Bayes classifier. Over 53,766 drug data are tested against whole data set from standard dataset i.e 161297 that resulted with an Accuracy of 100% for Naives Bayes classifier compared to existing system [1] where the authors experimented with 15714 datasets with Deep neural network like CNN, LSTM, CLSTM classifiers and obtained an accuracy of 93%.

II. RELATED WORK

In previous literature survey **Boshu Ru [1]** have proposed Serendipity- A Machine-Learning Application for Mining Serendipitous Drug Usage from Social Media by using the classifiers such as Deep neural networks. Drug repositioning reduces safety risk and development cost, compared to developing new drugs. The gold-standard dataset based on filtered drug reviews from WebMD is considered. Dataset includes 15,714 sentences, Finally the system with CNN, CLSTM, LSTM classifiers were designed and evaluated their classification performance. The model achieved an AUC score of 0.919 on the independent test dataset. The future scope of their work was to use optimal features, developing more effective methods to handle imbalance data, and verifying prediction results using other existing methods. **M.Huang[2**] have proposed Mapping client messages to a unified data model with mixture feature embedding convolutional neural network by using mixture feature embedding convolutional neural network. Data mapping among different data standards in health institutes is often a necessity when data exchanges occur among different institutes. Multi-modal features were extracted from different semantic space with a medical NLP package and powerful feature embedding were generated by MfeCNN. Experimental results show that proposed MfeCNN achieved best results than traditional state-of-the-art machine learning models. **M.Huang[3]** have proposed MfeCNN: Mixture Feature Embedding Convolution Neural network for Data Mapping. Data mapping plays an important role in data integration and exchanges among institutions and organizations with different data standards. The MfeCNN model converts the data mapping task to a multiple classification problem. In the model, it incorporated multimodal learning and multi view embedding into a CNN for mixture feature tensor generation and classification prediction. Multimodal features were extracted from various linguistic spaces with a medical natural language processing package. The combination of mixture feature embedding and a deep neural network can achieve high accuracy for data mapping and multiple classification. **Y.Jia[4**] have proposed An Empirical study of using an ensemble model in e-commerce taxonomy. Their approach was based on deep convolutional neural networks to predict product taxonomies using their descriptions. The classification performance of the work is further improved with oversampling, threshold moving and error corrected output coding. The best classification accuracy is obtained through ensembling multiple networks trained differently with multiple inputs comprising of various extracted features. **Sijia Liu[5]** have proposed on Mapping Textual Queries to a Common Data Model-. In this work the adoption of Electronic Health Records (EHRs) has enabled data-driven approaches to clinical care and research. However, the performance having lack of syntactic and semantic interoperability of EHR data across institutions. So to resolve this problem, Common Data Models (CDMs) can be used to standardize the clinical data in clinical repositories. The first step describes the mapping of entity mention types from patient-level information retrieval queries to an empirical subset of Observational Medical Outcomes Partnership (OMOP) CDM data fields, then investigated the empirical data model by annotating multi-institutional clinical data model fields.

III. METHODOLOGY

Serendipity is an occurrence and development of medicines by chance in a fortunate or beneficial way, such types of accidentally found drugs are serendipitous drugs. Usage of Serendipitous drugs refers to the unexpected relief of comorbid diseases or symptoms when taking a medication for a different known indication. From WebMD the total drug data of 161297 is

ISSN: 2233-7857 IJFGCN

considered for experimentation, it includes the features like drugID, drug name, condition, review, ratings, useful count, Date of manufacture the drug. **Apriori** algorithm works well to find association rules among the data inside database or dataset and the rules are based on transaction and items inside database, it reduces the space search. In the proposed system by applying Naives Bayes and SVM classifiers have been used and obtained Optimal results. Over 53766 drug data are tested against whole data set from standard dataset i.e. 161297 that resulted with an Accuracy of 100% for Naives Bayes classifier against existing system [1] where the authors experimented with 15714 datasets with Deep neural networks classifiers and obtained 91%. Data sets are used to integrate drug-review sentences with the patient's basic demographic information, ratings for the drug, drug therapeutic areas, and outputs from the filtering tools. For making predictions it is necessary to Combine social media WebMD text and context information. Incorporated natural-language processing and machine-learning methods help scientists and software developers to scan social media for serendipitous drug use. Much of the future efforts will be focused in creating more descriptive functionality, enhancing precision in the mapping of diseases, managing imbalanced data, and combining social media results with other data sources to create genuinely usable applications for drug repositioning. Two core model of the system includes implementing Apriori algorithm and Naives bays algorithm.

