ISSN: 2456-8848, Vol. I, Issue. I, June-2017

International Journal
of
Linguistics & Computing Research

PREDICTIVE ANALYSIS USING CLASSIFICATION TECHNIQUES IN HEALTHCARE DOMAIN

Shweta Sharma
B. Tech(CSE):
Amity school of engineering
Amity University
Uttar Pradesh, India
shwetasharma1295@gmail.com

Sahil Anand
B. Tech(CSE)
Amity school of engineering
Amity University
Uttar Pradesh, India
sanand95@yahoo.co.in

Anant Kumar Jaiswal
Assistant Professor
Amity school of engineering
Amity University
Uttar Pradesh, India
akjayswal@amity.edu

Abstract— *The main objective behind data mining applications is to specify that data, a fact, number, text etc. can be processed by a software system which results out as a useful knowledge. Data mining is interactive and iterative process. It is a discovery of association changes, automatic and semi-automatic patterns, anomalies, different structures and also events in data. The main purpose behind the implementation of data mining classification techniques on mental health care data set is to develop an automated tool for recognition, identification and publication of relevant mental health care information. This paper aims to help experts in healthcare domain in decision making by doing predictive analysis on mental healthcare dataset using classifiers in weka. We have mainly applied 3 classifiers- Naïve Bayes, J48 and Multilayer Perceptron. Naïve Bayes is an advanced form of Bayesian's theorem, J48 is a decision tree based approach and Multilayer Perceptron is the simplest form in Neural networks. Dataset to be supplied to weka is Mental Healthcare survey with respect to IT industry all around the world. Data mining thus improves the quality of decision making process in its various applicative domains. Finally, this paper concludes by determining the major objective by illustrating data mining techniques and processes, methodologies and also the performance and accuracy observed in determining the best possible result from each existing technique so as to get the authentic information from the data set that we have supplied.*

Keywords— *Predictive analysis, Comparative study, Weka, Naïve Bayes, J48, Neural Network, Mental healthcare dataset.*

I. APPLICATION IN HEALTHCARE DOMAIN

Data mining extracts the most significant information from gross amount of data. Varieties of approach in data mining automatically analyse, classify and summarize the data into useful information. Data mining go hand in hand with some database techniques like statistics and visualization.[13] The types of data that could be mined using data mining is flat files, relational data, transactional data, special temporal data, time series data, images, videos, mixture, sequence and survey data.[10] Coming onto the application part -

[11] For physicians, data mining benefits are – that they would be able to identify effective treatments and best practices and on the other hand, patients receive more affordable and better healthcare services when they get to know themselves and their disorders.

Healthcare industries generates huge amount of complex data, which includes data of patients, hospital resources, disease diagnosis, survey records etc.

Physician needs to analyse and diagnose the symptoms through mental healthcare survey record using data mining tools and classifiers.

II. CLASSIFICATION AND PREDICTION

A. Predictive model or prediction method

[9] It uses few variables to predict anonymous values of other variables. It includes regression, time series analysis and

prediction. There are many tasks and their techniques that could be applied in different domains to achieve successful analysis or desired results.

It consists of activities like clustering, summarization, association rules, and sequence discovery. Some algorithms that are used in it are- decision trees, neural networks, regression analysis.

B. Descriptive model or classification method

This method works on the assumption that the set of objects or instances portrayed by some attributes corrosponds to distinct classes. The purpose is to build the classification model which will assign the correct class to formerly unseen, hidden and unlabelled object(s).

Some algorithms used in classification- discriminant analysis, decision tree, rule induction methods and genetic algorithms.

Classification works on discrete unordered data. Prediction works on continuous data.

The classifiers can be applied to the healthcare dataset with the intent of finding and making valuable predictions and important conclusions.

Moreover accuracy plays a vital role in order to offer prediction and conclusions. Accuracy depends on various conditions such as size of data set, number of attributes, the classifier used, mean absolute error, time to build model etc.

This paper gives the accuracy of Naïve Bayes, J48 and Neural Network classifiers when applied on the healthcare data set with their required number of attributes.

III. USE OF TOOL AND DATA

So much of availability of so many tools for data prediction and classification but here weka is going to cover up the entire dataset and the classifiers that will be used along with it. So it is important to have some basic knowledge of weka.

