Information function of the heart:
Discrete and fuzzy encoding of the ECG-signal for multidisease diagnostic system

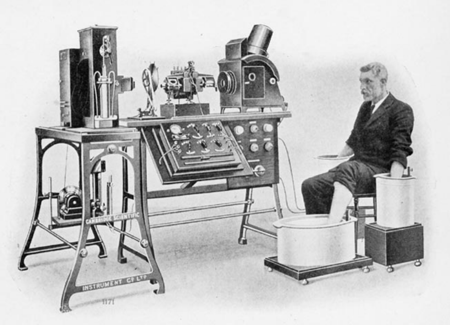
Vyacheslav Uspenskiy • P. V. Mandryka Medical Education and Research Clinical Center of the Ministry of Defence of RF
Konstantin Vorontsov • MIPT, CC RAS, Moscow, RF
Vlada Tselykh, *Vasiliy Bunakov* • MIPT, Moscow, RF

International Conference AMCTM-2014
Advanced Mathematical and Computational Tools in Metrology and Testing

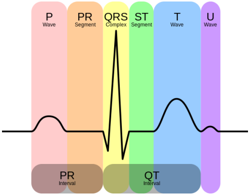
11 September 2014

Contents

Electrocardiography

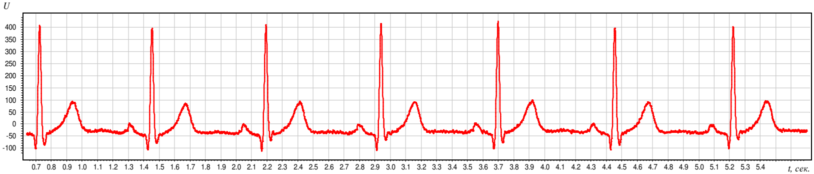


Photograph of a Complete Electrocardiograph, Showing the Manner in which the Electrodes are Attached to the Patient, In this Case the Hands and One Foot Being Immersed in Jars of Salt Solution

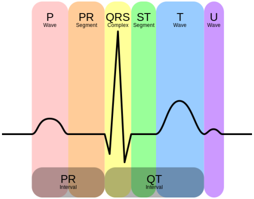


1872 — first record of the electrical activity of the heart
1911 — an early commercial ECG device (photo)
1924 — Nobel Prize in Medicine for the description of the ECG features of a number of cardiovascular disorders (Willem Einthoven)

Classical approach vs. Uspenskiy's Informational Analysis



The classical diagnosis of *heart disorders* is based on PQRST-complex analyzing



The diagnosis of *many diseases* proposed by prof. V.Uspenskiy is based on variations of *amplitudes* and *intervals* of cardiac cycles

Theory of Information Function of the Heart

Main theoretical assumptions:

- ECG signal carries information about the functioning of not only the heart, but all the systems of the body
- Each disease exhibits a specific modulation of the amplitudes and intervals of cardiac cycles
- Information about the disease can be detected at any stage including latent and preclinical stages

Thus, an *early diagnosis of many diseases* from one ECG is possible

---

V. Uspenskiy. Information Function of the Heart. *Clinical Medicine*, vol. 86, no. 5 (2008), pp. 4–13.

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.

Multidisease Diagnostic System «Skrinfaks» (2-nd generation)

- more than 30 years of research (from 1978)
- more than 10 years of operation
- more than 20 000 cases (ECG record + diagnosis)
- more than 40 internal diseases can be detected

Technology of ECG Informational Analysis

ECG Preprocessing Stage:

1. *Demodulation* gives amplitudes and intervals of 600 subsequent cardio cycles
2. *Discretization* gives a *codogram* — a 599-character string in a 6-letter alphabet
3. *Vectorization* gives a vector of $6^3 = 216$ triplet frequencies

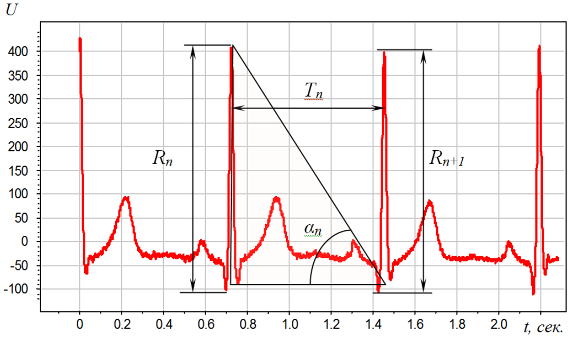
Machine Learning Stage:

1. Building a *classification model*
2. Model *optimization* from cases with known diagnosis
3. Model *evaluation* by other cases with known diagnosis

Preprocessing step 1: Demodulation

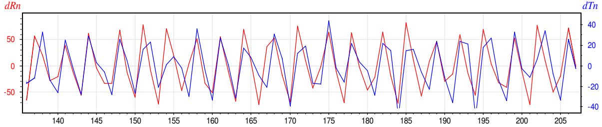
Input: a detailed raw ECG signal (3Mb file)
Output: a sequence of increment signs (225b — 10^4 compression!)

amplitude	$dR_n = R_{n+1} - R_n$
interval	$dT_n = T_{n+1} - T_n$
angle	$d\alpha_n = \alpha_{n+1} - \alpha_n$, where $\alpha_n = \mathrm{arctg}\,\frac{R_n}{T_n}$

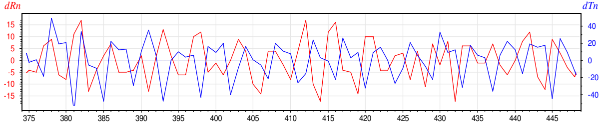


Variation of increments dR_n and dT_n for ill and healthy persons

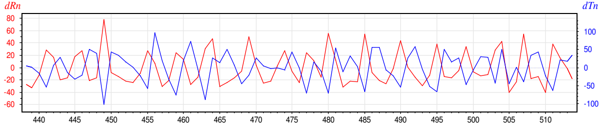
healthy:



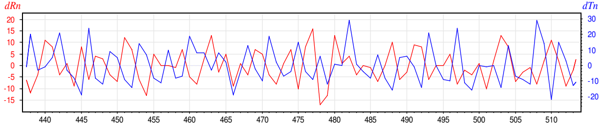
peptic ulcer:



hypertension:



cancer:



Preprocessing step 2: Discretization

Input: intervals and amplitudes $(T_1, R_1), \dots, (T_N, R_N)$
Output: *codogram* $x = (s_1, \dots, s_{N-1})$ — a sequence of symbols from the alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$

$$
\begin{array}{lllllll}
\text{if} & R_n < R_{n+1}, & T_n < T_{n+1}, & \alpha_n < \alpha_{n+1} & \text{then} & s_n = \mathtt{A} \\
\text{if} & R_n \geqslant R_{n+1}, & T_n \geqslant T_{n+1}, & \alpha_n < \alpha_{n+1} & \text{then} & s_n = \mathtt{B} \\
\text{if} & R_n < R_{n+1}, & T_n \geqslant T_{n+1}, & \alpha_n < \alpha_{n+1} & \text{then} & s_n = \mathtt{C} \\
\text{if} & R_n \geqslant R_{n+1}, & T_n < T_{n+1}, & \alpha_n \geqslant \alpha_{n+1} & \text{then} & s_n = \mathtt{D} \\
\text{if} & R_n < R_{n+1}, & T_n < T_{n+1}, & \alpha_n \geqslant \alpha_{n+1} & \text{then} & s_n = \mathtt{E} \\
\text{if} & R_n \geqslant R_{n+1}, & T_n \geqslant T_{n+1}, & \alpha_n \geqslant \alpha_{n+1} & \text{then} & s_n = \mathtt{F}
\end{array}
$$

Preprocessing step 3: Vectorization

Input: *a codogram* $x = (s_1, \dots, s_{N-1})$ as a text string

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

Output: triplet frequency $f_j(x)$ — how many times the triplet j appears in the codogram x, $j = 1, \dots, n$, $n = 6^3 = 216$

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

Modeling diagnostic rule

x_i — a training set of cases (codograms), $i = 1, \dots, \ell$
y_i — diagnosis for the i-th case: 0 = healthy, 1 = ill
$f_j(x_i)$ — a frequency of triplet j in the codogram

Assumption: for each disease there are triplets, which are significantly frequent in codograms of ill people

Linear model of classification:

$$a(x) = \bigl[\langle x, w\rangle \geqslant w_0\bigr], \qquad \langle x, w\rangle = \sum_{j=1}^{n} w_j \bigl[f_j(x) \geqslant \theta\bigr],$$

where w_j is the weight of triplet j:

- $w_j > 0$, if the triplet is more specific for ill people
- $w_j < 0$, if the triplet is more specific for healthy people
- $w_j = 0$, if the triplet is irrelevant for a given disease

Machine Learning

Linear model of classification:

$$a(x) = \big[\langle x, w\rangle \geqslant w_0\big], \qquad \langle x, w\rangle = \sum_{j=1}^{n} w_j \big[f_j(x) \geqslant \theta\big],$$

There are a number of classification algorithms to learn optimal weights w_j from training sample (x_i, y_i), $i = 1, \dots, \ell$:

- NB — Naïve Bayes
- SVM — Support Vector Machine
- LR — Logistic Regression
- RLR — Regularized Logistic Regression
- LASSO — Least Absolute Shrinkage and Selection Operator
- etc.

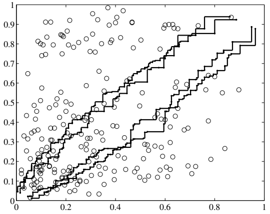
Permutational test

Points at these charts correspond to triplets $j = 1, \dots, 216$

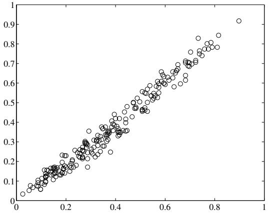
X-axis: $\frac{1}{\ell_0} \sum\limits_{y_i=0} \bigl[f_j(x_i) \geqslant \theta\bigr]$ — healthy people with frequent triplet j

Y-axis: $\frac{1}{\ell_1} \sum\limits_{y_i=1} \bigl[f_j(x_i) \geqslant \theta\bigr]$ — ill people with frequent triplet j

Disease: necrosis of the femoral head



true y_i classifications



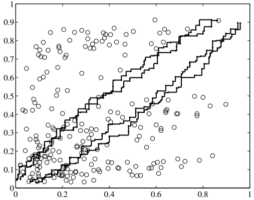
randomly permuted classifications

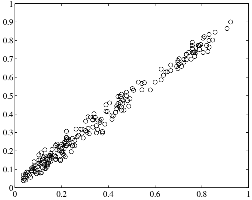
Significant triplets are outside of 90% or 99.8% confidence region (estimated from 20 and 1000 random permutations respectively)

Permutational test

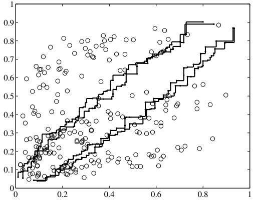
For each disease there are specifically frequent and unfrequent triplets

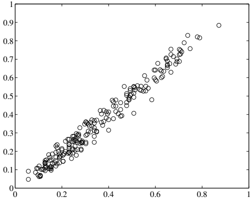
Disease: coronary heart disease





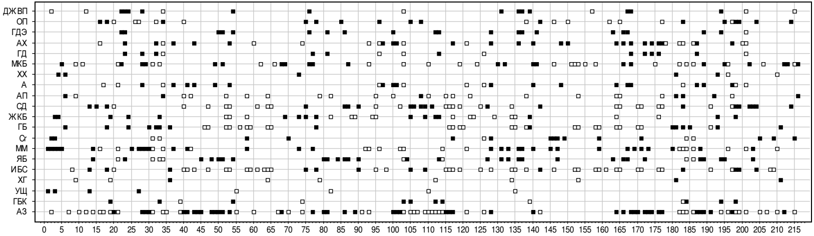
Disease: nodular goiter thyroid





How different specific patterns of diseases are?

