مقارنة خوارزميات التصنيف باستخدام أداة التنقيب عن البيانات Orange

حميدة أوشاح[1] سميرة الشفح[2] نجوى الذيب[2]
[1]- جامعة صبراتة كلية الهندسة - قسم الهندسة الكهربية والالكترونية
[2]- جامعة الزاوية / كلية التربية / قسم الحاسوب

الملخص :

تساعد تقنيات التنقيب عن البيانات في العثور على المعرفة المخفية ضمن مجموعات بيانات الأمراض التي يمكن استخدامها لتحليل سلوك المرض المستقبلي والتنبؤ به، حيث تتوفر تقنيات وخوارزميات مختلفة لاستخراج البيانات، منها التصنيف وهو الأسلوب الأكثر شيوعًا لاستخراج قواعد التنقيب من مجموعات البيانات الضخمة. في هذه الورقة، تم استخدام الخوارزميات: شجرة القرار، الانحدار اللوجستي، الشبكة العصبية، نييف بييز والجار الأقرب، والمقارنة بين ادائها، باستخدام أداة تنقيب البيانات Orange، لتصنيف البيانات الطبية للتنبؤ بألم الظهر والرقبة.

النتيجة في مرحلة التدريب: كانت المنطقة الواقعة تحت منحنى (AUC) لشجرة القرار 0.983 ،الشبكة العصبية 0.844 والجار الأقرب 0.839. بينما أعلى دقة(0.930) حققتها خوارزمية شجرة القرار

في مرحلة التنبؤ: كانت معدلات الاداء Calcification Accuracy = 0.767، Area Under Curve =0.686، accuracy = 0.733، Recall = 0.765 لخوارزمية نييف بييز، ودقة 0.750 للانحدار اللوجستي.

ففي مرحلة التدريب، تم الحصول على أداء أفضل بواسطة الخوارزميات شجرة القرار، الجار الأقرب والشبكة العصبية، بينما لوحظ أقل أداء من قبل نييف بييز.

في مرحلة التنبؤ كان أفضل أداء من قبل نييف بييز، ثم الانحدار اللوجستي.

Comparison of Classification Algorithms using the Orange Data Mining Tool

Hamida Oushah[1] Samira Ahshafah[2] Najwa Adeeb[2]

1-Electrical And Electronic Engineering Department / Engineering Faculty/ Sabratha University

2-Computer Department / Faculty of Education / University of Zawia

Abstract

Data mining techniques help to find hidden knowledge within disease datasets that can be used to analyze and predict future disease behavior, various techniques and algorithms are available for data mining. The classification is the most common technique for extracting mining rules from huge datasets. The algorithms: Decision Tree, Logistic Regression, Neural Network, Naive Bayes, and K-Nearest Neighbor will be presented in this paper, along with comparison among these algorithms, by using Orange data mining tool to classify medical data for back and neck pain prediction.

The result in the training phase:

The Area Under the Curve of Decision Tree 0.983, Neural Network 0.844 and K-Nearest Neighbor 0.839. The highest Precision and Recall were achieved with the Decision Tree algorithm; 0.930, 0.927.respectivly.

In the prediction phase:

The Calcification Accuracy, Area Under the Curve, accuracy, and Recall of Naive Bayes algorithm (0.767, 0.686, 0.733, 0.765) respectively, and the Calcification Accuracy of Logistic Regression was 0.750.

In the training phase the better performance was obtained by Decision Tree, K-Nearest Neighbor and Neural Network, whereas the lowest performance was noted by Naive Bayes.

In the prediction phase the best performance was by Naive Bayes, while less performance was noticeable by Logistic Regression.

Keywords: Data Mining, Orange mining tool, Classification algorithm, Decision Tree, Logistic Regression, Neural Network, Naive Bayes, K-NN.