Weka stands for Waikato environment for knowledge analysis, developed at the University of Waikato, New Zealand. It is a set of machine learning software written in Java. The Weka tool is a collection of visualization tools and various algorithms for data analysis. In Weka dataset should be formatted to either ARFF or CSV file format. [5]

Some basic terms in weka-

Supervised and unsupervised learning: Supervised learning is learning from prototypes which help the software to construct models by assigning classes and supplying prototypes of each class.
Unsupervised learning can be seen as acquiring knowledge from observation & discovery. [6]

Training and testing:
Full data set is trained using remove percentage filter in the pre- process panel. Full data set is again loaded for testing the data set.
Testing set is prepared using invert selection property to true values by applying the correct percentage filter. Remove Useless filter: it removes the large and less varying data in the entire data set. The considered attributes are name, average weight, average size, life span, ori gin . Remove percentage filter is used to split
the overall data set into training and tested data set. In our data set, name is the largely varying attribute. Remove useless filter to remove the name attribute in the data set. [3]

The testing information is often created by specifying the proper share for the split of training set and test data. Training set of data is a part of data set to be trained.
In this paper, we have chosen techniques which train the classifier and model the dataset. Training specifies which instance is correlated to which attribute in data.

The test data shows the success of classifier upto what extent.
By default, training set is always higher in percentage so as to give the best results possible. [13]The testing information is created by clarifying the proper share for the split.

Use Training set: In this mode, the classifier is evaluated on the basis that how well it predicts the class of the instances on which it gets trained upon. [8]

Percentage split: Percentage split means that technique we have selected would be able to make the required training set and test data by itself. Percentage split is set to make 66% training set by default because it is good to analyse data when training set is higher in percentage, in order to get satisfying results. [8]

Mean absolute error: It is calculated to measure the predictions of conditional outcomes.
Mathematically,

$$MAE = 1/n \sum_{i=1--n} | f_i - y_j| = 1/n \sum_{i=1--n} | e^i|$$

Mean absolute error= average of the absolute errors = $|f_i - y_j|$,
Where,
f_i = prediction y_i= true value. [3]

False positive and true positive rate: When the predicted outcome and the actual outcome is equal in value then the prediction is called true positive (TP) and if the values are different then the prediction is false positive (FP).

Precision: Precision is a fragment of fetched instances that are relevant.

$$precision=tp/tp+fp$$

Recall: Recall is the fragment of relevant instances that are fetched.

$$recall=tp/tp+fn$$

Thus, precision and recall are based upon the measure of relevancy.
Precision is an exactitude or quality, while recall is the degree of integrity or quantity. [4]

Confusion matrix: It is a method of visualization typically used in more advanced form of supervised learning but this

paper is highlighting the classifiers and the data in basic form of both supervised and unsupervised learning features. One can do more with supervised learning feature if needed.
Learn more about false positive, true positive rates and the confusion matrix in [4].

Performance metrics: It is the study of analyzing classifier accuracy.
The performance evaluation is done using accuracy parameters or criteria.

Accuracy= 100-(numberofincorrectsample*100/totalnumberofsample)

This numeric value is usually equal or close to the percentage of correctly classified instances.

Data information- Dataset is an accumulation of statistical data, where each predictor impersonate distinct variable and each instance has its own portrayal. [7].

For this paper- Mental healthcare dataset has been taken for the study of classifiers & data mining in weka. Mental healthcare is the largest survey done on mental health in IT or technical industry in 2014. This survey aims to measure attitude towards mental health in technical workplace and examine the frequency of mental health disorders among technical workers.

We are basically interested in gauging how mental health is viewed within the technical or IT workplace and the prevalence of certain mental health disorders within the technical industry. We have chosen this so as to drive our work in raising awareness and improving conditions for those with mental health disorders in IT sector.

Attributes or Predictors in dataset are-

Timestamp,
Age,
Gender,
country,
State,
Self-employed?,
Family_history(family history of mental illness),
treatment?(took any treatment or not),
work_interference?(interference in work due to mental illness),
no_of_employees,
remote_work(also work outside the office),
tech_company?(IT or another?),
benefits(healthcare benefits your company provide),
care_options(are you aware of the care options available around you?),
wellness_program(ever happened in office?),
seek_help(how often you can seek help of another person?),
anonymity(are you protected?),
leave(how often you can take leave from office?),
coworkers_help(are you comfortable in sharing your mental condition with any of your coworker?),
mental_health_consequences
physical_health_consequences
obsv_consequnces?(This is the outlook of the entire data, means – do I believe that the patient with the above symptoms or attributes has any kind of serious mental disorder or consequnces of mental health disease?)
Comments.

