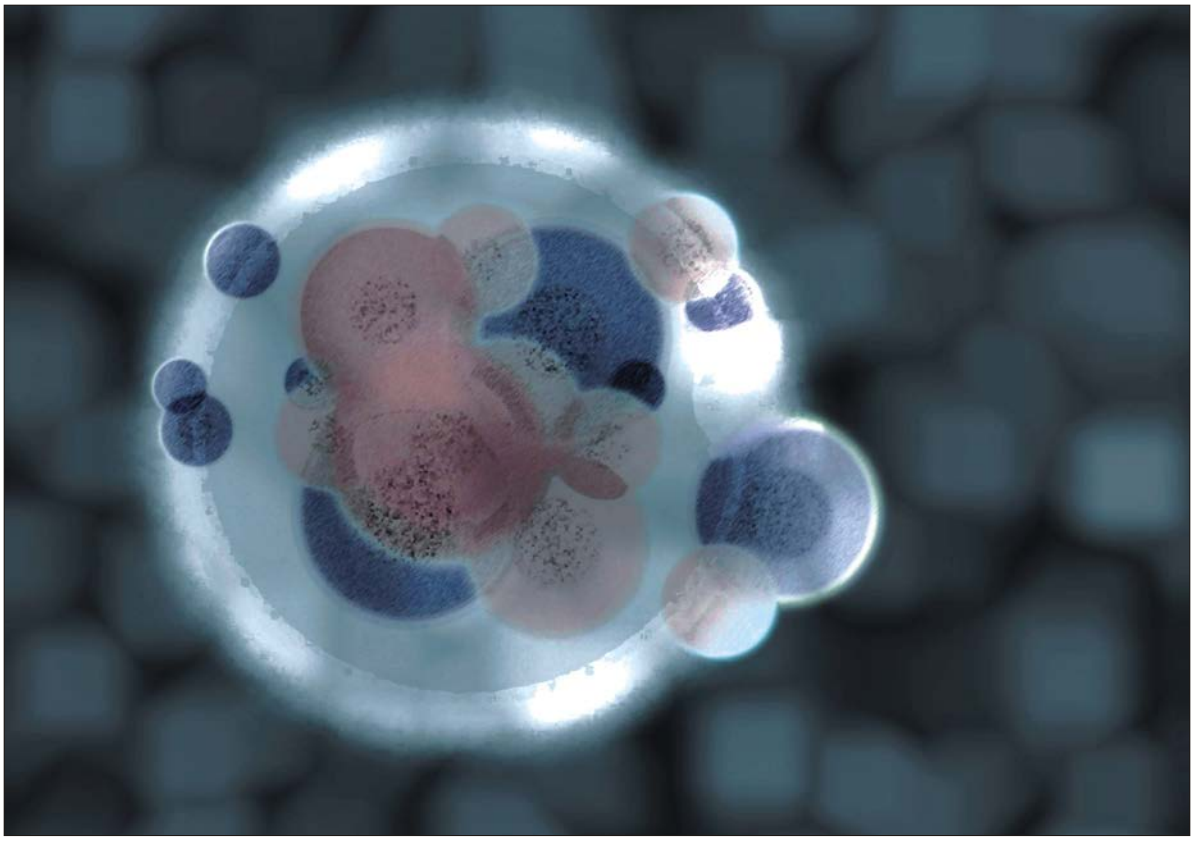

Journal of Biological Sciences

ISSN 1727-3048

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Biological Sciences 13 (4): 293-297, 2013
ISSN 1727-3048 / DOI: 10.3923/jbs.2013.293.297

Ranking Drugs in Spontaneous Reporting System by Naive Bayes

[1]A. Bazila Banu, [2]S. Appavu Alias Balamurugan and [3]P. Thirumalaikolundu Subramanian
[1]Faculty of Information Technology, Velammal College of Engineering and Technology, Madurai-625009, Tamilnadu, India
[2]Prof of Electronics and Communication Engineering, K.L.N College of Information Technology, Madurai-630611, Tamilnadu, India
[3]Director, Professor and Head (Retd), Institute of Internal Medicine, Madras Medical college, Chennai, Tamilnadu, India

