Тематические модели классификации

Шапулин Андрей
Воронцов Константин Вячеславович
ВМК МГУ

Научный семинар MLIR • 30 сентября 2014

Содержание

1. **Тематическая модель классификации**
   - Классификация документов
2. **Информационный анализ ЭКГ-сигналов**
   - Метод В.М.Успенского
   - Задача тематического моделирования

Тематическая модель классификации

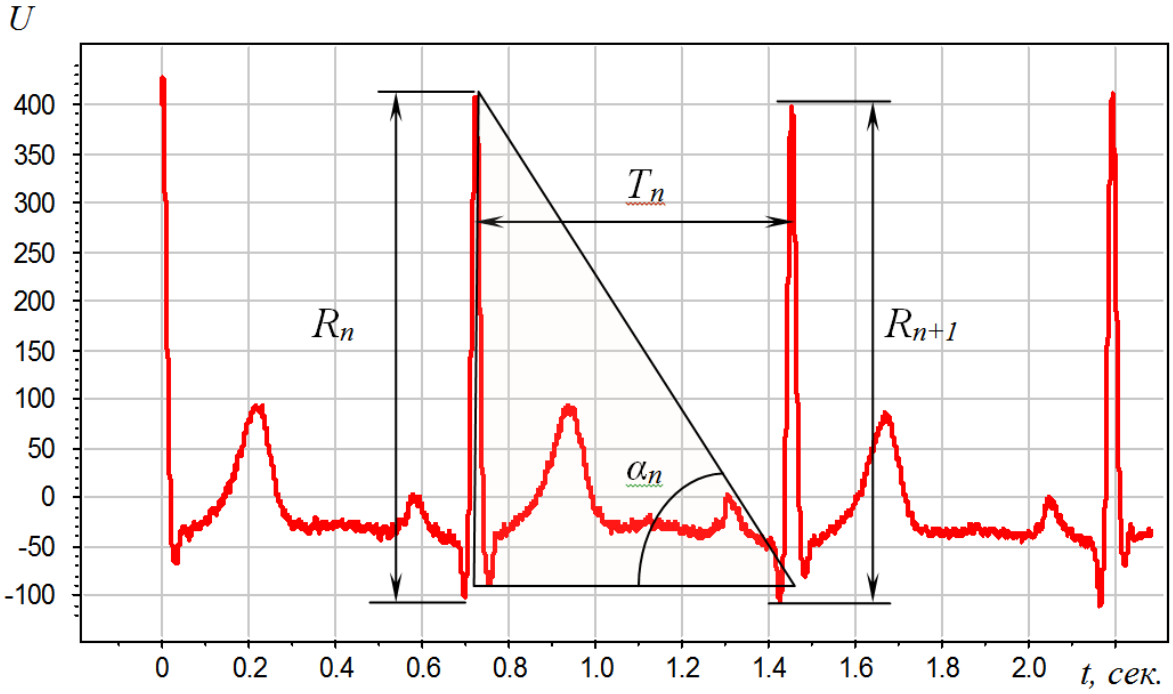
Модель позволяет решать задачу классификации документов по классам $c \in C$, имея обучающую информацию о принадлежности документа классу.

Основная статья:
Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.
"A drawback of discriminative modeling techniques such as support vector machines is that performance rapidly drops off as the total number of labels and the number of labels per document increase".

Выводы:
Тематическая модель классификации работает лучше многих других алгоритмов, особенно в задачах с несбалансированными классами.