Apriori Algorithm: This algorithm works well to find association rules among the data inside database or dataset. These rules are based on transaction and items inside a database. In this work, items or drugs refer to a set of inter related drug data, which conveys a concept of object or entity among which some associations are to be found. Item or drug is a single member and only include one piece of data. A set of items which are put beside each other and construct a work unit is called transaction**.** For e.g. In a store a person purchase portfolio of combining social media WebMD text and context information. Here, customer from the store is a transaction and the purchased items inside the purchase portfolio are its items or drugs. Each of these drug items contains one or more pieces of data, such as drug number, drug price, Drug name for merchandise inside the shop.

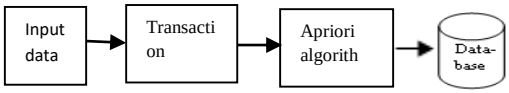

Fig-1: Balancing and managing drug data using Apriori algorithm.

Fig-1 works in the following two steps:

Drug and its prescriptions are collected from WebMD, then some of the association rules are identified by applying Apriori algorithm. So the most inter related drugs are identified and are grouped based on prescription. Following are the two important steps considered in this algorithm.

Step-1: Finding conventional Drug Item sets.

Step-2: Constructing association rules based on the found sets.

The two following hypotheses are considered in this algorithm:

a. Each subset of an iterative Drug Item set is iterative. If set {Drug1, Drug2, Drug3} is assumed to be iterative, then set {Drug1, Drug2} is also iterative;

b. Each hyper set of a non-iterative item set is non-iterative. If set {Drug1, Drug2} is assumed to be non-iterative, then set {Drug1, Drug2, Drug3} is also non-iterative.

Likewise, If Vitamin D3, then Calcium D3 tablet.

If Calcium-D3, then vitamin D3. These drugs are mentioned as antecedent and subsequent. So two drugs are prescribed together.

Apriori algorithm constructs a series of large item-set with length of K + 1 from the selected item-sets with length of K in each time and continues until an item-set with the longest length is achieved, provided that its support exceeds the required threshold.

Naive Bayes Algorithm:

ISSN: 2233-7857 IJFGCN

Naive Bayes algorithm is applied on the dataset for predicting whether the drug has side effects or not.

Algorithm to predict side effects of the drug.

Start

Step-1: Retrieve all the attribute values of drugs, ratings and counts from database.

Step-2: Calculate the sum of relative molecular mass attribute values for all the drugs.

Step-3: Compute the mean of relative molecular mass using the step-1 for medicines. Where n= number of drugs

Step-4: Computing standard deviation for relative molecular mass attribute for class using equation.

Step-5: Computing probability density function for relative molecular mass attribute using step-2 and 3. Where t = relative molecular mass of drug

Step-6: Repeat these steps for rest five features and calculate the probability of each feature.

Step-7: Total probability of side effect drug class determined.

Stop.

The main components of the algorithm are:

A. Drug Details

The drug details where collected from WebMD to add the details of the drugs in database. WebMD is one of the medical dataset which gives us all the details of the drug which includes the features like drugID, Drug Name, Condition, Review, Ratings, Useful count, Date of manufacture the drug.

B. Dataset

In this proposed work the standard datasets are considered. These dataset which consists of the attributes such as drug name, condition, review, ratings for identifying serendipitous drug usages, it would be ideal. WebMD gives information of drugs such as drug use, side effects, interactions, overdose, etc. These data are used as the benchmark for known drug usages in this work.

C. Preprocessing

Data **Preprocessing** is that step in which the raw data from WebMD is collected and transformed to understandable format. This raw data may be incomplete, inconsistent, may include missing drug details and may consist errors. So in this work, data is preprocessed to carry out training and testing for developing prediction model.

