IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 11, November 2015.
www.ijiset.com
ISSN 2348 – 7968

A Model for Predicting Ischemic Stroke Using Data Mining Algorithms

BALAR Khalid[1] and NAJI Abdelwahab[2]

[1] Department of Computer Science, Hassan II University-Faculty of Medicine
Casablanca, Morocco

[2] Department of Computer Science, Hassan II University-ENSAT
Mohammedia, Morocco

Abstract

Data mining appears in the middle of the 1990s in the United States as a new discipline at the interface of Statistics and Information Technology: databases, artificial intelligence, machine learning. Along with advanced researches in health Domain monstrous of data are available, but the main difficulty is how to cultivate the existing information into a useful practices. Data mining has a great potential to enable healthcare systems to use data more effectively. Hence, DM improves care and reduces costs. In this paper, we used the various Data Mining techniques such as classification, logistic regression in health domain to predict Ischemic Stroke.

Keywords: *Data Mining, Classification, Logistic regression, Ischemic Stroke.*

1. Introduction

In the early seventies, it was very expensive to store data or information. But due to the advancement in the field of information gathering and Internet tools, we saw huge amount of information and data are available in electronic format. David Hand (1998) gives the following definition: consists in the discovery of interesting, unexpected, or valuable structures in large data sets [1]. To store this amount of data or information databases sizes are increased very rapidly. This information may be very useful for decision-making process in any field. It becomes possible with the help of data mining or Knowledge Discovery in Databases. Data mining is the process of extracting the useful information from a large collection of data which was previously unknown [2]. A number of relationships are hidden among such a large collection of data [3].

With the help of figure 1 five stages are identified in knowledge discovery process [4], [5] and [6] Selection, Preprocessing, Transformation, Data mining and Interpretation and evaluation.

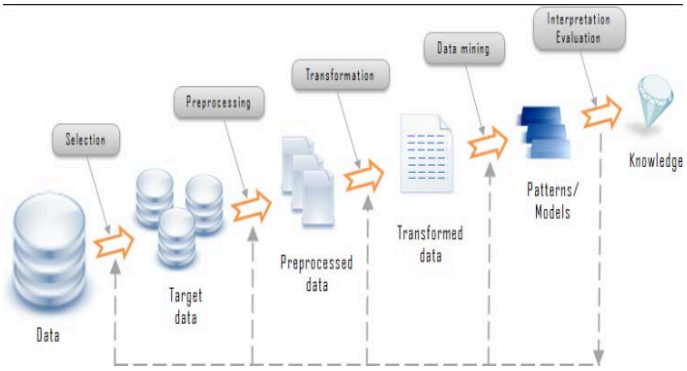

Fig. 1 Knowledge Discovery Process

With regard to these findings and emphasis on prediction of stroke incidence to reduce complications, disabilities and healthcare costs, this study was aimed to investigate some risk factors for ischemic stroke. After that, for collecting, pre-processing and data cleaning, data software WEKA 3.6, the C4.5 algorithm (DT) and Logistic regression were used to analyze the data.

2. Application of Data Mining in Public Health

Major headings are to be column centered in a bold font without underline. They need be numbered. "2. Headings and Footnotes" at the top of this paragraph is a major heading.

The use of methods of "data mining" in public health is growing rapidly. As in other areas, it is the availability of large historical databases (we now talk about data warehouses).

To mention only two, the journal Artificial Intelligence in Medicine and the Journal of the American Medical Informatics Association devote more and more items.

ISSN 2348 – 7968

Most publications focus on decision trees and association rules [7]. Among the areas covered include the search for risk factors for stroke, diabetes, suicide, [8], detection of fraud. These publications often mention the discovery of unexpected and effective rules.