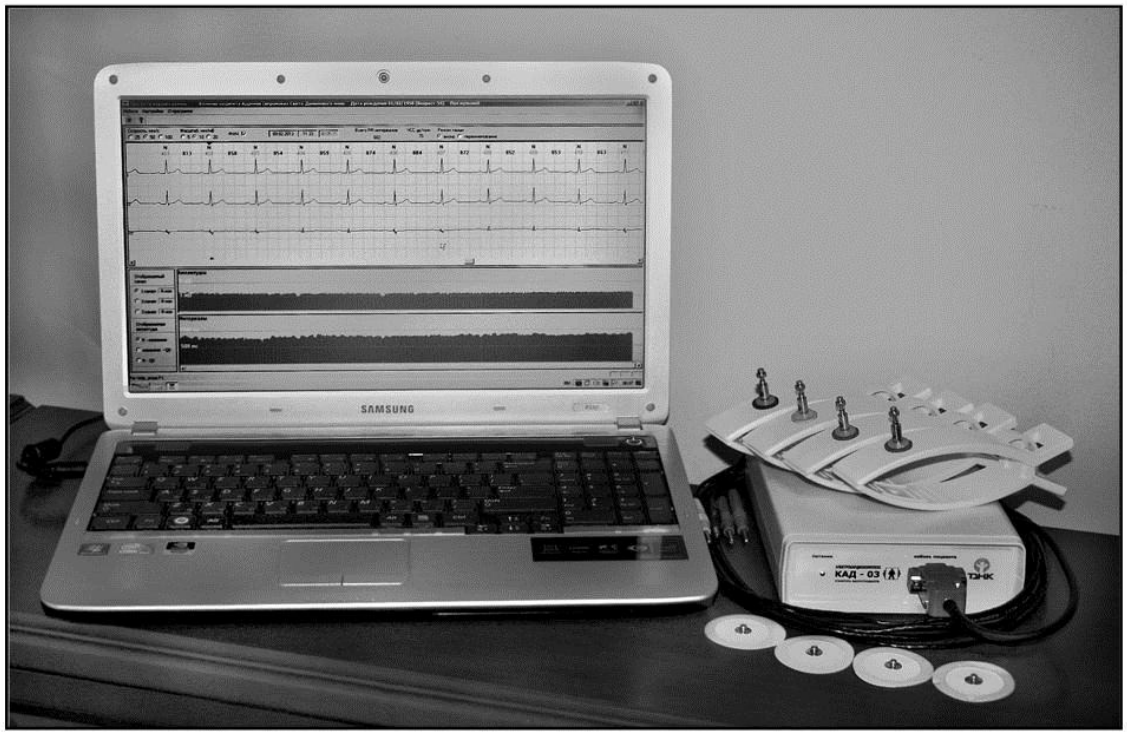
Информационный анализ электрокардиосигналов

Теория информационной функции сердца В.М.Успенского: амплитуды R_n и интервалы T_n кардиоциклов несут информацию о многих заболеваниях внутренних органов.



$$\alpha_n = \operatorname{arctg} \frac{R_n}{T_n}$$

Диагностическая система «Скринфакс» (2-е поколение)



- более 10 лет эксплуатации (начало исследований: 1978)
- более 20 тысяч прецедентов (кардиограмма + диагноз)
- более 40 заболеваний

Дискретизация ЭКГ-сигнала

Вход: последовательность интервалов и амплитуд $(T_n, R_n)_{n=1}^{N}$

Правила кодирования:

если	$R_n < R_{n+1}$,	$T_n < T_{n+1}$,	$\alpha_n < \alpha_{n+1}$	то	$s_n = \mathtt{A}$
если	$R_n \geqslant R_{n+1}$,	$T_n \geqslant T_{n+1}$,	$\alpha_n < \alpha_{n+1}$	то	$s_n = \mathtt{B}$
если	$R_n < R_{n+1}$,	$T_n \geqslant T_{n+1}$,	$\alpha_n < \alpha_{n+1}$	то	$s_n = \mathtt{C}$
если	$R_n \geqslant R_{n+1}$,	$T_n < T_{n+1}$,	$\alpha_n \geqslant \alpha_{n+1}$	то	$s_n = \mathtt{D}$
если	$R_n < R_{n+1}$,	$T_n < T_{n+1}$,	$\alpha_n \geqslant \alpha_{n+1}$	то	$s_n = \mathtt{E}$
если	$R_n \geqslant R_{n+1}$,	$T_n \geqslant T_{n+1}$,	$\alpha_n \geqslant \alpha_{n+1}$	то	$s_n = \mathtt{F}$

Выход: *кодограмма* $x = (s_n)_{n=1}^{N-1}$ — последовательность символов алфавита $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$:

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

Векторизация кодограммы ЭКГ-сигнала

Вход: *кодограмма* $x = (s_1, \ldots, s_{N-1})$ как текстовая строка

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

Выход: частоты триграмм $f_j(x)$ — сколько раз триграмма j появилась в кодограмме x, $j = 1, \ldots, n$, $n = 6^3 = 216$

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

Лингвистический анализ электрокардиосигналов

Дано:
20 тысяч кодограмм ЭКГ (строки в 6-буквенном алфавите), каждая отнесена к некоторым из 40 заболеваний,
важно учесть случаи сочетания заболеваний.

Найти:

- темы классов (диагностические эталоны заболеваний)
- алгоритм классификации (диагностики заболеваний)

Регуляризаторы:

- разреживание, сглаживание, антикоррелирование
- привязка документов к классам (категоризация)
- учёт различий в степени доверия диагнозам
- учёт несбалансированности классов

Регуляризатор для классификации документов

Пусть C — множество классов (для ЭКГ — заболевания, для текстов — категории, авторы, ссылки, годы, читатели)

Гипотеза:
классификация документа d объясняется его темами:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d) = \sum_{t \in T} \psi_{ct}\theta_{td}.$$

Минимизируем дивергенцию между моделью $p(c|d)$ и «эмпирической частотой» классов в документах m_{dc}:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct}\theta_{td} \to \max.$$

Процесс обучения и контроля

Этап обучения: $p(c|t) = \psi_{ct}^{train}$, $p(t|d) = \theta_{td}^{train}$, $p(w|t) = \varphi_{wt}^{train}$

Этап контроля:
Искомое распределение:

$$p(c|d)^{test} = \sum_{t \in T} p(c|t)^{train} p(t|d)^{test} = \sum_{t \in T} \psi_{ct}^{train} \theta_{td}^{test}.$$

Главные вопросы:

- Как выбирать начальное приближение в EM-алгоритме?
- Как без регуляризатора классификации строить распределение θ_{td}^{test}?
- Каким выбрать число диагностических эталонов (тем) и сколько тем отнести к фону?

Текущие эксперименты

Классические алгоритмы машинного обучения:

- Naive Bayes
- SVM
- и т.д.

Точность диагностики более 90%!

Тематической модели кассификации пока не удается побить Naive Bayes.

Основная идея:
Использовать информацию полученную Naive Bayes в инцализиции матриц Φ, Ψ.

Текущие эксперименты

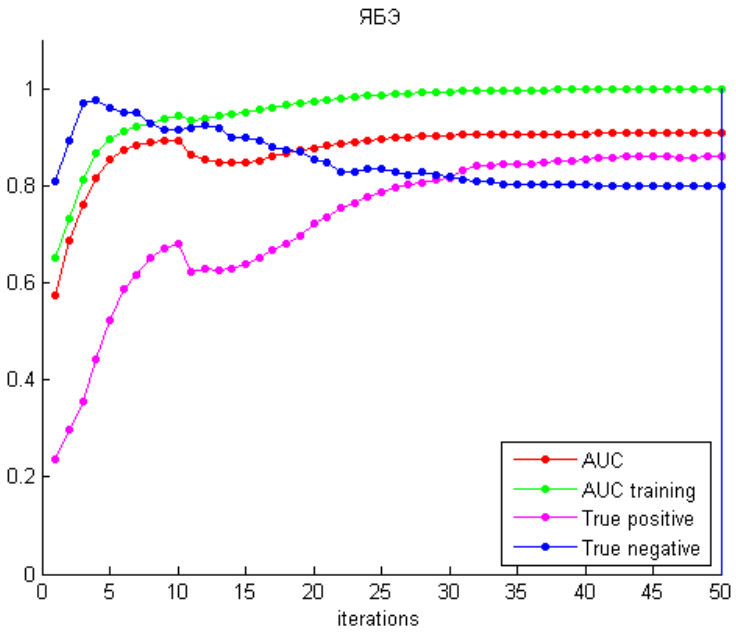
Качество классификации оценивается при помощи следующих показателей:

Доля больных с верным положительным диагнозом:
чувствительность: $\frac{1}{l_1}\sum_{i:y_i=1}[a(x_i)=1]$

