

DISCRIMINATION ABILITY OF TIME-DOMAIN FEATURES AND RULES FOR ARRHYTHMIA CLASSIFICATION

Umut Arıkan and Fikret Gürçen

Boğaziçi University, Dept. of Computer Eng. Bebek, 34342 Istanbul/Turkey
umut.arikan@boun.edu.tr, gurgun@boun.edu.tr

Abstract- This study investigates relevant diagnosis information for arrhythmia classification from previously collected cardiac data. Discrimination ability of various time-domain attributes and rules were discussed for automatic diagnosis of arrhythmia using electrocardiogram (ECG) signals. Naive Bayes, C4.5, multilayer perceptron (MLP) and support vector machines (SVM) algorithms were tested on a number of the input features selected by correlative feature selection (CFS) method. Hot Spot algorithm was employed to extract a number of rules that is useful in diagnosing cardiac problems from ECG signal. 257 time domain features of 452 cases from a cardiac arrhythmia database [1] were used. Various testing configurations and performance measures such as accuracy, TP and FP rates, precision, recall and AUC were considered. The discrimination ability of selected-features and the extracted-rules were demonstrated.

Key Words- Arrhythmia, ECG, Rule extraction, Hot Spot algorithm, Classification, Naive Bayes, C4.5, multilayer perceptron (MLP) and support vector machines (SVM).

1. INTRODUCTION

Automatic diagnosis of cardiac arrhythmia via electrocardiogram (ECG) signals is quite responsible process and cost of faulty diagnosis is quite high [4]. It is generally popular to employ data mining techniques [2-4] for medical purposes both to reduce the error potential and to discover relevant information and/or hidden patterns from previously collected data. In ECG arrhythmia classification area, a number of studies were available [4-10]. They typically deal with extracting various features [2, 7, 9] improving the accuracy of classification performance or reducing the error rate and realizing real-time implementations [8-10].

In this study, we investigate the discrimination ability of various features and rules from time domain ECG signals and to measure the effectiveness of certain factors in the arrhythmia decision. It is also a verification process to highlight observable feature and a number of extracted rules derived from a time domain ECG signal with a cardiologist's view during making the arrhythmia decision. For example, in ECG holtering applications, it is necessary to distinguish normal and arrhythmia cases by a set of rules from ECG waveforms and intervals.

Among the contributions of our study, the performances of Naive Bayes, C4.5, multilayer perceptron (MLP) and support vector machines (SVM) were tested on a number of input features selected by correlative feature selection (CFS) method to validate the effectiveness of individual components. Then, we count the extraction of various rules affecting the discrimination of healthy and unhealthy ECG cases through

the Hot Spot algorithm [6]. A number of performance measures such as accuracy, precision, recall and F-measure, true positive (TP) and false positive (FP) rates, ROC and area under ROC (AUC) were employed to support the experimental results.

This paper is organized as follows: Section 2 briefs on dimension reduction. Section 3 summarizes the rule extraction algorithm and Naive Bayes, C4.5, multilayer perceptron (MLP) and support vector machines (SVM) classifiers. A brief on the performance measures including accuracy, precision, recall and F-measure, true positive (TP) and false positive (FP) rates, ROC and area under ROC (AUC) were given in Section 4. The experimental evaluation and the decision results are given and are discussed in Section 5. Section 6 provides a discussion of results and conclusion.

2. DIMENSION REDUCTION

2.1. Feature extraction and selection

The time and space complexity of a classifier or regressor primarily depends on the input data size. Thus dimensionality reduction is applied to the data to reduce input size without losing the integrity of it [3]. Two groups of dimensionality reduction techniques are: Feature Selection which aims to obtain a subset of the original dataset without losing any information and Feature Extraction which aims to find a new set of features that are the combinations of the original attributes.

2.2. Correlative feature selection (CFS)

Here we apply correlative feature selection (CFS) method to select a small number of time domain attributes [3]. Feature Selection is also called as *Subset Selection*. Among various approaches CFS is employed to select fewer number of features from the time domain ECG signal including Q, R, S, T intervals etc. CFS is a correlation-based filter method which finds out the greatly connected factors to the class result. It assesses the significance of an attribute subset by judging the individual predictive potential of each attribute as well as the redundancy amount within the subset. Subgroups correlated greatly with the results and low inter-correlation is determined. A heuristic to assess the worth or merit of a subset is used which deals with the usefulness of individual features to foresee the class label along with the level of inter-correlation among them. CFS algorithm quickly identifies and screens irrelevant, redundant, and noisy features.