2.1 Classification

Classification is one of the most popularly used methods of Data Mining in public Health. The classification technique predicts the target class for each data points. With the help of classification approach a risk factor can be associated to patients by analyzing their patterns of diseases.
In binary classification, only two possible classes such as, "high" or "low" risk patient may be considered while the multiclass approach has more than two targets for example, "high", "medium" and "low" risk patient. Data set is partitioned as training and testing dataset. It consists of predicting a certain outcome based on a given input. Training set is the algorithm, which consists of a set of attributes in order to predict the outcome.
In order to predict the outcome it attempts to discover the relationship between attributes. Goal or prediction is its outcome. There is another algorithm known as prediction set. It consists of same set of attributes as that of training set. But in prediction set, prediction attribute is yet to be known. In order to process the prediction it mainly analyses the input. The term, which defines how "good" the algorithm, is its accuracy. Consider a medical database of LGPM, training set consists all the information regarding patient, which were recorded previously. Whether a patient had a Ischemic Stroke or not is the prediction attribute there. With the help of table 1 and 2 given below we demonstrates the training sets of such database.

Table 1: Training set for LGPM database

Age	***MTHFR***	***High Blood Pressure***	***Ischemic Stroke***
38	CC	Yes	No
40	TT	No	Yes
52	CT	No	No
67	TT	Yes	Yes

Table 2: Prediction set for LGPM database

Age	***MTHFR***	***High Blood Pressure***	***Ischemic Stroke***
39	CC	Yes	?
45	TT	No	?
57	CT	Yes	?
70	CC	No	?

Prediction rules are divulged in the form of IF-THEN rules. With the help of above example, a rule predicting the first row in the training set may be represented as follows:

IF (Age=45 AND MTHFR=TT) OR (Age>44 AND High blood pressure=Yes) THEN Ischemic Stroke=Yes.

Following are the various classification algorithms used in health domain:

- **Decision Tree:** DT is considered to be one of the most popular approaches for representing classifier. We can construct a decision tree by using available data which can deal with the problems related to various research areas. It is equivalent to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. Root node is the top most node of a decision tree. For example, with the help of medical readmission decision tree we can decide whether a particular patient requires readmission or not. Knowledge of domain is not required for building decision regarding any problem. The most common use of Decision Tree is in operations research analysis for calculating conditional probabilities [9]. Using Decision Tree, decision makers can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain [10]. Decision Tree is widely used by many researchers in healthcare field.

Several advantages of decision tree as follows: Decision trees are self–explanatory and when compacted they are also easy to follow. Even set of rules can also be constructed with the help of decision trees. Hence, representation of decision tree plays a very important role in order to represent any discrete-value classifier because it can be capable to handle both type of attributes, nominal as well as numeric input attributes. If any datasets have missing or erroneous values, such type of datasets can be easily handled by decision trees.
Due to this reason decision tree can be considered to be nonparametric method. The meaning of above sentence is that there is no need to make assumptions regarding distribution of space and structure of classifier. Decision

ISSN 2348 – 7968

trees have several disadvantages. These are as follows: Most of the algorithms (like ID3 and C4.5) require that the target attributes have only discrete values because decision trees use the divide and conquer method. If there are more complex interactions among attributes exist then performance of decision trees is low. Their performance is better only when there exist a few highly relevant attributes. One of the reasons for this is that other classifiers can compactly describe a classifier that would be very challenging to represent using a decision tree. A simple illustration of this phenomenon is replication problem of decision trees [11], and the greedy characteristic of decision trees leads to another disadvantage. This is its over-sensitivity to the training set, irreverent attributes and to noise [12].

- **Neural Network:** In the early 20th century it was developed [13]. Before the introduction of decision trees and the Support Vector Machine (SVM) it was regarded as the best classification algorithm [14]. This was one of the reasons which encouraged NN as the most widely used classification algorithm in various biomedicine and healthcare fields [15], [16] and [17]. For example, NN has been widely used as the algorithm supporting the diagnosis of diseases including cancers [18] and [19] and predict outcomes [20] and [21]. In NN, basic elements are neurons or nodes. These neurons are interconnected and within the network they worked together in parallel in order to produce the output functions.

2.1 Regression

Regression is very important technique of data mining. With the help of it, we can easily identify those functions that are useful in order to demonstrates the correlation among different various variables. It is mainly a mathematical tool. With the help of training dataset we can easily construct it. Consider two variables ‘P’ and ‘Q’. These two types of variables are mainly used in the field of statistics. One of them is known as dependent and another one is independent variables. The maximum number of dependent variables cannot be more than one while independent can be exceeds one. Regression is mostly used in order to inspect the certain relationship between variables.

