ISSN:0975-9646
P. Jyotsna et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (5) , 2016, 2265-2269

Design and Development of Efficient Drug Reposition Scheme with Probabilistic Kernel based Text Mining Classification Model

P. Jyotsna [1], Dr. P. Govindarajulu[2]

Research Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India[1].
Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India[2].

Abstract: The work done for the research aims at characterizing these various modes and mechanisms of action for standard drugs, using a mathematical framework called description logics. The performance of a solution required first some theoretical basis to be set, namely the black box model of the cell, specified using description logics. Drug repositioning also referred as drug repurposing, re-profiling, therapeutic switching and drug re-tasking is the identification of new therapeutic indications for known drugs. For this we proposed probabilistic kernel based aspect model for text classification based on drugs data for registered users with probabilistic kernel based edge classification model. The Scheme is based on the Common Neighbors, Adamic-Adar Index, Jaccard-Coefficient, Edge Weight models with our proposed model for better classification accuracy and performance.

Keywords: Text Mining, Drug Repositioning, Machine Learning, Classification, Probabilistic Kernel based Edge Classification Model, Probabilistic Kernel based Aspect Model.

I. INTRODUCTION

A. Text Mining on Drugs

Text mining has been shown to be useful for discovering of new, previously unknown information, by automatically extracting information from different text resources. However, from a knowledge discovery science perspective, there is still a lack of research studies investigating the literature based approach towards link discovery and the specifics of such practices. It would thus be interesting to make a systematic examination of alternative techniques for discovering links and relations in literature through text mining. In this regard, the analysis below first identifies the repositioning hypotheses present in the FTC(Federal Trade Commission) and discusses the relevance of function in the process. Secondly, the extracted hypotheses are examined in depth and interpreted from a biological viewpoint. Because the FTC provides systemic insight on the drug repositioning topic, it is therefore possible to explore the broad relationship between therapeutic areas as well as the connection between drug repositioning and off-label uses[3,5]. These pathologies will serve to demonstrate how different methodologies can be applied over the FTC to extract hypotheses, the similarity-based and the computational approaches.

B. Relevance of Drug Repositioning

Drug repositioning can be defined as renewing failed drugs and expanding successful ones. One motivation behind drug repositioning is the possibility to further market and extend the application line or patent life of a drug, therefore increasing the revenue stream generated from it. Another aim is the treatment of rare or neglected diseases; usually such conditions are difficult to address for financial reasons, yet there might exist some safe and active molecules already developed for other indications, deemed suitable for this scenario.

C. Drug Repositioning faces legal and Efficient Challenges

In hypothesis, it is possible for a chemical to be active for multiple therapeutic indications. Yet in practice, several obstacles can impair the development of a potential new usage. The first factor to handle is serendipity. Biology is complex and capricious, exemplified by diseases such as cancers or dementias. A drug rarely keeps its original, it gets reoriented throughout the years when more data becomes available and it's in vivo pharmacology better understood. To conclude, the molecular opportunities for drug repositioning are contrasted by practical challenges[7,9]. Even if a compound is found to be active and safe for a new indication, additional factors, in particular legal issues and intellectual property, have to be considered in order to successfully bring the molecule to the market.

II. COMPUTATIONAL APPROACHES TOWARDS DRUG REPOSITIONING

The hypothesis and the clinical cases presented the reality of drug repositioning. I briefly commented on the fundamental reasons enabling new usages and stressed the importance of serendipity in this process. Now the fantasy of many scientists working in the drug discovery domain is to be able to formally predict such repositioning scenarios and unveiling new pharmacology in an automated fashion. In order to reach this distant goal or at least get closer to it, several computational approaches have been developed throughout the years. This section summarizes the previous work done on the topic and motivates the novel way I explored, based on a formal representation of the mode of action. I chose to classify the different computational

approaches based on the biomedical concepts used as the Centre of the methodology[2,6].

1. Chemical Structure-Based Approaches

Traditionally speaking, orally active drugs are mostly small lipophilic molecules. It therefore intuitively makes sense to directly look at the chemical structure to compare the similarity among drugs: similar structures are deemed to lead to similar biological outcomes. This rule of thumb goes by the name of similar property principle and is at the core of any quantitative structure-activity relationship (QSAR). A variety of methodologies exist to calculate the structural similarity between two chemicals, such as fingerprints or clustering algorithms. These methods can be used to perform ligand-based virtual screenings from a set of known active ligands, trying to find in a database of interest the structurally related molecules, supposedly bioactive too. In the context of drug repositioning, one can search only among approved compounds for instance[9,13]. This approach was successfully for Implementing an unsupervised machine learning algorithm in order to cluster chemicals based on their structure.

