

Métodos para Classificação de Texto de Etiqueta Única

(Improving Methods for Single-label Text Categorization)

Ana Cardoso Cachopo

Instituto Superior Técnico — Universidade Técnica de Lisboa / INESC-ID

8 de Outubro de 2007



Estrutura

- 1 Introdução
- 2 Ambiente Experimental
- 3 Comparação dos Métodos Existentes
- 4 Combinações entre Métodos
- 5 Utilização de Documentos Não Etiquetados
- 6 Contribuições

Outline

- 1 Introdução
- 2 Ambiente Experimental
- 3 Comparação dos Métodos Existentes
- 4 Combinações entre Métodos
- 5 Utilização de Documentos Não Etiquetados
- 6 Contribuições

Introdução e Objectivos

Introdução

- Classificação de Texto
- Documentos de Etiqueta Única
- Classificação Semi-supervisionada

Objectivos

- Melhorar a qualidade dos resultados através da combinação de classificadores.
- Reduzir a quantidade de dados pré-processados que é necessária através da utilização de documentos não etiquetados.

Introdução e Objectivos

Introdução

- Classificação de Texto
- Documentos de Etiqueta Única
- Classificação Semi-supervisionada

Objectivos

- Melhorar a qualidade dos resultados através da combinação de classificadores.
- Reduzir a quantidade de dados pré-processados que é necessária através da utilização de documentos não etiquetados.

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.
 - Pesos binários
 - $tfidf$

$$w_{ij} = \frac{freq_{ij}}{\max_j(freq_{ij})}$$

Os pesos são normalizados por documento e por termo.

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.
 - Pesos binários
 - *tfidf*

$$w_{ij} = \frac{freq_{ij}}{\max_l(freq_{lj})}$$

Os pesos binários são calculados de acordo com a seguinte fórmula:

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.
 - Pesos binários
 - *tfidf*

$$w_{ij} = \frac{freq_{ij}}{\max_l(freq_{lj})} \times \log \frac{|D|}{n_t}$$

w_{ij}

$$w_{ij} = w_{ij} \times \log \frac{|D|}{n_t}$$

Os pesos são normalizados para garantir comparabilidade entre os documentos.

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.
 - Pesos binários
 - *tfidf*

$$w_{ij} = \frac{\text{freq}_{ij}}{\max_l(\text{freq}_{lj})} \times \log \frac{|D|}{n_{t_i}}$$

* *td*

$$wtd_{ijk} = w_{ij} \times \text{icsd}_k^2 = \text{icsd}_k^2 \times w_{ij}$$

Os pesos são atribuídos aos termos de acordo com a sua importância

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.
 - Pesos binários
 - *tfidf*

$$w_{ij} = \frac{freq_{ij}}{\max_l(freq_{lj})} \times \log \frac{|D|}{n_{t_i}}$$

- *td*

$$wtd_{ijk} = w_{ij} \times ic_{sd_i}^{\alpha} \times csd_{jk}^{\beta}$$

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.
 - Pesos binários
 - *tfidf*

$$w_{ij} = \frac{freq_{ij}}{\max_l(freq_{lj})} \times \log \frac{|D|}{n_{t_i}}$$

- *td*

$$wtd_{ijk} = w_{ij} \times icd_i^\alpha \times csd_{ik}^\beta \times sd_i^\gamma$$

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.

- Pesos binários
- *tfidf*

$$w_{ij} = \frac{freq_{ij}}{\max_l(freq_{lj})} \times \log \frac{|D|}{n_{t_i}}$$

- *td*

$$wtd_{ijk} = w_{ij} \times icsd_i^\alpha \times csd_{ik}^\beta \times sd_i^\gamma$$

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.

- Pesos binários
- *tfidf*

$$w_{ij} = \frac{freq_{ij}}{\max_l(freq_{lj})} \times \log \frac{|D|}{n_{t_i}}$$

- *td*

$$wtd_{ijk} = w_{ij} \times icsd_i^\alpha \times csd_{ik}^\beta \times sd_i^\gamma$$

- Os vectores são normalizados para terem comprimento unitário.