IV. PREDICTIVE STUDY

All the working is done under experimenter window of weka tool because it is the most simplest of approaches to be used in order to analyse the data clearly and more precisely .This is the reason why it is widely used to analyse data in most of the industries.

A. Naïve Bayes

This algorithm is based upon Baysian's theorem along with the independent assumptions between predictors. It makes sure that all attributes shall contribute independently and equally. All attributes are statistically independent means that by knowing the value of one attribute says nothing about the value of other. Identical attributes may cause problem.

Naïve Bayes model is easy to build. It has no complicated iterative criteria to be estimated. These features make it useful for huge datasets classification. Apart from the simplicity, the Bayesian's classifier frequently works exceptionally and surprisingly wonderful. It is used universally as it performs much complex and sophisticated classification approaches.

Weka output and observations-

Total number of instances in data- 1259

Test mode used: Percentage split, 85% to train the classifier and rest 15% to test the data. 85% has been chosen to train the classifier as it gave the most accurate result, particularly in this technique. The correctly classified instances were 167 among 189 i.e. 88.3598%. That means only 22 instances were left incorrectly classified.(we are only left with 189 instances in total because 85%of data is used to train the classifier). Also, the mean absolute error was 0.132. So, this classifier is pretty good in giving accurate results.

Naïve Bayes working-

It provides an approach to calculate posterior probability. The posterior probability of assumptions could be estimated by Naïve Bayes reasoning in the presence of given data or knowledge. Generally, this kind of relevant and crucial knowledge is present in the predictors. Prediction forecast by identifying the symptoms in patients and then identification of sick patients from a block of sick and healthy ones is done. [7]

Following equation specifies the posterior probability-

$$P(C/X) = \frac{P(X/C)*P(C)}{P(X)}$$

Where,
P(C/X): Posterior probability of the specific class attributes.
P(X/C): Likelihood of predictor (or an attribute).
P(C): Prior probability of class.
P(X): Prior probability of predictor.

Now we are going to apply naïve bayes upon our dataset-

1. The posterior probability is calculated by making a frequency table for each attribute separately across the outlook. The target or outlook of the dataset is the consequences_observed or disorders.
2. Converting frequency table into multiple tables of likelihood instances. After that using Bayesian's equation, find out the posterior probability of each attribute.
3. The class with the maximum posterior probability would be the outcome of the prediction.

Applying it to first attribute of the data-

Attribute 1 is family_history and outlook of the dataset is consequences (disorder) of mental health disease.
That means it would help me to take a decision as-
Do I believe that the patient with the following symptoms has any kind of mental disease or consequences of mental health disease?

We simply have to calculate the posterior probability (pp) for both.

pp(yes_conseq_obv/no_family_history)=posterior probability=(0.0025*0.11)/0.6090=0.0004515

pp(no_conseq_obv/yes_family_history)=posterior probability=(0.2039*0.11)/0.6090=0.03682

Doing this to all the other remaining attributes concerning the respective class,
we saw that pp(all_no)=0.25 whereas pp(all_yes)=0.012

P(all_yes) is bit lesser than P(all_no). , which means that most of the people in technical or IT industry doesn't have any serious sort of mental disorder but sometimes it depends upon the various geographical locations all around the world which we will see further.

Performance metrics of naïve bayes=100-22*100/189
=100-2200/189
= (18900-2200)/189
=16700/189
=88.35%

Naive Bayes classifier obtained 88.35% accuracy which shows it's strong and independent assumptions among independent features provided in the class.

B. J48

J48 is top-down induction method of decision tree. J48 is the decision tree based algorithm and an extension to C4.5 algorithm. In this classifier a tree is constructed to model the prediction formation. The nodes of the tree denote a test on an attribute, branches personify the sequel of the test, leaf node occupies the label class and the topmost node represents a root of the tree. Model developed by decision tree classifier helps to predict new instances from the dataset. [5] Decision trees aims to get the smallest and most simplified tree. If we consider the information gain from the dataset, it is the amount of information gained by knowing the value of an attribute. Choose the attribute which produces the purest node which tends to provide greatest information gain. Simplifying the decision tree by implementing remove or add features in weka upon the dataset attributes, sometimes best results could be generated.

