Research Article
Volume 9:2, 2020

Telecommunications System & Management

Open Access

Risk Prediction of Osteoarthritis using Data Mining Classification Techniques

OLAYEMI, Olufunke Catherine

Department of Computer Science, Joseph Ayo Babalola University, Ikeji- Arakeji, Osun State, Nigeria

Abstract

Osteoarthritis (OA) is the most common reason of disability among the ageing population. The awareness of machine learning as a tool in medicine is growing rapidly and has provided new avenues for research into a number of diseases and infections. Creating better predictive models for these diseases could provide opportunities for better care, which we have applied to osteoarthritis, a degenerative disease that affects a large number of both gender in older population. A number of studies have been undertaken in order to understand the prediction of Osteoarthritis risks using data mining techniques. Hence, this study is focused at using two different types of s data mining techniques to predict Osteoarthritis risks in Nigerian patients using the Naïve Bayes' and the K nearest neighbor algorithms. The performances of these two classification techniques was evaluated in order to determine the most efficient and effective model. To achieve this, a dataset containing patients who have participated in an osteoarthritis treatment program was used and analyzed. The Naïve Bayes' showed a higher accuracy with lower error rates compared to that of the KNN method while the evaluation criteria proved the Naïve Bayes' to be a more effective and efficient classification techniques for the prediction of Osteoarthritis risks among patients of the study location. Our results shows that it is possible to predict an efficient and effective classifier for Osteoarthritis risks.

Keywords: Osteoarthritis, classification, prediction, risk factors, naïve bayes, K nearest neighbour

Introduction

Osteoarthritis (OA) is a deteriorating sickness that usually affects the human knee joints. OA is the most common form of arthritis and one of the leading causes of disability globally, affecting 3.8% of the global population [3]. It causes painful joint locking. This breakdown usually affects the daily functional activities of an affected individual. This type of health breakdown challenge frequently happens to middle-aged and elderly person due to breakdown of cartilage. It is one of the leading causes of disability among the elderly people [5]. It was estimated that more than 27 million Americans have this condition, which primarily affects people who are 60 years of age or older Osteoarthritis often times involves the joints that bear most of the body weight (weight-bearing joints), such as the knees or hips. In many cases, only one joint aches. OA can also occur in any other joint, such as the middle and lower spine or the joints in the hands and fingers. [3]. It has been estimated by the Center for Disease Control that nearly 1 in every 2 people develop symptoms of OA of the knee by age 85. The symptoms may vary from person to person. In most people, the joint damage occurs gradually over many years, and as it does, the pain usually increases. At times, the pain may progress rapidly. In some people, OA is relatively mild and does not interfere much with daily life. Others may experience significant pain and disability. It ranks as the fifth highest cause of years lost to disability in the whole population in high-income countries, and the ninth highest cause in low- and middle-income countries [10]. It accounts for 50% of the entire musculoskeletal disease burden, and thus is considered the highest-burden condition within the musculoskeletal group of diseases, which also includes rheumatoid arthritis and osteoporosis. Radiographic evidence of knee osteoarthritis is present in approximately 30% of men and women over the age of 65.2 Worldwide estimates are that 9.6% of men and 18.0% of women over the age of 60 years have symptomatic osteoarthritis. Approximately 80% of those with OA will have limitations in movement, and 25% cannot perform their major activities of daily life [11]

The various ways of early detecting of Osteoarthritis is by identifying the risk factors and guiding against those ones that can be guided. Some of the factors that increases the risk of Osteoarthritis are, older age; because this disease increases with age, sex; women are more likely to develop osteoarthritis according to [11] Obesity, joint pains and joint injuries, repeated stress on the joint, Genetics, bone deformities, certain metabolic diseases etc. [12]. Osteoarthritis conditions can usually be managed, although the damage to joints can't be reversed. Staying active, maintaining a healthy weight and some treatments might slow progression of the disease and help improve pain and joint function. Osteoarthritis symptoms often develop slowly and worsen over time. Signs and symptoms of osteoarthritis are, Pains: affected joints might hurt during or after movement. Stiffness; Joint stiffness might be most noticeable upon awakening or after being inactive. Tenderness, loss of flexibility, Grating sensation, Bone spurs, swelling, etc. [12]

Data mining can be a useful tool in the health sector and healthcare. Organizations that perform data mining are better positioned to meet their long-term needs. Data can be a great advantage to healthcare organizations, but they have to be first changed into information. Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. Classification is a data mining technique used to predict group membership for data instances [2] . This study aims at using data mining techniques to classify Osteoarthritis risks using datasets of patients' information from Federal teaching hospital, Ido-Ekiti, Ekiti State, which contains the risk factors of Osteoarthritis and Osteoarthritis classes (yes

*Address for Correspondence: By Ken Perez, Vice President of Healthcare Policy, Omnicell, Inc.

