Избор на признаци за on-line подписи чрез прилагане на статистически методи

Десислава Бояджиева

докторант в ОСРО-ИИКТ

Разпознаване на подписи

- Дефиниция;
- Методи – статичен (off-line) и динамичен (on-line);
- Стъпки;
- Видове признаци;
- Видове подправени подписи (forgeries) – със или без познаване на подписа (съответно skilled и random).

Бърз преглед

- Избор на признаци;
- Входни данни;
- Статистически методи
  - Редукция на признаците след анализ на корелационната матрица;
  - Избор на променливи в регресионния модел.
- Експеримент
  - Постановка;
  - Графичен потребителски интерфейс;
  - Резултати.

Избор на признаци

- Това е процесът по намирането на оптимално подмножество от *k* на брой информативни признаци от първоначалните *p* на брой $k \leq p$,същевременно се цели откриването и премахването на повтарящата се и излишна информация.
- Методи за избор на променливи

Множествена регресия

- Нека разполагаме с $n \geq k+1$ наблюдения над *k*-мерен вектор от независими променливи $x^t = (x_1, ..., x_k)$ и зависимата променлива *y*. Моделът на множествената регресия се представя с израза:

$$y_j = \beta_0 + \sum_{i=1}^{k} \beta_i x_{ij} + e_j, \qquad j = 1, ..., n$$

$$y = X\beta + e \qquad \text{(матричен вид)}$$

- Регресионните коефициенти $\beta_i, i = 1, ..., k$ се оценяват чрез МНМК

$$b = (X^T X)^{-1} X^T Y.$$

- β_0 е изключен от модела

- *k* независими променливи $\Rightarrow 2^k$ възможни регресионни модела

Избор на променливи в регресионния модел

- Критериите за избор са функции са на остатъчната сума от квадрати (RSS)
- Критерий *Cp* на Mallows

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p)$$

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{j=1}^{n} (y_j - \sum_{i=1}^{k} \beta_i x_{ij})^2$$

$$RSS_p = \sum_{j=1}^{n} (y_j - \sum_{i=1}^{p} \beta_i x_{ij})^2$$

- “Добри” са подмножествата от променливи, за които стойността на *Cp* е по-малка от *p* и близка до *p*.

Метод на LaMotte и Hocking (1)

- Базиран е на критерия на Mallows. Разглеждат се само малка част от всичките $\binom{k}{p}$ подмножества с размер *p*;
- Подмножество от *r* променливи се премахва от модела, а подмножество от *p* променливи остава, $p+r=k$;
- Редукцията в RSS от премахването на подмножество от *r* променливи се дефинира така:

$$\mathrm{Red}_r = RSS_p - RSS_k$$

- Подмножество от *r* променливи, за което тази редукция е най-малка определя съответното подмножество от *p* променливи, за което RSS е минимална;
- Статистиката *Cp* се пресмята чрез тази редукция по следния начин:

$$C_p = \frac{\mathrm{Red}_r}{\hat{\sigma}^2} - (2p-k).$$

Метод на LaMotte и Hocking (2)

Стъпки:

- Намиране на регресионните коефициенти за модела с всичките *k* на брой променливи;
- За всяка променлива x_i се пресмятат редукцията от премахването ѝ θ_i (univariate reduction) и променливите се нареждат по нарастващ ред на тези редукции:

$$\theta_i = \hat{\sigma}^2 t_i^2$$

$$t_i^2 = \frac{b_i^2}{\hat{\sigma}_{b_i}^2}$$

- Пресмятат се т.нар. *m*-редукции от премахването на подмножества от *m* променливи от модела. Подмножествата се означават се чрез $(i_1, i_2, ..., i_m)$, $1 < i_j < k$ и $i_1 < i_2 < ... < i_m$;

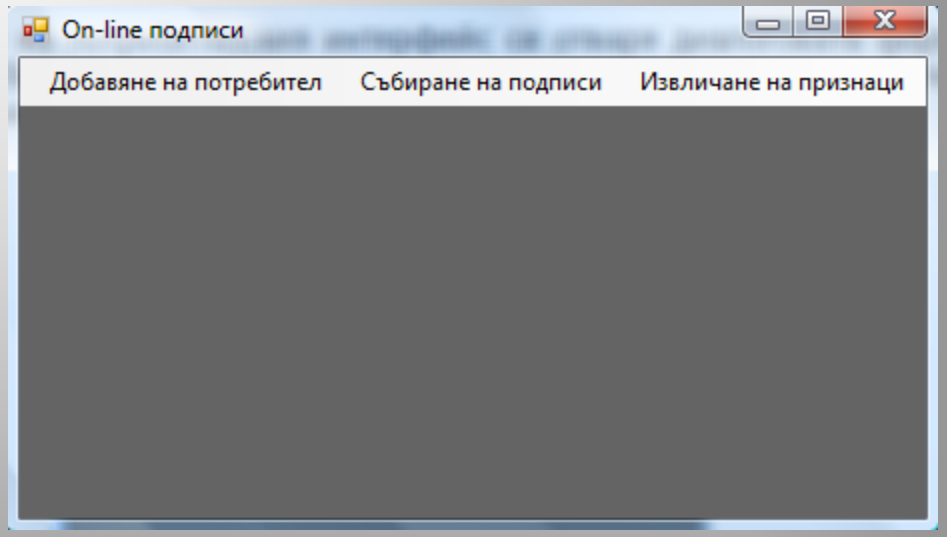
Метод на LaMotte и Hocking (3)

- Така пресметнатите *m*-редукции се сортират в нарастващ ред;
- Дефинират се т.нар. *етапи* от подмножествата от *m* елемента с първи индекс $\geq r-m+1$;
- На всеки *етап* се разглеждат подмножества от *r* елемента (*m* на брой елемента от съответното *m*-подмножество и *r-m* на брой индекси, по-малки от първия индекс в *m*-подмножеството);
- Изобщо, на *q*-тия *етап* се изчисляват редукциите от всички дефинирани на съответния етап подмножества от *r* елемента, и, ако най-малката редукция от всички етапи $1,\ldots,q$ е по-малка от *m*-редукцията от премахването на подмножеството *m* променливи, дефиниращо *етап q+1*, се счита, че е открито най-доброто *r*-подмножество, което да се премахне от изходния модел.

Извличане на признаци от графичен таблет

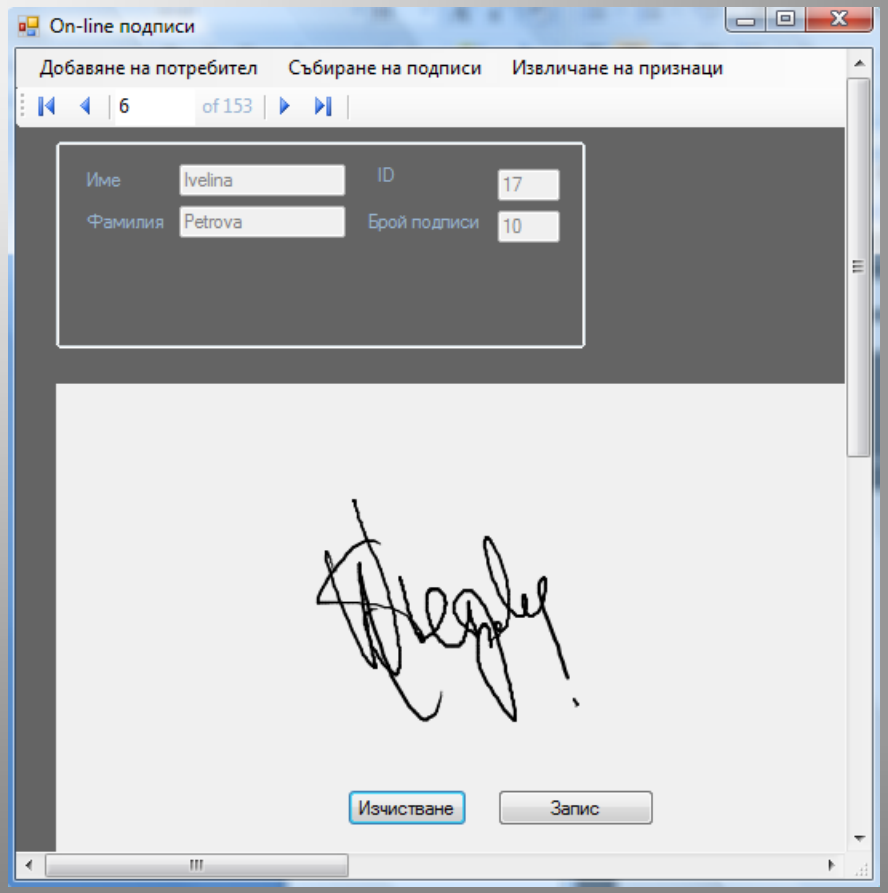
- Технология
  - Графичен таблет Wacom Intuos3 A5 PTZ-630;
  - Библиотека Microsoft .NET Tablet PC SDK 1.7.
- Графичен потребителски интерфейс
  - Реализиран на C#;
  - Възможности.
- Процес .

Графичен потребителски интерфейс (1)



Фиг. 1 Начален екран

Графичен потребителски интерфейс (2)



Фиг. 2 Екран събиране на подписи

Графичен потребителски интерфейс (3)

- спомага бързото събиране на подписи и извличането на признаци;
- предоставя възможност на потребителя да положи подписа си по естествен за него начин.

Експеримент

- 140 подписа от 14 души;
- Признаци k=20 и n=20 (десет оригинални и десет подправени). Ролята на подправени играят произволно взети подписи от останалите участници;
- След анализ на корелационната матрица остават k=15 признака;
- За всеки участник се създава текстов файл
  - Всеки ред съдържа стойностите на признаците, разделени от “;“ и завършва с -1 или 1 (съответно за подправен и оригинален подпис);

    Пример:

    157;0,97;414;64,2;62,65;198,15;24,52;2,91;93,91;206,57;35,99;31,83;165,68;223,73;31,68;1

    175;0,93;452;79,85;86,12;202,07;12,75;183,98;113,88;210,62;188,7;36,88;182,04;229,4;1,26;-1
  - Броят на редовете е n=20.
- Статистическият метод се прилага върху всеки от 14-те файла за $3<r<13$. За целта е създадено приложение на C#.

Признаци

A1 - дължина	A2 = височина/дължина - сплеснатост	A3 - брой точки	A4- разстояние от началната точка до центъра
A5 - разстояние от крайната точка до центъра	A6- ъгъл на правата от центъра до началната точка	A7 - ъгъл на правата от центъра до крайната точка	A8- ъгъл на правата от началната до крайната точка
A9- разстояние от най-лявата точка до центъра	A10 - разстояние от центъра до най- лявата точка	A11 - ъгъл на правата от центъра до най-лявата точка	A12 - разстояние от най-лявата до началната точка
A13 - разстояние от най-дясната до крайната точка	A14 - ъгъл на правата от най-лявата до началната точка	A15 - ъгъл на правата от крайната до най-дясната точка	

Табл. 1 Признаци за подписи

Резултати

Участник	Размерност на *p*-подмножество	Най-добро *p*-подмножество
1	6	A2;A5;A7;A12;A13;A15
2	9	A3;A5;A6;A7;A8;A9;A10;A11;A12
3	11	A2;A4;A5;A6;A7;A8;A9;A10;A13;A14;A15
4	11	A1;A3;A4;A6;A7;A8;A9;A11;A12;A13;A15
5	5	A1;A2;A5;A9;A15
6	6	A4;A8;A9;A10;A12;A14
7	7	A1;A2;A5;A8;A11;A12;A13
8	9	A1;A2;A3;A4;A5;A8;A9;A10;A15
9	11	A2;A5;A6;A7;A8;A9;A11;A12;A13;A14;A15
10	8	A4;A6;A9;A10;A11;A12;A13;A14
11	9	A1;A3;A4;A5;A6;A7;A9;A11;A13
12	11	A1;A2;A3;A4;A5;A6;A7;A9;A10;A11;A15
13	10	A1;A2;A3;A5;A6;A8;A10;A11;A12;A14
14	10	A1;A2;A3;A4;A5;A7;A8;A9;A10;A11

Табл. 2 Най-добри *p*-подмножества

Интерпретиране на резултатите

- Не може да се идентифицират общи за всички участници признаци, които да се премахнат;
- За всеки участник би могло да се определят информативните признаци и те да бъдат обект на разглеждане в процеса на верификацията на подпис.

Настояща работа

- Обобщаване на резултатите от експеримент върху публична база от данни и други признаци;
- Експериментиране в областта на класификацията (чрез различни НМ):
  - Тестване на различни класификатори върху едни и същи признаци
  - Разделяне на признаците в подмножества и за всяко подмножество се тества различен класификатор
  - Комбиниране изходите от класификаторите
  - Многоетапна класификация (последователно)

Благодаря за вниманието !