Decision tree - J48 working algorithm-

1. Basic algorithm
   - Tree can be composed in a topdown recursive aspect. Attributes or predictors are categorically
   - If the values are continuas, they are not considered in the beginning.
   - Test attributes are selected on the basis of a heuristic or statistical magnitude.
   - Information could be gained from predictors or attributes.

2. Extracting classification rules from trees
   - Rules are easier for humans to understand.
   - Trees represent the knowledge in the form of IF-THEN-ELSE rules.
   - One rule is created for each path from the root node to a leaf node.
   - The prediction class is in the leaf node. [4]

Weka output and observations-

Correctly classified instances are 1127 among 1259, which means that 89.515% of instances were correctly classified. Mean absolute error came out 0.1789.Incorrectly classified instances were 132 that is 10.48%. Test mode used is Use training set on checking 22 attributes. Sum of their weights-1259

Following is the tree visualization in weka using J48 algorithm. This graph is independent of geographical locations globally around the world.

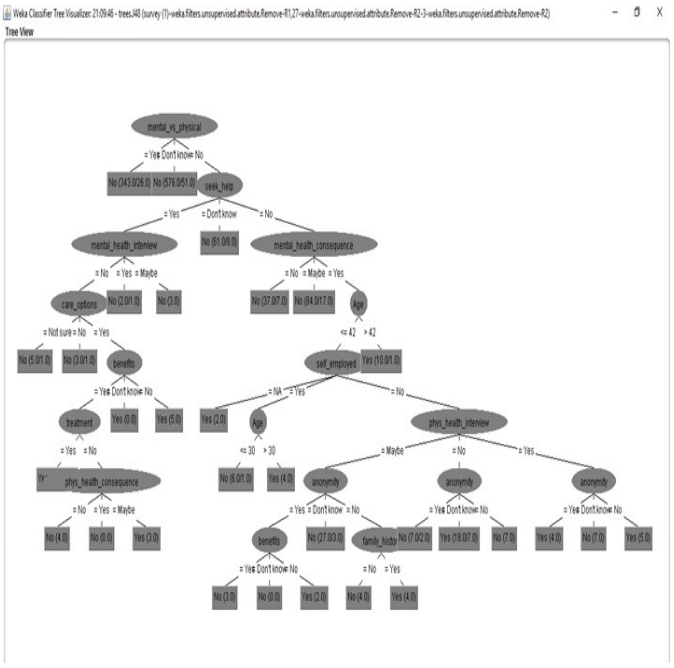

Figure 1 J48 Decision tree- top down induction

Considering the above J48 visualization of decision tree-
{
IF mental_vs_phy→Yes THEN seek_help.
IF seek_help→Yes THEN mental_health_interview.
IF mental_health_interview→Yes THEN care_options.
IF care_options→Yes THEN benefits.
IF benefits→Yes THEN treatment.
IF treatment→No THEN phys_health_consequence
}
ELSE IF
{
IF seek_help→No THEN mental_health_consequence.
IF mental_health_consequnce→Yes THEN age.
IF age<=42 THEN self_employed.
IF self_employed→No THEN phys_health_interview.
IF phys_health_interview→Maybe THEN anonymity.
IF anonymity→No THEN family_history ELSE benefits.
}
ELSE terminate.

Performance metrics of J48 classifier=100-13200/1259
= (125900-13200)/1259
=112400/1259
=89.28%

C. Neural Network

Neural network and it's 3 layers are so vast in their own respective fields. Neural network uses a set of processing elements called nodes correspondent to neurons in human brain. These nodes are interconnected & mutually dependent in a network. This eventually identifies the important patterns in data. This differentiates neural network from traditional computing algorithms which simply pursue instructions in a tenacious subsequent manner. [6]

Here in this paper you may get to know the basics about neural network and its simplest form of classification called Multilayer perceptron.

The classifier used in this paper to implement neural network is multilayer perceptron which is also known as artificial neural network.