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.

- Pesos binários
- *tfidf*

$$w_{ij} = \frac{freq_{ij}}{\max_l(freq_{lj})} \times \log \frac{|D|}{n_{t_i}}$$

- *td*

$$wtd_{ijk} = w_{ij} \times icsd_i^\alpha \times csd_{ik}^\beta \times sd_i^\gamma$$

- Os vectores são normalizados para terem comprimento unitário.

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.

- Pesos binários
- *tfidf*

$$w_{ij} = \frac{freq_{ij}}{\max_l(freq_{lj})} \times \log \frac{|D|}{n_{t_i}}$$

- *td*

$$wtd_{ijk} = w_{ij} \times icd_i^\alpha \times csd_{ik}^\beta \times sd_i^\gamma$$

- Os vectores são normalizados para terem comprimento unitário.

Peso dos Termos dos Documentos

- Os documentos são representados por vectores p -dimensionais.
- Os termos são pesados de acordo com a sua importância.

- Pesos binários
- *tfidf*

$$w_{ij} = \frac{freq_{ij}}{\max_l(freq_{lj})} \times \log \frac{|D|}{n_{t_i}}$$

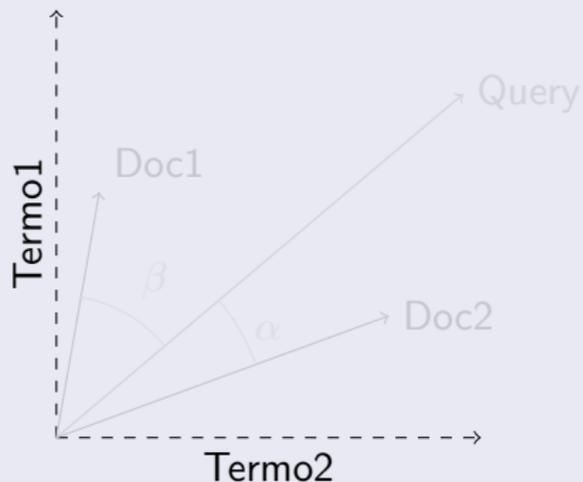
- *td*

$$wtd_{ijk} = w_{ij} \times icsd_i^\alpha \times csd_{ik}^\beta \times sd_i^\gamma$$

- Os vectores são normalizados para terem comprimento unitário.

Métodos de Classificação

Vector

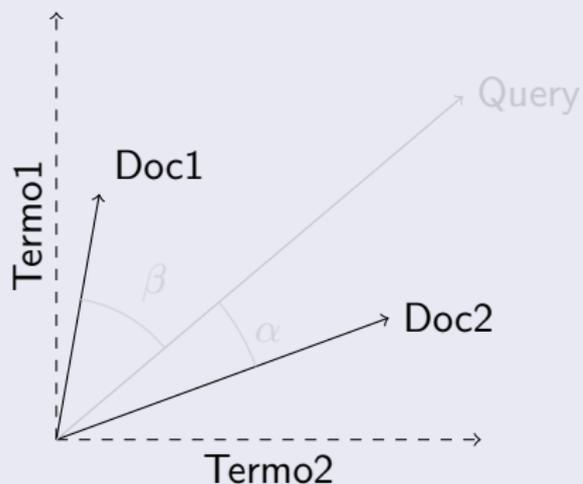


$$\text{sim}(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{\|\vec{d}_j\| \times \|\vec{q}\|}$$

Considera a semelhança entre os vectores que representam os documentos.

