# Information Function of the Heart: Discrete and Fuzzy Encoding of the ECG-Signal for Multidisease Diagnostic System

V. Uspenskiy

*The 2-nd Central Military Clinical Hospital, Moscow, Russia*
*E-mail: medddik@yandex.ru*

K. Vorontsov and V. Tselykh and V. Bunakov

*Moscow Institute of Physics and Technology, Moscow, Russia*
*Dorodnicyn Computing Centre of RAS, Moscow, Russia*
*E-mail: voron@forecsys.ru, celyh@phystech.edu, va.bunakov@gmail.com*

Discrete and fuzzy encodings of ECG signal are proposed for multidisease diagnostic system. Cross-validation experiments on 5000 ECGs and 11 internal organ diseases show that sensitivity 93% and specificity 90% can be augmented by 1% with fuzzy encoding.

*Keywords*: information function of the heart, multidisease diagnostic system, ECG signal, digital signal processing, rule-based learning, cross-validation.

## 1. Introduction

The information analysis of signals generated by the heart reveals entirely new opportunities in the assessment of the human health and the disease diagnosis[1]. The informational analysis of ECG includes measuring the amplitudes and the intervals of cardial cycles, the ECG signal discretization, and the induction of diagnostic rules[2,3]. This technology is uniquely positioned to diagnose dozens of internal organs diseases using a single ECG record. In this paper we investigate the ways to improve the diagnosis accuracy by means of fuzzy encoding. This type of encoding allows to smooth noise and to decrease uncertainties in the ECG signal.

## 2. Discrete and Fuzzy Encoding

The informational analysis of ECG is based on the measurement of the amplitudes $R_n$ and intervals $T_n$ for each cardial cycle, $n = 1, \ldots, N$. The sequence $R_1, \ldots, R_N$ represents the *amplitudogram*, and the sequence $T_1, \ldots, T_N$ represents the *intervalogram* of the ECG.

**Discrete Encoding.** In successive cardial cycles, the signs of increments of $R_n$, $T_n$ and $\alpha_n = \frac{R_n}{T_n}$ contain important information for the diagnosis[1,2]. Only 6 combinations of increment signs are possible. They are encoded by the letters of a 6-character alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$:

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| $R_{n+1} - R_n$ | $+$ | $-$ | $+$ | $-$ | $+$ | $-$ |
| $T_{n+1} - T_n$ | $+$ | $-$ | $-$ | $+$ | $+$ | $-$ |
| $\alpha_{n+1} - \alpha_n$ | $+$ | $+$ | $+$ | $-$ | $-$ | $-$ |

Thus, the ECG can be encoded into a sequence of characters from $\mathcal{A}$ called a *codegram*, $S = (s_1, \ldots, s_{N-1})$. Define a frequency $p_w(S)$ of a *trigram* $w = (a, b, c)$ of three symbols $a, b, c$ from $\mathcal{A}$ in the codegram $S$:

$$p_w(S) = \frac{1}{N-3} \sum_{n=1}^{N-3} [s_n = a][s_{n+1} = b][s_{n+2} = c].$$

Denote by $p(S) = \big(p_w(S) \colon w \in \mathcal{A}^3\big)$ a frequency vector of all $|\mathcal{A}|^3 = 216$ trigrams $w$ in the codegram $S$. The informational analysis of ECG is based on the observation that each disease has its own *diagnostic subset* of trigrams significantly frequent in the presence of the disease[1,3].

**Fuzzy encoding.** There are two reasons to consider a smooth variant of the discrete encoding. First, ECG may have up to 5% of outliers among the values $R_n$, $T_n$. In discrete encoding each outlier distorts 4 neighboring trigrams; so, the total number of distorted trigrams may reach 20%. Second, equalities $R_n = R_{n+1}$, $T_n = T_{n+1}$ counts up to 5% of data. In such cases it is more natural to consider $s_n$ as several equiprobable characters.

In a general way we propose to replace each character $s_n$ with a probability distribution $q_n(s)$ over $\mathcal{A}$. This distribution depends on $R_n$, $R_{n+1}$, $T_n$, $T_{n+1}$ values. Then we redefine the frequency of a trigram $w = (a, b, c)$ as a probability of $w$ averaged over the codegram $S$:

$$p_w(S) = \frac{1}{N-3} \sum_{n=1}^{N-3} q_n(a)\, q_{n+1}(b)\, q_{n+2}(c).$$

To estimate how probabilities $q_n(s)$ depends on $R_n$, $R_{n+1}$, $T_n$, $T_{n+1}$ we apply a Monte-Carlo simulations over ECG data set. We simulate a true amplitude $R_n^0$ for each observation $R_n$ from a normal model of measurements with rounding to the nearest integer $R_n = \mathsf{round}\,\mathcal{N}(R_n^0, \sigma_R^2)$, where $\mathcal{N}(R_n^0, \sigma_R^2)$ is a normal random variable with mean $R_n^0$ and variance $\sigma_R^2$.

For intervals $T_n$ we use a similar model with a variance parameter $\sigma_T^2$.

**Rule induction** is a machine learning technique that discovers general rules of classification from a sample of classified cases. We learn diagnostic rules for a disease from a two-class sample: healthy persons and ill patients, each represented by its ECG trigram frequency vector. We use two rule induction algorithms, both applicable for both discrete and fuzzy encoding.

The first algorithm ($A_1$) sorts the sample by $\Gamma(S) = \frac{N_S(\mathtt{A,B,E,F})}{N_S(\mathtt{C,D})}$, where $N_S(X)$ is a frequency of symbols $X \subseteq \mathcal{A}$ in a codegram $S$. Then it divides the range of $\Gamma$ values into intervals with boundaries 0, 1, 1.4, 2, 3. For each interval it finds a *diagnostic subset* of trigrams which contains all trigrams that co-occurs in the codegrams of ill people and never co-occur in the codegrams of healthy people. The diagnosis is positive if the codegram of the person contains the diagnostic subset of the disease.

The second algorithm ($A_2$) sorts the trigrams by their frequency among ill people. The diagnosis is positive if the codegram contains any $\beta K$ of the $K$ most frequent trigrams. Parameters $K, \beta$ are optimized by a full search.

The sensitivity and the specificity of diagnostic rules are estimated by a standard $t \times k$-fold cross-validation procedure. A two-class sample of codegrams is randomly divided into $k$ equi-sized blocks. Each block is used in turns as a testing sample. All but one blocks are used as a training sample to learn a classifier. Then a fraction of erratic testing diagnoses for healthy ($E_1 = 1 - \text{specificity}$) and for ill ($E_2 = 1 - \text{sensitivity}$) persons are calculated. This procedure is repeated $t$ times add the results are averaged.

## 3. Experiments and Results

In the experiment we use 5000 ECGs, $N = 600$ cardial cycles each. 198 ECGs were registered from healthy persons, others had reliable diagnoses of one or more of 11 diseases: (1) necrosis of the femoral head, (2) nodular goiter, (3) chronic gastritis, (4) coronary heart disease, (5) cancer, (6) hypertension, (7) cholelithiasis, (8) diabetes, (9) benign prostatic hyperplasia, (10) gastroduodenitis, (11) biliary tract dyskinesia.

Table 1 shows the errors $E_1$, $E_2$ for 11 diseases and two learning algorithms $A_1$, $A_2$ applied after the discrete encoding.

Fig. 1 shows how the error $\frac{1}{2}(E_1 + E_2)$ depends on the variance parameters $\sigma = \sigma_R = \sigma_T$ of the fuzzy encoding. When $\sigma \approx 5$, the error rate is reduced in average by 1% compared to the discrete encoding (horizontal lines). Error rate does not significantly increase up to $\sigma \approx 10$, thus indicating a satisfactory accuracy of the measurer. The proximity of the training and testing errors indicates that overfitting is negligible.

Table 1. Errors $E_1$ (1 – specificity) and $E_2$ (1 – sensitivity) for 11 diseases, %.

| disease: | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| cases: | 327 | 750 | 698 | 1262 | 267 | 1891 | 277 | 868 | 257 | 321 | 714 |
| $E_1(A_1)$ | 8.59 | 33.1 | 15.2 | 10.2 | 7.32 | 24.3 | 9.19 | 10.4 | 9.80 | 8.13 | 28.0 |
| $E_2(A_1)$ | 8.78 | 11.8 | 20.4 | 15.3 | 28.3 | 10.6 | 10.4 | 20.8 | 24.0 | 21.9 | 14.5 |
| $E_1(A_2)$ | 5.15 | 8.99 | 23.8 | 6.36 | 14.1 | 10.0 | 4.55 | 10.1 | 8.33 | 7.27 | 15.2 |
| $E_2(A_2)$ | 3.79 | 9.29 | 6.63 | 6.09 | 14.6 | 7.69 | 4.01 | 6.91 | 9.03 | 6.64 | 9.54 |

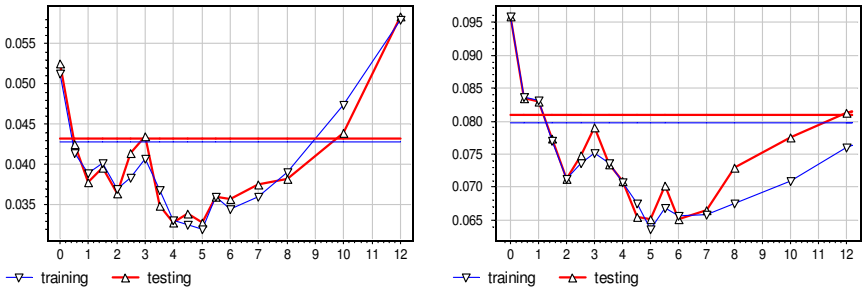


Fig. 1. The error over $\sigma$ for cholelithiasis (left) and diabetes (right).

**Conclusion.** The multidisease diagnostic system based on informational analysis of ECG signals reaches a high level of average sensitivity (93%) and specificity (90%) in cross-validation experiments. Fuzzy encoding helps to improve it by 1% in average. Future research will benefit from more accurate model selection and advanced machine learning techniques.

The work was supported by the Russian Foundation for Basic Research grants 14-07-00908, 14-07-31163.

## References

1. V. Uspenskiy, Information Function of the Heart. *Clinical Medicine*, vol. 86, no. 5 (2008), pp. 4–13.
2. V. Uspenskiy, Information Function of the Heart. A Measurement Model. *Measurement 2011, Proceedings of the 8-th International Conference* (Slovakia, 2011), p. 383–386.
3. V. Uspenskiy, Diagnostic System Based on the Information Analysis of Electrocardiogram. *MECO 2012. Advances and Challenges in Embedded Computing* (Bar, Montenegro, June 19-21, 2012), pp. 74–76.