PREDICTIVE ANALYTICS IN HEALTHCARE SYSTEM USING DATA MINING TECHNIQUES

Basma Boukenze[1*], Hajar Mousannif[2] and Abdelkrim Haqiq[3]

[1]Computer, Networks, Mobility and Modeling laboratory
FST, Hassan 1st University, Settat, Morocco
basma.boukenze@gmail.com
[2]LISI Laboratory,FSSM Cadi Ayyad University,Marrakech 40000,Morocco
mousannif@uca.ac.ma
[3]Computer, Networks, Mobility and Modeling laboratory
FST, Hassan 1st University, Settat, Morocco
e-NGN Research Group, Africa and Middle East
ahaqiq@gmail.com

ABSTRACT

The health sector has witnessed a great evolution following the development of new computer technologies, and that pushed this area to produce more medical data, which gave birth to multiple fields of research. Many efforts are done to cope with the explosion of medical data on one hand, and to obtain useful knowledge from it on the other hand. This prompted researchers to apply all the technical innovations like big data analytics, predictive analytics, machine learning and learning algorithms in order to extract useful knowledge and help in making decisions. With the promises of predictive analytics in big data, and the use of machine learning algorithms, predicting future is no longer a difficult task, especially for medicine because predicting diseases and anticipating the cure became possible. In this paper we will present an overview on the evolution of big data in healthcare system, and we will apply a learning algorithm on a set of medical data. The objective is to predict chronic kidney diseases by using Decision Tree (C4.5) algorithm.

KEYWORDS

big data, big data analytics, machine learning, Healthcare, learning algorithm, C4.5

1. INTRODUCTION

There is an almost universal definition shared with proponents of the ideology of big data, it is that "Big Data sets a situation in which data sets have increased at such huge sizes that conventional technologies of information, can no longer manage them effectively, either the size or the extent and the growth of the data set" [1].

The world has become submerged by a large amount of data. Every moment is equivalent to the generation of thousands data. All sectors and all their activities are involved due to digitization,

David C. Wyld et al. (Eds) : DBDM, CICS, CSIP, AI&FL, SCOM, CSE, CCNET-2016
pp. 01–09, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60501

the introduction of information technology as an effective tool, and the Internet which is becoming a very important user interface for daily interactions [2]. However, these generated data become more and more difficult to manage in terms of volume, variety and velocity [3]. This gave birth to a new domain named big data. In 2008, Gartner used for the first time the term "Big Data", in reference to the explosion of digital data and quoted it will impact the way we work [4].

"Big Data" and "analysis of big data" are inseparable. This reflects the common opinion that "Big data" does not refer to the problem of information overload, but refers also to the analytical tools used to manage the flow of data and transform the flood in a source of useful information.

The medical field has its great contribution in this deluge of data because of some technological innovations in the field like cloud computing which has relocated the tests of care beyond the four walls of the hospital, and has made them available anywhere and anytime [5], laparoscopic surgery and robotic surgery, which replaced classical surgery [6], also smart homes which allow patients self-care and monitoring using simple devices that deliver results on specific physiological conditions. There are also smart applications or software that can analyze the body signals using integrated sensors with the aim of monitoring [7], as well as mHealth technologies that support new methods of biological, behavioural and environmental data collection. These include sensors that monitor the phenomena with a higher accuracy [8].

All these innovations participated to the explosion of medical data, by multiplying data sources and electronic medical records containing diagnostic Images, lab results, and biometric information that are generated and stored [8.9.10].

Researchers have deduced that this explosion of medical data has the potential to improve clinical decisions at the point of care. Doctor will become able to extract relevant knowledge for each patient, which gives better decisions and results [11].

On most of this, the term "analyzing medical data" and "predictive analytics" in Google Trends showed an impressive growth of interest from 2011 [12], because the process of analysis in the medical sector does not stop just at the level of the ability to manage large databases, but it exceeds this to the ability to retrieve future knowledge, which is encouraged by many researchers and experts. Seen that an analysis of the big data shows itself as the only solution able to solve all the problems of the medical sector [13] by:

- ✓ Providing best Service
- ✓ Monitoring quality in hospital
- ✓ Improving treatment processes
- ✓ Detection of disease earlier

There are many algorithms for classification and prediction applied to predict the most killer diseases like breast cancer, heart disease, motor neuron, and diabetes. In this present paper, we apply a decision tree classifier (C4.5) [14], which is among the most influential data mining algorithm in the research community and among the top 10 data mining algorithms. Our aim is to predict chronic kidney disease by this learning algorithm.

The rest of this paper is organized as follows:

- Section 2 is about related work.
- Section 3 presents the context of the experiment, metrics and research hypothesis.
- Section 4 presents Experimental results
- Section 5 discusses the results

Finally, section 6 concludes the paper.

2. RELATED WORKS

Andrew Kusiak et al [15] have used data pre-processing, data transformations, and a data mining approach to elicit knowledge about the interaction between many of measured parameters and patient survival. Two different data mining algorithms were engaged for extracting knowledge in the form of decision rules. Those rules were used by a decision-making algorithm, which predicts survival of new unseen patients. Important parameters identified by data mining were interpreted for their medical significance. They have introduced a concept in their research work have been applied and tested using collected data at four dialysis sites. The approach presented in their paper reduces the cost and effort of selecting patients for clinical studies. Patients can be chosen based on the prediction results and the most important parameters discovered.

Abhishek et.al [16] have used two neural network techniques, Back Propagation Algorithm (BPA), Radial Basis Function (RBF) and one non-linear classifier Support Vector Machine (SVM) and compared in accordance with their efficiency and accuracy. They used WEKA 3.6.5 tool for implementation to find the best technique among the above three algorithms for Kidney Stone Diagnosis. The main purpose of their thesis work was to propose the best tool for medical diagnosis, like kidney stone identification, to reduce the diagnosis time and improve the efficiency and accuracy. From the experimental results they concluded, the back propagation (BPA) significantly improved the conventional classification technique for use in medical field.

Andrew Kusiak et al [17] have used data pre-processing, data transformations, and data mining approach to elicit knowledge about the interaction between many of measured parameters and patient survival. Two different data mining algorithms were employed for extracting knowledge in the form of decision rules. Those rules were used by a decision-making algorithm, which predicts survival of new unseen patients. Important parameters identified by data mining were interpreted for their medical significance. They have introduced a new concept in their research work, it has been applied and tested using collected data at four dialysis sites. The approach presented in their paper reduced the cost and effort of selecting patients for clinical studies. Patients can be chosen based on the prediction results and the most significant parameters discovered.

Ashfaq Ahmed K et.al, [18] have presented a work using machine learning techniques, namely Support Vector Machine [SVM] and Random Forest [RF]. These were used to study, classify and compare cancer, liver and heart disease data sets with varying kernels and kernel parameters. Results of Random Forest and Support Vector Machines were compared for different data sets such as breast cancer disease dataset, liver disease dataset and heart disease dataset. The results with different kernels were tuned with proper parameter selection. Results were better analyzed to establish better learning techniques for predictions. It is concluded that varying results were observed with SVM classification technique with different kernel functions.

Sadik Kara et.al [19] had concentrated on the diagnosis of optic nerve disease through the analysis of pattern electroretinography (PERG) signals with the help of artificial neural network (ANN). Implemented Multilayer feed forward ANN trained with a Levenberg Marquart (LM) back propagation algorithm. The end results were classified as healthy and diseased. The stated results shown that the proposed method PERG could make an effective interpretation.

With respect to all related work mentioned above, our work is predicting disease using chronic kidney failure datasets by C4.5 algorithm.

3. EXPERIMENT

In this work, we will apply C4.5, a learning algorithm that will make classification and prediction on a database to extract knowledge and classify patients into two categories: chronic kidney disease (ckd) and not chronic kidney disease (notckd).

3.1 Experiment environment:

In this study, we use the Waikato Environment for Knowledge Analysis (Weka). It is a comprehensive suite of Java class libraries that implement many algorithms for data mining clustering, classification, regression, analysis of results. This platform gives to researchers a perfect environment to implement and evaluate their classification model comparing to TANAGRA or ORANGE **[20]**

3.2 Chronic kidney disease dataset

We used the database Chronic Kidney Disease Dataset from UCI Machine Learning Repository **[21]**. This database contains 400 instances and 24 integer attributes, two class (chronic kidney disease (ckd), not chronic kidney disease (notckd)). Table 1 describes the attributes of the database, while Table 2 describes the distribution of classes.

Table1 : Information attribute

Attribute	**Representation**	**Information attribute**	**Description**
Age	Age	Numerical	Years
Blood pressure	Bp	Numerical	Mm/Hg
Specific gravity	Sg	Nominal	1.005,1.010,1.015,1.020,1.025
Albumin	Al	Nominal	0.1.2.3.4.5
Sugar	Su	Nominal	0.1.2.3.4.5
Red blood cells	Rbc	Nominal	Normal, abnormal
Pus cell	Pc	Nominal	Normal, abnormal
Pus cell clumps	Pcc	Nominal	Present, notpresent
Bacteria	Ba	Nominal	Present, notpresent
Blood glucose random	Bgr	Numerical	Mgs/dl
Blood urea	Bu	Numerical	Mgs/dl
Serum creatinin	Sc	Numerical	Mgs/dl
Sodium	Sod	Numerical	mEq/L
Potassium	Pot	Numerical	mEq/L

Haemoglobin	Hemo	Numerical	Gms
Packed cell volume	Pcv	Numerical	
White blood cell count	Wc	Numerical	Cells/cumm
Red blood cell count	Rc	Numerical	Millions/cmm
Hypertension	Htn	Nominal	Yes, no
Diabetes mellitus	Dm	Nominal	Yes, no
Coronary artery disease	Cad	Nominal	Yes, no
Appetite	Appet	Nominal	Good, poor
Pedal edema	Pe	Nominal	Yes, no
Anemia	Ane	Nominal	Yes, no
Class	Classe	Nominal	Ckd notckd

Table 2 : Class distribution

	Class	Distribution
1	Ckd	250 (62.5%)
2	Notckd	150 (37.5%)

3.3 Metrics and research hypotheses

To understand classifier's behaviour we should calculate those metrics, we use hypothesis below:

- True positive (TP) is the number of positive samples correctly predicted.
- True negative (TN) is the number of negative samples correctly predicted
- False negative (FN) is the number of positive samples wrongly predicted.
- False positive (FP) is the number of negative samples wrongly predicted as positive.

Table 3. Metric and research hypotheses

Metric	Description	Formula	
Accuracy	Number of correct predictions from all predictions made.	$\frac{TP + TN}{TP + FP + TN + FN}$	(1)
Sensitivity	Proportion of positives predictions that are correctly identified.	$\frac{TP}{TP + FN}$	(2)
Specificity	Proportion of negatives predictions that are correctly identified	$\frac{TN}{FP + TN}$	(3)
Precision	Positive predictive values	$\frac{TP}{TP + FP}$	(4)
Mean Absolute Error (MAE)	Comparison between forecasts or predictions and the eventual outcomes	$\frac{FP + FN}{\mathrm{TP + FP + TN + FN}}$	(5)
F-measure	Combination of precision and recall.	$\frac{2 * \mathrm{Precision} * \mathrm{Sensitivity}}{\mathrm{Precision} + \mathrm{Sensitivity}}$	(6)

Another important metric Confusion Matrix is taken into account. It is a visualization tool which is commonly used to present the accuracy of the classifiers in classification. The columns represent the predictions, and the rows represent the actual class as shown in Table 4

Table 4. Confusion matrix description.

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

4. EXPERIMENTAL RESULTS

In order to apply our classifier and evaluate its performance, we apply the 10-fold cross validation test which is a technique that splits the original set into a training sample to train the model, and a test set to evaluate it. After applying the pre-processing and preparation methods, we try to analyse the data visually and figure out the distribution of values in terms of performance and accuracy of the model.

Table 5 : C4.5 performance

Evaluation criteria	C4.5
Time to build model (s)	0.08
Correctly classified instances	396
Incorrectly classified instance	4
Accuracy	63%
Error	0.37

Table 6 : Simulation error

Evaluation criteria	C4.5
Kappa statistic	0.97
Mean absolute error	0.02
Root mean squared error	0.08
Relative absolute error %	4.79
Root relative squared error %	16.66

Table 7 : Accuracy measures by class

	TP	FP	precision	recall	F-measure	Class
C4.5	0.99 0.98	0.02 0.004	0.98 0.99	0.99 0.98	0.99 0.98	Ckd Notckd

Table 8 : Diffusion Matrix

	Ckd	NotCkd	
C4.5 (J48)	249 3	1 147	**Ckd** **Notckd**

5. DISCUSSION

we can say that the C4.5 classifier is powerful, according to the number of correctly classified instances (396) and just 4 instances misclassified, This is mentioned with a low error rate (0.37) (see Table 5) . We can also deduce that this algorithm is excellent because of the value of KS = 0.97, which reflects the performance and accuracy of the classifier (see table 6). Fromtable 7, we deduce that C4.5 achieves best results regarding precision (0.98 ckd and 0.99 notckd), and sensitivity or (recall) (0.99ckd and 0.98 notckd).

C4.5 has proved its performance as a powerful classifier in term of accuracy and the minimum execution time, which makes it a good classifier to be used in the medical field for classification and prediction.

6. CONCLUSION

As conclusion, the application of data mining techniques for predictive analysis is very important in the health field because it gives us the power to face diseases earlier that threaten the human being; child, young and old people, through the anticipation of cure and helping in decision-making. In this work we used a learning algorithm C4.5 to predict patients with chronic kidney failure disease (ckd), and patients who are not suffering from this disease (notckd). The classifier used proved its performance in predicting with best results in terms of accuracy and minimum execution time.

This is a major challenge in the medical field and pushes us to increase our efforts to develop machine learning methods, to exploit information intelligently and extract the best knowledge.

REFERENCES

[1] franck Ohlhorst, January 2013 ' Big Data Analytics: Turning Big Data into Big Money', ISBN: 978-1-118-14759-7, pp 176 .

[2] Samson Oluwaseun, F., Serdar , S., and Vanduhe,V ., (2014)," Advancing big data for humanitarian needs ", Procedia Engineering, vol . 78,N., pp 88-95

[3] Amir, G., Murtaza, H., (2015),"Beyond the hype: Big data concepts, methods, and analytics», International journal of Information Management, vol . ,pp 137-144.

[4] H., Chen, H. L., Chiang, C., Storey, (2012), 'BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT', MIS Quarterly, Vol. 36 ,No. 4, pp. 1165-1188.

[5] Jonathan Northover, Brian McVeigh, Sharat Krishnagiri. Healthcare in the cloud: the opportunity and the challenge. MLD.
Available at http://www.sunquestinfo.com/images/uploads/CMS/445/mlo_02-12014_ healthcare_in_the_cloud.pdf

[6] Gabriel I. Barbas, Sherry A. Glied, (2010),'' New Technology and Health Care Costs — The Case of Robot-Assisted Surgery''; the new England journal of medicine , N°, 363, pp 707-704 . Available at http://www.nejm.org/doi/full/10.1056/NEJMp1006602

[7] Marianthi Theoharidou, Nikos Tsalis, ''Smart Home Solutions for Healthcare: Privacy in Ubiquitous Computing Infrastructures''.
Available online at http://www.cis.aueb.gr/Publications/Smart%20Home%20-%20Site%20TR.pdf

[8] Steve G. Peters, James D. Buntrock,(2014),''Big Data and the Electronic Health Record'', Ambulatory Care Manage , Vol. 37, No. 3, pp. 206–210

[9] R. Weil, (2014),'' Big Data In Health: A New Era For Research And Patient Care Alan R. Weil'', Health Affair, Vol. 33, N° 7, pp 1110.

[10] Peter Groves; Basel Kayyali, (2013),'' The 'big data' revolution in healthcare'', McKinsy and Company. Center for US Health System Reform Business Technology Office. Available at http ://digitalstrategy.nl/wp-content/uploads/E2-2013.04-The-big-data-revolution-in-US-health-care-Accelerating-value-and-innovation.pdf.

[11] T., Huang, L., Lan, (2015), ''Promises and Challenges of Big Data Computing in Health Sciences'', Big Data Research vol. 2, pp 2-11 available at http://dx.doi.org/10.1016/j.bdr.2015.02.002

[12] Khurshid R., G., Kai, Z., John T., W., and Charles P., F., (2014), ''Harnessing Big Data for Health Care and Research Are Urologists Ready? ", Journal of European Urology, vol. N., pp 1-3

[13] Wullianallur Raghupathi,Viju Raghupathi, (2014),''Big data analytics in healthcare: promise and Potential'', Health Information Science and Systems.
Available at http://www.biomedcentral.com/content/pdf/2047-2501-2-3.pdf

[14] Rashedur M. Rahman, Fazle Rabbi Md. Hasan 'Using and comparing different decision tree classification techniques for mining ICDDR, B Hospital Surveillance data'', Elsevier, Vol. 38, pp 11421–11436

[15] AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005),'' Predicting survival time for kidney dialysis patients: a data mining approach'', Elsevier Publication, Computers in Biology and Medicine, Vol. 35, pp 311–327

[16] Abhishek, Gour Sundar Mitra Thakur, Dolly Gupta, (2012) "Proposing Efficient Neural Network Training Model for Kidney Stone Diagnosis", International Journal of Computer Science and Information Technologies, Vol. 3 (3), pp 3900-3904

[17] AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005) “Predicting survival time for kidney dialysis patients: a data mining approach”, Elsevier Publication, Computers in Biology and Medicine ,Vol.35, pp 311–327

[18] Ashfaq Ahmed K, Sultan Aljahdali and Syed Naimatullah Hussain, (2013) “Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques”, International Journal of Computer Applications Vol. 69, No.11, pp 12-16

[19] Sadik Kara, Aysegul Guvenb, Ayse OztUrk Onerc, (2006) “Utilization of artificial neural networks in the diagnosis of optic nerve diseases”, Elsevier Publication, Computers in Biology and Medicine,Vol. 36, pp 428–437

[20] M Hall, E Frank, G Holmes, B Pfahringer,(2009), ‘The WEKA data mining software: an update’, Volume 11, Issue 1, pp 10-18

[21] “UCI Machine Learning Repository: Kidney failure Data Set [Online].
Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease#

AUTHOR

Basma BOUKENZE, PhD student at the Faculty of Science and Technology in Settat, Morroco. after obtaining a degree in computer Genie in 2009, and a master degree in engineering network and System in 2011, has continued studies and currently registered in doctoral science and technology formation at the Faculty of Science and technology Settat, Moroco, member of The Mathematical Research structure and applied computing. Laboratory Computer Networks, Mobility and modelling.