Discretization of Continuous ECG based Risk Metrics Using Asymmetric and Warped Entropy Measures

A Singh, J Liu, JV Guttag

Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Abstract

We investigate several entropy based approaches to finding cut points for discretizing continuous ECG-based risk metrics. We describe two existing approaches, Shannon entropy and asymmetric entropy, and one new approach, warped entropy. The approaches are used to find cut points for the end point of cardiovascular death for three risk metrics: heart rate variability (HRV LF-HF), morphological variability (MV) and deceleration capacity (DC). When trained on multiple instances of training set containing 2813 patients, warped entropy yielded the most robust cutoffs.

The performance of the cutoffs obtained using warped entropy from the training sets was compared with those in the literature using a Naive Bayes classifier on corresponding test sets. Each test set contained 1406 patients. The resulting classifier resulted in a significantly ($p<0.05$) improved recall rate at the expense of a lower precision.

1. Introduction

In medicine, multiple risk metrics are used to evaluate the risk profile of a patient. These risk metrics consist of both continuous (e.g. age) and categorical (e.g. history of diabetes) variables. From a clinical perspective, categorization of continuous variables into high risk, medium risk and low risk is useful since it offers a simple risk stratification tool for both physicians and patients. Furthermore, many machine learning algorithms generate better models when discretized variables are used [1].

The purpose of discretization is to identify cutoffs that partition a sequence of values into subsequences that exhibit good class coherence. Supervised discretization methods use class information during the discretization process [1]. Several supervised discretization approaches are based on measuring class entropy of a sequence. Class entropy of a sequence is a measure of uncertainty of the class labels of the examples that belong to the sequence. It is a measure of information where a lower value of entropy corresponds to higher amount of information. Entropy based discretization algorithms evaluate each candidate cut point based on a joint measure of the entropy of the two resulting subsequences generated by the cut point.

For the endpoint of cardiovascular death (CVD), deaths (positive outcome) are much less represented in the datasets than non-deaths (negative outcome). If one uses the traditional Shannon entropy for discretization in such highly unbalanced datasets, a subsequence with an equal class distribution of positive and negative outcomes is assigned the maximum uncertainty value of 1. However, for highly unbalanced datasets, such a subsequence actually contains a lot of information. It suggests that patients who belong to the subsequence are at extremely high risk of cardiovascular death.

In this paper we present a novel supervised entropy based discretization method that handles unbalanced data. In Section 2, we present a general outline of our proposed discretization algorithm. In Section 3, we discuss different types of entropy measures. In addition to Shannon entropy, we present asymmetric entropy and a new measure, warped entropy. In Section 4, we evaluate each of the entropy measures in terms of stability of the cutoffs identified. The evaluation is for the endpoint of CVD in a population of roughly $4,000$ patients. Next, we compare the cutoffs derived from warped entropy with those found in the literature based on the performance of a Naive Bayes classifier on recall and precision. The Naive Bayes classifier built using cutoffs derived from warped entropy yielded a significantly ($p<0.05$) higher recall and lower precision than the classifier built using the literature cutoffs. Finally, in Section 5, we present some conclusions and recommendations for further research.

2. Proposed discretization algorithm

Let V be a continuous variable and Ω be a sequence of N examples sorted in an ascending order of the continuous variable. Each example is a pair $[v, l]$ where v is a value of the continuous variable and l is the class label.

We find the midpoint of the value of variable V for each successive pair of examples in Ω. These midpoint values are the candidate cut points. Each candidate cut point C

partitions Ω into two sequences, Ω_1 and Ω_2, where Ω_1 contains examples with $v < C$ and Ω_1 contains examples with $v > C$. Next, we find the class entropy of each subsequence Ω_j using an entropy measure. We then use **W**eighted **J**oint **E**ntropy (WJE) to evaluate the quality of the partition generated by a candidate cut point C:

$$\mathbf{WJE}(C, \Omega) = \frac{|\Omega_1|}{|\Omega|} H(\Omega_1) + \frac{|\Omega_2|}{|\Omega|} H(\Omega_2) \qquad (1)$$

where H is an entropy measure.

The C that minimizes $\mathbf{WJE}(C, \Omega)$ is selected as the cut point for binary discretization for Ω.

Equation 1 can be easily generalized to generate n cutoffs. However, the size of the set of candidate cutoffs is of size $O(N^n)$. However, since N is typically very large, we instead use a greedy approach. To find n cutoffs for $n > 1$, we first perform a binary split on the entire sequence Ω to identify the first cutoff. To find the next cutoff, we identify the subsequence Ω_{MaxEnt} of Ω which has the maximum class entropy (uncertainty). A binary split is then performed on the subsequence by picking the cutoff that minimizes $\mathbf{WJE}(C, \Omega_{MaxEnt})$. This process is repeated until n cutoffs are found.

We use a Bagging (**B**ootstrap **agg**regat**ing**) algorithm in an attempt to avoid over-fitting. Given a sample sequence of size N, bagging generates r new training sequences, also called replicates, of size N by uniformly sampling examples with replacement from the original sample sequence [2]. Once the cutoffs for all the replicates are identified for a fixed number of cutoffs n, we take the median of the distribution to identify the final cut point.

3. Entropy measures

We tried this approach using a symmetric entropy measure (Shannon entropy) and two different types of non-symmetric entropy measures (asymmetric entropy and warped entropy). The entropy measures can be generalized to the case of k class labels, however, we restrict our discussion to two class labels for ease of explanation.

3.1. Shannon entropy

Shannon entropy [3] is the most commonly used entropy measure. Let the class-label variable L take two different values, with probabilites p_1 and p_2 respectively. The Shannon entropy of a subsequence S, with class distribution $p_1(S)$ and $p_2(S)$ is given by

$$H(S) = -p_1(S).log_2 p_1(S) - p_2(S).log_2 p_2(S) \qquad (2)$$

Since Shannon entropy is a symmetric measure, it is maximized when the two classes in a S are present in equal proportions (Figure 1).

3.2. Asymmetric entropy

For the binary class case, the asymmetric entropy measure of a subsequence S derived from a parent sequence P is given by,

$$H(S, P) = \frac{p_1(S).p_2(S)}{(-2.z_1(P) + 1).p_1(S) + (z_1(P))^2} \qquad (3)$$

where, $p_1(S)$ and $p_2(S)$ are defined as in Section 3.1. The variable z_1 is a function of the parent sequence P such that $z_1(P) = p_1(P)$. The value of z_1 determines the asymmetry of the entropy measure. Specifically, for a given parent sequence P, the function $H(S, P)$ is maximized when $p_1(S) = z_1(P)$.

By setting z_1 to the probability of class 1 in the parent sequence, we are essentially considering the distribution of the parent sequence to be the most uninformative. Any subsequence with $p_1(S) = z_1(P)$ has the same distribution as the parent sequence. Therefore, it does not provide any additional information and is assigned the maximum entropy value of 1 (Figure 1).

The concept of asymmetric entropy was first introduced by Zighed et.al. [4].

3.3. Warped entropy

This entropy measure is a modified version of Shannon entropy (Section 3.1). In Section 1, we motivate the concept of asymmetry based on the fact that the prior distribution of classes is highly unbalanced in many medical datasets. One way to deal with the class imbalance is to assign greater weights to examples from the minority class than those from the majority class so that the distribution of the weighted samples is balanced.

The warped entropy measure of subsequence S derived from a parent sequence P is given by,

$$H(S, P) = -\sum_{l=1}^{2} p_l^*(S, P).log_2 p_l^*(S, P) \qquad (4)$$

where,

$$p_l^*(S, P) = \frac{p_l(S).w_l(P)}{w_1(P).p_1(S) + w_2(P).p_2(S)} \qquad (5)$$

The variables w_1 and w_2 are weights assigned to examples of class 1 and 2 respectively. Specifically, $w_l(P) = \frac{z_1(P)}{z_l(P)}$ where $z_l(P) = p_l(P)$, as defined in Section 3.2.