1. Introduction

Nowadays, there are many available data and collections saved on the web, as well as there are databases that contain a lot of data and information that a person cannot analyze manually to discover knowledge and benefit from it, hence the need for automated tools that can help us convert those huge amounts of data into Useful information and knowledge. Data mining techniques and their applications are one of the most helpful tools in this field **[1]**.

The evolution of data extraction in different fields has led to the emergence of many algorithms, making it important to choose a suitable mining algorithm to obtain better results due to the difference and diversity of data, so what works well on certain data may not work like other data mining [2].

In this research, orange was chosen as a data mining tool to classify the prevalence of neck and shoulder pain in elementary school students and its relationship with school bags [3]. Five different classification techniques and methods were compared to predict neck and shoulder pain.

2. Related works:

- For the prediction of cardiovascular problems, (Weka 3.8.3) tools for this analysis are used for the prediction of data extraction algorithms like sequential minimal optimization (SMO), multilayer perceptron (MLP), random forest and Bayes net. The data collected combine the prediction accuracy results, the receiver operating characteristic (ROC) curve, and the PRC value. The performance of Bayes net (94.5%) and random forest (94%) technologies indicates optimum performance rather than the sequential minimal optimization (SMO) and multilayer perceptron (MLP) methods **[4]**
- Data mining tools were compared on the basis of their classification accuracy. According to the result of three data mining tools used in this paper, it has been observed that different data mining tools give different results on the same data set using different classification algorithm. WEKA shows the best rating accuracy when compared to Rapidminer and Orange **[5]**
- Researchers used Orange data mining tool to classify two types of selected medical data (Breast cancer and heart-disease) by applying decision tree,

Naïve Bayes and K-nearest neighbor (KNN) classification algorithms. The accurately of KNN classifier was more efficient in accuracy for the both given data set while the NB classifier was the lowest efficient from the selected data classifier **[2]**.

- Using information-mining methods requires some investment to predict disease more accurately. They assert that data-mining tools can tackle business addresses that are usually a lot of pain to identify. The use of information mining account gives effective results. The application of information mining systems to coronary artery disease treatment information can lead to a reliable implementation like that achieved in coronary artery disease diagnosis **[6]**.

3. Methodology

3.1 Data Mining Classification Algorithms:

This study focuses on the following classification algorithm for comparison in Orange data mining tool:

Decision Tree (DT): Decision tree algorithm is a classification technique that helps in decision making. Its structure is similar to a tree with axes that create a known tree, which means that it is a tree coordinated by a node called the "root". The inner node "root" contains partitions and partition properties. It is a trait test. The parentheses between the inner node and its followers contain the test results. Each leaf node is associated with a class label. A decision tree is generated from the training set. This decision tree is then used to classify groups with an unknown class label **[8,9]**. Decision tree development involves recursive partitioning of much of the preparation information, which is part of progressively homogeneous subsets based on tests related to at least one of the item estimates. These tests are spoken by the interviewer. The signs are distributed to the terminal (paper) axes by ways of making the part, for example, the bulk of casting votes **[10].**

Logistic regression (LR)

Logistic Regression (LR) is one of the most important statistical and data mining techniques employed by statisticians and researchers for the analysis and classification of binary and proportional response data sets .

Some of the main advantages of LR are that it can naturally provide probabilities and extend to multi-class classification problems[11]

Neural Network (NN):

Artificial Neural Networks (ANNs) are a Machine Learning paradigm inspired by the way biological neural networks in the nervous system process information. An ANNs is an interconnected system of collaborative interacting neurons (vector variables known as "nods") that produce an output (prediction, decision, forecast, classification or data abstraction) from a given input **[12]**.

Naïve Bayes (NB):

Naive Bayes (NB) is a classification algorithm for multiclass classification problems. It is called Naive Bayes because the calculations of the probabilities for each class are simplified to make their calculations tractable and it rely on Bayes's theorem, equation describing the relationship of conditional probabilities of statistical quantities.