Métodos de Classificação

Vector

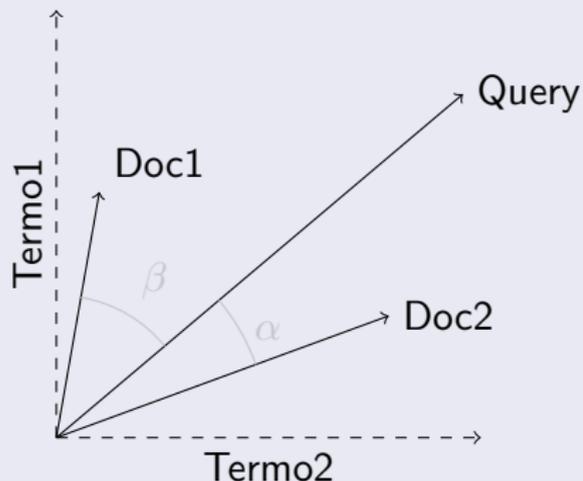


$$\text{sim}(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{\|\vec{d}_j\| \times \|\vec{q}\|}$$

Considera a semelhança entre os vectores que representam os documentos.

Métodos de Classificação

Vector

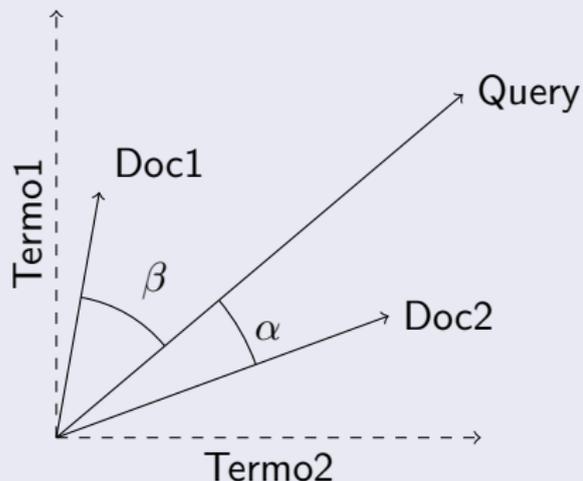


$$\text{sim}(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{\|\vec{d}_j\| \times \|\vec{q}\|}$$

Considera a semelhança entre os vectores que representam os documentos.

Métodos de Classificação

Vector

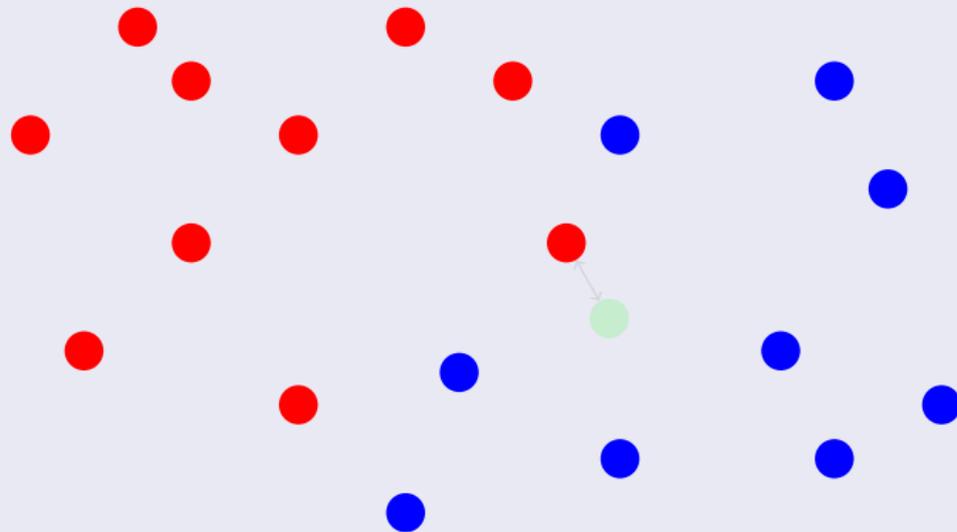


$$\text{sim}(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{\|\vec{d}_j\| \times \|\vec{q}\|}$$

Considera a semelhança entre os vectores que representam os documentos.