ANN or artificial neural network or multilayer perceptron is an adjoin group of nodes analogous to the humongous network of neurons in human brain.

[6]

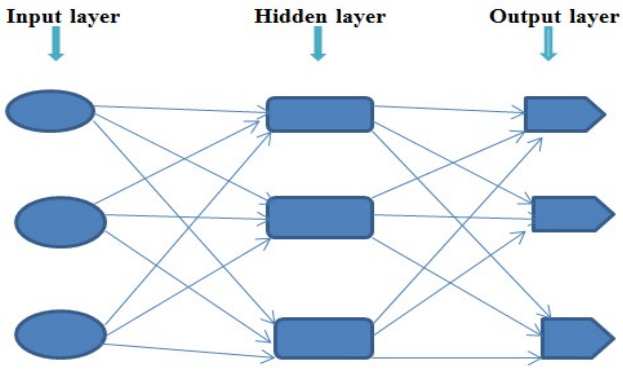

Figure 2 Signal flow diagram of multilayer perceptron

Work Flow-

Multilayer perceptron has the same kind of structure as that of single layer perceptron but including one extra layer called hidden layer. Each layer is made up of multiple units. [Fig.2] Inputs, as a data are dispensed simultaneously into the units which tend to develop the input layer.

These inputs are passed, weighted and then transferred simultaneously to the next layer, as a bunch of units, called hidden layer.

The outputs of the hidden layer are inputs to either output layer or any further hidden layer. Each output unit takes an input and the weighted sum of the outputs in the previous layer. [2]

A network of neurons is called as perceptrons in Multilayer Perceptron (MLP). Perceptron enumerates an individual output out of various real valued inputs by forming a linear combination in accordance with input weights and then probably placing the output through an activation function which is non-linear. [9]

The perceptron contains weights, summation processor and a function called activation function. Perceptron takes a weighted sum of inputs and outputs:

IF sum> any adjustable threshold value ($) THEN 1
ELSE 0.
Following are the equations:-
$W_1x_1+w_2x_2+\ldots\ldots w_nx_n >$ $ THEN output is 1.
$W_1x_1+w_2x_2+\ldots\ldots w_nx_n <=$ $ THEN output is 0.
W and x consists of real values. The input values are granted to the perceptron. But following is the condition to be kept in mind strictly while performing,
If the predicted output is same as that of expected output then the performance is considered as adequate or satisfactory, provided no changes in the weights are done.

In vice-versa situation, weights are needed to be changes in order to reduce error.

In this paper an artificial neural network is trained to extract the decision maker information and predictions out of the dataset. First, the MLP algorithm translates nominal data into numerical values and then gives an output. The output generated in the output layer from the instances and the attributes in the hidden layer and the input layer respectively, is like complex network visualization with numerous numerical values as a data. Neural network techniques are too outstanding to generate most precise and accurate results even though it takes little time to do so.

Weka output and observations-

Test option used:-Use training set
Correctly classified instances:-1246 out of 1259, means that 98.9674% instances were successfully classified. This is astonishingly good in analyzing data. Incorrectly classified instances were only 13 i.e. 1.0326%.This percentage would hardly affect the analysis of large dataset.

Performance metrics of neural network classifier
=100-1300/1259

= (125900-1300)/ 1259
=124600/1259
=98.96%

V. PREDICTION RESULTS AND PERFORMANCE EVALUATION

PREDICTION RESULTS: As per the above classifiers and their classification results, it's been predicted that, employers in IT companies relate their mental sickness to the symptoms they already have but, majority of people hardly requires doctor's consultation. They experience mental illness, stress, tensions, and workload etc. but at supreme and extreme level. Very few of them decided that they should ask doctor regarding their mental illness as, if they would not do that, then they might fall under worse conditions, such as depression, mental instability etc.

This signifies that IT industry generally provides good physical as well as mental health benefits to their employees especially in specific areas like US, Canada, UK, France, Netherlands, Australia and Ireland .

Being trapped in huge workloads on their head, employers yet somehow manage to deal with their insecurities and illness quite well.

That is why we have seen this outcome that employers in IT companies rarely observe any serious mental health issue.

After comparing the prediction results and performance in 3 different classifiers, following is the most refined output.

