Comparative Analysis of Data Mining Classifiers in Analyzing Clinical Data

Dr. Adetunji A.B[1]Ayinde A.Q[2], Odeniyi O.A[3] and Adewale J.A[4]
Ladoke Akintola University of Technology Ogbomoso, Oyo State, Nigeria[1]
Osun State College of Technology Esa Oke, Osun State, Nigeria[2,3,4]

Abstract

Health-care providers know there's a wealth of valuable information trapped in the hand-written notes on patients' charts. But the challenge of collecting and interpreting the data on a large scale remains to be solved. Now, researchers have taken a step forward in mining patient-based information by using existing language-analysis methods to identify drug use side effects in advance of the NAFDAC (National Agency for Food and Drug Administration and Control) issuing official alerts. Bioinformatics is an interdisciplinary field that develops and improves on methods for storing, retrieving, organizing and analyzing biological data. A major activity in bioinformatics is to generate useful biological knowledge. In experimental molecular biology, bioinformatics techniques such as image and signal processing allow extraction of useful results from large amounts of raw data. This research is streamlined to the biological and clinical data. It plays a role in the textual mining of biological literature and computer science literature to organize, query and mine biological data. Biological data obtained is subjected to cross industry standard process for data mining and WEKA will serve as the bioinformatics tools for biological knowledge extraction. In this research work K-Nearest Neighbour (K-NN) and Classification and Regression Tree (CART) algorithms will be apply to biological data via the WEKA tool with aim to predict the effect of drug used considering the most probable target which is the patient symptoms after using the anti-malaria drugs.

1. INTRODUCTION

Data mining has attracted a great deal of attention in the information technology industry, due to availability of large volume of data which is stored in various formats like files, texts, records, images, sounds, videos, scientific data and many new data formats. There is imminent need for turning such huge data into meaningful information and knowledge. The data collected from various applications require a proper data mining technique to extract the knowledge from large repositories for decision making. Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large volume of data [1].

Data mining and knowledge discovery in databases are treated as synonyms, but data mining is actually a step in the process of knowledge discovery.

The main functionality of data mining techniques is applying various methods and algorithms in order to discover and extract patterns of stored data. These interesting patterns are presented to the user and may be stored as new knowledge in knowledge base. Data mining and knowledge discovery applications have got a rich focus due to its significance in decision making. Data mining has been used in areas such as database systems, data warehousing, statistics, machine learning, data visualization, and information retrieval. Data mining techniques[21],[22] have been introduced to the following areas including neural networks, patterns recognition, spatial data analysis, image databases and many application fields such as business, economics and bioinformatics, education, banking. fraud detection, etc. The main objective of this paper is to predict the side effect of drug usage by patients taking into consideration their blood group and genotype, by application of classifiers to biological data from the Health Centre of Ladoke Akintola University of Technology, Ogbomoso, Oyo State, Nigeria.

.

2. REVIEW OF BIOLOGICAL DATA MINING

2.1 Data Mining Unfairly Exploits Patient-Physician and Patient-Pharmacist Relationships

If data mining is problematic because it intensifies the problems created by drug detailing, why not regulate drug detailing directly? There are two important reasons that can explain why the legislative efforts have focused on data mining.

First, pharmaceutical companies manufacture legal — indeed socially important — products, and they are entitled to cultivate potential customers of those products. The Supreme Court has recognized this interest through a first amendment right for businesses to advertise and solicit clients for their goods and services (the "commercial speech" doctrine). If marketing activities have harmful effects, first amendment principle instructs society to counter the harmful effects with counter-speech, not by prohibiting pharmaceutical companies from promoting their drugs[2].Indeed, some medical schools and health care organizations have done exactly that. By using "academic detailing," universities, professional societies and others can encourage physicians to base their prescribing on medical evidence rather than drug company promotions [3]

Legislative efforts have focused on data mining also because data mining directly invades the interests of patients in a way that drug detailing does not. In particular, it involves an expropriation of information created in the privacy of patient-physician and patient-pharmacist relationships [4].

2.2 Confidentiality Interests of Patients

Information about a patient's health is highly sensitive. It can cause embarrassment and result in stigmatization and discrimination. Consider in this regard the implications when family, friends, acquaintances, or employers find out that a person has a drug abuse problem, a sexually-transmitted disease, a mental illness, or a cancer. While a prescription may provide only indirect evidence of a patient's health, it can provide fairly clear evidence of illness. If a patient fills prescriptions for efavirenz (Sustiva) and tenofovir/emtricitabine (Truvada), people can readily conclude that the patient is being treated for an HIV infection [5].If a patient fills a prescription for olanzapine (Zyprexa), others can reasonably suspect that the patient is being treated for mental illness [6].