Métodos de Classificação

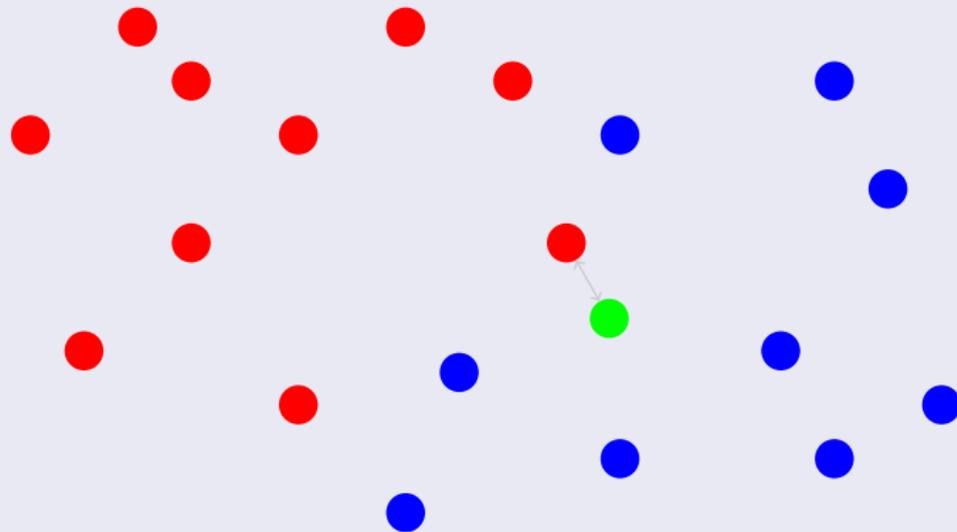
Vector



Considera a classe do documento mais próximo.

Métodos de Classificação

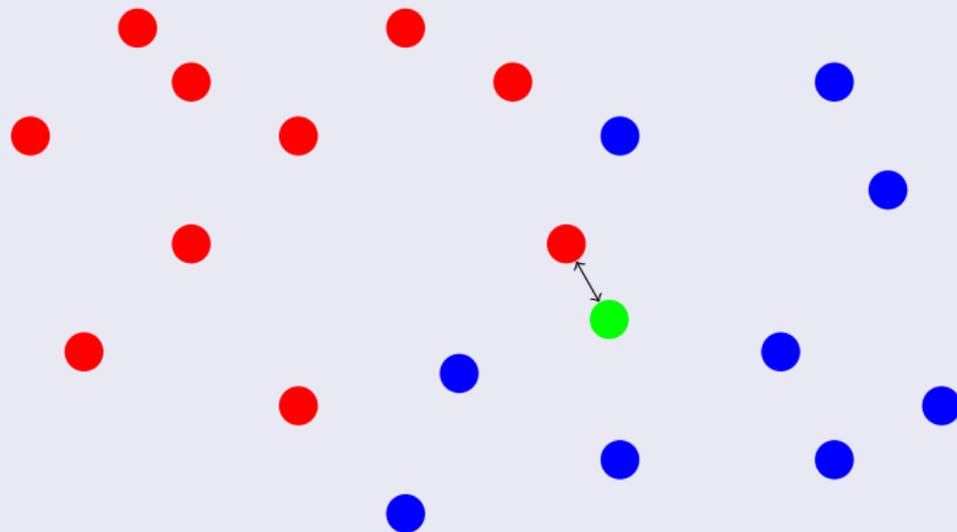
Vector



Considera a classe do documento mais próximo.