D. Feature Extraction

In this proposed system, drug details are extracted and considered as optimal features. Features like drugID, drug name, condition, review, ratings, useful count, date of manufacture the drug are considered in this paper. Feature construction and selection is an important part of data mining analysis, in which the drug data is processed and presented in a way understandable by machine learning algorithms.

E. Classification

Naives Bayes classifiers is used in this proposed work, their performance is compared with Deep neural networks, convolution neural network, long short term memory, convolution long short term memory in the experimentation.

IV. Experimental Results and Discussion.

In this proposed work, drug details are collected from webMD and considered the standard dataset for implementation of the work. Modules of the proposed work has been implemented using Python.

ISSN: 2233-7857 IJFGCN

The following Fig.-2 shows whole dataset is divided into test and train data. Total of 161297 dataset are trained and 53766 are tested with seven features mentioned earlier.

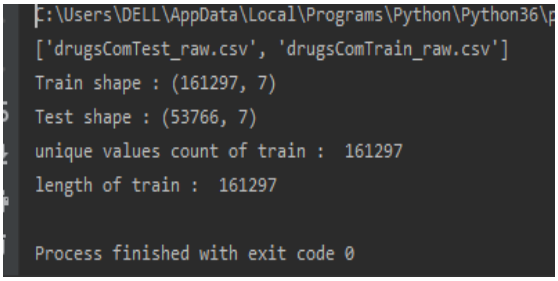

Fig.-2: Sample results of Training and Testing.

Fig.-3 shows the performance of the accuracy rate of proposed Naives Bayes system with 1.00 AUC, precision is 1.00 and recall is 1.00 which is better than the existing system.

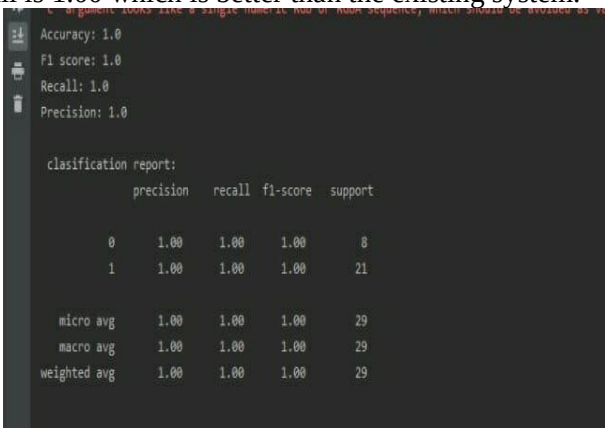

Fig.-3: Accuracy rate of proposed Naives Bayes Algorithm.

Table-1 shows accuracy of all the four classifiers, the result of Naive Bayes Classifier out performs compared to other classifiers.

Table-1: Performance Comparison of existing and proposed system.

Model	Accuracy (%)
Adaboost	93.7
SVM	90.0
Random Forest	92.6
Naive Bayes	100

The graphical representation of accuracy for existing system (Adaboost, SVM, Random forest) and proposed system (Naives Bayes) is shown in Fig.-4. In this experimentation the proposed model Naives Bayes has reached to the range 1.00 of accuracy.

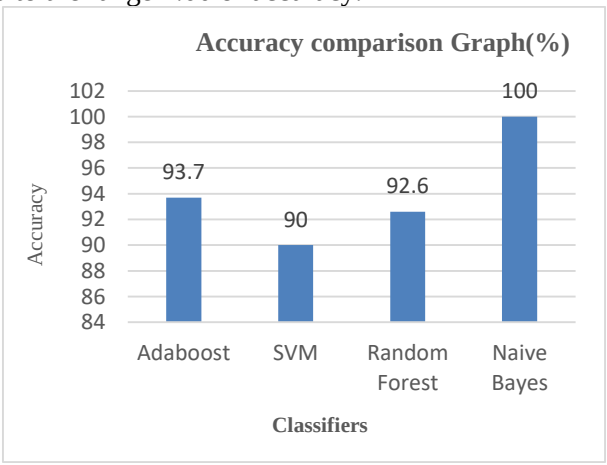

Fig.-4: Graphical representation of accuracy for existing and proposed Naives Bayes system.