Abstract: In this study detection of association between drugs and Adverse Drug Reactions (ADRs), is carried out by using Naive Bayes method. Adverse event reports submitted to the United States Food and Drug Administration (FDA) were reviewed to find top 10 drugs causing frequent ADRs for a particular period. The main objective of this paper is to evaluate drugs associated with list of outcomes provided by FDA. For a particular category of disease, drugs creating outcomes are ranked using Naive Bayes method. FDA represents ADRs in Preferred Terms(PT) by referring Medical Dictionary for Regulatory Activities (MedDRA).To create conceptual hierarchy System Organ Class (SOC) present in MedDRA is mapped with low level Preferred Terms (PT) in FDA dataset. For each SOC the drugs are ranked based on posterior probability obtained by Naive Bayes method. Data mining model has been built to analyse drugs associated with outcome for a disease category in SOC level. The newly designed tool is user friendly and applicable to pharmaceutical industries, policy makers and practitioners.

Key words: Adverse drug reaction, food and drug administration, medical dictionary for regulatory activities, system organ class, preferred terms, spontaneous reporting system

INTRODUCTION

Adverse Drug Reaction (ADR) is defined as all toxic and inadvertent response to a drug with any dosage level. ADRs remain an important health issue to be investigated in earlier stage to avoid accidental effects. There are two types of ADR.Type 'A' or dose related creates highest percentage of adverse reactions than Type 'B' reactions (Talbot and Waller, 2005). Type 'A' reactions are considered for this study. To detect ADR signals, data mining methods like Proportional Reporting Ratio and Naive Bayes are used in the Spontaneous Reporting System (SRS) (Wang *et al*., 2011; Hauben *et al*., 2007).

United States Food and Drug Administration (US FDA) plays an important role in pharmacovigilance activity especially in providing drug safety. Since 1969 FDA has maintained a computerized repository for storing and retrieving all ADR reports (USFDA, 2011).This repository can also be referred as SRS. It stores all reported ADRs by adopting the protocols of standardized international terminology, Medical Dictionary for Regulatory Activities (MedDRA).It represents disease names in Preferred Terms (PT) level (Pearson *et al*., 2009). MedDRA is a medical dictionary for describing adverse events, with five levels: the highest level is System Organ Class(SOC), followed by High Level Group Term (HLGT), Higher Level Term (HLT), Preferred Term (PT) and Lowest level Term (LLT) (Henegar *et al*., 2006). List of outcomes as given by FDA like Life-Threatening (LT), Hospitalization - Initial or Prolonged (HO), Disability (DS), Congenital Anomaly (CA), Required Intervention to prevent permanent impairment/damage (RI), Death (DE) are ranked by Naive Bayes method based on the attributes drug, category of the diseases.

In data mining, supervised learning methods such as decision trees, association rules, Naive Bayes and neural networks are used in the field of ADR to analyse the reports (Chazard *et al*., 2011). The Naive Bayes classifier (NB) is one among the statistical classifier used to predict class membership probability. It detects the class based on the maximum probability obtained for the given tuple to a particular class. It has been widely used by

Corresponding Author: A. Bazila Banu, Faculty of Information Technology, Velammal College of Engineering and Technology, Madurai-625009, Tamilnadu, India Tel: 9940161706

researchers for classification. It assumes all variables participating in the classification as independent and produces good results for prediction. Zhang and Su (2004) have proved that apart from classification Naive Bayes can be used for ranking and it outperforms other traditional decision tree algorithms like c4.5. The limitations intrinsic in SRS like FDA are studied by Sakaeda *et al.* (2011a, b). Aim of this study is to design an user friendly tool to be used by pharmaceutical industries and practitioners to analyse ADRs based on disease category.

