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, 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 the ECG signal, discrete and fuzzy variants of signal encoding are compared for multidisease diagnostic system. Cross-validation experiments on more than 10 000 ECGs and 18 internal diseases show that the AUC performance criterion can be improved by up to 1% with fuzzy encoding.

Keywords: electrocardiography, information function of the heart, multidisease diagnostic system, signal discretization, machine learning, cross-validation.

1. Introduction

Heart rate variability (HRV) is the physiological phenomenon of variation in the time interval between heartbeats, or, more precisely, between R-peaks (see Fig. 1). *HRV analysis* is widely used to diagnose cardiovascular diseases[1,3]. HRV reflects many regulatory processes of the human body and therefore has a high potential to contain valuable diagnostic information about many internal diseases, not only related to heart problems.

The *information analysis of ECG signals*[4], instead of averaging time interval variability around the signal, discovers patterns of variability for both intervals and amplitudes of consecutive R-peaks. It was found that some of these patterns are significantly correlated with various diseases[5,6]. This approach has been implemented in the multidisease diagnostic system which permits a diagnosis of a multitude of internal diseases through a single ECG record. This diagnostic technology is based on the encoding of the electrocardiogram into a symbolic string with each cardiac cycle

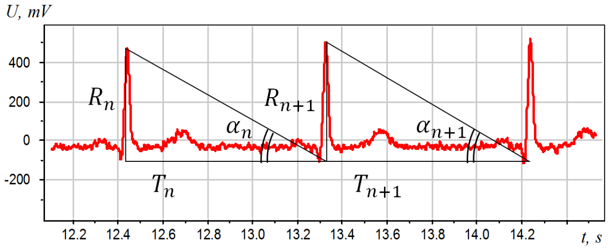


Fig. 1. Three consecutive R-peaks of the ECG signal determine two full cardiac cycles with amplitudes R_n, R_{n+1}, intervals T_n, T_{n+1}, and "phase angles" α_n, α_{n+1}.

corresponding to one symbol. Subsequently, computational linguistics and machine learning techniques are used to infer diagnostic rules from a training sample of ECGs collected from healthy and sick persons.

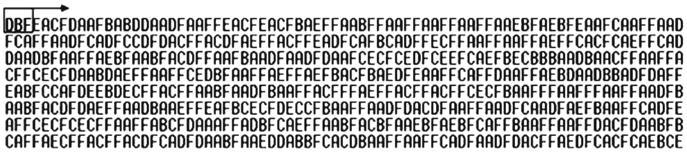
In this paper, we improve the diagnostic performance by means of fuzzy encoding. Note that we use the term "fuzzy" only in its intuitive sense, without regard to the fuzzy logic. *Fuzzy encoding* aims to smooth out the noise and decrease uncertainties in the ECG signal. To do this, we introduce a simple two-parametric probabilistic model of measurements. We make an extensive cross-validation experiment to estimate the model parameters and to show that fuzzy encoding improves the performance.

2. Discrete and Fuzzy Encoding

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

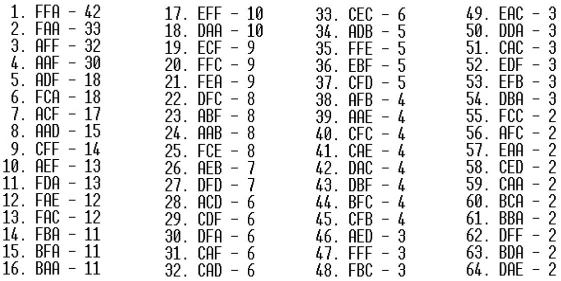
Discrete Encoding. In successive cardiac cycles, we take the signs of increments ΔR_n, ΔT_n and $\Delta\alpha_n$, where $\alpha_n = \arctan \frac{R_n}{T_n}$. Only six of the eight combinations of increment signs are possible. They are encoded by the letters of a six-character alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$:

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



```
DBFEACFDAAFBABDDAADFAAFFEACFEACFBAEFFAABFFAAFFAAFFAAFFAAEBFAEBFEAAFCAAFFAAD
FCAFFAADFCADFCCDFDACFFACDFAEFFACFFEADFCAFBCADFFECFFAAFFAAFFAEFFCACFCAEFFCAD
DAADBFAAFFAEBFAABFACDFFAAFBAADFAADFDAAFCECFCEDFCEEFCAEFBECBBBAADBAACFFAAFFA
CFFCECFDAABDAEFFAAFFCEDBFAAFFAEFFAEFBACFBAEDFEAAFFCAFFDAAFFAEBDAADBBADFDAFF
EABFCCAFDEEBDECFFACFFAABFAADFBAAFFACFFFAEFFACFFACFFCECFBAAFFFAAFFFAAFFAADFB
AABFACDFDAEFFAADBAAEFFEAFBCECFDECCFBAAFFAADFDACDFAAFFAADFCAADFAEFBAAFFCADFE
AFFCECFCECFFAAFFABCFDAAAFFADBFCAEFFAABFACBFAAEBFAEBFCAFFBAAFFAAFFDACFDAABFB
CAFFAECFFACFFACDFCADFDAABFAAEDDABBFCACDBAAFFAAFFCADFAADFDACFFAEDFCACFCAEBCE
```

Fig. 2. An example of a 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. 3. Vector representation $n_w(S)$ of the codegram S shown in Fig. 2. Only 64 of 216 trigrams with frequency $n_w(S) \geq 2$ are shown.

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. 2. We define a frequency $p_w(S)$ of a *trigram* $w = (a, b, c)$ with 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 transform 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. 3. The informational analysis of the ECG is based on the idea that each disease has its own *diagnostic subset* of trigrams frequently observed in the presence of that disease[4,6].

Fuzzy encoding. There are two reasons to consider a smooth variant of discrete encoding. First, the ECG may contain up to 5% of outliers among the values R_n and T_n. In discrete encoding, each outlier distorts four neighboring trigrams; accordingly, the total number of distorted trigrams may reach 20%. Second, the discreteness of the ECG digital sensor results in uncertainties $\Delta T_n = 0$ and $\Delta R_n = 0$ in 5% of cardiac cycles. In such cases, it is appropriate to consider the increment as positive or negative with equal probabilities. In general, the smaller the increment, the greater the uncer-

R_n, mV	313	343	343	318	344	350	327	321	340	340
T_n, ms	843	843	865	828	865	880	861	808	825	825
α_n, °	33.4	36.6	35.7	34.6	35.8	35.8	34.2	35.8	37.1	37.1
ΔR_n, mV	30	0	-25	26	6	-23	-6	19	0	
ΔT_n, ms	0	22	-37	37	15	-19	-53	17	0	
$\Delta\alpha_n$, °	3.2	-0.9	-1.1	1.2	0.0	-1.6	1.6	1.3	0.0	
s_n	C	D	F	A	A	F	B	A	F	
$q_n(A)$, %	50	6	0	93	39	0	0	84	11	
$q_n(B)$, %	0	2	8	0	0	3	87	0	14	
$q_n(C)$, %	50	3	0	1	11	0	10	10	25	
$q_n(D)$, %	0	47	2	0	8	8	0	1	25	
$q_n(E)$, %	0	41	0	6	41	0	0	5	14	
$q_n(F)$, %	0	1	90	0	1	89	3	0	11	

Fig. 4. An example of discrete and fuzzy encoding.

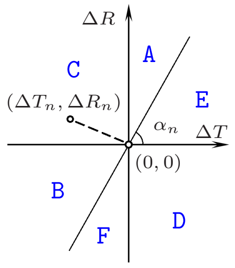


Fig. 5. Six sectors.