X-axis: all triplets $j = 1, \dots, 216$
Y-axis: diseases (A3 = absolutely healthy)



□ — significantly low triplet frequency
■ — significantly high triplet frequency

Conclusion 1. Each disease has its own *specific pattern*
— a set of triplets that discriminates ill and healthy persons well

Conclusion 2. Diseases differ significantly by their specific patterns

Sensitivity, Specificity & AUC (the higher, the better)

Sensitivity is a ratio of ill people with true positive diagnosis

$$\text{Sensitivity} = \frac{1}{\ell_1} \sum_{i:\, y_i=1} \bigl[a(x_i) = 1\bigr]$$

Specificity is a ratio of healthy people with true negative diagnosis

$$\text{Specificity} = \frac{1}{\ell_0} \sum_{i:\, y_i=0} \bigl[a(x_i) = 0\bigr]$$

AUC (Area Under Curve) is a ratio of truly ordered pairs of cases

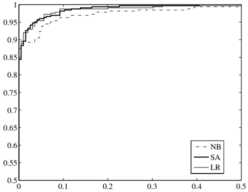
$$\text{AUC} = \frac{1}{\ell_0 \ell_1} \sum_{i:\, y_i=0} \; \sum_{k:\, y_k=1} \bigl[\langle x_i, w\rangle < \langle x_k, w\rangle\bigr]$$

Cross-validation experiments

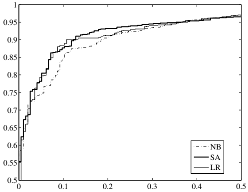
Training set — for learning model parameters w_j, $j = 1, \ldots, 216$
Testing set — for evaluating sensitivity, specificity and AUC
$40{\times}10$-fold cross-validation to build 95% confidence intervals

disease	cases	AUC, %	spec, % (sens=95%)
femoral head necrosis	327	99.19 ± 0.10	96.6 ± 1.76
cholelithiasis	277	98.98 ± 0.23	94.4 ± 1.54
coronary heart disease	1262	97.98 ± 0.14	91.1 ± 1.86
gastritis	321	97.76 ± 0.11	88.3 ± 2.64
hypertensive disease	1891	96.76 ± 0.09	84.7 ± 1.99
diabetes	868	96.75 ± 0.19	85.3 ± 2.18
benign prostatic hyperplasia	257	96.49 ± 0.13	80.1 ± 3.19
cancer	525	96.49 ± 0.28	82.2 ± 2.38
nodular goiter thyroid	750	95.57 ± 0.16	73.5 ± 3.41
chronic cholecystitis	336	95.35 ± 0.12	74.8 ± 2.46
biliary dyskinesia	714	94.99 ± 0.16	70.3 ± 4.67
urolithiasis	649	94.99 ± 0.11	69.3 ± 2.14
peptic ulcer	779	94.62 ± 0.10	63.6 ± 2.55

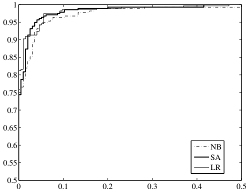
ROC-curves: X-axis is (1−specificity), Y-axis is sensitivity



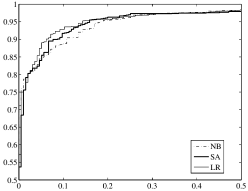
femoral head necrosis



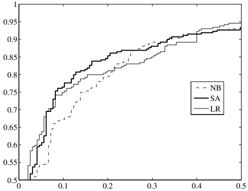
peptic ulcer



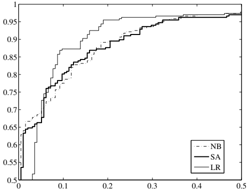
cholelithiasis



diabetes



anemia



canser

NB — Naïve Bayes, SA — Syndrome rule algorithm, LR — Logistic Regression

Two problems that motivate the usage of fuzzy encoding

1. **The problem of outliers:**
   ECG may have up to 5% of outliers among the values R_n, T_n
2. **The problem of noise:**
   sign dR_n, sign dT_n become uncertain when $dR_n \to 0$, $dT_n \to 0$