To identify the drugs causing outcomes of an ADR, this paper aims at ranking drugs associated with list of outcomes for a particular disease category. To design a user friendly tool for analyzing ADRs.To Map the adverse event representation of FDA from PT level to MedDRA SOC level. Researchers suggested that it may be more beneficial to perform data mining, using highest level adverse event representation like SOC than the MedDRA PT level (Pearson *et al.*, 2009). The depth of the PT level is high and hence difficult to evaluate the disease associated with particular outcome. Mapping of PT with SOC will produce better clarity in evaluating the drug outcome association.

Klementiev *et al.* (2007) has proposed an unsupervised learning algorithm for rank aggregation. Zhang *et al* (2005) has done extensive work in applying Naive Bayes for ranking. He has proposed a method named as Augmenting Naive Bayes, suitable for applying in limited training data. Jiang *et al.* (2005a, b) has proposed K-nearest neighbor Naive Bayes for ranking, however the performance is unknown for ranking items. Jiang *et al.* (2005a) also proposed Tree Augmented Naive Bayes for ranking, but the performance is poor when compared to Naive Bayes.

MATERIALS AND METHODS

ADRs in FDA 2011 are considered as case set for the study. PTs used in FDA are mapped with SOC of MedDRA by referring cancer therapy evaluation program simplified disease classification v4.0 (MedDRA v 12.0) (National cancer institute, United States). Data is extracted from SRS provided by FDA. Then the duplicate reports were deleted according to FDA's suggestion of using recent case number as described in the file "Asc-nts.doc' from the website of the FDA. SRS such as US FDA provide an opportunity to study drugs causing side effects. Once the schema is designed and the associated database is constructed, the data is loaded from FDA's text file to oracle database, using ETL (Extract, Transform and Load) tools. Indices are constructed using patient identifier. Xml Mapping is created by constructing SOC as the root node and PT as the child nodes. Based on XML Sql Utility (XSU), the root nodes are injected into the database for the equivalent PTs to create disease categorical mapping. Figure 1 represents the hierarchical design of XML for SOC and PT. XML Sql Utility (XSU) in oracle 11 g is used to extract the data from an x mL document and to inject the data in to the database (Yu and Stahnikam, 2011).

First step is to establish Java Database Connection (JDBC) then an instance of OracleXmlQuery is created to pass sql query which contains the table name and column name to be updated from xml data. Next step is to obtain Document Object Model (DOM) by calling get method of OracleXmlQuery.XSU converts the elements to sql types and binds them to the appropriate statement. At this step the disease category of SOC level is obtained by mapping the PT of FDA with MedDRA PT. This step creates concept hierarchy. For ranking the drugs based on outcome Naive Bayes method is used because it is based on the assumption of class conditional independence of all attributes. If a learning algorithm is able to estimate accurate class probability, it certainly produce precise ranking (Gouthami *et al.*, 2012).

Data mining model

Schema: Description of database schema is shown in Fig. 2. The central entity is that of a patient. Patient is uniquely identified by ISR (unique number for identifying an AERS report). For mapping PT with SOC, each ISR is assigned to only one SOC. All other data in the schema is adverse event based. For data mining phase, the data

<Root>
<SOC> Bladder disorders NEC
<PT> Bladder dilatation </PT>
<.Root>

Fig. 1: Hierarchical design of XML for SOC and PT

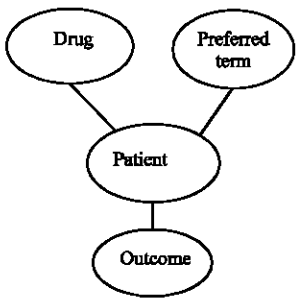

Fig. 2: Description of database schema

have to be simplified. For instance the duration and dose of medications have been ignored. Three distinct entities considered are drug, preferred term and outcome.

Data mining prototype: Computational steps for data storage and retrieval is shown in Fig. 3. A web based system has been designed to analyse the various levels of disease categories for a particular period. First step of Fig. 3 describes how the data from FDA are transformed into analytical schema using ETL workflows. The data presented in the. txt format are transformed and loaded in to the schema by using SQL Loader. PTs used in FDA are mapped with SOC of MedDRA to create ADR schema.