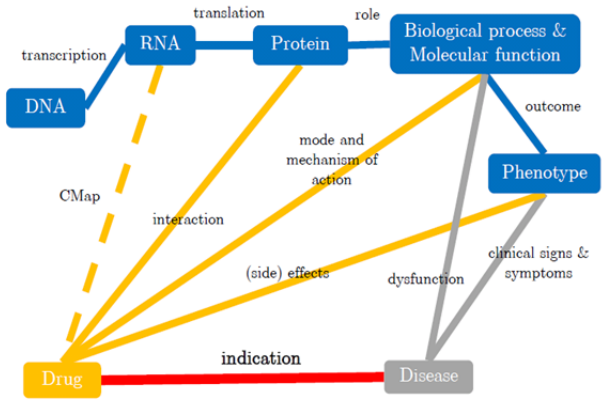

Fig.1: Conceptual map of the relationship between the different biomedical concepts.

Relation related to the drug and its action are in orange, diseases in grey and genetic concepts are in blue. Computational drug repositioning methods are based on either one or a series of such concepts in order to forward new indications for a drug, ultimate goal (red edge). Another interesting approach related to structural similarity for off-target identification comes. For this research, known ligands were grouped based on their known target binding partners and chemical features. The method is called similarity ensemble approach and calculates whether a molecule will bind to a target based on the chemical features it shares with those of known ligands, using a statistical model to control for random similarity. In the case of drug repositioning, the molecules tested were only approved drugs. The results revealed a series of off-target cases from the similarity analysis. A retrospective investigation showed the validity of the approach; then some predicted off-target bindings were experimentally validated, providing insightful clues about the pharmacological mechanism of some drugs. In some cases, such as for fabahistin, the off-target affinity (5-HT5A) was even better than for the known canonical receptor (H1), opening doors for meaningful alternative indications as shown below.

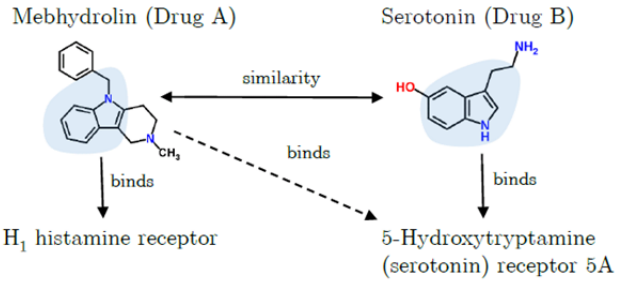

Fig.2: Drug repositioning using the chemical structure.

Compounds with similar structures have similar biological activities (similarity principle). Drug A shares some similarity with molecule B, indicated by the blue areas. This observation leads to the conclusion that molecule A could be active on the canonical target of molecule B, and indicated accordingly[13,15].

2. Machine Learning and Concepts Combination Approaches

The approaches presented before mostly focus on one of the concepts of the map shown and orient their analysis around it. It is also perfectly possible to use a combination of these biomedical descriptors to train a machine-learning algorithm and then generate predictions out of the statistical model. The recent analysis address drug repositioning from this perspective. In both cases, first a series of biomedical heuristics is defined, then the model is trained on known data and predictions are made.

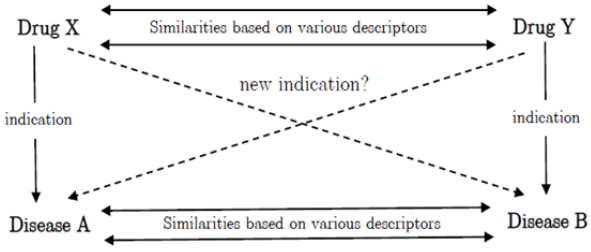

Fig.3: Drug repositioning using a combination of descriptors.

A machine learning algorithm is trained over a series of features, such as chemical similarity, shared target proteins, etc [10,14]. After evaluation of the model, some repositioning predictions can be generated from the statistical learning.

III. Design And Implementation Of The Proposed Models

A. Probabilistic Kernel based Edge Classification Model

Edge Classification Model, The aim of the model is to classify each edge E in the graph as ADE or not ADE. To enable the classification of this graph we extract a five

topological features used, along with time difference between the symptom and drug usage.