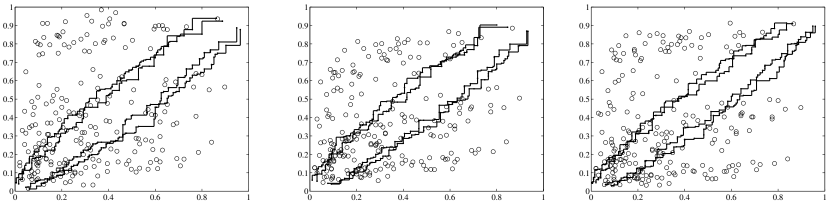
tainty in their sign. We can replace each character s_n with a probability distribution $q_n(s)$ over $\mathcal{A}$ (see Fig. 4) and redefine the frequency of a trigram $w = (a, b, c)$ as a probability of w averaged across the codegram S:

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

To estimate the probability $q_n(s)$ from R_n, R_{n+1}, T_n, and 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 the RMS error parameter σ_T. Subsequently, we calculate probabilities $q_n(s)$ analytically by integrating a two-dimensional probability distribution centered at a point $(\Delta T_n, \Delta R_n)$ over six sectors corresponding to symbols A, B, C, D, E, F shown at Fig. 5.

Machine learning techniques are designed to learn a classifier automatically from a sample of classified cases[2]. We learn a diagnostic rule for each disease from a two-class training sample that contains both healthy persons and patients, each represented by its trigram frequency vector.

In this work we compare three classification models: NB — Naïve Bayes with greedy feature selection, LR — Logistic Regression after dimensionality reduction via Principal Components Analysis, and RF — Random Forest, which is known as one of the strongest classification model. For all classifiers we use binary features $\big[p_w(S) \geq \theta\big]$ instead of frequencies $p_w(S)$, and optimize threshold parameter θ experimentally.



necrosis of the femoral head toxic nodular goiter coronary heart disease

Fig. 6. The result of permutational tests for three diseases. Points indicates trigrams. The X-axis and the Y-axis indicate the proportion of healthy and sick people correspondingly, with two or more occurrences of the trigram in their codegram. The trigrams located in the region of acceptance near the diagonal are likely to have occurred by chance (the significance level equals 10% for the narrow region and 0.2% for the wider one). The trigrams located in the critical region far above the diagonal are specific to the disease, and the trigrams far below the diagonal are specific to a healthy condition.

This approach is motivated by an empirical observation that each disease induces a diagnostic subset of trigrams that are significantly more frequent in the codegrams of sick people. Also, there are trigrams that are highly specific to the codegrams of healthy people. Fig. 6 shows the results of permutational statistical tests for three diseases. If the frequency of the trigram and the class label were independent random variables, then all trigrams would be close to the diagonal of the chart. However, many trigrams are located far away from the chart diagonal. This fact means that for each disease the diagnostic subset of highly specific trigrams exists and can be reliably determined.

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

We measure the diagnostic rules performance using a standard 40×10-fold cross-validation procedure. During procedure, a two-class sample of codegrams are randomly divided into 10 equi-sized blocks 40 times. Each block is used in turns as a testing sample, while the other nine blocks are used as a training sample in order to learn a classifier.

For each partitioning, we calculate three performance measures, for both training and testing samples. *Sensitivity* is the proportion of sick people with true positive diagnosis. *Specificity* is the proportion of healthy people with true negative diagnosis. AUC is defined as the area under the curve of specificity as a function on sensitivity. For each of three performance measures the higher the value, the better. From all 40 cases of partitioning we estimate the mean AUC values as well as their confidence intervals.

