Discrete and Fuzzy Encoding of the ECG-Signal for Multidisease Diagnostic System

V. Uspenskiy

Federal Medical Educational-Scientific Clinical Center n. a. P. V. Mandryka of the Ministry of Defence of the Russian Federation, 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

In information analysis of ECG signals the discrete and fuzzy encodings of ECG signal are proposed for multidisease diagnostic system. Cross-validation experiments on more than 10 000 ECGs and 18 internal diseases show that the accuracy of diagnostics can be augmented up to 1% with fuzzy encoding.

Keywords: electrocardiography, heart rate variability, information function of the heart, multidisease diagnostic system, digital signal processing, signal discretization, linear classifier, naïve Bayes, feature selection, cross-validation.

1. Introduction

Heart rate variability (HRV) analysis is widely used to diagnose cardiovascular diseases[1,2]. HRV reflects many regulatory processes of the human body and therefore has a high potential to contain a valuable diagnostic information about many internal diseases, not only cardiac diseases. HRV analysis is usually based only on the temporal variation between sequences of consecutive heart beats. On a standard electrocardiogram (ECG), the maximum upwards deflection of a normal QRS complex is at the peak of the R-wave, and the duration between two adjacent R-wave peaks is termed as the RR-interval.

The *information analysis of ECG signals*[3] is based on the measurement and joint analysis of both RR-intervals and amplitudes of adjacent R-wave peaks. Further data processing stages includes discretization, vectorization, and learning diagnostic rules[4,5], which are also different from usual statistical, geometric or spectral methods used if HRV analysis. Discretization

encodes the electrocardiogram into a *codegram* — a sequence of symbols, each cardiac cycle corresponding to one symbol. After that the standard techniques from computational linguistics and machine learning are used to build a diagnostic rule from a training sample of ECGs collected from healthy persons and ill patients. This approach is used in the Multidisease Diagnostic System which allows to diagnose dozens of internal diseases by a single ECG record.

In this paper we propose to improve the diagnostics accuracy by means of fuzzy encoding. Fuzzy encoding aims to smooth noise and to decrease uncertainties in the ECG signal. To do this we introduce a simple probabilistic model of measurements with two unknown parameters: the RMS error of RR-interval and the RMS error of R-peak amplitude. Then we encode the electrocardiogram into a sequence of fuzzy symbols, each represented by a distribution on the alphabet.

To estimate the unknown RMS error parameters from the training sample of ECGs we maximize the area under ROC-curve (AUC). Finally, we make an extensive cross-validation experiment to show that fuzzy encoding improves the accuracy of diagnostics.

2. Discrete and Fuzzy Encoding

The informational analysis of ECG is based on the measurement of the interval T_n and amplitude R_n for each cardiac cycle, $n = 1, \ldots, N$. The sequence $T_1, \ldots, T_N$ represents the *intervalogram* of the ECG, and the sequence $R_1, \ldots, R_N$ represents the *amplitudogram* of the ECG. Note that in HRV analysis only intervals T_n are used, while we analyze jointly the variability of intervals T_n and amplitudes R_n.

Discrete Encoding. In successive cardiac cycles, we take the signs of increments ΔR_n, ΔT_n and $\Delta\alpha_n$, where $\alpha_n = \frac{R_n}{T_n}$. Only 6 from 8 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
$\Delta R_n = R_{n+1} - R_n$	+	−	+	−	+	−
$\Delta T_n = T_{n+1} - T_n$	+	−	−	+	+	−
$\Delta\alpha_n = \alpha_{n+1} - \alpha_n$	+	+	+	−	−	−

Thus, the ECG is encoded into a sequence of characters from $\mathcal{A}$ called a *codegram*, $S = (s_1, \ldots, s_{N-1})$, see Fig. 1. Define a frequency $p_w(S)$ of

Fig. 1. An example of codegram with a sliding window of three symbols.

1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EDF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

Fig. 2. Vector representation $n_w(S)$ of the codegram S shown at Fig. 1. Only 64 of 216 trigrams with frequency $n_w(S) \geq 2$ are shown.

a *trigram* $w = (a, b, c)$ of three symbols a, b, c from $\mathcal{A}$ in the codegram S:

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

where brackets transforms logical values false/true into numbers 0/1.

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, see Fig. 2. 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[3,5].

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.

Therefore we propose to replace each character s_n with a probability distribution $q_n(s)$ over $\mathcal{A}$, see Fig. 3. The distribution $q_n(s)$ 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:

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

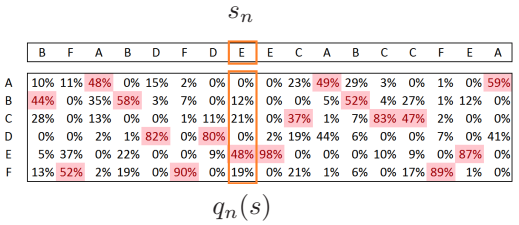


Fig. 3. An example of discrete and fuzzy encoding.