Step 1. Common Neighbors

$$score(d,s) = |T(d) \cap T(s)|$$

where T(d) and T(s) are the set of neighboring concepts within two hops for concepts d and s respectively.

Step 2. Adamic-Adar Index

The measure uses the common neighbors between two nodes and weights each of the common neighbors. It gives higher score for nodes with low degree.

$$score(d,s) = \sum_{z \in (d) \cap T(s)}^{n} \frac{1}{log|T(z)|}$$

Step 3. Jaccard-Coefficient

$$score(d,s) = \frac{|T(d) \cap T(s)|}{|T(d) \cup T(s)|}$$

Step 4. Preferential Attachment

$$score(d,s) = |T(d)| \times |T(s)|$$

Step 5. Edge weight between both nodes.

Step 6. Average temporal distinction between both nodes.

Each edge e=(d,s) is labeled as Adverse Drug Effects(ADE) or non-Adverse Drug Effects using sider. If the drug d, symptom s combinations are mentioned as a side effect in sider, it is labeled as a ADE and non-ADE otherwise. A decision tree based classifier is trained using this labeled data and topological features for each edge that are extracted from the graph. Enhanced EM algorithm is used to generate a decision making based classifier to classify the drug-symptom interaction. Since the classifier identifies a drug-symptom interaction as an ADE or not, the performance of the model can be measured using standard metrics like precision, recall, F-Score and Accuracy.

B. Probabilistic Kernel based Aspect Model for Text Mining

Algorithm:
1: Compute the empirical mean for $\{(x_n)\}N_n = 1$ (i.e. μ).
2: Center the data by $x_n \leftarrow (x_n - \mu)$ for n=1,…..,N,
3: Initialize the entries of W randomly to small positive Numbers.
4: repeat
5: {E-step}
6: for n=1 to N do
7: Calculate z* n
8: end for
9: {M-step}
10: for i=1 to M do
11: Update Wi ,
12: end for
13: until Change of ||W|| in consecutive EM iterations $< \delta$
14: return W.

C. Extracting Drug Names and Symptoms

To identify drug names and symptoms within drug database, n grams are generated from each reply after eliminating special characters. Each and every reply is tokenized into multiple tokens. A series of n-grams are generated where each n-gram is a series of n consecutive tokens, n<S:4. All of these n-grams are compared against a lexicon to identify drug names and symptoms. This lexicon is generated by combining terms from two resources. The Medical Subject Headings is vocabulary in life sciences, created and controlled by United States National Library of Medicine.

D. Graph Generation

Once the symptom and drug name concepts are extracted from the data for a particular user, we associate the timestamp of each review reply to the concepts extracted from it. This information is used to generate the medical profile graph of the user. The graph G(V,E), for a user u each node is a concept (drug or symptom) and edge E exists between a drug and symptom pair whose weight is the temporal difference between the mentioning of the drug and the symptom measured. All user graphs are them aggregated into a single graph G(V,E) where each vertex V is a drug or symptom concept and each edge e = (d,s) exists between a drug d and symptom s concepts with two weights, a total number of users with that drug symptom pairs in their history and the average temporal weight of that edge.

V. RESULTS

The classifier identifies a drug-symptom interaction as an Adverse Drug Effect or not, the performance of the model can be measured using standard metrics like precision, recall, F-Score and Accuracy. In order to compare the performance of the classifiers a 5-fold and 10-fold cross-validation is performed on the drug data. Using the 5-fold cross validation generally resulted in the significance value indicating the accuracy difference is less significant, it additionally resulted in a slightly smaller variation in the significance values. The smaller variation was generally advantageous, in that changing the partition has less impact on the conclusion drawn. When using 10-fold Cross Validation, changing the partition may result is more accurate. Determining which validation is more accurate on 90% of the available instances is closer to the ideal than determining which validation is more accurate on 50% of the available instances. Thus, if using 50% of the available instances yields accuracy values comparable to using 90%, 5-fold Cross Validation seems preferable to 10-fold Cross Validation, as it results in decreased variation in the significance values. When this is not the case, 10-fold Cross Validation seems preferable, because it results in accuracy values closer to what one would expect over the population.