Métodos de Classificação

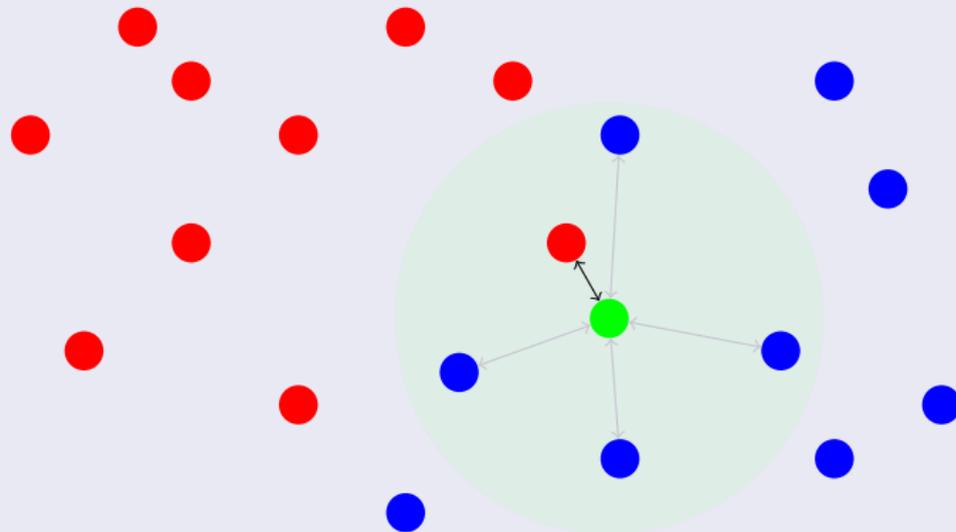
Vector



Considera a classe do documento mais próximo.

Métodos de Classificação

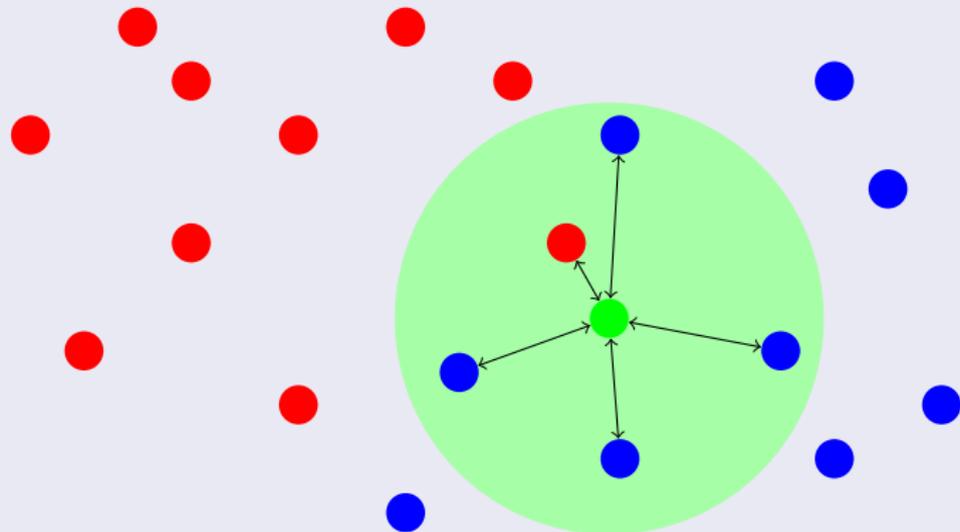
k-NN



Considera os k vizinhos mais próximos.