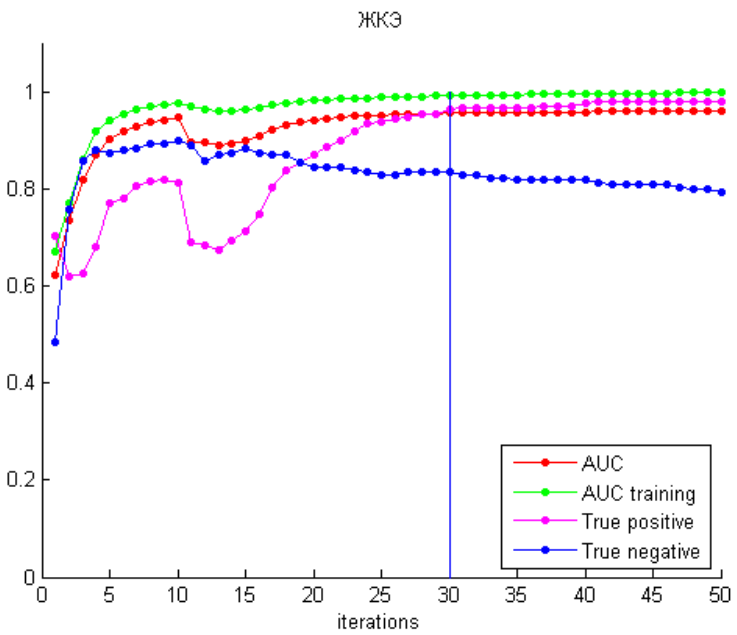
Доля здоровых с верным отрицательным диагнозом:
специфичность: $\frac{1}{l_0}\sum_{i:y_i=0}[a(x_i)=0]$