Table 1:Results of 10 Fold Cross-Validation for Different Models

Model	Precision	Recall	F-Score	Accuracy
Common Neighbors	0.63	0.55	0.57	0.80
Adamic-Adar index	0.59	0.54	0.49	0.77
Jaccard Co-efficient	0.54	0.56	0.58	0.89
Edge Weight	0.64	0.57	0.57	0.58
Proposed Model	0.78	0.76	0.77	0.86

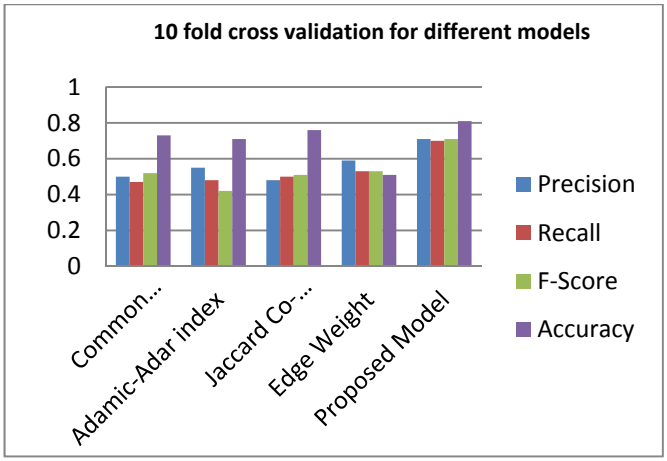

Fig.4: Graphical Representation of 10 fold on Different Models

Table 2: Results of 5 Fold Cross-Validation for Different Models

Model	Precision	Recall	F-Score	Accuracy
Common Neighbors	0.50	0.47	0.52	0.73
Adamic-Adar index	0.55	0.48	0.42	0.71
Jaccard Co-efficient	0.48	0.50	0.51	0.76
Edge Weight	0.59	0.53	0.53	0.51
Proposed Model	0.71	0.70	0.71	0.81

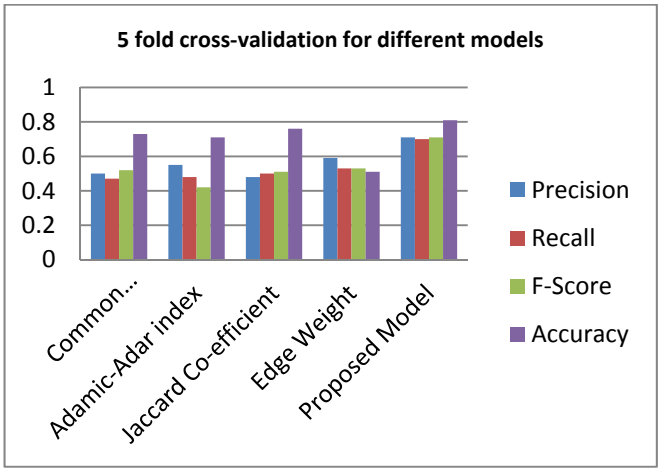

Fig.5: Graphical Representation of 5 fold on Different Models

Use case Diagram:

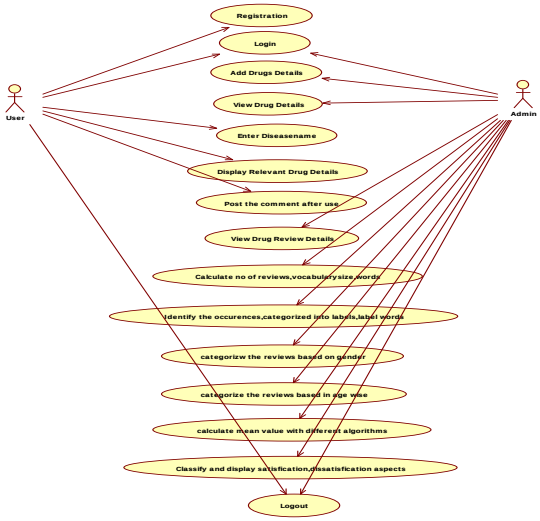

Fig.6 Representation of Use case Diagram

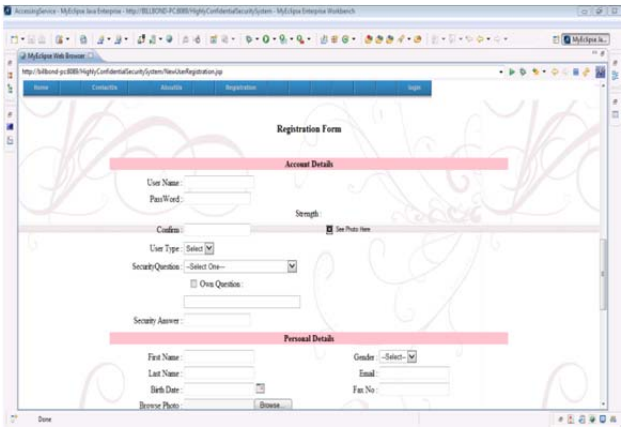

Fig.7: Registration Form for the user

VI. CONCLUSION

Computational drug repositioning offers secure for discovering latest uses of existing drugs, as drug associated molecular, chemical and clinical information has improved over the past decade and become broadly accessible. In this research, we developed a systematic scheme for mining probable new drug indications by exploring both chemical and molecular features in similar drugs. The proposed Probabilistic Kernel based Aspect Model uses and finds the aspects that are helpful in identifying the target class. As human lifespan becomes longer and living environment becomes increasingly polluted, medicinal area of data mining becomes one of the focused research area. The Proposed model is for mining aspects relating to specified labels or groupings of drug data. The Novel probabilistic kernel based aspect model gives better performance and accuracy for text classification based on drugs data for registered users with probabilistic kernel based edge classification model and the scheme is based on the Common Neighbors, Adamic-Adar Index, Jaccard-Coefficient, Edge Weight models. In future work we will design and apply Multi View Aspect Mining model with Machine Learning on clustering mechanism using string kernel to evaluate mean PMI of the derived aspects.

REFERENCES

1) Ashburn, T. T. and Thor, K. B." Drug repositioning: identifying and developing new uses for existing drugs". Nature reviews Drug discovery, 3(8):673-683, 2004.
2) A Lew and H. Mauch, "Introduction to Data Mining Principle" ,SCI, springer, 2006.
3) Benton A., Ungar L., Hill S., Hennessy S., Mao J., Chung A., & Holmes J. H, "Identifying potential adverse effects using the web: A new approach to medical hypothesis generation" Journal of biomedical informatics, 44(6), pp. 989-996, 2011.
4) Berry Michael W., "Automatic Discovery of Similar Words in Survey of Text Mining: Clustering,
Classification and Retrieval", Springer Verlag, New York, LLC, 24-43, 2004
5) C.-P. Wei, K.-A. Chen, and L.-c. Chen, "Mining Biomedical Literature and Ontologies for Drug Repositioning Discovery," in Advances in Knowledge Discovery and Data Mining", Springer International Publishing, pp. 373-384, 2014.
6) G. Forman., "An extensive empirical study of feature selection metrics for text classification", The Journal of Machine Learning Research, 3:1289–1305, 2003.
7) L. Yang, J. Chen, L. Shi, M.P. Hudock, K.Wang and L. He, "Identifying unexpected therapeutic targets via chemical-protein interactome", PLoS One, 5(3): e9568, 2010.
8) Liritano S. and Ruffolo M., "Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining", IEEE, 454-458 , Italy, 2001.
9) Liu, X., & Chen, H., "AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. In Smart Health (pp. 134-150). Springer Berlin Heidelberg, 2013.
10) Mitrofanova, V. Pavlovic, and B. Mishra, "Prediction of Protein Functions with Gene Ontology and Interspecies Protein Homology Data," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 8, no. 3, pp. 775-784, 2011.
11) M. Seno and G. Karypis, "Slpminer: An Algorithm for Finding Frequent Sequential Patterns Using Length-Decreasing Support Constraint," Proc. IEEE Second Int'l Conf. Data Mining, pp. 418-425, 2002.
12) Ning Zhong, Yuefeng Li, Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE Transact ions on Knowledge and Data Engineering. 2010.
13) Radovanovic M., Ivanovic, M., CatS: "A classification-powered meta-search engine." In: Last, M., et al., editors, Advances inWeb Intelligence and Data Mining, pages 191–200, Springer-Verlag, 2006.
14) Torkkola K., Linear discriminant analysis in document classification. In: IEEE ICDM Workshop on Text Mining, pages 800–806, 2001.
15) Witten, I. and Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, San Francisco, second edition, 2005.
16) Y. Yamanishi et al., "Drug-Target Interaction Prediction from Chemical, Genomic and Pharmacological Data in an Integrated Framework," Bioinformatics, vol. 26, pp. i246-i254, 2010.
17) Y. Yamanishi, M. Kotera, M. Kanehisa and S. Goto, "Drugtarget interaction prediction from chemical, genomic and pharmacological data in an integrated framework", Bioinformatics, 26(12): i246–54, 2010.