With the help of regression technique we can easily entrenched the addiction of one variable upon others [22]. Regression can be classified into linear and non-linear on the basis of certain count of independent variables. In order to appraisal associations between two types of variables in which one is dependent variable and another one is independent variables (one or more), linear regression used. In order to construct the linear model, linear function is utilized by linear regression. But there is limitation while we use linear approach because both types of variables are known already and hence, its main purpose is to trace a line that correlates between both these variables [23].

We cannot use linear regression for categorized data. It is restricted only to numerical data. With the help of logistic regression the categorical data can be used. Such type of data is used by non-linear regression and logistic regression is basically a type of non-linear regression. Logistic regression with the help of logit function can predict the probability of occurrence. However, between variables logistic regression cannot consider linear relationship [24]. Due to all these reasons regression is widely used in medical field for predicting the diseases or survivability of a patient.

3. Data Mining in Ischemic Stroke: Logistic Regression

The data from our sample are analyzed using the most complete statistical software of Microsoft "XLSTAT" which is based on the Visual Basic language. The XLSTAT code uses both the code (VBA for display) and C ++ code (for calculations), compatible with Windows and Mac platforms.

We proceeded to the construction of a logistic regression model [25], [26], [27]. It predicts the probability that happens Stroke (1 value) or not (value 0) from the optimization of the regression coefficients. The result is always between 0 and 1. [28], [29]. When the predicted value is greater than 0.5, the event (Ischemic Stroke) is likely to occur, whereas when this value is less than 0.5, it is the 'is not.

The logistic regression model included several steps:

- Extensive bibliographic research in advance is mandatory. Indeed the quality of a logistic regression based above all on the choice of explanatory variables [30], [16] that one is likely to incorporate the model.

- It was then necessary to study and analyze the links between each of the explanatory variables and the dependent variable was performed univariate analysis; the calculated odds ratios are gross

- We were forced to try several strategies to reach a final model that should include as much information while having a limited number of variables, in order to facilitate interpretation: the most used are the purported "step-down or step upward.

ISSN 2348 – 7968

3.1 Logit Model

This is to model the probability of the occurrence of cardiovascular disease in an individual i as a function of risk factors.

Ischemic Stroke is a latent variable that can be written as the sum of a linear combination of the characteristics of each individual and a random term.

Ischemic Stroke = β xi + εi

- X is a vector of explanatory variables; - β is the vector-associated parameters; - ε is random.

To calculate the probability, it was necessary to specify a statistical distribution for εi. The two most commonly used statistical laws are the logistic and the normal distribution, which then give the binary logit qualitative model called "The logit model offers an advantage in terms of the estimation technique parameters and its mathematical foundation is relatively simple "[33].

The logit model assumes that X follows a logistic law. Under these conditions, the probability that an individual has an Ischemic Stroke is:

In logistic regression, the dependent variable is binary i.e. it only contains data coded as 1 (TRUE, stroke.) or 0 (FALSE, failure, no-stroke.).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

Where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$Odds = P/(1-P)$$

And

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

$$p = \frac{1}{1 + e^{-logit(p)}}$$

We have chosen as hypothesis testing, testing on the parameters (Wald test) and test specification (contingency table).

Wald's test, close the test score, is used specifically to test the nullity of one or more factors, in particular of all but constant.

H0 = all coefficients are zero / H1 = at least one of the coefficients is not 0.

We applied the Wald test for all variables in the model one by one and it was concluded at the H0 rejection, since the coefficients of our variables are all zero. [cf. Table 3].

Table 3: Testing of The Null Hypothesis H0 : Y=0,500 (Stroke Variable)

STATISTIC	***DDL***	***Khi2***	***Pr > Khi2***
2Log(likelihood)	22	190,799	< 0,0001
Score	22	144,511	< 0,0001
Wald	22	81,947	< 0,0001

Once the prediction model was designed. We evaluate the efficacy and adjustment.

Compare the observed values of the dependent variable with the predictions [cf. Table 4].

Using the specification test [34], which allowed us to ensure the quality of fit of the model and its degree of prediction and calculation thereafter although the percentage of predicted observations gives a performance criterion the model.

Table 4: Classification for the estimation sample (Stroke Variable)

	PREDICTION		
OBS	***0***	***1***	***Total***
0	137	28	165
1	37	128	165
Total	174	156	330

We set a threshold prediction, the share of people who have strokes (165/300 = 0.5) which we had as result, 128 people who have a stroke have been well over 156 predict with a positive predictive value of 82.05% and 137 which do not predict stroke were well over 174.

The rate prediction of our model is 80.30% ((128 + 137/330 * 100). [see Table 5].

ISSN 2348 – 7968

Table 5: Indicators of our Data Mining model

Error rate	*Hit rate*	*Sensitivity*	*Specificity*
19,70%	81,30%	77,58%	83,03%

The results obtained with the "XLSTAT" software show a very good model (sensitivity 77.58% and specificity of 83%).

The ROC curve, evaluating the results of classification according to the decision threshold is sensitivity according to specificity.

This shows ROC curve, AUC area under the ROC curve (Area Under Curve) equals 0.89 (which leads to good sensitivity) as we pointed out, and the AUC is, the better the test.

4. Conclusions

Modeling results have shown that there are strong risk factors for ischemic stroke by applying the DATAMINING technique (Regression).

It was observed that the model of logistic regression in our case study Witnesses, allowed us to analyze the correlation between the occurrence of an ischemic stroke and its risk factors (genetic and clinical). The computer tool and its applications have enabled us to achieve more easily this analysis. However, in the confusion matrix, we concluded, the prediction model achieves 28 + 37 = 65 bad predictions. The error rate is 65/330 = 19.7%

For any algorithm its accuracy and performance is of greater importance. But due to presence of some factors any algorithm can greatly lost the above-mentioned property of accuracy and performance. Classification is also belongs to such an algorithm. Classification algorithm is very sensitive to noisy data. If any noisy data is present then it causes very serious problems regarding to the processing power of classification.

In order to achieve better accuracy in the prediction of diseases, improving survivability rate regarding serious death related problems etc. various data mining techniques must be used in combination.

To achieve medical data of higher quality all the necessary steps must be taken in order to build the better medical information systems which provides accurate information regarding to patients medical history rather than the information regarding to their billing invoices. Because high quality healthcare data is useful for providing better medical services only to the patients but also to the healthcare organizations or any other organizations who are involved in healthcare industry.

Takes all necessary steps in order to minimize the semantic gap in data sharing between distributed health domain databases environment so that meaningful patterns can be obtained. These patterns can be very useful in order to improve the treatment effectiveness services.

References

[1] J.H Friedman, Data mining and statistics: what's the connection?(1997). http://www stat.stanford.edu/~jhf/ftp/dm-stat.ps
[2] D. Hand, H. Mannila and P. Smyth, "Principles of datamining", MIT, (2001).
[3] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, (2005).
[4] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "The KDD process of extracting useful knowledge form volumes of data.commun.", ACM, vol. 39, no. 11, (1996), pp. 27-34.
[5] J. Han and M. Kamber, "Data mining: concepts and techniques", 2nd ed. The Morgan Kaufmann Series, (2006).
[6] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases", Commun. ACM, vol. 39, no. 11, (1996), pp. 24-26.
[7] N. Lavrac (1999) Selected techniques for data mining in medicine, Artificial Intelligence in Medicine, 16, 3
[8] S.E, Brossette, & al. (1998) Association rules and data mining in hospital infection control and public health surveillance, J Am Med Inform Assoc. 5(4):373-81.
[9] Goharian & Grossman, Data Mining Classification, Illinois Institute of Technology, http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM- Classification.pdf, (2003).
[10] Apte & S.M. Weiss, Data Mining with Decision Trees and Decision Rules, T.J. Watson Research Center, http://www.research.ibm.com/dar/papers/pdf/fgcsapteweissue_with_cover.pdf, (1997).
[11] Agallo, G. and Huassler, D., Boolean feature discovery in empirical learning, Machine Learning, 5(1): 71-99,1990.
[12] Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos, 1993.
[13] Anderson, J. A., and Davis, J., An introduction to neural networks. MIT, Cambride, 1995.
[14] Obenshain, M. K., Application of data mining techniques to healthcare data. Infect. Control Hosp. Epidemiol. 25(8):690–695, 2004.
[15] Bellazzi, R., and Zupan, B., Predictive data mining in clinical medicine: current issues and guidelines. Int. J. Med. Inform. 77:81–97, 2008.
[16] Übeyli, E. D., Comparison of different classification algorithms in clinical decision making. Expert syst 24(1):17–31, 2007.

ISSN 2348 – 7968

[17] Kaur, H., and Wasan, S. K., Empirical study on applications of data mining techniques in healthcare. J. Comput. Sci. 2(2):194–200 2006.
[18] Romeo, M., Burden, F., Quinn, M., Wood, B., and McNaughton, D., Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer. Cell. Mol. Biol. (Noisy-le-Grand, France) 44(1):179, 1998.
[19] Kononenko, I., Bratko, I., and Kukar, M., Application of machine learning to medical diagnosis. Machine Learning and Data Mining: Methods and Applications 389:408, 1997.
[20] Sharma, A and Roy, R. J., Design of a recognition system to predict movement during anesthesia. IEEE Trans. Biomed. Eng.44(6):505–511, 1997.
[21] Einstein, A. J., Wu, H. S., Sanchez, M., and Gil, J., Fractal characterization of chromatin appearance for diagnosis in breast cytology. J. Pathol. 185(4):366–381,1998.
[22] J. Fox, "Applied Regression Analysis, Linear Models,and Related Methods", (1997).
[23] Gennings, R. Ellis and J. K. Ritter, "Linking empirical estimates of body burden of environmental chemicals and wellness using NHANES data", http://dx.doi.org/10.1016/j.envint.2011.09.002,2011.
[24] P. A. Gutiérrez, C. Hervás-Martínez and F. J. Martínez-Estudillo, "Logistic Regression by Means of Evolutionary Radial Basis Function Neural Networks", IEEE Transactions on Neural Networks, vol. 22, no. 2, (2011), pp. 246-263.
[25] Aminot I, Damon MN The Use of Logistic Regression in the Analysis of Data Concerning Good Medical Practice
[26] J. Jaccard, Intercation E_ects in Logistic Regression, Series: Quantitative Applications in the Social Sciences, n0135, Sage Publications, 2001.
[27] D. Garson, Logistic Regression, http://www2.chass.ncsu.edu/garson/PA765/logistic.htm
[28] P.L. Gonzales, "Modèles à réponses dichotomiques", in Modèles statistiques pour données qualitatives, Droesbeke, Lejeune et Saporta Editeurs, Chapitre 6, pages 99-136, Technip, 2005.
[29] Gennings, R. Ellis and J. K. Ritter, "Linking empirical estimates of body burden of environmental chemicals and wellness using NHANES data", http://dx.doi.org/10.1016/j.envint.2011.09.002,2011.
[30] M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST ISSN: 2229- 4333, vol. 2, no.2, (2011) June.
[31] S. Menard, Applied Logistic Regression Analysis (Second Edition), Series: Quantitative Applications in the Social Sciences, n0106, Sage Publications, 2002.
[32] A.A. O'Connell, Logistic Regression Models for Ordinal Response Variables, Series: Quantitative Applications in the Social Sciences, n0146, Sage Publications, 2006.
[33] R. Rakotomalala, Régression logistique - Une approche pour rendre calculable P (Y/X), http://eric. univ-lyon2.fr/~ricco/cours/supports_data_mining.html
[34]D.W.Hosmer,S.Lemeshow,AppliedLogisticRegression,SecondEdition, Wiley,2000.
[35] A. Aljumah, M. G.Ahamad and M. K. Siddiqui, "Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia", Intelligent Information Management, vol. 3, (2011), pp. 252-261.

BALAR Khalid[1], PhD in Computer Science, Hassan II University- Faculty of Medicine and Pharmacy of Casablanca, 19 Rue Tarik Ibnou Ziad, B.P. 9154, Casablanca, Morocco.
Email: balarkhalid@gmail.com

NAJI Abdelwahab[2], Assistant Prof in Computer Science, Hassan II University- Superior Normal School of Technical Education, Rue Boulevard Hassan II Mohammedia, Morocco.
Email: abdelwahab.naji@gmail.com