Título do Trabalho de Projecto

Fernando Pessoa
Ricardo Reis

Licenciatura em Engenharia Informática e de Computadores
Projecto e Seminário

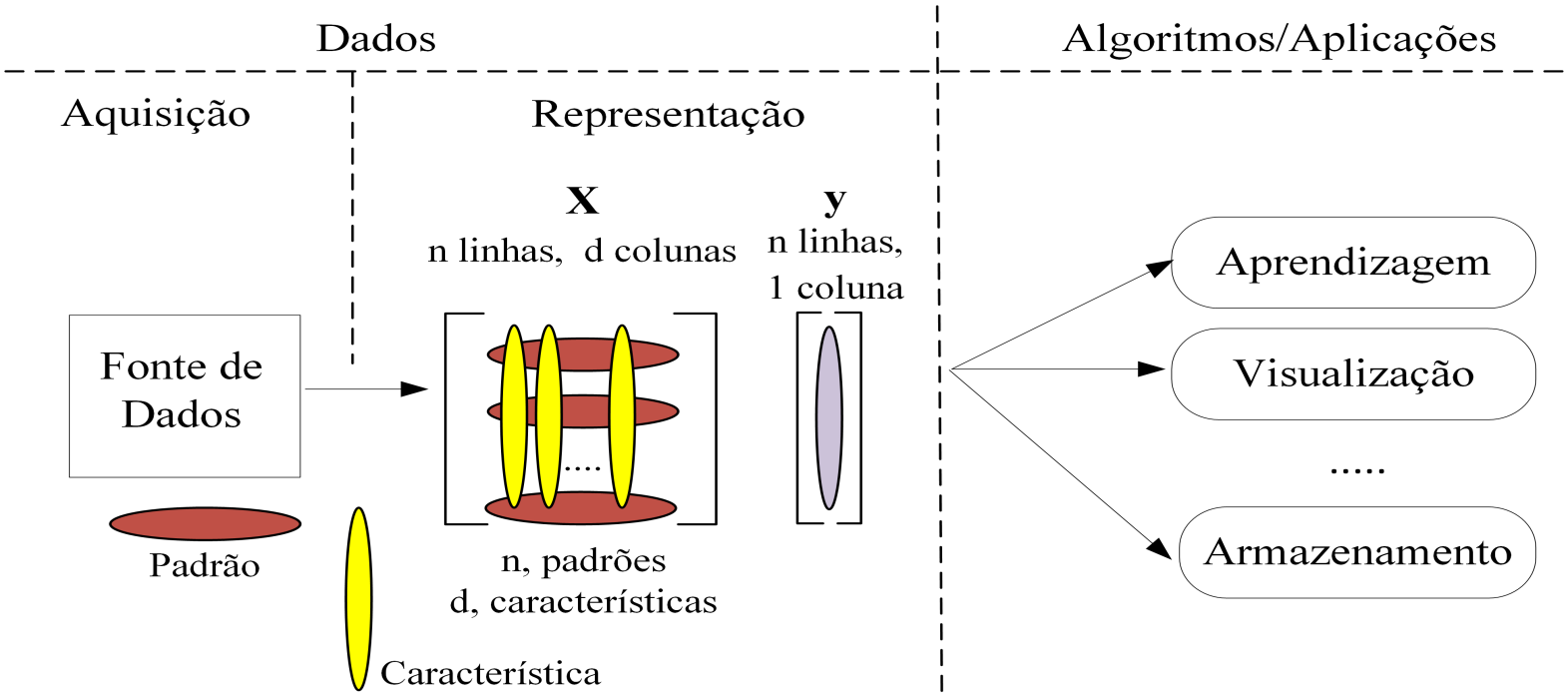
Orientadores: *Álvaro de Campos*
Alberto Caeiro, SoftCompany

Apresentação * *
Maio de 2015

Sumário

Introdução

Cenários típicos de manipulação de dados (figura)



Os dados (lista enumerada):

- organizados em *n* padrões, exemplos ou instâncias
- cada padrão tem *d* características (dimensionalidade)
- podem ser manipulados com diferentes propósitos/objetivos

O Problema: conceitos

Exemplos de dados de baixa dimensionalidade e com AD (tabela):