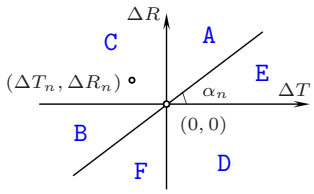


Fig. 4. Six sectors.

To estimate the probability $q_n(s)$ from R_n, R_{n+1}, T_n, T_{n+1} we introduce a probabilistic model of measurement. We assume that each amplitude R_n comes from Laplace distribution with a fixed but unknown RMS error parameter σ_R, which is the same for all ECGs. For intervals T_n we introduce a similar model with RMS error parameter σ_T. Then we calculate probabilities $q_n(s)$ analytically by integrating the two-dimensional probability distribution centered at a point $(\Delta T_n, \Delta R_n)$ over six sectors corresponding to six alphabet symbols $s \in \mathcal{A}$, see Fig. 4.

Machine learning techniques are designed to induce diagnostic rules automatically from a sample of classified cases[6]. We learn diagnostic rules for each disease from a two-class training sample: healthy persons and ill patients, each represented by its ECG trigram frequency vector.

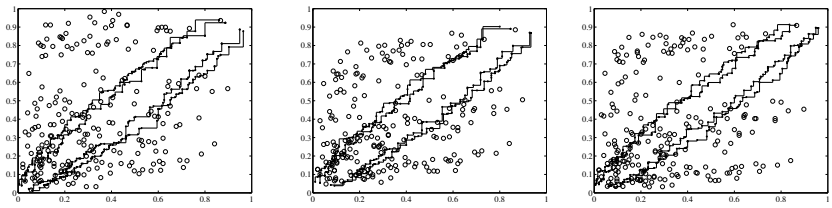
To build a diagnostic rule we use a linear classifier with feature selection:

$$c(S) = \big[\beta(S) \geq \beta_0\big], \qquad \beta(S) = \sum_{w \in \mathcal{A}^3} \beta_j \big[p_w(S) \geq \theta\big],$$

where $\beta(S)$ is a score function, β_0 is a score threshold, β_w is the weight of trigram w learned automatically from the training sample; $\beta_w > 0$ means that the trigram is specific for a disease, $\beta_w < 0$ means that it is specific for health, $\beta_w = 0$ means that the trigram is not used for the diagnostic rule.

Note that both discrete and fuzzy encoding can be used to calculate features $p_w(S)$, thus enabling the comparative study of two types of encoding with the same criterion of diagnostic accuracy.

The linear classification model is motivated by an empirical observation that each disease is characterized by a set of trigrams that are significantly more frequent in codegrams of ill people. Also, there are a set of trigrams that are highly specific for codegrams of healthy people. Fig. 5 shows the result of permutational statistical tests. If the frequency of the trigram and



necrosis of the femoral head nodular goiter thyroid coronary heart disease

Fig. 5. The result of permutational test for three diseases. Points correspond to trigrams. The X-axis indicates the proportion of healthy people, and the Y-axis indicates the proportion of ill people that have a trigram in their codegram more than once. Trigrams located in the region of acceptance near to the diagonal are likely to have occurred by chance (two regions are shown: with significance level 10% and 0.2%). Trigrams located in the critical region far above the diagonal are specific for the disease, and trigrams far below the diagonal are specific for the health.

the class label are independent random variables, then such a trigram will be close to the diagonal of the chart. The results of the test encourage that there are many trigrams far away from the diagonal, and that for each disease the diagnostic subset of highly specific trigrams can be reliably determined.

Cross-Validation. We measure the quality of diagnostic rules by three estimates: the sensitivity, the specificity, and AUC (area under ROC-curve) using a standard 40×10-fold cross-validation procedure. A two-class sample of codegrams is 40 times randomly divided into 10 equi-sized blocks. Each block is used in turns as a testing sample, while other 9 blocks are used as a training sample to learn a classifier. For each partitioning we calculate two AUC values, for both training and testing. From all 40 partitioning we estimate the mean AUC values and their confidence intervals.

3. Experiments and Results

In the experiment we use more that 10 000 ECG records, $N = 600$ cardial cycles each. 193 ECGs were registered from healthy persons, others had reliable diagnoses of one or more of 18 diseases: (1) necrosis of the femoral head, (2) cholelithiasis, (3) coronary heart disease, (4) chronic hyperacidic gastritis (gastroduodenitis), (5) diabetes, (6) hypertension, (7) cancer, (8) benign prostatic hyperplasia, (9) nodular goiter, (10) chronic hypoacidic gastritis (gastroduodenitis), (11) biliary tract dyskinesia, (12) urolithiasis, (13) chronic cholecystitis, (14) peptic ulcer, (15) hysteromyoma,

Table 1. AUC, specificity Sp_1 with sensitivity $Sen_1 = 95\%$, specificity Sp_2 with equal sensitivity $Sen_2 = Sp_2$ of linear classifier for discrete and fuzzy encoding for 18 diseases.

disease	data	discrete			fuzzy
	size	AUC, %	Sp_1, %	$Sp_2 = Sen_2$, %	AUC, %
(1)	324	**99.23 ± 0.05**	97.4 ± 1.0	95.8 ± 0.9	99.01 ± 0.05
(2)	278	98.90 ± 0.02	95.3 ± 0.5	95.5 ± 0.5	**99.00 ± 0.03**
(3)	1265	97.84 ± 0.03	91.8 ± 0.4	93.3 ± 0.0	**98.52 ± 0.03**
(4)	324	97.84 ± 0.09	89.4 ± 1.3	93.0 ± 0.8	**98.20 ± 0.10**
(5)	871	96.66 ± 0.05	84.0 ± 0.9	91.2 ± 0.6	**97.17 ± 0.02**
(6)	1894	96.60 ± 0.05	81.6 ± 1.8	91.5 ± 0.4	**97.31 ± 0.04**
(7)	530	95.81 ± 0.14	80.2 ± 3.0	90.5 ± 0.8	**96.45 ± 0.04**
(8)	260	96.59 ± 0.10	79.8 ± 3.7	91.2 ± 0.7	**96.96 ± 0.04**
(9)	748	95.17 ± 0.10	66.7 ± 2.2	90.4 ± 0.6	**95.72 ± 0.03**
(10)	700	94.77 ± 0.11	71.7 ± 2.8	88.8 ± 1.0	**95.85 ± 0.06**
(11)	717	95.14 ± 0.08	70.9 ± 2.2	89.1 ± 1.0	**95.82 ± 0.10**
(12)	654	95.17 ± 0.07	69.0 ± 4.2	89.0 ± 0.3	**96.03 ± 0.05**
(13)	340	95.51 ± 0.10	76.3 ± 1.9	90.1 ± 0.5	**96.44 ± 0.05**
(14)	785	94.67 ± 0.05	64.3 ± 2.5	89.6 ± 0.5	**95.09 ± 0.04**
(15)	781	93.37 ± 0.10	59.0 ± 2.1	87.6 ± 1.0	**94.28 ± 0.03**
(16)	276	**91.90 ± 0.29**	49.0 ± 3.4	85.6 ± 1.0	91.50 ± 0.23
(17)	260	89.27 ± 0.28	35.9 ± 6.1	83.0 ± 1.2	**90.34 ± 0.10**
(18)	694	86.35 ± 0.24	39.5 ± 4.5	77.9 ± 1.0	**86.50 ± 0.21**

(16) chronic adnexitis, (17) iron deficiency anemia, (18) vasoneurosis.

Table 1 shows the AUC of linear classifier for discrete and fuzzy encoding calculated on testing data for 18 diseases. Fuzzy encoding gives better results for 16 of 18 diseases.

Fig. 6 shows the testing AUC averaged over all diseases depending on the RMS error parameters σ_R and σ_T. From these charts we select the optimal values of parameters $\sigma_R = 5$ and $\sigma_T = 15$.

Note that zero values $\sigma_T = \sigma_R = 0$ correspond to the discrete encoding and are evidently far from optimality.

Fig. 7 shows the testing AUC depending on the frequency threshold parameter $\theta(N-3)$. Its optimal value $\theta = \frac{2}{N-3}$ means that trigrams that occur less than twice in a codegram are not meaningful for the diagnosis.

Fig. 8 shows the testing AUC depending on the RMS error parameters σ_R and σ_T for 2 of 18 diseases.

At all charts the proximity of the training and testing errors indicates that overfitting is small and optimal parameters could be obtained from training set, even without cross-validation.

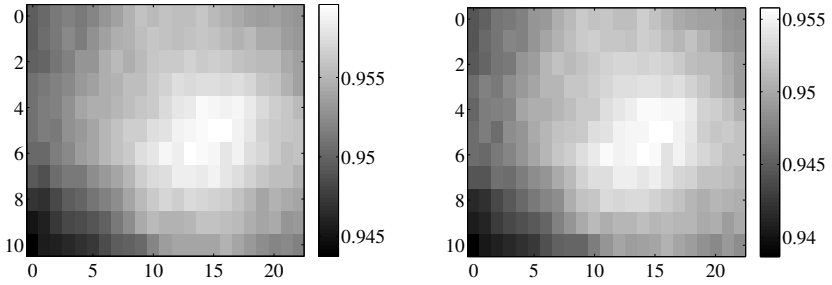


Fig. 6. AUC averaged over all diseases on training set (left-hand chart) and testing set (right-hand chart) depending on $\sigma_T = 0, \ldots, 22$ (X-axis) and $\sigma_R = 0, \ldots, 10$ (Y-axis).

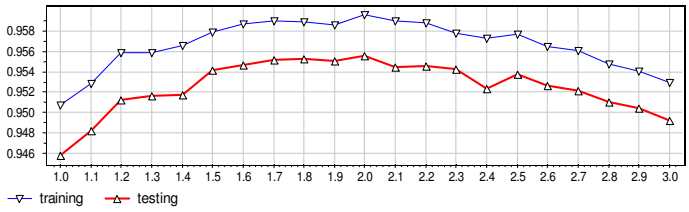


Fig. 7. AUC averaged over all diseases depending on $\theta(N-3)$.

4. Conclusion

The information analysis of ECG signals is a further development of the HRV analysis by two directions. First, it uses not only the RR-intervals but also the amplitudes of R-peaks. Second, it encodes the sequence of intervals and amplitudes into a text string, thus enabling the usage of well established techniques from computational linguistics, text classification, and machine learning. Our experiments show that the information analysis of ECG signals reaches a high level of sensitivity and specificity (90% and higher) in cross-validation experiments. Fuzzy encoding helps to improve this level by 0.5% 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. A. J. Camm, M. Malik, J. T. Bigger, et al. Heart rate variability — standards of measurement, physiological interpretation, and clinical

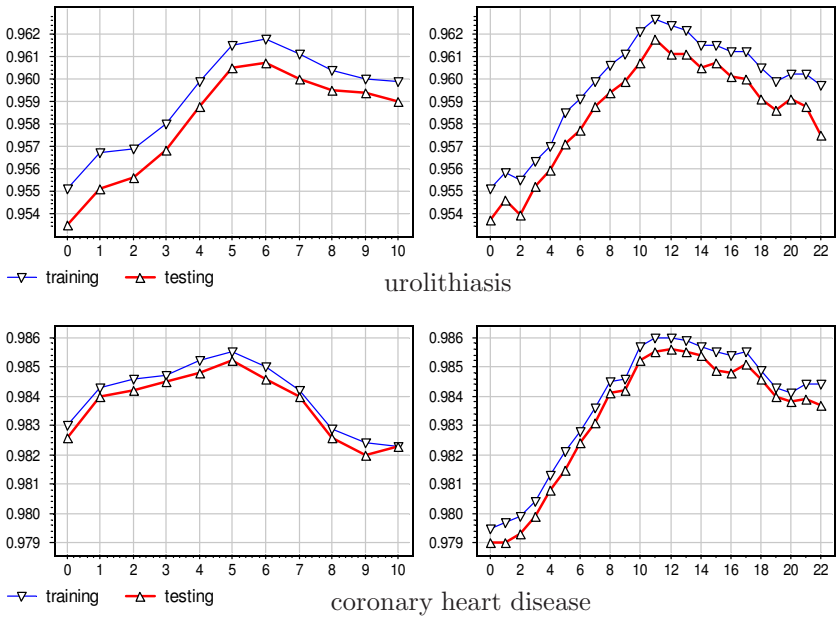


Fig. 8. AUC on training and testing set depending on $\sigma_R = 0, \ldots, 10$ at fixed $\sigma_T = 15$ (left-hand charts) and depending on $\sigma_T = 0, \ldots, 22$ at fixed $\sigma_R = 5$ (right-hand charts) for two of 18 diseases.

use. *Circulation*, vol. 93 (1996), pp. 1043–1065.

2. M. Malik, A. J. Camm. Components of heart rate variability. What they really mean and what we really measure. *Am. J. Cardiol*, vol. 72 (1993), pp. 821–822.
3. V. Uspenskiy, Information Function of the Heart. *Clinical Medicine*, vol. 86, no. 5 (2008), pp. 4–13.
4. V. Uspenskiy, Information Function of the Heart. A Measurement Model. *Measurement 2011, Proceedings of the 8-th International Conference* (Slovakia, 2011), p. 383–386.
5. 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.
6. T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2nd edition. Springer, 2009. 533 p.