Reviewer's report

Title: Using n-gram analysis to cluster heartbeat signals

Version: 1 Date: 8 April 2012

Reviewer: Chunhua Bian

Reviewer's report:

The manuscript presents a method named AIIA to use the Simple K-Means algorithm for symbolization, which offers a new way to represent subtle variations between two interbeat intervals. Then the authors classify, i). the patients with Atrial Fibrillation (AF), Congestive Heart Failure (CHF) and healthy people, ii). the patients with apnea with AIIA method. Six classifiers, including Bayesian Network, Logistic, Naïve Bayesian, Neural Network, Support Vector Matrix (SVM) and Tree-J48 are used.

The present version of the manuscript still needs some revision. I would recommend publication provided the authors have considered the remarks I give below.

Major Compulsory Revisions:

1. On page 4, "At present, there are three different approaches for using non-linear symbolic sequences to represent heart beat time series." "The first approach is based on the deviation of the heart rate time series from the local mean", "The second approach is to symbolize the increase or decrease of the momentary heart rate by two different symbols.""The third approach is to divide the range between minimum and maximum heart rate into a few equidistant intervals, each interval is denoted by a unique symbol."

The reference was not summarized completely. In Ref(1) and Ref(2) permutation entropy and modified PE method has been studied, which is symbolized by the order of the interval value.


2. The databases used in the present paper are from physionet, which have been studied by many researchers. The authors should have a comparison between the results in the presented paper with those of other methods with the same database.

3. What is the data length in the analysis of this paper?
4. On page 8, “This research uses 1-gram, 2-gram and 3-gram for analysis, which includes 18,278 different kinds of string combinations.” How does the number 18,278 been calculated? And does it have relevance with the cluster number k?

5. In the paper the authors declares that both experiments acquired the best category results when using the Bayesian Network. While from the accuracy value in table 2 and table 5, we find no Essential difference between different classifiers.

6. The English expression in the paper need to be improved.

**Level of interest:** An article of importance in its field

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests