Prediction and Control of Stroke by Data Mining

Leila Amini, Reza Azarpazhouh[1], Mohammad Taghi Farzadfar[1], Sayed Ali Mousavi[2], Farahnaz Jazaieri[3], Fariborz Khorvash[2], Rasul Norouzi[2], Nafiseh Toghianifar[2]

Department of Computer Engineering and Information Technology, Payam Noor University, Tehran, Iran, [1]Department of Neurology, Medical University of Mashhad Sciences, Mashhad, Iran, [2]Department of Neurology, Isfahan Neurosciences Research Center, Isfahan University of Medical Sciences, Isfahan, Iran, [3]Department of Pharmacology, School of Medicine, Tehran University of Medical Sciences Tehran, Iran

Correspondence to:
Dr. Fariborz Khorvash,
Isfahan Neurosciences Research Center,
Isfahan University of Medical Sciences,
Isfahan, Iran.
E-mail: fkhorvash@med.mui.ac.ir

Date of Submission: Feb 23, 2013

Date of Acceptance: Feb 23, 2013

How to cite this article: Amini L, Azarpazhouh R, Farzadfar MT, Mousavi SA, Jazaieri F, Khorvash F, *et al*. Prediction and control of stroke by data mining. Int J Prev Med 2013;Suppl 2: S245-9.

ABSTRACT

Background: Today there are abounding collected data in cases of various diseases in medical sciences. Physicians can access new findings about diseases and procedures in dealing with them by probing these data. This study was performed to predict stroke incidence.

Methods: This study was carried out in Esfahan Al-Zahra and Mashhad Ghaem hospitals during 2010-2011. Information on 807 healthy and sick subjects was collected using a standard checklist that contains 50 risk factors for stroke such as history of cardiovascular disease, diabetes, hyperlipidemia, smoking and alcohol consumption. For analyzing data we used data mining techniques, *K*-nearest neighbor and C4.5 decision tree using WEKA.

Results: The accuracy of the C4.5 decision tree algorithm and *K*-nearest neighbor in predicting stroke was 95.42% and 94.18%, respectively.

Conclusions: The two algorithms, C4.5 decision tree algorithm and K-nearest neighbor, can be used in order to predict stroke in high risk groups.

Keywords: Data mining, decision tree, *K*-nearest neighbor, prediction, stroke

Original Article

INTRODUCTION

Based on studies of more than 56 million deaths in 2001, it was found that 7.1 million cases were due to heart disease and 5.4 million were also due to stroke.[1] This indicates that stroke – after heart disease – is the second major cause of death in the world that is nearly 10% of all deaths reported. Stroke is the third leading cause of death in the United States, and about 137,000 Americans die due to this disease each year. In 2006, 6 out of every 10 deaths from stroke had occurred in women.[2] In the United States, one suffers from stroke every 40 second and every 3–4 minute one dies from stroke. The cost of this disease – only in America – has been estimated about 73.7 million dollars.[3] Stroke is one of the major causes of disability in the world. According to the reports published in 2005, close to 1.1 million people have survived a stroke but live with some problems in their daily activity.[4] In Iran, Based on researches conducted on a population of 450,229 people in the city of

Mashhad, it was seen that the stroke occurred nearly a decade sooner than in the Western countries and the incidence rate in Iran was also higher than in most of them.[5] Most studies performed on automated diagnosis of stroke and its subtypes were on the image processing techniques and computerized tomographic scan and magnetic resonance imaging.[6-8] For example, computerized tomographic scan images have been used for diagnosis of stroke and its subtypes. After improvement of images and noise reduction, the skull line of symmetry is determined and then a histogram chart is created for the brain hemispheres. Hemorrhagic and chronic stroke are distinguished by the histogram chart. We used wavelet features for diagnosis of acute stroke and normal images.[6] Precision and Recall obtained were 90% and 100%, respectively.[6] In another study of a mining algorithm, classification rules were used to analyze data from stroke patients. The NGTS (New General-To-Specific) algorithm, which is a sequential covering algorithm for extracting classification rules, has been applied for 162 specimens. Total number of extracted rules was 84% and 84.8% classification accuracy has also been achieved.[9] The T3 algorithm, which provides a decision tree with a maximum depth of 3, has been investigated to construct the decision tree from the data of stroke patients. The results obtained from comparison with the C4.5 algorithm show that the accuracy of the T3 classification algorithms for training and test data sets was higher and overall displayed better performance. A data set contains 795 records and 37 attributes per record. The best error classification for the T3 algorithm was 0.4%, whereas the best value for the C4.5 algorithm was 33.6%.[10] There are several factors that play a role in stroke incidence, some of which are heredity, age, gender and race, certain medical conditions such as high blood pressure, hypercholesterolemia, heart disease and diabetes. Overweight, past history of stroke can also increase the incidence risk of stroke. No smoking and no alcohol consumption and daily activities can also be effective to reduce the risk of stroke. By use of the aforementioned risk factors, and techniques of data mining, decision support system can be designed that besides knowledge and experience of a physician, can be used to predict stroke. Owing to the human need of knowledge and increasing data volume, technique development for automated extraction of knowledge from these data is inevitable. Data mining is extraction of knowledge and attractive patterns from a large volume of data.[11] Data mining techniques based on knowledge that can be extracted are divided into three major groups: Pattern classification, data clustering and association rule mining.