Physicians, pharmacists and other health care professionals therefore promise strict rules of confidentiality. Indeed, the duty to protect patient confidentiality has been a hallmark of medical codes of ethics throughout history [7]. Moreover, the government reinforces ethical principles of confidentiality with legal safeguards, including the Health Information Portability and Accountability Act (HIPAA) and state law provisions [8].While concerns about patient confidentiality have been voiced about data mining[9] data mining companies strip their records of information that can identify patients and indeed are required to do so by HIPAA[10].

Advocates of data mining laws have invoked the privacy interests of physicians, but identifying the prescribing practices of physicians does not entail disclosure of the kind of sensitive information that privacy safeguards are designed to protect. Accordingly, even when a federal court of appeals upheld New Hampshire's data mining law, it did not invoke the privacy concerns of physicians in doing so[11].

2.3 Property Interests of Patients

As the entire data mining enterprise reflects, information can have substantial economic value. Data mining companies pay pharmacies for the right to extract information about prescriptions, and pharmaceutical companies pay the mining companies for information about individual physicians' prescribing practices.

This use of prescription information for economic purposes has troubling implications for patients. In effect, data mining companies exploit the relationships between patients and their physicians or pharmacists for the pecuniary benefit of pharmacies, pharmaceutical manufacturers, and the mining companies themselves. To be sure, the practice of medicine regularly entails the realization of profit by pharmacies and pharmaceutical manufacturers, as well as physicians and hospitals. But ethical principle justifies the ability of health care providers and companies to profit from patient care because economic incentives are necessary to ensure high quality care. Capable people will not pursue the practice of medicine or pharmacy if they are not compensated for doing so, and industry will not develop new therapies if they are not compensated for doing so.

It is difficult, however, to justify the mining of prescription data in terms of the interests of patients. As discussed above, drug detailing encourages physicians to prescribe a drug even when scientific evidence indicates that the prescription is not desirable. Patients may receive a drug when one is not needed, they may all tables and figures will be processed as images. You need to embed the images in the paper itself. Please don’t send the images as separate files.

2.4 Professional Response to Data Mining

All of these considerations suggest that as an ethical matter, physicians should eliminate the incentive for data mining by refusing to meet with drug company sales representatives[12].Medical schools do not permit sales representatives to participate in the instruction of students; physicians also should not turn to sales representatives for their post-graduate education.Even if they continue to meet with sales representatives, physicians can prevent the representatives from using information about their prescribing practices from the data mining companies. The American Medical Association established the Physician Data Restriction Program in May 2006, and any physician can opt out of data mining by registering with the Program [13]. For registered physicians, pharmaceutical companies still have access to prescriber data for marketing and research purposes and also for making compensation decisions for their employees. However, the companies agree to withhold individual prescriber information from their sales representatives [14]. Pharmacy companies also should take steps to prevent prescription information from being used to enhance drug detailing. To comply with state and federal privacy laws, pharmacies require patient identifiable information to be stripped from prescription records before the records are retrieved by health information organizations [15]. The pharmacies also should require physician-identifiable information to be stripped.

2.5 Legal Response to Data Mining

While professional self-regulation can increase adherence to ethical norms, legal mandates often are necessary to ensure that professionals meet their moral obligations [16]. Accordingly, it is important to consider the role of legislative action to regulate data mining.The empirical evidence suggests that legislation may be needed to prevent the use of prescription information for drug detailing activities. Through April 2009, the AMA’s Physician Data Restriction Program had enrolled 22,000 of roughly 650,000 actively prescribing physicians, or less than 4 percent of those who can enroll [17]. According to Dr. Robert Musacchio, who oversees the Program, the AMA has strongly promoted the opt-out option, but few physicians have demonstrated interest [18].This supports the view of the courts that physician privacy is not a serious concern with data mining[19]. However, it also may reflect the fact that physicians tend not to appreciate the extent to which they are influenced by the promotional activities of sales representatives [20]. In any event, while it makes sense to rely on physicians to protect their own privacy interests, they should not have sole authority for protecting the privacy interests of patients.