Métodos de Classificação

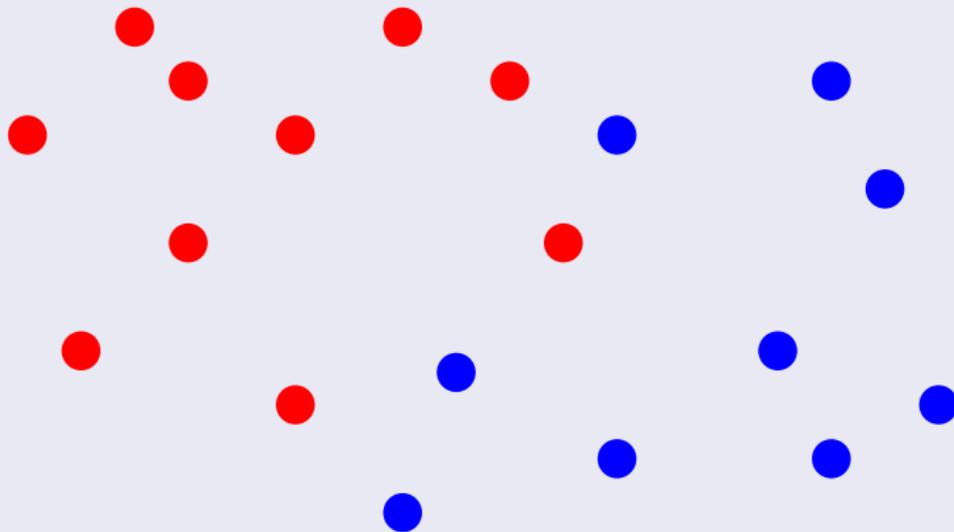
k-NN



Considera os k vizinhos mais próximos.

Métodos de Classificação

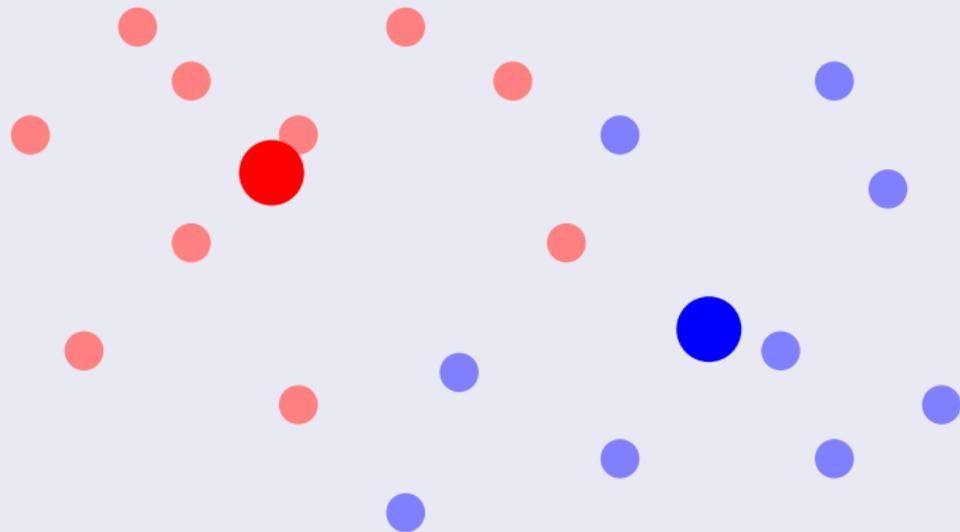
Centroid



Considera o centroide de cada classe.