Performance evaluation: The parameters below are taken to configure out the working of each classifier in the paper and also to recognize which classifier gave better results on providing the dataset in weka so as to evaluate the performance of each. These results are only confined to weka. Change in tool may lead to variation in results.

Evaluation parameters	Naïve Bayes	J48	Neural Network
Correctly classified instances	167/189	1127/1259	1246/1259
Incorrectly classified instances	22/189	132/1259	13/1259
Mean absolute error	0.132	0.1798	0.013
Time to build model	0 seconds	0 seconds	0.09 seconds
Prediction Accuracy	88.35%	89.28%	98.96%

Table 1 Performance evaluation of naïve bayes, J48 & neural network

Multilayer perceptron has the maximum percentage of prediction accuracy. So we can say that neural network classifier (ANN) has predicted most accurate and precise results with minimum error intake. But if we consider the other side of Multilayer perceptron or ANN, it is bit slower than Naïve Bayes and J48. To make it better we can use some sophisticated and advanced data structure algorithms.

CONCLUSION

The objective of this paper was help experts in healthcare domain in decision making by doing predictive analysis on mental healthcare dataset using classifiers in weka, Data mining. In the end, we discussed about the data and classifiers in section V where we can see how well classifiers and data are inter-relating with each other and giving their best possible, precise and accurate results. Also we can see that how the frequency of mental health illness and attitude towards mental health vary by geographical location and what were the strongest predictors of mental health disorders.

However, the results generated and the experiments done are eventually helping in decision making process in mental healthcare domain.

Future work-

i. Implementing advanced neural network mechanisms to get more precise output.
ii. Visualize boundary of each classifier as every classifier creates boundary in instance space.
iii. Apply essemble learning methods. Essemble learning in data mining includes 4 major techniques- Bagging, Randomization, Boosting, and Stacking. They majorly improve prediction performance and produces reliable results and output that is hard to analyse.
iv. Confusion matrix calculations, FP, TP rates, precision and recall calculation
v. Cost analysis of each classifier.

REFERENCES

[1] Mental healthcare dataset- https://www.kaggle.com/osmi/mental-health-in-tech-survey

[2] L. L. Dhande and Dr. Prof. G. K. Patnaik, "*Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier*" , International Journal of Emerging Trends & Technology in Computer Science , Volume-3, Issue 4 July-August 2014, pg:313-319

[3] E. Bhuvaneswari, V. R. Sarma Dhulipala,"*The Study and Analysis of Classification Algorithm for Animal Kingdom Dataset*",Information Engineering Volume 2, Issue 1, March 2013. Pg:6-12

[4] A.Goyal and R.Mehta, "*Performance Comparison of Naïve Bayes and J48 Classification Algorithms*", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012), pg:1-5

[5] S. Joshi, R. Pandey and N. Joshi, *"Comparative analysis of Naive Bayes and J48 Classification Algorithms*", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 12, December 2015, pg: 813-817

[6] K. Amarendra, K.V. Lakshmi and K.V. Ramani, "*Research on Data Mining Using Neural Networks*", Special Issue of International Journal of Computer Science & Informatics, Vol.- II, Issue-1, 2, pg:1-8

[7] K. Ara Shakil, S. Anis and M. Alam, "*Dengue Disease Prediction Using Weka Data Mining Tool*", pg:1-26

[8] R. Kirkby & E. Frank, "*WEKA Explorer User Guide for* Version *3-4-3*", November 9, 2004, pg:1-13

[9] J. Jackson, "*Data Mining: A Conceptual Overview* ",Management Science Department University of South Carolina, Communications of the Association for Information Systems, Volume 8, 2002, pg: 267-296

[10] R. Gehrke, "*Database Management Systems"*, 3rd Edition, 2007, pg: 1-15

[11] O.R. Zaïane , "*Principles of Knowledge Discovery in Databases*" University of Alberta , 1999, pg:4-5

[12] M. Durairaj, V. Ranjani, "*Data Mining Applications In Healthcare Sector: A Study"* , International journal of scientific and technology research, Volume 2, issue – 10 october 2013, pg: 29-35

[13] G.K Gupta, "*Introduction to data mining with case studies"*, Monash University, Clayton, Australia, Prentice Hall of India pvt ltd., 3rd edition, 2011, pg: 1-2

.