3. TABLES

TABLE 1: CHLOROQUINE AND ARTESURNATE USAGE

	PRECISION		FPR		CE	
Symptoms	KNN	CART	KNN	CART	KNN	CART
Headache	0.875	0.986	0.764	0.876	0.915	0.892
Nausea	0.653	0.785	0.956	0.945	0.879	0.565
Vomitting	0.785	0.478	0.436	0.786	0.765	0.876
Itching	0.457	0.764	0.987	0.764	0.876	0.875
A.P	0.775	0.786	0.975	0.786	0.923	0.876
H.P	0.897	0.764	0.876	0.965	0.872	0.643
Diarrohea	0.674	0.721	0.853	0.764	0.987	0.786
Blurring	0.986	0.765	0.876	0.876	0.864	0.876

TABLE 2: FANSIDAR AND ARTESURNATE USAGE

	PRECISION		FPR		CE	
Symptoms	KNN	CART	KNN	CART	KNN	CART
Headache	0.877	0.786	0.864	0.473	0.615	0.591
Nausea	0.453	0.387	0.856	0.745	0.871	0.573
Vomitting	0.684	0.576	0.836	0.481	0.163	0.875

Itching	0.659	0.764	0.987	0.774	0.276	0.477
A.P	0.575	0.785	0.932	0.886	0.623	0.279
H.P	0.897	0.561	0.874	0.867	0.677	0.843
Diarroh ea	0.674	0.621	0.853	0.564	0.687	0.384
Blurring	0.886	0.765	0.966	0.779	0.567	0.675

TABLE 3: QUININE USAGE

	PRECISION		FPR		CE	
Symptoms	KNN	CART	KNN	CART	KNN	CART
Headache	0.478	0.754	0.764	0.231	0.976	0.453
Nausea	0.865	0.564	0.764	0.756	0.994	0.621
Vomitting	0.731	0.264	0.436	0.481	0.969	0.578
Itching	0.943	0.854	0.180	0.766	0.979	0.879
A.P	0.865	0.888	0.835	0.986	0.929	0.972
H.P	0.994	0.762	0.984	0.868	0.777	0.748
Diarroh ea	0.694	0.926	0.657	0.664	0.987	0.984
Blurring	0.986	0.765	0.966	0.979	0.867	0.975

TABLE 4: CAMOQUINE USAGE

	PRECISION		FPR		CE	
Symptoms	KNN	CART	KNN	CART	KNN	CART
Headache	0.675	0.832	0.465	0.438	0.879	0.959
Nausea	0.767	0.668	0.769	0.859	0.898	0.827
Vomitting	0.938	0.868	0.839	0.480	0.796	0.854
Itching	0.943	0.854	0.180	0.766	0.979	0.879
A.P	0.768	0.987	0.836	0.788	0.788	0.778
H.P	0.798	0.869	0.987	0.969	0.979	0.749
Diarroh ea	0.978	0.827	0.856	0.984	0.889	0.685
Blurring	0.986	0.765	0.966	0.979	0.867	0.975

TABLE 5: QUININE AND ARTESURNATE USAGE

	PRECISION		FPR		CE	
Symptoms	KNN	CART	KNN	CART	KNN	CART
Headache	0.875	0.843	0.868	0.921	0.876	0.858
Nausea	0.969	0.777	0.966	0.876	0.195	0.628
Vomitting	0.831	0.824	0.637	0.243	0.667	0.678
Itching	0.949	0.923	0.782	0.466	0.447	0.979
A.P	0.467	0.218	0.938	0.586	0.679	0.772
H.P	0.697	0.724	0.784	0.768	0.457	0.849
Diarroh ea	0.884	0.763	0.747	0.964	0.677	0.989
Blurring	0.887	0.823	0.396	0.978	0.967	0.575

FPR= FALSE POSITIVE RATE CE= CLASSIFICATION ERROR

CART=Classification and Regression Tree A.P=Abdominal Pain H.P= Hearing Problems

4. EQUATIONS

$$d(x_1 , x_2) = \sqrt{\sum_{i=1}^{m}} \ (f_i(x_1) - \sqrt{\sum_{i=1}^{m}} (f_i (x_2) \quad (1)$$

$$F(T) = \sum_i \alpha_i \ y_i (X_i \ .T) + \ b \quad (2)$$

Equation (1) above is for K-Nearest Neighbor and the function of the parameters is stated below.The equation (1) is obtained from the standard Euclidean distance.
Where, m= symptoms to describe the drug used (x)
f_i(x) = denoted the values of the feature (i=H.P, A.P,…m)

Equation (2) above is for Support Vector Machine and the parameters are explained below:

Xi=supporting attributes
y_i= are the class labels (which are greater ≤ 1)
T= Testing sets
(Xi.T) = Is the dot product of the test sample T with one of the support vector Xi
αi and b are string parameters.