Received 15 June 2020; Accepted 22 June 2020; Published 29 June 2020

and no). Naïve Bayes' and KNN classification of Osteoarthritis was performed using the WEKA software.

Related Works

A number of papers have been documented and published on the use of data mining techniques in the classification of Osteoarthritis risks. Some of such works are reviewed in the following paragraphs.

[8] presented a paper titled Identification of knee Osteoarthritis based on Bayesian Network: Pilot study. The aim of the paper was to propose a Bayesian network (BN)-based classification model to classify people with knee OA. A total of 249 elderly people between ages 60 and 80 years living in the Konggiang community were recruited for the research work. I57 patients were later adopted for the osteoarthritis research work after the data preprocessing. The results after the evaluation shows that their proposed model gave a higher result than the existing models been used. A method was used in the study.

[10] worked on Predicting and Analyzing Osteoarthritis Patient Outcomes with Machine Learning. The aim of their work was to answer two questions. The questions are; ''Is it possible to predict Osteoarthritis patient outcomes?" and "What factors contribute to the Osteoarthritis patient outcome?" In their work, construction and evaluation of machine learning models was done. The dataset containing 75,366 patients who have participated in an osteoarthritis treatment program was used and analyzed. The selection of models used in the work included neural networks, logistic regression and gradient boosting machines among others in order to capture the performance of several types of machine learning models. Their results show that it is possible to predict patient outcomes on a test set with 60% accuracy. Future enhancement of the work will require the improvisation of algorithm to improve classification rate to achieve greater accuracy.

[1] worked on Magnitude of knee osteoarthritis and associated risk factors among adult patients presented in a family practice clinic in Nigeria. The study used a semi-structured questionnaire to interview Four-hundred (400) respondents. Knee osteoarthritis was diagnosed clinically using the American College of Rheumatology (ACR) criteria. The Four-hundred patients were aged 18 years and above. Only those who gave their informed consent were included, while those who were too ill to participate in the study were excluded. Descriptive statistics were employed for the socio-demographic, lifestyle, and self-reported health status of the respondents. Chi-square statistics was used to assess the association between categorical variables. The P-value of significance was set at < 0.05. Logistic regression was used to explore the relationship between the socio-demographic, lifestyle and other risk factors associated with knee OA. The result shows that, the point prevalence of knee osteoarthritis was 11.5%. Increasing age, female gender, marital status, low educational status, financial dependency, poor income, obesity, previous knee injury, epigastric pain, peptic ulcer disease, varus deformity of the knee, and poor health status were significantly associated with knee osteoarthritis. Data mining technique was not used for classification in the study.

[9] presented a paper titled" Prediction model for knee osteoarthritis based on genetic and clinical information". The aim of the paper was that, the current association studies have revealed the hereditary factors behind OA, with its susceptibility inheritable factors. This will enable the researchers to predict disease occurrence based on genotype knowledge. The method used was that the genotyped risk alleles of the three susceptibility genes were statistically analyzed with their effects. They later constructed prediction models by using the logistic regression analysis. The result of this work shows that Individuals with five or six risk alleles showed significantly higher susceptibility when compared with those with zero or one risk alleles.

Data Mining Techniques

Data mining is the process of extracting patterns from data; these patterns may be discovered depending on the data mining tasks that are applied on the dataset. The two types of data mining tasks are: descriptive and predictive data mining task. The descriptive data mining task help to understand the characteristic properties of dataset and predictive data mining tasks are used to perform predictions based on available dataset. Predictive data mining is the chosen data mining task for this study. According to data mining applications can used for different parameters to examine data which includes; association (patterns that define the relationship between data), sequence/pattern analysis (patterns where one event leads to another), classification (identification of new patterns with predefined targets) and clustering (grouping of identical of smaller objects). The basic steps include:

- Problem definition is the definition of the goals and objectives and the identification of tools to
- be used to build the defined model.
- Data exploration is the recommendation for useful dataset if the existing dataset does not
- meet the required need for analysis.
- Data preparation is the process of cleaning and transforming data to remove missing and
- invalid data and validation of data for robust analysis.
- Modeling is based on the desired outcomes and data. This involves the use of data mining
- algorithms (for this study; naïve bayes, decision trees and multi-layer perceptron) in meeting
- the necessary objectives-which for the purpose of this study is classification.
- Evaluation and deployment is the analysis and interpretation of the results of analysis to create
- recommendations for consideration.