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 50 risk factors for brain stroke. After that, for collecting, pre-processing and data cleaning, data software WEKA 3.6, the C4.5 algorithm (version 8) and the *K*-nearest neighbor algorithm were used to analyze the data.

METHODS

In the pattern classifications – which have been used in this article – based on a set of attributes, one class label was assigned to one sample of data.[12] Classification is a two-step process. In the first phase, which is called the Learning Phase, a clustering algorithm makes a model from analysis of a training data set that describes a set of class labels and predefined concepts. In the second phase, which is called the Test Phase, the classification accuracy of the model is measured using a test data set.

In this investigation, which lasted from August 1387 to March 1389, at first after studying sources and texts written on science data mining, in order to extract the concepts, structures and algorithms, 50 risk factors that were effective in stroke incidence were provided for a healthy community and a population with stroke. A total 807 checklists were collected then the samples were formed using Excel files.

Meanwhile, some records were unspecified values, therefore the following techniques were used.[11]

1. Use of average property values: Using the mean values of a feature to fill unspecified values of that features. This method is commonly used for continuous data such as height or weight.
2. Using mean values in lines so that their class is equal to lines with unspecified values. This methods used for continuous features.

After that data mining, software WEKA 3.6 was used for data analysis. The J48 and IBK algorithms that are for implementation of version 8 of the C4.5 algorithm and the *K*-nearest neighbor algorithm, respectively, in the JAVA language were applied on the data.[13]

RESULTS

The Sensitivity, Specificity, Precision and Accuracy criterion of the two algorithms, *K*-nearest neighbor and C4.5, have been shown for different values in Table 1.

Also three criterions of Accuracy, Sensitivity and Precision were perused and compared in two methods in Figure 1.

The Specificity criterion is shown in Figure 2.

The figures show that in the *K*-nearest algorithm, Specificity and Precision decreased after *K* increment, and then Accuracy decreased a little. In addition, after *K* increment, Sensitivity increased a little and the number of patients who were truly diagnosed increased but simultaneously false-positive patients also increased. This issue resulted in Precision reduction as *K* increased. The best results are pertaining to the C4.5 algorithm that outruns the *K*-nearest neighbor algorithm in the Accuracy, Precision and Specificity criteria by a small difference.

DISCUSSION

Most studies performed on stroke diagnosis and its species differentiation focused on image-processing techniques and no research studies have been conducted on the C4.5 decision tree and *K*-nearest neighbor methods. In this study, 50 risk factors such as gender, age, hours of activity, sleep duration, body mass index, hypertension, diabetes, hyperlipidemia, smoking, alcohol, narcotics, stimulants and other risk factors that have not been considered previously were extracted then the C4.5 and *K*-nearest neighbor algorithms in data mining software WEKA 3.6 were used to analyze stroke data.[14]

After applying the algorithms and performing a comparison and computation of accuracy using the decision tree and *K*-nearest neighbor methods, the best results are those pertaining to the C4.5 algorithm that outruns the *K*-nearest neighbor algorithm in the Accuracy, Precision and Specificity criteria by a small difference. In medical diagnosis systems, even a small difference in classification is important since right prediction of illness is vital and very important.

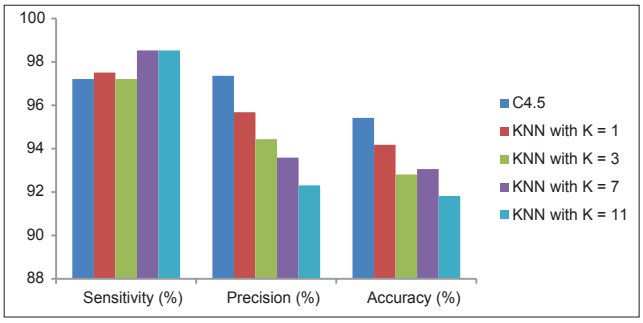

Figure 1: Comparison of the C4.5 and KNN algorithms based on the Accuracy, Sensitivity and Precision criteria

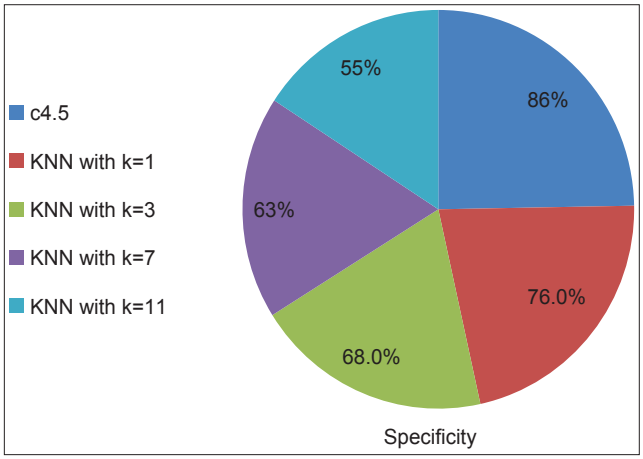

Figure 2: Comparison of specificity based on the C4.5 and KNN criterion

Table 1: Results of the classification algorithms using data sets on stroke

Algorithm	TN	FP	FN	TP	Sensitivity (%)	Specificity (%)	Precision (%)	Classification accuracy (%)
C4.5	107	18	19	663	97.21	85.6	97.36	95.42
KNN with K=1	95	30	17	665	97.51	76.0	95.68	94.18
KNN with K=3	86	39	19	663	97.21	68.8	94.44	92.81
KNN with K=7	79	46	10	672	98.53	63.2	93.59	93.06
KNN with K=11	69	56	10	672	98.53	55.2	92.31	91.82