- d dimensões, c classes, n padrões
- em baixa dimensionalidade, $n > d$ ou $n \gg d$
- com AD, tipicamente tem-se $d \gg n$

Conjunto	d	c	n	**Tipo de Dados**	**Problema**
Car	6	4	1728	Automóveis	Possibilidade de Compra
Pima	8	2	768	Clínicos	Deteção da Diabetes
Contraceptive	9	2	1473	Clínicos	Método de Contraceção
Wine	13	3	178	Químicos	Tipo de Vinho
Hepatitis	19	2	155	Clínicos	Deteção de Hepatite
Dermatology	34	6	358	Clínicos	Doença de Pele
Colon	2000	2	62	Exp. Genética	Deteção de Cancro
SRBCT	2309	4	83	Exp. Genética	Deteção de Cancro
TOX-171	5748	4	171	Exp. Genética	Deteção de Cancro
Example1	9947	2	2000	Texto	Assunto da Notícia
ORL10P	10304	10	100	Faces	Identificação
11-Tumors	12553	11	174	Exp. Genética	Deteção de Cancro
Lung-Cancer	12601	5	203	Exp. Genética	Deteção de Cancro
Dexter	20000	2	2600	Texto	Assunto da Notícia
GLI-85	22283	2	85	Exp. Genética	Deteção de Cancro
Dorothea	1000000	2	1950	Clínicos	Deteção de Composto

Propostas para Seleção: Complexidade

Em termos de complexidade, tem-se para o RFS

$$C_{RFS} = \underbrace{O(nd)}_{Relevancia} + \underbrace{O(d \log d)}_{Ordenacao}$$

RRFS apresenta a complexidade adicional de calcular as semelhanças entre pares

$$C_{RRFS} = \underbrace{O(nd)}_{Relevancia} + \underbrace{O(d \log d)}_{Ordenacao} + \underbrace{O(nm)}_{Redundancia}$$

- RRFS é mais rápido do que outros filtros (FCBF e MRMR)
- Nalguns casos, RRFS é o mais rápido com menos erro
- CFS é o mais lento (inadequado para dados com AD)
- Medidas MAD e MM adequadas para todos os tipos de dados
- Medida AMGM é mais adequada para dados esparsos

Representação de Dados

Assim, identifica-se a necessidade de:

- encontrar novas formas de representação dos dados
- representar de forma independente da tarefa a jusante
- facilitar a visualização e análise de dados com AD

As técnicas de **seleção** (*feature selection* - FS) e **discretização** (*feature discretization* - FD) realizam essa representação

- **Seleção**

→ escolha de sub-conjuntos de características adequados

- **Discretização**

→ representações discretas de características numéricas
→ com informação suficiente para aprendizagem
→ ignora ruído e flutuações irrelevantes

Seleção: Taxonomia e Opções Tomadas

Algoritmos de seleção são categorizados como:

i) filtros (*filters*)

ii) envolvimento (*wrapper*)

iii) embebidos (*embedded*)

iv) híbridos (*hybrid*)

Escolha inicial $\rightarrow$ filtros não supervisionados e supervisionados

- Baixa complexidade, eficiência e interpretabilidade
- Independentes da tarefa de mineração de dados
- Alguns filtros existentes são:
  - ineficientes (tempo) em dados com AD[1]
  - sensíveis ao problema de elevado d, baixo n[2]

[1] CFS-Correlation-based Feature Selection, MRMR-Maximum Relevance Minimum Redundancy, RELIEF-Recursive Elimination of Features

[2] FCBF - Fast Correlation-Based Filter

O algoritmo PS

Dados: X: $n \times d$ matrix, n patterns of a d-dimensional training set.
m ($\leq d$):, maximum number of features to keep.
M_S: maximum allowed similarity between pairs of features.

Resultado: $\mathrm{FeatKeep}$: an $m'-$dimensional array (with $m' \leq m$) containing the indexes of the selected features.
$\widetilde{X}$: $n \times m'$ matrix, reduced dimensional training set, with features sorted by decreasing relevance.