Implementation: Naive Bayes algorithm is used to obtain the top 10 drugs for each year based on the attributes drug, category and the outcome. Using Naive Bayes, model can be built with different prior probability assumptions. The outcome of the adverse events is considered as major class and the other attributes like category and drugs are considered as subset for ranking. First, the fundamental assumption of attribute independence is considered for this study. Naive Bayes theorem given in formula 1 is used to calculate the probability of an outcome:

$$P(H/X) = \frac{P(X/H)^* P(H)}{P(X)} \quad (1)$$

where, P(H), the probability that the hypothesis H holds for the observed data tuple X. This is the prior probability that any patient may get an outcome regardless of the drug and category. The posterior probability P(H/X) is based on more patient information like drug and disease category. All the possible outcomes given in FDA were considered for the P(H/X). Where, X = (Disease Category = Gastrointestinal disorders |Any of 26 disorders, Code of drug=1|2, outcome = LT| HO| DS| CA| RI |DE).Code 1 denotes valid trade name of the drug and 2 denotes verbatim name of the drug.

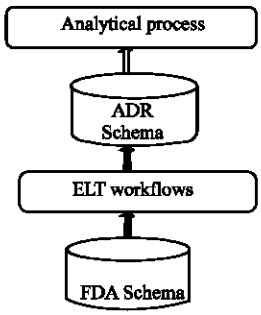

Fig. 3: Computational steps for data storage and retrieval

The outcomes LT, HO, DS, CA, RI and DE are used for computing the values of P(X), where P(X) denotes prior probability of X. Based on the outcome, the probabilities of category and drug are evaluated. By using the posterior probabilities for corresponding category and outcome, the top 10 drugs are listed for ranking. The algorithm applies disease category as a variant. Disease categories are chosen for an outcome and the posterior probabilities are estimated for ranking the drugs.

RESULTS

The experimental result helps the medical practitioners to identify the drugs creating particular outcome based on the category of the disease in the SOC level. It is tedious to evaluate the outcomes based on PT level. Posterior probability obtained from Naive Bayes is used to rank the drugs .Among all the drugs Aspirin occur in four outcomes like DE, DS, HO and LT. It has the probability of 1 for HO. Lasix is the second drug to have three outcomes like HO, LT, DE. It produces the probability of .6158(HO), .0009(LT), .1410(DE). Drugs like Humira and Remicade occur in multiple outcomes like HO and LT with probability of .8908(HO), .7598(HO), .0006(LT), .0009(LT).Drugs like Yaz and Yasmin occur in HO with probability of, .9142 and .9216. Heparin Sodium Injection and Revlimid occurs in multiple outcomes with less probability of .0008(LT), .0009(LT), .0309(DE), .0181(DE).Other drugs like Trasylol, Dianeal, Avandia, Tracleer, Acetaminophen, Heparin, Dexamethasone and Coumadin produces DE alone. Outcomes like CA and RI are very less. Drugs like Zolosoft have the probability of .000006(CA) and Actos have the probability of .00013(RI). List of SOC's are given in Table 1.

Table 1: Representation of disease categories interns of SOC level

Name of disease category in System Organ Class level	
Blood and Lymphatic system disorders	Musculoskeletal and connective tissue disorders
Cardiac disorders	Neoplasms benign, malignant
Congenital ,family and genetic disorders	Nervous system disorders
Ear and labyrinth disorders	Pregnancy, puerperium and perinatal conditions
Endocrine disorders	Psychiatric disorders
Eye disorders	Renal and urinary disorders
Gastrointestinal disorders	Reproductive system and breast disorders
General disorders and administration site conditions	Respiratory, thoracic and mediastinal disorders
Hepatobiliary disorders	Skin and subcutaneous tissue disorders
Immnne system disorders	Social circumstances
Infections and infestations	Surgical and medical procedures
Injury, poisoning and procedural complication	Vascular disorders
Investigations	Metabolism and nutrition disorders

SOC: System organ class

Table 2-5 represent the ranking of the drugs for the disease category Gastrointestinal Disorder, with respect to outcomes like HO, LT, DE and DS. As per the results of Table 2 and 3, it is observed that Aspirin holds the 1st rank to cause Gastrointestinal Disorder with outcome as HO and LT. For outcome HO, drugs like Yasmin, Yaz and Humira acquire 2nd , 3rd and 4th ranks with probability greater than 0.8. For outcome, LT drugs like Prednisolone, Revlimid, Lasix acquire 2nd, 3rd and 4th ranks with minimal probability. Table 4 represents one of the severe outcomes DE and the drug Revlimid acquires the 1st rank. Other drugs like Trasylol, Dianeal and Aspirin acquires 2nd, 3rd and 4th ranks with nominal probability. Table 5 represents hazardous outcome DS and the drug Metoclopramide acquires the 1st rank and Asprin acquires the 4th position but the probability is nominal.

Table 2: Top 10 drugs for the period 2011(Q1-Q4) with outcome as HO in gastrointestinal disorder

Rank	Name of the drug	Probability
1	Aspirin	1.0000
2	Yasmin	0.9216
3	Yaz	0.9142
4	Humira	0.8908
5	Remicade	0.7598
6	Avandia	0.6872
7	Forteo	0.6807
8	Dianeal	0.6620
9	Avonex	0.6566
10	Tysabri	0.6543

[1]HO: Hospitalization

Table 3: Top 10 drugs for the period 2011(Q1-Q4) with outcome as LT in gastrointestinal disorder

Rank	Name of the drug	Probability
1	Aspirin	0.0018
2	Prednisolone	0.0010
3	Revlimid	0.00095
4	Lasix	0.00094
5	Dexamethasone	0.00093
6	Omeprazole	0.00086
7	Heparin Sodium Injection	0.00084
8	Acetaminophen	0.00083
9	Methotrexate	0.00082
10	Remicade	0.0007

LT: Life threatening

Table 4: Top 10 drugs for the period 2011(Q1-Q4) with outcome as DE in gastrointestinal disorder

Rank	Name of the drug	Probability
1	Revlimid	0.0309
2	Trasylol	0.0264
3	Dianeal	0.0224
4	Aspirin	0.0206
5	Lasix	0.141
6	Heparin Sodium Injection	0.0181
7	Acetaminophen	0.0167
8	Avandia	0.0147
9	Tracleer	0.01244
10	Dexamethasone	0.01241

DE: Death

Table 5: Top 10 drugs for the period 2011(Q1-Q4) with outcome as DS in gastrointestinal disorder

Rank	Name of the drug	Probability
1	Metoclopramide	0.0089
2	Fosamax	0.0060
3	Reglan	0.0037
4	Aspirin	0.0013
5	Simvastatin	0.00075
6	Omeprazole	0.00072
7	Fosamax Plus D	0.00071
8	Prednisolone	0.00066
9	Lipitor	0.00065
10	Prednisone	0.00062

DS: Disability

DISCUSSION

Detection of ADR depends on FDA's assessment of patient cases. Determining precise knowledge from large volume of medical prescription dataset has come to realism for decision making as well as to avoid hazards in the drugs (Xiuzhen *et al*., 2011). Post marketing surveillance system is responsible for providing drug safety. As a result it may necessitate years to identify difficult drugs from the promoters (Ji *et al*., 2011). A list of outcomes and their link between drug and categories is used for analysis. In this paper 6 outcomes enable us to trace ADRs. Among the 26 categories listed in the Table 1, Disease Category Gastrointestinal Disorder (SOC) is taken to evaluate the six outcomes. The drugs are ranked based on the probabilities obtained by Naive Bayes method. Performing causality assessment in pharmacovigilance may help decision making in single ADR pertaining to the drug.

CONCLUSION

In this study we presented our knowledge of designing an user friendly environment for conducting Adverse Drug Reaction studies (ADR) based on mining large scale primary care database FDA. We have used Naive Bayes theorem's posterior probabilities to rank the drugs. This approach serves as a reusable environment for ranking drugs based on disease categories and outcomes. In this paper the calculations of Naive Bayes theorem are presented for ADR study. In Future the SOC level of MedDRA can be further tuned to lower levels like HLGT and HLT for analysis and decision making at different levels of disease classification.