Графики по итерациям EM-алгоритма

Язвенная болезнь

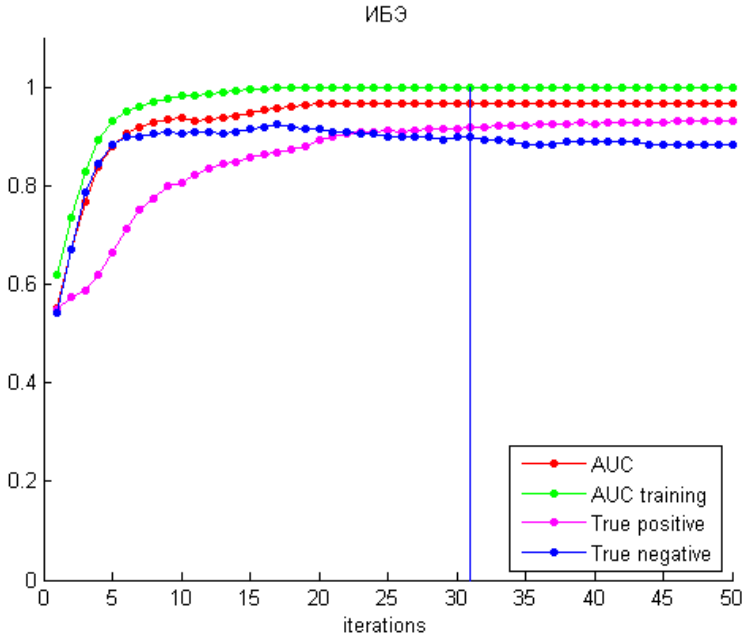


Желчнокаменная болезнь

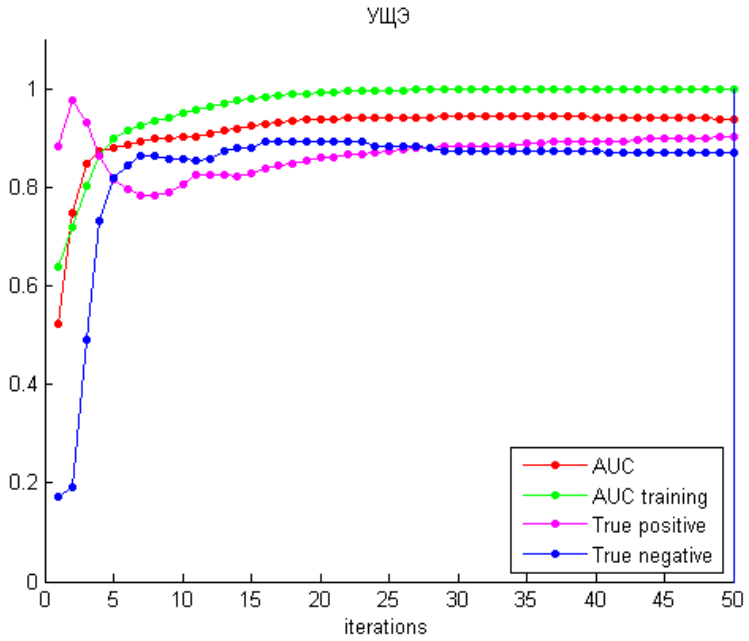


Графики по итерациям EM-алгоритма

Ишемическая болезнь сердца

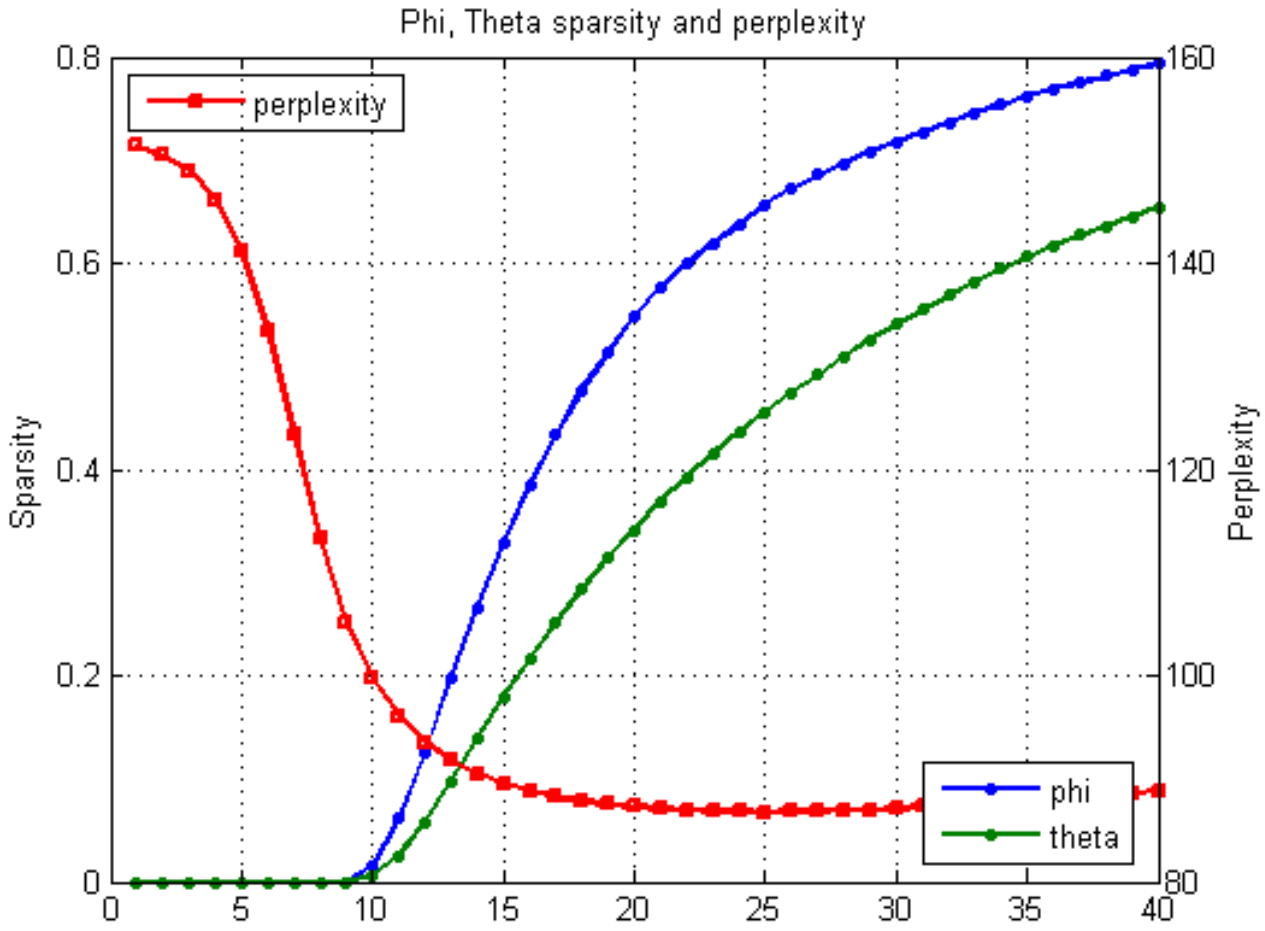


Узловой зоб щитов. железы



Графики по итерациям EM-алгоритма

График разреженности матриц и перплексии модели Язвенная болезнь



Спасибо за внимание!