3. RULE BASE DEVELOPMENT AND ARRHYTHMIA CLASSIFICATION SYSTEM

3.1. Rule-Base Development from ECG signal for arrhythmia classification

The rule extraction algorithm learns a set of rules displayed like tree structure that optimizes a variable/value of interest. Here, Hot Spot algorithm [6] learns a set of rules for target value optimization. To be more specific, HotSpot learns a set of rules for maximizing the number of correctly classified instances count. Since the target is a numeric attribute which is a number for each possible arrhythmia case; then looking for

the subgroups of data where the occurring probability is more than the overall average probability was considered. As a result, we obtain a set of rules derived from the ECG waveform that is effective for cardiac decision and rule-base development is an important issue of in arrhythmia decision to support practitioner.

3.2.Classifiers: Naive Bayes, C4.5, multilayer perceptron (MLP) and support vector machines (SVM)

A number of decision algorithms were employed in our study [5]. These are Naive Bayes, C4.5, multilayer perceptron (MLP) and support vector machines (SVM) algorithms. Naive Bayes is a simple and common probabilistic decision approach. C4.5 algorithm uses a number of rules based on input features. MLP is known as a neural approach for a number of nonlinear elements and connections. SVM describes an optimized decision boundary margin between classes using support vectors.

The probabilistic naive Bayes (NB) classifier learns and classifies fast as well but due to its ad-hoc restrictions placed on the graph, it might be hard to interpret the results. In NB, corresponding probabilities are obtained by the Bayes' Rule (Equation 2.2.1):

$$P(H / E) = \frac{P(E / H) * P(H)}{P(E)} \quad (3.2.1)$$

where H is the hypothesis and E is the evidence about hypothesis. As a result, NB must be applied with care to provide understandable results.

Decision trees (DT) are commonly used in data analysis with visualization properties and usage with noisy data points. In the process of building a decision tree, choosing the best attributes with maximum entropy is essential. For this purpose various methods can be used such as calculation of entropy thus the amount of information that is carried by it is maximized. Overtraining might be a problem in some scenarios such as too long learning phase (too deep tree) or rare training examples. Pruning might be a solution to overtraining. The C4.5 algorithm is a well known algorithm which can determine how deeply a tree may grow thus avoids data overfitting.

Multilayer Perceptron (MLP) is a common, neural algorithm in data analysis. Generally artificial neural networks (ANN's) were inspired from observations of human brain which is a quite sophisticated network of neurons. Analogically, the ANN is an interconnected set of 3 types of neurons: input, hidden and output ones. The input attributes (symptoms and measurements) form a first layer: the set S. Output is the diagnosis: the set D. The hidden neurons process the outcomes of previous layers. Each connection has an assigned weight whose values are adjusted with the use of appropriate algorithms, like back propagation (BP). BP algorithm minimizes the error at the output by changing the connection weights. The hidden layers add nonlinear features to the network. The aim of learning of an ANN is to solve a task T, having a set of observations and a class of functions F which is the optimal solution of the task T.

Support vector machines (SVM) find the optimal hyper-plane with the maximum margin between the classes by the solution of a global maximum or optimum for the quadratic optimization problem. In the linear separable vector space (x_i, y_i) case, the hyper-plane decision function can be written as

$$f(x) = \text{sgn} \left(\sum_{i=1}^m y_i \alpha_i \cdot (x \cdot x_i) + b \right). \quad (3.2.2)$$

Here the training samples with nonzero α_i 's are called "support vectors" (SV) and the decision function is constructed by only these vectors. The fewer number of SVs indicates better generalization capability.