Table 1. The AUC (in percents) on testing data for three types of classifiers (RF, LR, NB) and two types of encoding (.d for discrete and .f for fuzzy). Confidence intervals are: ±0.26 for RF, ±0.19 for LR, and ±0.08 for NB.

disease	cases	RF.d	RF.f	LR.d	LR.f	NB.d	NB.f	RF-2	RF-4
(1)	278	98.72	99.00	**99.00**	98.94	98.96	99.00	95.16	94.49
(2)	324	99.24	98.86	**99.26**	99.07	99.24	99.01	98.11	95.49
(3)	1265	98.43	**98.75**	98.21	98.70	97.85	98.52	91.68	92.72
(4)	530	97.15	**97.99**	96.79	97.42	96.03	96.45	93.09	93.43
(5)	700	97.74	97.95	97.64	97.67	97.81	**98.20**	82.54	87.14
(6)	871	97.34	**97.79**	97.10	97.74	96.68	97.17	91.05	92.73
(7)	260	96.65	**97.55**	96.64	97.38	96.61	96.96	89.33	90.59
(8)	1894	97.13	97.49	96.87	**97.68**	96.59	97.31	87.43	90.12
(9)	748	96.07	**96.90**	95.73	96.04	95.17	95.72	85.56	88.10
(10)	324	95.53	**96.37**	95.20	95.98	94.79	95.85	88.95	92.17
(11)	340	95.21	96.25	95.06	96.17	95.51	**96.44**	86.29	87.60
(12)	717	95.29	**96.20**	95.13	96.12	95.13	95.82	86.92	87.86
(13)	654	95.09	**96.16**	95.14	95.94	95.14	96.03	87.80	86.90
(14)	785	94.99	**95.58**	94.74	95.33	94.68	95.09	86.60	89.17
(15)	781	94.43	**95.26**	93.58	94.74	93.38	94.28	84.06	85.97
(16)	276	92.37	**92.65**	92.44	92.32	91.88	91.50	81.49	84.96
(17)	260	90.03	**91.82**	90.03	91.07	89.56	90.34	79.39	81.77
(18)	694	88.07	**88.63**	87.70	87.65	86.59	86.50	76.48	82.39

3. Experiments and Results

In the experiment, we used more that 10 000 ECG records with $N = 600$ cardiac cycles in each. 193 ECGs were taken from healthy participants, while the others were taken from patients who were reliably diagnosed with one or more of the 18 diseases: (1) cholelithiasis, (2) AVN, necrosis of the femoral head, (3) coronary heart disease, (4) cancer, (5) chronic hypoacidic gastritis (gastroduodenitis), (6) diabetes, (7) BPH, benign prostatic hyperplasia, (8) HTN, hypertension, (9) TNG, toxic nodular goiter or Plummer syndrome, (10) chronic hyperacidic gastritis (gastroduodenitis), (11) chronic cholecystitis, (12) biliary dyskinesia, (13) urolithiasis, (14) peptic ulcer, (15) hysteromyoma, (16) chronic adnexitis, (17) iron-deficiency anemia, (18) vasoneurosis.

Table 1 compares the performance of three classifiers (Random Forest, Logistic Regression and Naïve Bayes) on testing data for discrete and fuzzy encoding. Fuzzy encoding gives better results for 16 of the 18 diseases. Random Forest is usually the best choice. Nonetheless, Naïve Bayes with feature selection is not much worse. Two additional columns RF-2 and RF-4 show the performance of Random Forest for two simplified discrete

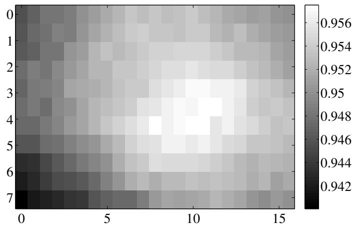


Fig. 7. The AUC on testing set averaged across all diseases depending on σ_T (X-axis) and σ_R (Y-axis).

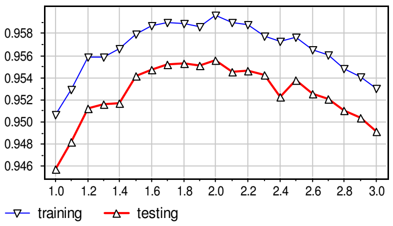


Fig. 8. The AUC on training and testing set averaged across all diseases depending on threshold parameter $\theta(N-3)$.

encodings. RF-2 uses a two-character alphabet for ΔT_n signs. RF-4 uses a four-character alphabet for ΔT_n and ΔR_n signs. From the comparison we conclude that the six-character encoding gives significantly better results.