4. Experimental evaluation

We tested our method on data from the MERLIN-TIMI 36 trial [5]. We used data from 4219 non-ST elevation

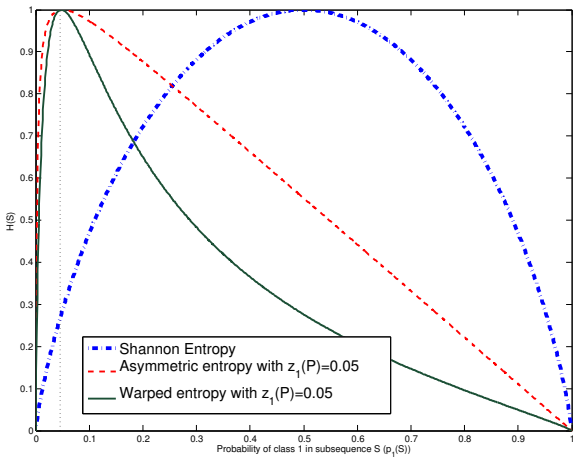

Figure 1. Entropy measures.

acute coronary syndrome (NSTEACS) patients and considered cardiovascular death within 90 days as an endpoint. There were 82 (2%) cardiovascular deaths within 90 days. Three electrocardiographic (ECG) risk metrics: heart rate variability (HRV) [6] , deceleration capacity (DC) [7] and morphological variability (MV) [8] were computed from the first 24 hours of ECG recording. For HRV, we computed HRV-LF/HF [6].

In our experiments, each training set contains 2813 patients; its corresponding test set contains a disjoint set of 1406 patients.

First, we evaluated the stability of the cutoff value generate using different entropy measures. We generated 100 instances of disjoint training and test sequences. We implemented the discretization algorithm on each training sequence using r=100 replicates for bagging to generate a binary split. The coefficient of variance (COV) for a single cutoff using different entropy measures are shown in Table 1. The worst (highest) COV among the three ECG measures for each entropy measure is highlighted in the table.

Table 1. Coefficient of variance for a single cutoff using different entropy measures

	Coefficient of variance		
Risk Metric	Shannon	Asymmetric	Warped
DC	0.11	**0.36**	**0.15**
HRV LF-HF	**0.31**	0.18	**0.15**
MV	0.06	0.06	0.14

The high instability exhibited by the asymmetric entropy measure was caused by its sensitivity to outliers. The outlier sensitivity can be explained by the shape of the asymmetric entropy function when a dataset is highly unbalanced, as shown in Figure 1. The asymmetric entropy curve falls sharply as $p_1(S)$ approaches 0 from $p_1 = z_1 = 0.05$, but the rate of decrease is slow when we move away from the maximum entropy point towards $p_1 = 1$. The latter property causes the entropy (uncertainty) to still be high for $p_1 > z_1$. Therefore, the evaluation function favors cutoffs where one of the subsequences has $p_1 < z_1$. This makes this measure susceptible to outliers from the minority class.

The high instability shown by cutoffs derived from Shannon entropy measure (Table 1) can be attributed to the fact that it places equal weights on both minority and majority examples despite their unbalanced prior distribution.

Next, we compared the cutoffs found using an entropy measure with those found in the literature (Table 2). Because of the robust performance of warped entropy compared to other entropy measures, only the warped entropy measure is used for this experiment. For this experiment, we used 100 instances of disjoint training and test sequences. For discretization, we used $r = 100$ replicates of a training set.

Table 2. Cutoffs for the ECG based risk metrics

	Cutoffs	
Risk Metric	Literature	Warped
DC	2.5 , 4.5	4.0, 6.0
HRV LF-HF	0.95	2.0
MV	50	40

We built two Naive Bayes (NB) classifiers[1] [9] from each of the training sequences using the cutoffs from the literature for one and the cutoffs derived using the warped entropy measure for the other. A NB classifier is a probabilistic classifier. Therefore, for each example in the test set, it generates a probability of death given the values of all three risk metrics: DC, HRV and MV. We used the death rate of the population (2%) as the threshold such that patients with probability of death>2% were considered as high risk.

The performance of the NB classifiers built from the training sequences were evaluated on the corresponding test sequences based on recall and precision on the minority class:

$$Recall = \frac{\text{true positives}}{\text{true positives + false negatives}} \quad (6)$$

$$Precision = \frac{\text{true positives}}{\text{true positives + false positives}} \quad (7)$$

[1]The NB classifier was built using Bayes Net Toolbox by Kevin Murphy available at http://code.google.com/p/bnt/.