In the non-separable case, the solution is to allow errors of the SVM by introducing positive slack variables $\xi_i, i = 1, \dots, m$, then: $y_i(w \cdot x_i + b) - (1 - \xi_i) \geq 0, i = 1, \dots, m$. If an error occurs, $\sum_i \xi_i$ becomes an upper bound on the number of training errors. The decision boundary is then determined by minimizing $\|w\|^2 / 2 + C(\sum_i \xi_i)^k$, where C is a user-defined parameter indicating the degree of penalty to errors. Now, the only difference from the linear separable case is that the α_i have an upper bound of C , and the support vectors can lie on the margin or inside the margin. The common approach for separation in this case is to map the original input space to a higher dimensional feature space using kernel functions. As the key idea of non-linear SVMs, kernel functions, $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, describe the inner products between vectors x_i and x_j . Then the optimization follows the same procedure as the linear SVM independent from the feature space dimension. Including the kernel inner products, the decision boundary is:

$$f(x) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i K(x, x_i) + b \right) \quad (3.2.3)$$

The proper choice of the kernel function is critical for the success of the SVM classifier.

3.3. Arrhythmia classification system

A sub-group of ECG measurements of heart activities are inputted to the decision support system based on naive Bayes, C4.5, MLP and SVM to predict cardiac arrhythmias (Figure 1). The decision system distinguishes normal and arrhythmia cases in two or more classes base and also derives a number of features and rules to support practitioner's findings. The CFS and Hot Spot algorithm introduce a number of features and rules using ECG signal measurements.

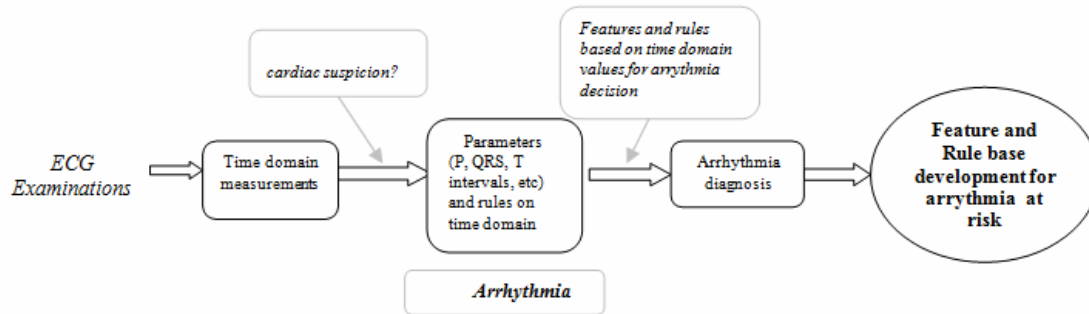


Figure 1. A Flowchart for cardiac readings at risk

As the first issue of our work, we search a number of time domain features that contribute to arrhythmia decision using CFS method. Trying to develop an accurate solution for a huge dataset may be quite exhaustive and inefficient. To provide accuracy, efficiency and optimization, a new and smaller data set that would be obtained from the initial data set was required because the factors do not affect the result equally most of the times. To get the most with less effort and time, contribution percentages of attributes have been investigated with CFS method. Attributes with the higher contribution percentages were selected using CFS method to reduce time and space complexity. Dealing with attributes that contribute at negligible amounts would cause inefficiency as well as higher costs.

During the early stages of this research a model of 452 samples with each having 279 features was employed. The contribution ratios were investigated to determine the features that contribute the most. We determined a number of features that requires attention in terms of arrhythmia decision procedure. After a number of experiments, the number of features was reduced to 22 which is almost 10 percent of the original input size. Also a smaller group of features up to size of 5 was formed based on the same principle to observe the effect of them on the decision procedure.

Near the input size of feature set above, 16 types of arrhythmia are considered as abnormal and a two-class decision problem is studied.

The other issue is to search a set of consistent rules for practical cases. The Hot Spot algorithm in the cardiac application gives a set of rules that help practitioner during its decision process. The accuracy of the derived rules are also tested by the performances of various classifiers. So the risk limits in automatic ECG decision may be adjusted by the choices of rules. Once the related rules are derived from ECG input space in time domain, a clear automatic decision on risky cases can be helpful.