REFERENCES

Chazard, E., G. Ficheur, S. Bernonville, M. Luyckx and R. Beuscart, 2011. Data mining to generate adverse drug events detection rules. IEEE Trans. Inform. Technol. Biomed., 15: 823-830.

Gouthami, S., G.J. Mary and P.S. Rao, 2012. Ranking popular items by naive Bayes algorithm. Int. J. Comput. Sci. Inform. Technol., 4: 147-163.

Hauben, M., S. Horn, L. Reich and M. Younus, 2007. Association between gastric acid suppressants and *Clostridium difficile* colitis and community-acquired pneumonia: Analysis using pharmacovigilance tools. Int. J. Infect. Dis., 11: 417-422.

Henegar, C., C. Bousquet, A. Lillo-Le Louet, P. Degoulet and M.C. Jaulent, 2006. Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. Comput. Biol. Med., 36: 748-767.

Ji, Y., H. Ying, P. Dews, A. Mansour, J. Tran, R.E. Miller and M.M. Massanari, 2011. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. IEEE Trans. Inform. Technol. Biomed., 15: 428-437.

Jiang, L., H. Zhang and J. Su, 2005a. Learning k-nearest neighbor naive Bayes for ranking. Proceedings of the 1st International Conference on Advanced Data Mining and Applications, July 22-24, 2005, Wuhan, China, pp: 175-185.

Jiang, L., H. Zhang, Z. Cai and J. Su, 2005b. Learning tree augmented naive Bayes for ranking. Proceedings of the 10th International Conference on Database Systems for Advanced Applications, April 17-20, 2005, Beijing, China, pp: 688-698.

Klementiev, A., D. Roth and K. Small, 2007. An unsupervised learning algorithm for rank aggregation. Proceedings of the 18th European Conference on Machine Learning, September 17-21, 2007, Warsaw, Poland, pp: 616-623.

Pearson, R., M. Hauben, D. Goldsmith and A.L Gouldf, 2009. Influence of the MedDRA® hierarchy on pharmacovigilance data mining results. Int. J. Med. Info., 78: e97-e103.

Sakaeda, T., K. Kadoyama and Y. Okuno, 2011a. Adverse event profiles of platinum agents: Data mining of the public version of the FDA adverse event reporting system, AERS and reproducibility of clinical observations. Int. J. Med. Sci., 8: 487-491.

Sakaeda, T., K. Kadoyama and Y. Okuno, 2011b. Statin-associated muscular and renal adverse events: Data mining of the public version of the FDA adverse event reporting system. PLoS ONE, Vol. 6. 10.1371/journal.pone.0028124

Talbot, J. and P. Waller, 2005. Stephens Detection of new Adverse Drug Reactions. 5th Edn., John Wiley and Sons, New York., USA., pp: 173-175.

USFDA, 2011. FDA adverse event reporting system. U.S. Food and Drug Administration. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm.

Wang, W., K. Haerian, H. Salmasian, R. Harpaz, H. Chase and C. Friedman, 2011. A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from pubmed citations. Proceedings of the Annual Symposium on Improving Health: Informatics and IT Changing the World, October 22-26, 2011, Washington, DC., USA., pp: 1464-1470.

Xiuzhen, F., H. Xiaohong and F. Bianling, 2011. A study on application of multidimensional association rule mining in adverse drug reactions. Energy Procedia, 11: 1720-1726.

Yu, T. and B. Stahnikam, 2011. Developing XML applications with oracle XML DB and Oracle XML developer's kit. http://www.oracle.com/technetwork/Database/features/xmldb/s311509-developingxmlapplicationswi-132942.pdf.

Zhang, H. and J. Su, 2004. Naive Bayesian classifiers for ranking. Proceedings of the 15th European Conference on Machine Learning, September 20-24, 2004, Pisa, Italy, pp: 501-512.

Zhang, H., L. Jiang and J. Su, 2005. Augmenting naive Bayes for ranking. Proceedings of the 22nd International Conference on Machine Learning, August 7-11, 2005, Bonn, Germany, pp: 1020-1027.