KNN=K Nearest Neighbor, TN=True Negative, FP= False Positive, FN=False Negative, TP=True Positive

Finally the "decision tree" was selected as the stroke-predicting algorithm because of its higher accuracy.

The efficiency, especially Sensitivity and Accuracy of classification, of the introduced algorithms was over 90%, showing that besides knowledge and experience of clinicians, we can take advantage of data mining techniques in diagnosis and management of patients with stroke.

The use of feature selection techniques to reduce data dimension and finding the effect of each of them on stroke incidence are very important in computational and medical vision. Also the effect of data dimension reduction on classification accuracy and other algorithm performance criteria should be examined. Designing a decision support system that not only diagnoses the species of stroke (hemorrhagic or ischemic), but also predicts the disease is very valuable. Generally, if we can design decision support systems that carefully define type and age of stroke in addition to stroke incidence that will affect an individual in future, many medical expenses will be saved. Since data mining is knowledge that is unfortunately unknown in our country, the numbers of people who have done good research in this area are lacking. Most investigations are presented theoretically and just one data mining conference is being held in Iran. This knowledge and its applications, which can be used in different fields, will gradually find place among professionals. But this scientific field is unfamiliar to medical specialists.

In fact, in Iran, one of the biggest problems with medical data mining is mistrust in the medical community because of lack of knowledge of this science. In fact, for many people the result of data mining is incredible. If you are looking to obtain interesting results on data mining, data must be strong and in large volume. Many medical centers do not provide data of their patients to data mining teams. Security, lack of confidence in the results of data mining and the desire to retain them exclusively for their next possible studies were the major problems in our study. Furthermore, if participants who will suffer a stroke in this project is determined during a multi-year process, more interesting results will be achieved.

CONCLUSIONS

The two algorithms, C4.5 decision tree algorithm and K-nearest neighbor, can be used in order to predict stroke in high risk groups.

REFERENCES

1. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ. Global and regional burden of disease and risk factors, 2001: Systematic analysis of population health data. Lancet 2006;367:1747-57.
2. Heron M, Hoyert DL, Murphy SL, Xu J, Kochanek KD, Tejada-Vera B. Deaths: Final data for 2006. Natl Vital Stat Rep 2009;57:1-134.
3. Lloyd-Jones D. Heart disease and stroke statistics-2010 update. A report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation 2010;121:e1-170.
4. Brault MW, Hootman J, Helmick CG, Theis KA, Armour BS. Prevalence and most common causes of disability among adult-United States, 2005. MMWR Morb Mortal Wkly Rep 2009;58:421-6.
5. Azarpazhooh MR, Etemadi MM, Donnan GA, Mokhber N, Majdi MR, Ghayour-Mobarhan M, *et al*. Excessive incidence of stroke in Iran, evidence from the mashhad stroke incidence study (MSIS), a population-based study of stroke in the middle East. Stroke 2010;41:e3-10.
6. Chawla M, Sharma S, Sivaswamy J, Kishore L. A method for automatic detection and classification of stroke from brain CT images. Conf Proc IEEE Eng Med Biol Soc 2009;2009:3581-4.
7. Przelaskowski A, Sklinda K, Bargieł P, Walecki J, Biesiadko-Matuszewska M, Kazubek M. Improved early stroke detection: Wavelet-based perception enhancement of computerized tomography exams. Comput Biol Med 2007;37:524-33.
8. Tang FH, Ng DK, Chow DH. An image feature approach for computer-aided detection of ischemic stroke. Comput Biol Med 2011;41:529-36.
9. Mroczek T, Grzymala-Busse J, Hippe ZS. A new machine learning tool for mining brain stroke data. 3rd International Conference on Human System Interactions (HSI) IEEE: Digital Object Identifier; 2010; p. 246-50.
10. Tjortjis C, Saraee M, Theodoulidis B, Keane JA. Using T3, an improved decision tree classifier, for mining stroke-related medical data. Methods Inf Med 2007;5:523-9.
11. Han J, Kamber M. Data Mining: Concept and Techniques. 2nd ed. California: Morgan Kaufmann Publishers; 2006.

12. Mitchell T. *Machine Learning*. New York: McGraw-Hill; 1997.
13. Witten IH, Frank E. Data mining, practical machine learning tools and techniques. 2nd ed. California: Morgan; 2005.
14. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An Update. SIGKDD Explorations 2009;11:10-8.

Source of Support: Nil, **Conflict of Interest:** None declared.