This algorithm belongs to the good algorithms in data mining. The naive Bayes algorithm is simple probabilistic classification. This algorithm calculates a set of probabilities by calculating the frequency and combination of values in a particular data set [13,14].

k-nearest neighbor (K-NN):

The k-Nearest Neighbors (K-NN) algorithm is a data classification method for estimating the probability that a data point will become a member of one group or another based on the group to which the data points closest to it belong. The K-NN classifier is the classification of unlabeled observations by assigning them to the class of the most similar sorted examples. The characteristics of the observations are collected for both the training and test data set.

The k-nearest neighbor algorithm is a type of supervised machine learning algorithm used to solve classification and regression problems. However, it is mainly used for classification problems. It is considered a non-parametric method because it does not make any assumptions about the underlying data distribution. Simply put, KNN tries to determine which

group a data point belongs to by looking at the data points surrounding it [15].

3.2 Orange is a library of C++ core objects and routines that includes a large variety of standard and not-so-standard machine learning and data mining algorithms, plus routines for data input and manipulation. This includes a variety of tasks such as pretty-print of decision trees, attribute subset, bagging and boosting, and alike. Orange also includes a set of graphical widgets that use methods from core library and Orange modules. Through visual programming, widgets can be assembled together into an application by a visual programming tool called Orange Canvas **[7].**

3.3 Dataset Description

A realistic dataset of students on musculoskeletal pain and school bag weight was used [3]. Table 1 presents the 8 attributes of the dataset. There are 409 records of students (204 male and 205 female) in the dataset. Their age range is between 8 and 16.

Table 1. Description of datasets

No	Attribute	Values
1	Age	Min Value:6 Max value:17
2	Sex	Male:204 Fmale:205
3	Class floor	Values: 1, 2, 3
4	Transportation	(1) Car (2) walk
5	Transport time in minutes	Min Value:0 Max value:60
6	Method of carrying the bag	Values: 1, 2, 3
7	Waist belt	(1)YES (2) NO
8	using of waist belt	(1) YES (2) NO
9	carrying other things	(1)YES (2) NO
10	Are parents help?	(1) YES (2) NO
11	Student weight in kg	Min Value:22.6 Max value:75.9
12	the bag weight in kg	Min Value:1.5 Max value:7.9
13	Backache or neck pain	(1)YES (2) NO

3.4 Classification model:

In order to achieve the objectives set, the classification model was designed using the free data mining tools which are oranges as shown in Figure 1.

The dataset was divided into two sets, training data made up approximately 75% and test data made up the remaining 25%. The first dataset was used to

train and learn model, and the last dataset was used to examine and evaluate the model