Instead of discretization $(T_n, R_n), (T_{n+1}, R_{n+1}) \to s_n$, $s_n \in \mathcal{A}$
we will estimate a distribution $q_n(s)$ over $s \in \mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$

s_n

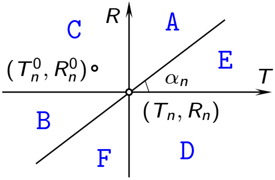
	B	F	A	B	D	F	D	E	E	C	A	B	C	C	F	E	A
A	10%	11%	48%	0%	15%	2%	0%	0%	0%	23%	49%	29%	3%	0%	1%	0%	59%
B	44%	0%	35%	58%	3%	7%	0%	12%	0%	0%	5%	52%	4%	27%	1%	12%	0%
C	28%	0%	13%	0%	0%	1%	11%	21%	0%	37%	1%	7%	83%	47%	2%	0%	0%
D	0%	0%	2%	1%	82%	0%	80%	0%	2%	19%	44%	6%	0%	0%	7%	0%	41%
E	5%	37%	0%	22%	0%	0%	9%	48%	98%	0%	0%	0%	10%	9%	0%	87%	0%
F	13%	52%	2%	19%	0%	90%	0%	19%	0%	21%	1%	6%	0%	17%	89%	1%	0%

$q_n(s)$

The model of measurements for fuzzy encoding

R_n comes from a Laplace distribution, $\mathsf{E}R_n = R_n^0$, $\mathsf{D}R_n = \sigma_R^2$
T_n comes from a Laplace distribution, $\mathsf{E}T_n = T_n^0$, $\mathsf{D}T_n = \sigma_T^2$

Geometric interpretation:
$q_n(a)$ is a probability that (T_n^0, R_n^0) belongs to the sector $a \in \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$

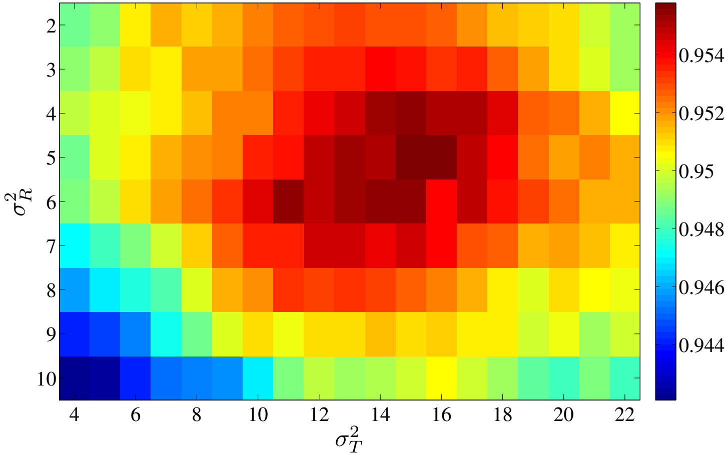


Fuzzy frequency of triplet j, consisting of three letters abc:

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

Outliers processing:
if R_n is outlier then $\mathsf{P}(R_{n-1} < R_n) = \mathsf{P}(R_n < R_{n+1}) = \frac{1}{2}$
if T_n is outlier then $\mathsf{P}(T_{n-1} < T_n) = \mathsf{P}(T_n < T_{n+1}) = \frac{1}{2}$