Methods

So as to classify the Osteoarthritis data collected form Federal Medical Center (FMC) Ido, with the aim of achieving high accuracy and precision; two supervised learning algorithms i.e., Naïve Bayes' are K-Nearest Neighbor (KNN) were used. The data preprocessing was performed in order to remove inconsistent data and the data converted into a format that is useful in the simulation environment. WEKA data mining software was the environment used for simulating the Osteoarthritis risk prediction model; which is an open-source data mining software used for academic purposes.

Training dataset description

Following the identification of the risk factors of Osteoarthritis from the review of literature and expert medical physicians, the case files of patients were used to collect information about the distribution of the risk factors of OA patients coming for treatment at the Federal medical Center Ido Ekiti, ekiti State, in the south-western Nigeria. The datasets collected from the patients records contains 102 instances with 15 attributes. The class distribution is framed as Yes or No. Hence there are 14 independent variables and 1 dependent variable. The nominal values are set for the independent variables and the dependent variable. A description of the attributes contained in the dataset is presented in Table 1: below, Gender is either male or female, Age; the ages of the patients included in the study ranges from twenty one (21) years to eighty five years (85), Family History is either yes or no, Hip ratio depends on the size of the hips of the patient, BMI is the Body Mass Index (BMI) which is the weight of the patient, it is Abnormal for the obsessed patients, Hypertensive Heart Disease (HHD) is whether the patient is having high blood pressure or not ,to just mention a few. The non- modifiable factors are the first seven variables while the modifiable variables are the next seven variables in the table. The Osteoarthritis is the last variable.

Table1: Distribution of Identified Features in the Original Datas

Types	Variable Names	Attribute Values
Input Variables	Gender	Male, Female
	Age (years)	Above 21 to 85 years
	Family History	Yes, No
	Waist Hip Ratio	Low, Normal
	BMI	Normal, Abnormal
	HHD	Yes, No
	Joint pains	Yes, No
	Cellulitis of Leg	Yes, No
	Seizure Disorder	Yes, No
	Ulcer of L/R Limb	Yes, No
	Septic Arthritis	Yes, No
	Repeated stress on Joint	Yes, No
	Bone Deformities	Yes, No
	Joint Injuries	Yes, No
	Osteoarthritis	Yes, No

A, Naïve Bayes' (NB) Classification

Naive Bayes' Classifier is a probabilistic model that depends on Baye's theorem. It is known as a statistical classifier. It is one of the habitually used methods for supervised learning. It provides a capable way of dealing with any number of attributes or classes which is purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data. Let X_ij be a dataset sample containing records (or instances) of i number of risks factors (attributes/features) alongside their respective Osteoarthritis, C (target class) collected for j number of records/patients and Hk = {H1 =Yes, H2 =No} be a hypothesis that X_ij belongs to class C. For the classification of the diagnosis of (OA) given the values of the risk factor of the jth record, Naïve Bayes' classification required the determination of the following Rupali Patil (2014)

P(H_k |X_ij) – Posteriori probability: is the probability that the hypothesis,H_k holds given the observed data sample X _ij for 1≤k≤2.

P(H_k) - Prior probability: is the initial probability of the target class 1≤k≤2;

P(X_ij) is the probability that the sample data is observed for each risk factor (or attribute), i;

P(|X_ij |H_k)is the probability of observing the sample's attribute, X_i given that the hypothesis holds in the training data X_ij.