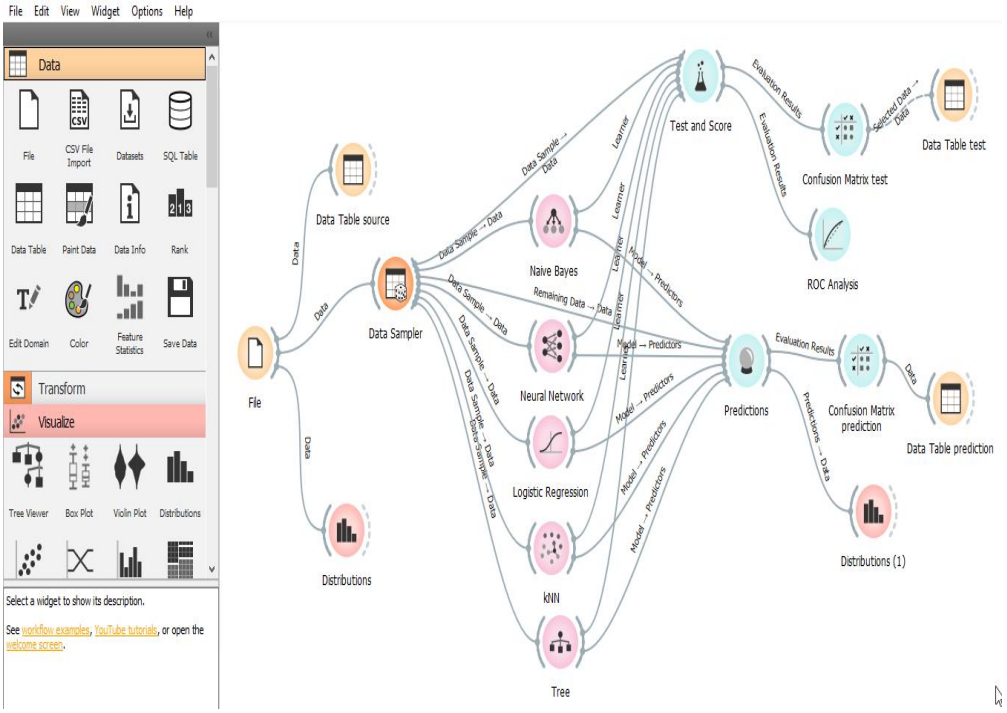

Figure 1. The classification

4. Results and discussion

The experiment for this research begins by opening the orange data mining tool and choosing the test data set to use. It then applies the selected classification algorithm and finally displays the evaluation results, as shown in Figure 2.

The CA (Classified Accurately) scores for DT, K-NN, and NN were 0.927, 0.800, and 0.780, respectively. LR was next, with 0.741 for Classified Accurately instances, while NB had the lowest CA value at 0.727.

Furthermore, the results in table 2, show a virtually optimal DT with an AUC (Area Under the Curve) value of 0.983, then a NN value of 0.844, and a K-NN value of 0.839. The analysis also shows that the highest Precision and Recall were achieved with the DT algorithm, the Precision value is 0.930 and the Recall value is 0.927. On the other hand, NB performed the worst in this case with an accuracy score of 0.689 and a Recall score of 0.727.

Furthermore, the analysis shows that the DT achieved his highest F1 measure of 0.928, while the LR achieved his lowest F1 measure of 0.696.

Table 2: The results obtained in the training Phase

Model	AUC	CA	F1	Precision	Recall
Tree	0.893	0.800	0.774	0.790	0.800
Logistic Regression	0.983	0.927	0.928	0.930	0.927
Neural Network	0.844	0.780	0.755	0.761	0.780
Naive Bayes	0.739	0.727	0.697	0.689	0.727
kNN	0.745	0.741	0.691	0.696	0.741

In the prediction stage, it can be seen that the NB algorithm is considered the best compared to the other algorithms where (CA = 0.767, AUC = 0.686, accuracy = 0.733, Recall = 0.765, and F1 = from 0.718), followed by the highest CA value is 0.750, which is considered the best, it was achieved by LR, the lowest CA value reached with the DT was 0.603, and the NN algorithm gave the highest values for F1-Meas and Recall (0.715, 0.745). Table 3 shows the prediction results for each algorithm. This comparison is presented graphically in Figure 2.

Table 3: Prediction results in the testing phase.

Model	AUC	CA	F1	Precision	Recall
Tree	0.538	0.603	0.620	0.643	0.603
Logistic Regression	0.636	0.750	0.698	0.703	0.750
Neural Network	0.631	0.745	0.715	0.709	0.745
Naive Bayes	0.686	0.765	0.718	0.733	0.765
kNN	0.527	0.716	0.674	0.659	0.716