5. RESULTS AND DISCUSSIONS

The most important step of the research is that it generated a new feature space (target variable) under which the original sequence was transformed to the format in which the WEKA tool can be easily applied.

To strictly compare with the results presented in [Table1, 2, 3, 4, 5] we conducted the same 3-fold cross validation. Tables showed our results for KNN and CART, using the feature selection by the entropy-based algorithm which is the CART and obtaining the nearest feature by the KNN**.** The precision of KNN and CART was 76% and 75.6% respectively in the predicting for any of the symptoms when the patient used chloroquine and artesurnate in treating malaria. Also, the result revealed that KNN has the highest precision of 98.6% for blurring and CART has highest precision of 98% for headache in taking any of the anti malaria drugs used in this research work. Have the lowest precision of 45.3% for nausea for KNN and the lowest precision of 21.8% for abdominal pain when taking any of the anti malaria drugs used in this research work.

Secondly, the precision of KNN and CART was 71.3% and 65.5% respectively in the predicting for any of the symptoms when the patient used fansidar and artesurnate in treating malaria. Also, the result revealed that KNN has the highest precision of 88.6% for blurring and the lowest precision of 45.3% for nusea when taking fansidar and artesunate.

Thirdly, the precision of KNN and CART was 81.9% and 72.2% respectively in the predicting for any of the symptoms when the patient used quinine in treating malaria. Also, the result revealed that KNN has the highest precision of 98.6% for blurring and the lowest precision of 47.8% for headache when taking quinine.

Fourthly, the precision of KNN and CART was 85.6% and 75.3% respectively in the predicting for any of the symptoms when the patient used camoqunine in treating malaria. Also, the result revealed that KNN has the highest precision of 98.6% for blurring and the lowest precision of 67.5% for headache when taking camoqunine.

Finally, the precision of KNN and CART was 81.9% and 73.6% respectively in the predicting for any of the symptoms when the patient used quinine and artesurnate in treating malaria. Also, the result revealed that KNN has the highest precision of 96.9% for nausea and the lowest precision of 45.3% for abdominal pain when taking qunine and artesunate. The results revealed the following similarities and difference in the performance of the both classifiers which are stated below:

KNN was able to predict for five symptoms namely: blurring, itching, nausea, headache and abdominal pain.

Also, CART was able to predict for five symptoms namely headache, vomiting, nausea, diarrhoea and abdominal pain regardless of the anti-malaria drugs the patient used.

The inability of KNN to predict for vomiting, diarrhoea and hearing problem is simply because just 10patients were statistical analyzed to have these symptoms from both the training set and the test set (dataset) therefore led to higher classification error of 70percent on the average.

CART inability to predict for blurring, itching and hearing problem is simply because the dataset contain 12 patients that have these symptoms which led to 78percent classification error for the classifier.

6. CONCLUSION AND FUTURE WORK

KNN and CART classifiers perform with similar overall accuracies and similar classification error on the clinical dataset, but they differ with respect to the symptoms. Both algorithms do not successfully predict all the symptoms which might be explained with the uneven construction of the target variable. Further research efforts will be directed at achieving higher accuracy of the classifiers' prediction by additional transformations of the dataset, reconstruction of the target variable, tuning of the classification algorithms' parameters, use of other data mining tools. Also, efforts should be directed towards achieving a lower classification error and researchers should correlate how biological data such as genotype and blood group can be use to predict the symptoms of the anti-malaria drugs used in this research work.

REFERENCES

[1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth,"From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, pp. 37-54, 1997.

[2] *New York Times v. Sullivan*, 376 U.S. 254, 270 (1964) (observing that "the fitting remedy for evil counsels is good ones").

[3] S. D. Graham, "Effect of an Academic Detailing Intervention on the Utilization Rate of Cyclooxygenase-2 Inhibitors in the Elderly," *Annals of Pharmacotherapy* 42, no. 6 (2008): 749- 756; S. B. Soumerai and J. Avorn, "Principles of Educational Outreach ('Academic Detailing') to Improve Clinical Decision Making," *JAMA* 263, no. 4 (1990): 549-556.

[4]Had the Supreme Court ruled in favor of federal pre-emption in *Wyeth v. Levine*, 129 S. Ct. 1187 (2009), that might have given state legislatures an additional reason to eschew direct regulation of drug detailing. In *Wyeth*, the Court rejected a claim that compliance with Food and Drug Administration (FDA) regulations should insulate pharmaceutical manufacturers from state tort law claims. *Id*., at 1190.

[5]Panel on Antiretroviral Guidelines for Adults and Adolescents,*Guidelines for the Use of Antiretroviral Agents in HIV- 1-infected Adults and Adolescents*, Washington D.C., Department of Health and Human Services, November 3, 2008,

[6] S. Dando and M. Tohen, "Olanzapine - Relapse Prevention Following Mania," *Journal of Psychopharmacology* 20, no. 2 Suppl. (2006): 31-38

[7] D. Orentlicher, "Genetic Privacy in the Patient-Physician Relatio ship," in M. Rothstein, ed., *Genetic Secrets: Protecting Privacy and Confidentiality in the Genetic Era* (New Haven: Yale University Press 1997)77-91, 77-78.

[8] M. A. Hall, M. A. Bobinski, and D. Orentlicher, *Health Care Law and Ethics*, 7th ed. (New York: Aspen Publishers, 2007): at 175-185.

[9] See Klocke, *supra* note 7, at 518-521; IMS Health, Inc., 490 F. Supp. 2d at 171.

[10] See Greene, *supra* note 4, at 747; Steinbrook, *supra* note 11, at 2746.

.[11] *IMS Health, Inc. v. Ayotte*, 550 F.3d 42, 55 (1st Cir. 2008).

[12] IMS Health Corp., 532 F. Supp. 2d at 163.

[13] The author relied on *The Medical Letter* during his years as

[14] See Brody, *supra* note 16.

[15] See Greene, *supra* note 4, at 742.

[16] R. A. Musacchio and R. J. Hunkler, "More Than a Game of Keep Away," *Pharmaceutical Executive*, May 1, 2006

[17] IMS Health, Inc., 490 F. Supp. 2d at 166.

[18]D. Orentlicher, "The Influence of a Professional Organization on Physician Behavior," *Albany Law Review* 57, no. 3 (1994): 583-605.

[19] Personal communication with Mark Frankel, American Medical Association (April 26, 2009) (reporting the 22,000 figure for enrollment); Greene, *supra* note 4, at 746 (estimating at 650,000 the number of physicians who actively prescribe drugs).

[20] Personal communication, *supra* note 11.

[21] Ayinde A.Q, Dr Adetunji A.B, Odeniyi O.A and Bello. M: Performance Evaluation of Naive Bayes and Decision Stump Algorithms in Mining Students' Educational Data (2013): IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No 1, July 2013 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784

[22] Q. A. AI-Radaideh, E. M. AI-Shawakfa, and M. I. AI-Najjar, "Mining Student Data using Decision Trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

- *Dr A.B Adetunji is a Senior Lecturer at the Department of Computer Science and Engineering Ladoke Akintola University of Technology, Nigeria.Her research area include Database, Artificial Intelligence and Data Mining08034858047. E-mail: abadetunji@yahoo.com*
- *Ayinde A.Q is currently pursuing masters degree program in Computer Science at the Department of Computer Scence and Engineering,LAUTECH, Nigeria. His research area include Data mining, Data base and ICT PH*

- *O.A Odeniyi obtained his B.Tech (Computer Science) from LAUTECH (1996). He held Post Graduate* Diploma *(PGD) in Education from National Teachers' Institute Kaduna (2006). He is presently a research student at the Department of Computer Science and Engineering, LAUTECH. He is a Lecturer I at Osun State College of Technology EsaOke. He is a member of Computer Professionals Registration Council of Nigeria (CPN) and Nigeria Computer Society (NCS).His research areas are soft computing, ICT, Data mining and Database.*

- *J.A ADEWALE lectures in the Department of Computer Science, Osun State College of Technology, Esa-Oke. He has B. Tech (Hons.) degree in Computer Science from Ladoke Akintola University of Technology, Ogbomoso and M. Sc. in Multimedia Applications and Virtual Environments from the University of Sussex, Brighton, United Kingdom. He also holds PGDE from the National Teachers' Institute (NTI).His major areas of expertise are in Multimedia Application, Computer Graphics, Modeling and Rendering, Animation, Human Computer Interaction and Web Development. He is a lecturer with many years experience.*