Therefore, the posteriori probability of an hypothesis H_k is defined according to Bayes' theorem as shown in equation (1) while the determination of OA class is in equation (2).

$$P(H_k|X_{ij}) = \frac{\prod_{i=1}^{n} P(X_{ij}|H_k)P(X_i)}{P(H_k)} \quad for\ k = 1,2 ----(1)$$

$$max.\left[P(H_1|X_{ij}), P(H_2|X_{ij})\right] -------(2)$$

K-Nearest Neighbor (KNN)

This can be described as learning by similarity, it is learnt by comparing a specific test tuple with a set of training tuples that are similar to it. It is classified based on the class of their closest neighbors, most times, more than one neighbor is taken into consideration hence, the name K-Nearest Neighbour (K-NN), the "K" indicates the number of neighbors taken into account in determining the class [6]. In this paper work, our data tuples are restricted to a patients with OA symptoms. The Euclidean distance between a training tuple and a test tuple can be derived as follows:

let p_(i) be an input turple with p features of OA (p_i1 p_i2 p_i3)

let n be the total numberof input tuples of OA i=1.2….,n)

let k be the total number of features of OA (j=1,2,…,k)

The euclidean distance between turple p_i and p_t (t=1,2,…,n) can be defined as:

The euclidean distance between turple p_i and p_t $(t = 1,2,\dots,n)$ *can be defined as:*

$$d(p_i p_t) = \sqrt{(p_{i1}-p_{t1})^2 + (p_{i2}-p_{t2})^2 + \cdots + (p_{in}-p_{tn})^2} ----------(3)$$

in general term: the euclidean diatance between two tuples for instances are

$$p_1 = (p_{11}, p_{12}, \dots p_{1n}) and\ p_2 = (p_{21}, p_{22}, \dots, p_{2n}) will\ then\ be:$$

$$dist p_1 p_2 = \sqrt{\sum_{i=1}^{n}(p_{1i}-p_{2i})^2} ----------(4)$$

Equation (3) is applicable to numeric attribute of OA, in which we take the difference between each corresponding values of attributes tuple P1 and P2, square the result and add them together to get the square root of the accumulated result ,this gives us the distance between the two points P1 and P2 . From equation (4), diagnosis of the input instance is based on the closest n neighbour

In order to evaluate the performance of the supervised machine learning algorithms used for the risk factor classification of the OA, there was the need to plot the results of the classification on a confusion matrix (Figure 2). The four parameters used to formulate the metrics are as follows:

- True positives (TP) are correctly classified Yes cases;

- False positives (FP) are incorrectly classified No cases;
- True negatives (TN) are correctly classified No cases; and
- False negatives (FN) are incorrectly classified Yes cases

Figure 1: Diagram of a Confusion Matrix

The true positive/negative and false positive/negative values recorded from the confusion matrix can then be used to evaluate the performance of the prediction model. A description of the definition and expressions of the metrics are presented as follows:

(a) True Positive (TP) rates (sensitivity/recall) – proportion of positive cases correctly classified.

$TP\ rate_{Yes} = TP/(TP+FN)$-----(5)

$TP\ rate_{No} = TN/(FP+TN)$-----(6)

(b) False Positive (FP) rates (1-specificity/false alarms) – proportion of negative cases incorrectly classified as positives.

$FP\ rate_{Yes} = FP/(FP+TN)$--- (7)

$FP\ rate_{No} = FN/(TP+FN)$ ---- (8)

Precision – proportion of predicted positive/negative cases that are correct.

$Precision_{Yes} = TP/(TP+FN)$--- -(9)

$Precision_{No} = TN/(TN+FP)$ --- -(10)

Accuracy – proportion of the total predictions that was correct

Results and Discussion

The results of the data mining process for the prediction of Osteoarthritis risk using NB classifier and KNN technique as discussed above was implemented using the WEKA software data mining tool. The 102 datasets that were collected from FMC, Ido in Ekiti State were divided into 70% for training sets and 30% for testing sets. The two techniques were used to evaluate the models using the testing data. From the results on Table 2 of the analysis made on the dataset using Naïve Bayes', It shows that Naïve Bayes' recorded the highest correctly classification of 81 instances and 21 incorrectly classified instances with accuracy of 93.14% while KNN recorded the least performance of 88.24% with the correctly and incorrectly classified instances respectively. From the Confusion matrix, the True Positive (TP) rate/recall which is the percentage of the actual number of positive that were classified as positive cases has an average of 93.14% and 88.24% for the naïve bayes' and KNN respectively. The False Positive (FP) rate which is the percentage actual number of positive cases that were misclassified also called false alarm has an average of 2.5% and 4.0% for the naïve bayes' and KNN respectively. From the above results shown, it shows very clearly that data mining techniques can be used in predicting Osteoarthritis risks and that the Naïve Bayes' classifier has a better accuracy than the KNN algorithm.