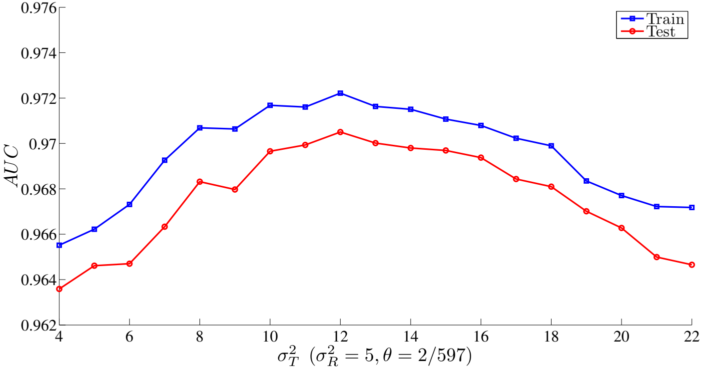
Model of measurement parameters optimization



Optimum found: $\sigma_T^2 = 15, \quad \sigma_R^2 = 5$

Cross-validated AUC

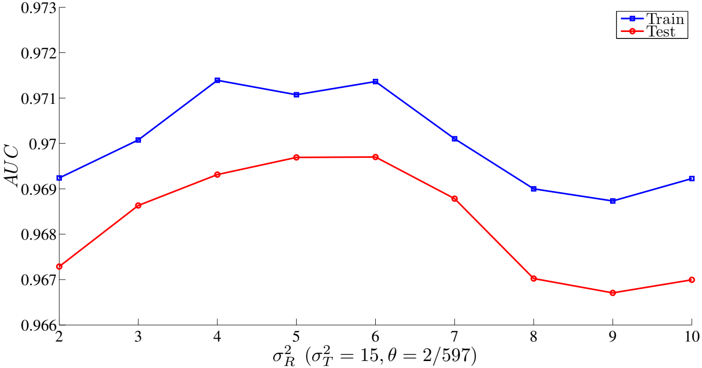
Disease: diabetes



Conclusion: Discrete encoding ($\sigma_T^2 = 0$) is not optimal!

Cross-validated AUC

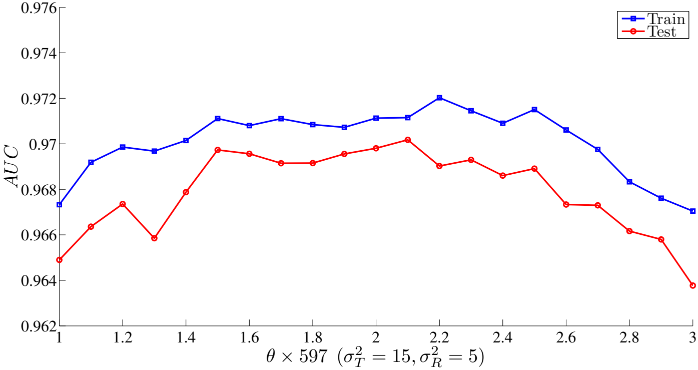
Disease: diabetes



Conclusion: Discrete encoding ($\sigma_R^2 = 0$) is not optimal!

Cross-validated AUC

Disease: diabetes



Conclusion: Triplets less frequent that $\theta = \frac{2}{597}$ are not significant

Summary

- A very promising innovative approach to noninvasive early diagnostics of many diseases from a single electrocardiogram
- Surprisingly high specificity and sensitivity!
- Fuzzy encoding further improves the diagnostic accuracy

[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. Information Function of the Heart. Biophysical substantiation of technical requirements for electrocardioblock registration and measurement of electrocardiosignals parameters acceptable for information analysis to diagnose internal diseases. *Joint International IMEKO TC1+TC7+TC13 Symposium* (Jena, Germany, August 31–September 2, 2011).

[4] 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.

Information function of the heart:
Discrete and fuzzy encoding of the ECG-signal
for multidisease diagnostic system

Vyacheslav Uspenskiy • medddik@yandex.ru
Konstantin Vorontsov • voron@forecsys.ru
Vlada Tselykh • celyh@phystech.edu
Vasiliy Bunakov • va.bunakov@gmail.com