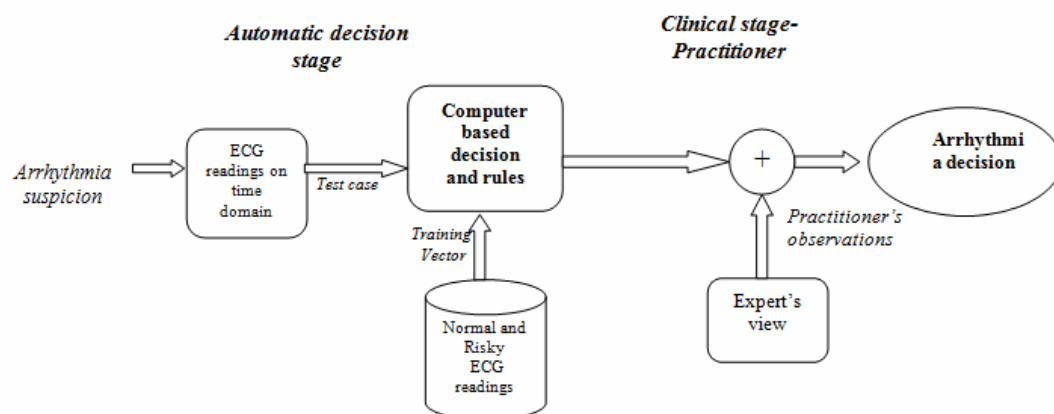


Figure 2. Arrhythmia Decision based on ECG readings

Noninvasive measurements of the time domain parameters to make dependable decision in a short time interval is a necessity and an instant minimum expected risk estimation is also vital point in cardiac cases. In this aspect, deriving a fast discriminative rule set becomes a fruitful line of enquiry.

4. PERFORMANCE MEASURES

Various performance metrics are employed in this research. To be able to evaluate discrimination performances of various classifiers with different subsets and make comparisons among them various metrics were used such as: *accuracy (Acc)*, *true positive (TP) and false positive (FP) rates of confusion matrix*, *precision*, *recall* and *F-measure*, *ROC* and *area under ROC (AUC)*.

For performance comparison, firstly, *accuracy (Acc)* is preferred which is the proportion of the total number of correct predictions. *Acc* ranges from 0 to 1 where *Acc* is closer to 1 then the performance of classifier increases.

For the classification results exhibition, a confusion matrix with the entries is a common description. The entries in the confusion matrix have the following meanings: *True Positive (TP)* is the number of correct predictions that an ECG reading is arrhythmic. *False Negative (FN)* is the number of incorrect predictions that a the reading is normal. *False Positive (FP)* is the number of incorrect predictions that the reading is arrhythmic. *True Negative (TN)* is the number of correct predictions that the reading is normal.

True Positive Rate (TPR) stands for the ratio of the number of correctly identified arrhythmic cases to the number of all arrhythmias in this work. It is also known as the *sensitivity*.

True Negative Rate (TNR) stands for the ratio of the number of correctly identified normal (healthy) cases to the number of all normal individuals. It is also known as the *specificity*.

So *False Positive Rate (FPR)* may also be calculated by subtracting specificity from 1 (i.e 100%).

Precision refers to the number of TPs divided by the total number of ECG readings labeled as belonging to the positive class or arrhythmic (including FPs or (TP+FP)). *Recall* is the number of TPs divided by the total number of cases in positive class (TP and FNs or (TP+FN)). Usually, *precision* and *recall* values are not considered in isolation. Instead, both are combined in a single measure such as *F-measure* which is a weighted form of them. The *F-measure* is given as: $(2 * recall * precision) / (recall + precision)$.

Receiver Operating Characteristic (ROC) curve is a two dimensional plot of *TPs* vs *FPs* for a binary classifier as its discrimination threshold changes. *ROC* analysis is related in a direct and natural way to cost analysis of diagnostic decision making. *ROC* is a more generic way than the error rate and helps to denote various classifiers on the same graph. The accuracy is measured by *area under the ROC curve (AUC)*. *ROC* and *AUC* have been widely accepted as the standards for describing and comparing the accuracy of diagnostic tests.

Another performance metrics we use when comparing the results is *practitioner's view*. In ECG readings, the aim is to diagnose arrhythmias from ECG recordings with a minimum risk. For this purpose various other tests are considered and a consensus of all

the tests is employed for a reliable diagnosis. From this point of view, the ratio of missed arrhythmia cases that are classified as normal or healthy becomes very important. In the computer based cardiac decision making, we use FP rate and FN rate. FP rate is false positives' rate which indicates the proportion of number of arrhythmic recordings who classified as normal recordings to total number of arrhythmic cases and FN rate is false negatives' rate which is the proportion of number of normal recordings who classified as arrhythmic to total number of normal cases.