Table 2: Results of the Correct and Incorrect Classification of Testing Datasets

Supervised machine Learning	Confusion Matrix		Classification Accuracy (%)	False Alarm Rate (%)
NB	TP =76 FP = 2	TN=19 FN = 5	93.14	2.5
KNN	TP = 72 FP =3	TN=18 FN=9	88.24	4.0

Figure 2: Results of the Correct and Incorrect Classification of Testing Datasets

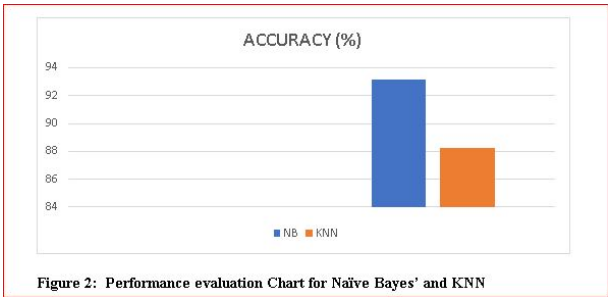

Figure 2: Performance evaluation Chart for Naïve Bayes' and KNN

Conclusion

In this study two different data mining classification techniques was used for the prediction of Osteoarthritis risk in adult population and their performances was compared in order to evaluate the best classifier. Experimental results shows that the Naïve Baye'classifier is a better model for the prediction of Osteoarthritis risks for the value of accuracy, recall, precision and error rates recorded for both models. Hence, an efficient and effective classifier for Osteoarthritis risks has been identified while the number of attributes covered by the classifier can be increased by increasing the sample size of the training set and hence the development of a more accurate model.

References

1. Adebusoye LA, Ogunbode A.M, Alonge T.O (2013) ''Magnitude of knee osteoarthritis and
2. associated risk factors among adult patients presenting in a family clinic in Nigeria". J, Med Trop 2013; 15; 144-50
3. Benko, A & Wilson, B. (2003) Online decision support gives plans an edge. Managed Healthcare Executive, Vol. 13 No. 5, p. 20.
4. Cross, E. Smith, D. Hoy, S. Nolte, I. Ackerman, M. Fransen, L. Bridgett, S. Williams, F. Guillemin, C. L. (2014) The global burden of hip and knee Osteoarthritis: estimates from the global burden of disease 2010 study. Annals of the Rheumatic Diseases, pages annrheumdis–2013, .
5. Gupta S., Kumar, D., Sharma, A (2011). Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering (IJCSE). Vol. 2 No. 2 pg 198-195, April, 2011. ISSN: 0976-5166.. Accessed on June 24, 2014.
6. Heidera B. (2011) Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part I. Caspian J Intern Med 2011;2:205-12.

7. Jiawei, H. and Micheline, K. (2006) Data Mining: Concepts and Techniques, Second Edition, Elsevier Inc.

8.

9. Rupali R Patil (2014)" Heart Disease Prediction System Using Naïve Bayes and Jelinek mercer Smoothing"

10.

11. Sheng B, Huang L, Wang X, Zhuang J, Tang L, Deng C, Zhang Y (2019) ''Identification of Knee Osteoarthritis Based on Bayesian Network: Pilot Study" JMIR Med Inform;7(3):e13562 URL: http://medinform.jmir.org/2019/3/e13562/ doi: 10.2196/13562 PMID: 31322132

12. Takahashi Hiroshi , Masahiro Nakajima , Kouichi Ozaki , Toshihiro Tanaka , Naoyuki Kamatani1 , Shiro Ikegawa (2010).: Prediction model for knee osteoarthritis based on genetic and clinical information. Arthritis Research & Therapy 2010 12:R187.

13. Teitel A.D, Zieve D. MedlinePlud Medical Encycolpedia. National Institutes of Health. "Osteoarthritis." Last updated: Sept 26, 2013. http://www.nlm.nih.gov/medlineplus/ency/article/000423.htm

14. World Health Organization. "Chronic Rheumatic Conditions." Chronic diseases and health promotion. 2014. http://www.who.int/chp/topics/rheumatic/en/

15. Web1 https://www.mayoclinic.org/diseases-conditions/osteoarthritis/symptoms-causes/syc-20351925 Accessed 17th oct, 2019

How to cite this article: OLAYEMI. Risk Prediction of Osteoarthritis using Data Mining Classification Techniques. Jtsm 9 (2020)