1: Compute the relevance r_i of each feature X_i, for $i \in \{1, \ldots, d\}$, using a dispersion measure (MAD, MM, AMGM, or IQR).
2: Sort the features by decreasing order of r_i. Let $i_1, i_2, ..., i_d$ be the resulting permutation of $\{1, ..., d\}$ (*i.e.*, $r_{i_1} \geq r_{i_2} \geq ... \geq r_{i_d}$).
3: $\mathrm{FeatKeep}[1] = i_1$; prev=1; next=2;
4: **for** $f = 2$ to d **do**
5: $s = S(X_{i_f}, X_{i_{\mathrm{prev}}})$;
6: **if** $s < M_S$ **then**
7: $\mathrm{FeatKeep}[\mathrm{next}] = i_f$; $\widetilde{X}_{\mathrm{next}} = X_{i_f}$; prev $= i_f$; next $=$ next $+ 1$;
8: **end if**
9: **if** next=m **then**
10: break; {/* We have m features. Break loop. */}
11: **end if**

Avaliação Experimental: Relevância

Algoritmo *relevance FS* (RFS):

- guarda as características com maior *relevância*
- relevância medida pela *dispersão*

Para o caso não supervisionado $\rightarrow$ medidas de *relevância* @*rel*:

- *Mean absolute difference* $\text{MAD}_i = \frac{1}{n}\sum_{j=1}^{n}\left|X_{ij} - \overline{X}_i\right|$
- *Mean-median* $\text{MM}_i = |\overline{X}_i - \text{median}(X_i)|$
- *Arithmetic mean geometric mean*

$$\text{AMGM}_i = \frac{1}{n}\sum_{j=1}^{n}\exp(X_{ij}) / \left(\exp\left(\sum_{j=1}^{n} X_{ij}\right)\right)^{\frac{1}{n}}$$

Critério de relevância cumulativa $\sum_{f=1}^{m} r_{i_f} / \sum_{i=1}^{d} r_i = c_m/c_d \geq L$, $L \in [0,8; 0,95] \rightarrow$ escolha do número de características m $(< d)$

Propostas para Seleção: Resultados Experimentais

- Classificação supervisionada (SVM linear), validação cruzada (*10-fold*)
- Percentagem de erro de generalização com filtros supervisionados

	RRFS, $M_S = 0.8$			**Filtros Supervisionados**[3]					**Base**
Conjunto	**MM**	**FiR**	**MI**	**RF**	**CFS**	**FCBF**	**FiR**	**MRMR**	**Sem FS**
Colon	24.2	22.6	24.2	**19.4**	25.8	22.6	**19.4**	21.0	21.0
Lymphoma	**2.2**	**2.2**	**2.2**	**2.2**	N/A	3.3	**2.2**	22.8	**2.2**
Leukemia1	5.6	**2.8**	6.9	6.9	N/A	5.6	4.2	9.7	5.6
B-Tumor1	13.3	12.2	13.3	11.1	N/A	18.9	11.1	25.6	**10.0**
Leukemia	**2.8**	12.5	**2.8**	**2.8**	N/A	4.2	4.2	8.3	**2.8**
Example1	2.3	2.2	2.2	3.7	N/A	6.3	**2.1**	28.3	2.4
B-Tumor2	34.0	**22.0**	30.0	**22.0**	N/A	36.0	24.0	42.0	26.0
P-Tumor	7.8	5.9	**4.9**	7.8	N/A	9.8	7.8	12.7	8.8
L-Cancer	5.9	6.4	4.9	4.9	N/A	6.4	5.4	11.8	5.9
Dexter	6.7	**6.0**	7.7	9.3	N/A	15.3	6.7	18.0	6.3
Dorothea	**25.0**	26.0	**25.0**	N/A	N/A	N/A	**25.0**	N/A	**25.0**

[3] Mutual Information (MI), RELIEF (RF), Correlation-based Feature Selection (CFS), Fast Correlation-based Filter (FCBF), Fishers Ratio (FiR) e Maximum Relevance Minimum Redundancy (MRMR)

Conclusões

- O projeto consistiu em....
- Atingiu-se uma solução ...
- Os métodos propostos foram avaliados:
  - sobre dados de domínio público
  - comparativamente com métodos existentes em implementações de domínio público
- Os métodos propostos complementam os existentes, podendo ser combinados entre si

Trabalho Futuro

Perspetivam-se as direções de trabalho futuro:

- Melhorar a proposta de ...
- Explorar o afinamento dos parâmetros ...
- ..

Slides elaborados em LaTeX com o package BEAMER.