ISSN: 2233-7857 IJFGCN

V. Conclusion

Drug repositioning is a significant but not yet completely used technique for enhancing the medicine's cost-effectiveness and reducing the production time. The dawn of social media brings with it large volumes of patient-reported medication outcome data, and thus creates an urgent need for drug repositioning to be examined. Proposed system aims to develop a whole computational pipeline focused on state-of-the-art machine learning and text mining methods. By using the machine learning classifiers like Apriori algorithm and Naives Bayes, it achieved the performance accuracy with 100% better than the existing system with accuracy of 93%. So approaches to machine learning seem feasible to tackle this question of locating a needle in the haystack. In future work, other classifiers can be used and that can distinguish drug traffickers and consumers.

References

[1] B.Ru, Dingcheng Li " Serendipity- A Machine-Learning Application for Mining Serendipitous Drug Usage from Social Media . 2019 , vol 18 pp 1-4.
[2] B.Ru, C.Warner-Hillard, Y.Ge "Identifying Serendipitous Drug Usages in Patient Forum Data", 2017 pp.106-108.
[3] D. Li, P. Liu, M. Huang, Y. Gu, Y. Zhang, X. Li, et al., "Mapping client messages to a unified data model with mixture feature embedding convolutional neural network", 2017, pp. 386-391.
[4] D. Li, M. Huang, X. Li, Y. Ruan, and L. Yao, "MfeCNN: Mixture Feature Embedding Convolutional Neural Network for Data Mapping," IEEE Transactions on NanoBioscience, 2018 pp. 165-171.
[5] Y. Jia, X. Wang, H. Cao, B. Ru, and T. Yang, "An Empirical Study of Using An Ensemble Model in E-commerce Taxonomy Classification Challenge", presented at the 2018 SIGIR Workshop On e-commerce (Accepted), Ann Arbor, MI, 2018 pp 1746-1751.
[6] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. Abbas, S. J. Hufeisen, et al., "Predicting new molecular targets for known drugs," Nature, vol. 462, 2009 pp. 175-181.
[7] P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon, et al., "Use of genome-wide association studies for drug repositioning," Nature Biotechnology, vol. 30 2012 , pp. 317-320.
[8] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," Molecular Systems Biology, vol. 7, 2011, pp 496.
[9] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," Nature Review Drug Discovery, vol. 3, 2004, pp. 673-683.
[10] D. Li, P. Liu, M. Huang, Y. Gu, Y. Zhang, X. Li, et al., "Mapping client messages to a unified data model with mixture feature embedding convolutional neural network", in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 386-391.
[11] L. Yao, Y. Zhang, Y. Li, P. Sanseau, and P. Agarwal, "Electronic health records: Implications for drug discovery," Drug Discovery Today, vol. 16, 2011 pp 673-683.
[12] C. Andronis, A. Sharma, V. Virvilis, S. Deftereos, and A. Persidis, "Literature mining, ontologies and information visualization for drug repurposing," Briefings in Bioinformatics, vol. 12, 2011 pp 357-368.
[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., " Scikit-learn: Machine Learning in Python," journal of machine learning research vol.12, 2011 pp. 2825-2830.
[14] G. E. Powell, H.A. Seifert, T. Reblin, P. J. Blowers, J, A. Menius, et .al, social media listening for routine post – marketing safety surveillance, "Drurg safety", vol. 39, 2016 pp. 443-454.
[15] P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon, et al., "Use of genome-wide association studies for drug repositioning, " Nature Biotechnology, vol. 30, 2012 pp 317-320.
[16] R .Eshleman and R.Singh, "Leveraging graph topology and semantic context for pharmacovigilance through twitter-treams , 2016 pp 335.

ISSN: 2233-7857 IJFGCN

[17] J. T. Dudley, T. Deshpande, and A. J. Butte, "Exploiting drug–disease relationships for computational drug repositioning," Briefings in Bioinformatics, vol. 12, 2011 pp. 303-311.

ISSN: 2233-7857 IJFGCN