The number of cutoffs derived using the entropy measure was the same as those used in the literature for risk stratification for cardiovascular deaths. Table 3 shows the mean performance of the NB classifier on the 100 instances of test sequences as measured by recall and mean precision. Based on a paired-sample t-test [10], the warped entropy cutoffs yielded a significantly ($p<0.05$) higher recall but a lower precision than the cutoffs in the literature.

Table 3. Mean±std(standard deviation) of Recall and Precision of the *NB classifier* built using the *same* number of cutoffs as in the literature. The percentage in the parentheses next to each method is the mean percentage of patients that were labelled as high risk in the test sequence.

	Recall	Precision
Method	Mean±std	Mean±std
Literature (27%)	0.59± 0.08	0.044±0.005
Warped (36%)	0.68±0.09	0.037±0.005

5. Conclusions

In this paper, we presented a discretization algorithm that uses weighted joint entropy as an evaluation function. In addition to Shannon entropy, we presented two alternative non-symmetric measures of entropy, asymmetric and warped. The non-symmetric measures take into account the imbalance in the prior distribution of classes in the dataset. We discretized three popular continuous ECG based risk metrics using each of the three entropy measures. Warped entropy yielded the most robust cutoffs.

We also compared the cutoffs derived using warped entropy with those found in the literature by evaluating the performance of the Naive Bayes classifier built from discretized training data. When the same number of cutoffs were found using warped entropy as those found in the literature, warped entropy yielded different cutoffs than reported in the literature. The NB classifier built using warped entropy cutoffs yielded a significantly ($p<0.05$) better recall rate than the classifier built using the literature cutoffs. But the improved recall rate was obtained at the expense of lower precision. Therefore, it is inconclusive whether cutoffs yielded using warped entropy are better or worse than the literature cutoffs.

It is important to note that the number of cutoffs used in the literature might not be optimal when using warped entropy for discretization. As future work, we plan to develop an appropriate stopping criteria that can be integrated with our proposed discretization algorithm so that it can automatically determine the appropriate number of cutoffs based on the characteristics of the data.

Acknowledgements

We would like to thank Collin Stultz, Gartheeban Ganeshpillai and Zeeshan Syed for their input throughout the course of this work. We also thank Ben Scirica for providing us with the data.

References

[1] Kotsiantis S, Kanellopoulos D. Discretization techniques: A recent survey, 2006.

[2] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. In Machine Learning. 1998; 105–139.

[3] Shannon CE. Prediction and entropy of printed english. Bell Systems Technical Journal 1951;30:50–64.

[4] Marcellin S, Zighed DA, Ritschard G. Detection of breast cancer using an asymmetric entropy measure. In Computational Statistics, volume 25. Springer, 2006; 975–982.

[5] Morrow DA, Scirica BM, Karwatowska-Prokopczuk E, Murphy SA, Budaj A, Varshavsky S, Wolff AA, Skene A, McCabe CH, Braunwald E, the MERLIN-TIMI 36 Trial Investigators F. Effects of Ranolazine on Recurrent Cardiovascular Events in Patients With Non-ST-Elevation Acute Coronary Syndromes: The MERLIN-TIMI 36 Randomized Trial. JAMA 2007;297(16):1775–1783.

[6] Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology. Circulation March 1996; 93(5):1043–1065.

[7] Bauer A, Kantelhardt JW, Barthel P, Schneider R, Mäkikallio T, Ulm K, Hnatkova K, Schömig A, Huikuri H, Bunde A, Malik M, Schmidt G. Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study. Lancet May 2006;367(9523):1674–81.

[8] Syed Z, Scirica BM, Mohanavelu S, Sung P, Michelson EL, Cannon CP, Stone PH, Stultz CM, Guttag JV. Relation of death within 90 days of non-st-elevation acute coronary syndromes to variability in electrocardiographic morphology. Am J Cardiol Feb 2009;103(3):307–11.

[9] Mitchell TM. Machine Learning. McGraw Hill, 1997.

[10] Kreyszig E. Introductory Mathematical Statistics. John Wiley, 1970.

Address for correspondence:

Anima Singh
32 Vassar Street, 32G-915, Cambridge, MA 02139, USA
anima@mit.edu