In knowledge extraction phase, performance measure to understand how applicable the obtained rule-set is defined as *precision*. Precision corresponds to accuracy in classification, which is the proportion of truly classified instances to total number of instances which meets rule's conditions.

5. EXPERIMENTAL RESULTS

452 ECG recordings each with 279 attributes were analyzed retrospectively [1]. A small set of the recordings with the same conditions were also collected to observe the attributes. Group 1 includes 245 normal recordings and group 2 includes 207 arrhythmic recordings with 16 different types. Two classes defined as "normal" (healthy) and "arrhythmic" (unhealthy).

To reduce the size of input space, a number of dimensionality reduction experiments were performed and 22 attributes were selected initially. The input size was reduced to less than %10 of the original size and was found to be successfully used for the experiments. The 22 attributes include age (years), sex (0 = male; 1 = female), height (centimeters), weight (kilograms), QRS duration (in msec), P-R, Q-T, T, P intervals (in msec), QRS, T, P, QRST vectors angles (in degrees), heart rate (beats per minute), Q, R, S wave quantities such as average width and intrinsic deflections. Furthermore we select 5 of the attributes such as QRS duration, Q-T and T interval, T vector angle and Q wave average width and performed the experiments with a smaller input size.

The reduced set of features from the ECG time waveform were the data chosen to demonstrate an ability for naive Bayes, C4.5, SVM and MLP classifiers to extract information for arrhythmia diagnosis in a two-class (normal and arrhythmic cases) problem. The classifier performances were optimized in terms of accuracy values with various possible parameters. For example, SVM besides their Lagrangian formulation, can be optimized in two aspects : (i) coefficient C controlled capacities (ii) the anisotropic radial basis in the Gaussian kernel transformation controlled classifier functions.

All data from 452 ECG recordings are divided into two halves: training and test data. 10 fold cross validation with a total of 45 values is constructed by a combination of 23 normal and 22 arrhythmic cases at each experiment is used for testing. The experiments are repeated 10 times for unseen test values and the average performance is computed. Two-class experiments were performed with 22 feature subset. Figure 3 shows the results.

Classifier	Naive Bayes	C4.5	MLP	SVM
Accuracy %	70.8	74.6	67.7	71.5
TP rate	0.708	0.746	0.677	0.715
FP rate	0.33	0.269	0.334	0.261
Precision	0.738	0.747	0.676	0.714
Recall	0.708	0.746	0.677	0.715
F-Mea	0.69	0.743	0.675	0.712
AUC	0.784	0.786	0.705	0.731

Figure 3. Automatic ECG classification performance results.

It is observed that all of the classifiers achieve well performance values for the experimented ECG data set. Accuracy and TP and FP rates were understandable for a classification task with certain difficulty. The other measures precision, recall, F-measure and AUC gives relevant performance values with the previous results.

Furthermore HotSpot rule-extraction algorithm was employed for observing effective rules on arrhythmia diagnosis. In this work, we used the *Hotspot* algorithm in *Weka* [16], which inspects the training data and generates the *association* rules corresponding to a class label in the form of a tree. In the clinical situation, an expert will probably look at the time waveform and make a decision about normal and arrhythmic cases (Figure 4). Here we produce a set of rules for

Figure 4: The HotSpot algorithm association rule example.

An example set of HotSpot association rules:

```

CarArrhythmiaType=1 (54.2% [245/452])
  QRSdur <= 87 (68.29% [168/246])
    / ChDI-Qwave <= 87 (75.24% [158/210])
      / | ChDI-Qwave > 58 (82.45% [155/188])
        / | | T <= 91 (86.21% [150/174])
          / | | | T-int > 117 (87.13% [149/171])
            ...

```

This rule can be explained partially as follows:

```

If QRSdur is less than or equal to 87
  if ChDI-Qwave is less than or equal to 87
    IF ChDI-Q wave is more than 58
      If T is less than or equal to 91
        If T-int is greater than 117
          .....

```


Example case with above values may contribute to a decision for a possible scenario for cardiac arrhythmia type to be case 1 (or abnormal).

6. DISCUSSION AND CONCLUSION

Prognostic role of ECG recordings in cardiac cases can not be overlooked. The decision and rule extraction system for arrhythmia detection is proposed. In the first stage time waveform features were automatically extracted and reduced and these features are retrospectively employed for making a two-class cardiac decision as “normal” and “arrhythmic”. In the second stage a number of rules were extracted. It is concluded that the proposed decision and rule extraction system may be a fruitful line of enquiry for distant heart monitoring and other automatic decision applications.

This study considers a number of ECG time waveform features. These features were also important for eye examination of the recordings by a practitioner to make a decision of normal and arrhythmia cases. Using the reduced set of time domain features, we demonstrate an ability for naive Bayes, C4.5, SVM and MLP classifiers to extract information for arrhythmia diagnosis in a two-class (normal and arrhythmia cases) problem. Furthermore a rule set is associated with for given time domain parameters. This also helps to understand the physician decision during a ECG inspection.

A small and effective input size (a total of 22 attributes) is used in the experiments. The parameters include age, physical quantities, QRST waveform intervals, angles, etc. The time domain discriminant features are supportive computation methods and may draw physician's attention to evidences of unexpected arrhythmia events that need further analysis of ECG waveforms.

Although the data given is extracted from a small population, the inducted rules may still be used for the other cardiac cases to help physician in a fast manner. The number of attributes also change up to the size of 22 and examined in terms of various performance values including *accuracy (Acc)*, *true positive (TP)* and *false positive (FP) rates of confusion matrix*, *precision*, *recall* and *F-measure*, *ROC* and *area under ROC (AUC)*. These evaluation methods are common and they indicate various aspects of classifiers and issues including accuracy, balance of given data and diagnosis efficiency such as miss of arrhythmic cases.

The probabilistic naive Bayes, rule-based C4.5, neural SVM and MLP classifiers were effectively employed in diagnosis of arrhythmia cases. It is found that they all performed reasonably well in terms of *accuracy*, *specificity* and *PPV*. It is observed in a small data set that they give a balanced pair of *sensitivity* and *specificity*.

As a conclusion, the above classifiers and rules may give an opportunity to obtain the results not very obvious at first glance and to easily tune with only a few parameters such as risk estimation in cardiac cases.

Using time domain ECG waveform in cardiac risk management is based on very few restrictive considerations and may reveal some important features and rules overlooked by many other methods. Therefore they may become an option of choice at risk decision of cardiac cases. Although a very limited data in cardiac cases are retrospectively analyzed, the results are meaningful. One may extend this study with new and large

sets of data and alternative rules and methods better suited for processing large data sets.

Authors' Contributions

UA carried out the numerical experiments. FG supervised and prepared the manuscript.

6. REFERENCES

1. <http://archive.ics.uci.edu/ml/datasets/Arrhythmia>.
2. N. Özlem Özcan, Fikret Gürgen, Fuzzy Support Vector Machines for ECG Arrhythmia Detection, Intl. Conf. of Pattern Recognition, 2010.
3. Witten I. H., Frank E., Data Mining, Practical Machine Learning Tools and Techniques, 2nd Elsevier, 2005.
4. Spiegelhalter D., Knill-Jones R., Statistical and knowledgebased approaches to clinical decision support systems, with an application in gastroenterology. Journal of the Royal Statistical Society, No 147, 35-77, 1984.
5. Jinkwon Kim, Hang Sik Shin, Kwangsoo Shin, Myoungho Lee, Robust algorithm for arrhythmia classification in ECG using extreme learning machine, *BioMedical Engineering OnLine*, 2009.
6. <http://rapid-i.com/wiki/index.php?title=Weka:W-HotSpot>.
7. Ghongade, R., & Ghatol, A., An Effective Feature Set for ECG Pattern Classification. *LNCS*, 4901, 25-32, 2007.
8. Soman, T., & Bobbie, P. O., Classification of Arrhythmia Using Machine Learning Techniques. *WSEAS Transactions on Computers*, 4(6), 548-552, 2005.
9. Neagoe, V. E., Iatan, I. F., & Grunwald, S., A neuro-fuzzy approach to classification of ECG signals for ischemic heart disease diagnosis. *AMIA Annu Symp Proc*, 494-498, 2003.
10. Zhang, H., & Zhang, L. Q., ECG analysis based on PCA and Support Vector Machines. *Neural Networks and Brain*, pp. 743-747, 2005.