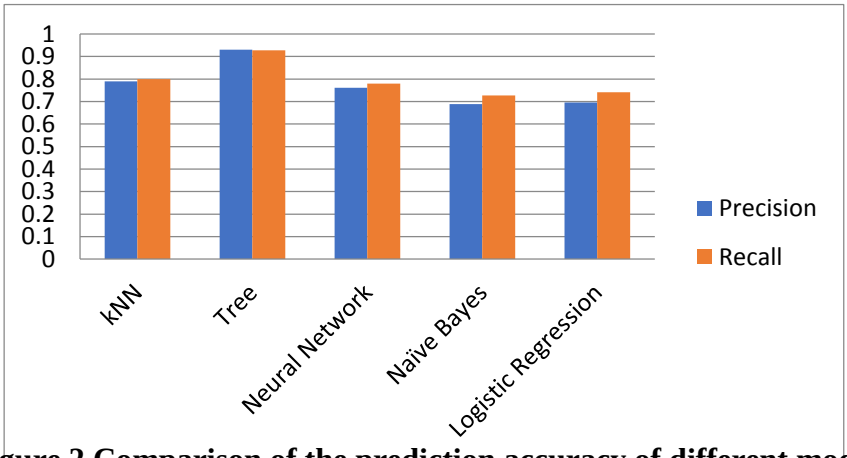

Figure 2 Comparison of the prediction accuracy of different model.

5. Conclusio

Our study demonstrate that, the better results in the training phase were achieved by using DT, LR, NN algorithms, while in the prediction phase the best achievement occurred by using LR, NN, NB algorithms compared to the rest of the algorithms.

6. References:

[1] Goebel, M., Gruenwald, L., A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations Newsletter, v.1 n.1 , June 1999, p.20-33.

[2] Zahraa Raji Mohi, " Orange Data Mining as a tool to compare Classification Algorithms ", Dijlah journal of Sciences and Engineering , Volume(3) Issue (3) Dec. 2020 , pp 13-23.

[3] Musculoskeletal Pain and School Bag Use: A cross-sectional Study Among Sabratha Students, Waheedah Aboulqasim Awushah,

[4] Rana T., Saleh B., Saedi A., Salman L., " Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: A case study in Iraq ", IJECE , Vol. 11, Dec 2021, pp. 5229-5239.

[5] Kauser Ahmed P, " Analysis of Data Mining Tools for Disease Prediction ", journal of pharmaceutical sciences & research, Vol. 9(10), Oct 2017, pp 1886-1888.

[6]. Beant K., Williamjeet S., "Review on Heart Disease Prediction System using Data Mining Techniques", IJRITCC , vol.2, 2014, pp.3003-3008.

[7] UCI Machine Learning Repository, Available at/http://archive.ics.uci.edu/ml/, (Accessed 22 April 2011).

[8] Davinder K., Rajeev B., Sunil G., " Review OF Decision Tree Data Mining Algorithms: ID3 AND C4.5 ", Proceedings of International Conference on Information Technology and Computer Science, July 11-12, 2015, pp 5-8.

[9] Shamrat F.M. , et al. " Performance Evaluation among ID3, C4.5, and CART Decision Tree Algorithms ", International Conference on Pervasive Computing and Social Networking (ICPCSN), March 2021, pp 19-20.

[10] Wahbeh Abd., et al, "A Comparison Study between Data Mining Tools over some Classification Methods", IJACSA, special issue on artificial intelligence, 2011, pp.18-26.

[11] Maalouf M., " Logistic Regression in Data Analysis: An Overview ", International Journal of Data Analysis Techniques and Strategy (IJDATS), July 2011, vol. 3(3), pp 281-299.

[12] Luna L., Artificial Neural Networks for Data Mining, July 2014.

[13]Kaviani P., Dhotre S., " Short Survey on Naive Bayes Algorithm ", International Journal of Advance Engineering and Research Development, Volume 4, Issue 11, November -2017, pp 607-611.
[14] Wibawa Aji et al. " Naïve Bayes Classifier for Journal Quartile Classification",International Journal of Advance Engineering and Research Development (IJAERD) Volume 4, Issue 11, November-2017, pp 607-611.
[15] Zhongheng Z. , " Introduction to machine learning: k-nearest neighbors", Annals of Translational Medicine , June 2016, vol.4(11), pp 218-218.