Fig. 7 shows the AUC on testing data averaged across all diseases as a function of the RMS error parameters σ_R and σ_T. Based on the charts we selected the optimal values of parameters $\sigma_R = 3.5$ mV and $\sigma_T = 10.6$ ms. Note that zero values $\sigma_T = \sigma_R = 0$, which corresponds to discrete encoding, are evidently far away from being optimal.

Fig. 8 shows how the average AUC for NB classifier on testing data depends on the frequency threshold parameter $\theta(N-3)$. Trigrams that occur less than twice in a codegram are not meaningful for the diagnosis.

Fig. 9 shows how the AUC for NB classifier on testing data depends on the RMS error parameters σ_R and σ_T for 2 of the 18 diseases.

The proximity of training and testing AUCs in all charts indicates that overfitting of NB classifier is minute, and optimal parameters could be obtained from the training set even without cross-validation.

4. Conclusion

The information analysis of ECG signals improves the HRV analysis by two directions. Firstly, it identifies patterns of joint variability of intervals and amplitudes of R-peaks specific to diseases. Secondly, this type of analysis is not restricted to cardiovascular diseases. Our experiments show that the information analysis of the ECG signals reaches a high level of sensitivity and specificity (90% and higher) in cross-validation experiments.

On average, fuzzy encoding helps to improve this level by 0.65%.

Future research will benefit from more accurate techniques for signal encoding, statistical modeling, and machine learning.

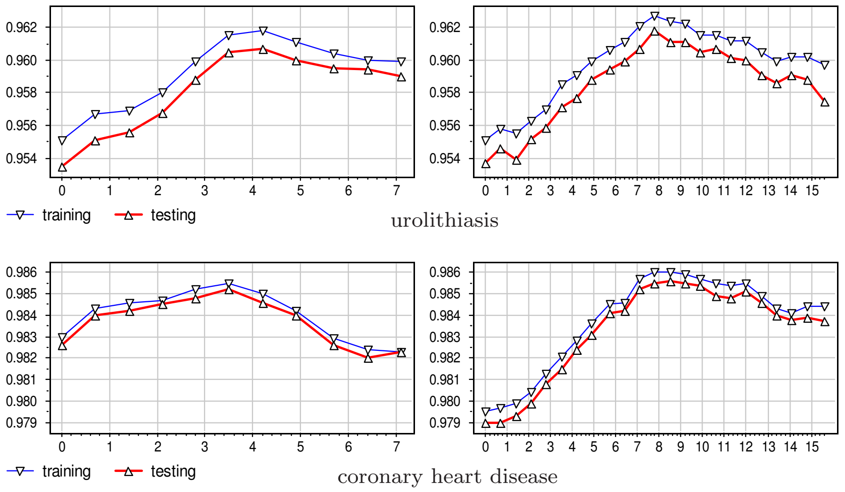


Fig. 9. AUC on training and testing set depending on σ_R at fixed $\sigma_T = 10.6$ (left-hand charts) and depending on σ_T at fixed $\sigma_R = 3.5$ (right-hand charts) for two of 18 diseases.

The work was supported by the Russian Foundation for Basic Research grants 14-07-00908, 14-07-31163. We thank Alex Goborov for his help with English translation and valuable discussion.

References

1. A. J. Camm, M. Malik, J. T. Bigger, et al. Heart rate variability — standards of measurement, physiological interpretation, and clinical use. *Circulation*, vol. 93 (1996), pp. 1043–1065.
2. T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2nd edition. Springer (2009), 533 p.
3. 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.
4. V. Uspenskiy. Information Function of the Heart. *Clinical Medicine*, vol. 86, no. 5 (2008), pp. 4–13.
5. V. Uspenskiy. Information Function of the Heart. A Measurement Model. *Measurement 2011, Proceedings of the 8-th International Conference* (Slovakia, 2011), p. 383–386.
6. 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.