Métodos de Classificação

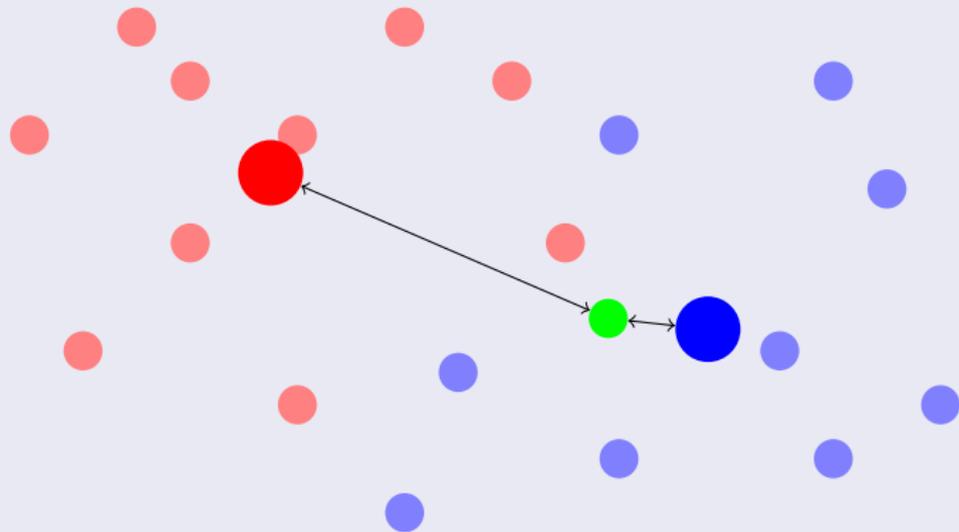
Centroid



Considera o centroide de cada classe.

Métodos de Classificação

Centroid



Considera o centroide de cada classe.

Métodos de Classificação

Centroid

$$\text{Sum: } \vec{c}_k = \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{Average: } \vec{c}_k = \frac{1}{|D_{c_k}|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{NormSum: } \vec{c}_k = \frac{1}{\left\| \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j \right\|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{Rocchio: } \vec{c}_k = \beta \cdot \frac{1}{|D_{c_k}|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j - \gamma \cdot \frac{1}{|D - D_{c_k}|} \cdot \sum_{\vec{d}_j \notin D_{c_k}} \vec{d}_j$$

Métodos de Classificação

Centroid

$$\text{Sum: } \vec{c}_k = \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{Average: } \vec{c}_k = \frac{1}{|D_{c_k}|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{NormSum: } \vec{c}_k = \frac{1}{\left\| \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j \right\|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{Rocchio: } \vec{c}_k = \beta \cdot \frac{1}{|D_{c_k}|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j - \gamma \cdot \frac{1}{|D - D_{c_k}|} \cdot \sum_{\vec{d}_j \notin D_{c_k}} \vec{d}_j$$

Métodos de Classificação

Centroid

$$\text{Sum: } \vec{c}_k = \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{Average: } \vec{c}_k = \frac{1}{|D_{c_k}|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{NormSum: } \vec{c}_k = \frac{1}{\left\| \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j \right\|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{Rocchio: } \vec{c}_k = \beta \cdot \frac{1}{|D_{c_k}|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j - \gamma \cdot \frac{1}{|D - D_{c_k}|} \cdot \sum_{\vec{d}_j \notin D_{c_k}} \vec{d}_j$$

Métodos de Classificação

Centroid

$$\text{Sum: } \vec{c}_k = \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{Average: } \vec{c}_k = \frac{1}{|D_{c_k}|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{NormSum: } \vec{c}_k = \frac{1}{\left\| \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j \right\|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j$$

$$\text{Rocchio: } \vec{c}_k = \beta \cdot \frac{1}{|D_{c_k}|} \cdot \sum_{\vec{d}_j \in D_{c_k}} \vec{d}_j - \gamma \cdot \frac{1}{|D - D_{c_k}|} \cdot \sum_{\vec{d}_j \notin D_{c_k}} \vec{d}_j$$

Métodos de Classificação

Naive Bayes

$$P(c_k|\vec{d}_j) = \frac{P(c_k)P(\vec{d}_j|c_k)}{P(\vec{d}_j)} \approx \sum_{i=1}^{|T|} w_{ij} \log \frac{P_{ik}(1 - P_{i\bar{k}})}{P_{i\bar{k}}(1 - P_{ik})}$$

Considera a probabilidade de um documento pertencer a uma determinada classe.

Métodos de Classificação

Naive Bayes

$$P(c_k | \vec{d}_j) = \frac{P(c_k)P(\vec{d}_j | c_k)}{P(\vec{d}_j)} \approx \sum_{i=1}^{|\mathcal{T}|} w_{ij} \log \frac{P_{ik}(1 - P_{i\bar{k}})}{P_{i\bar{k}}(1 - P_{ik})}$$

Considera a probabilidade de um documento pertencer a uma determinada classe.

Métodos de Classificação

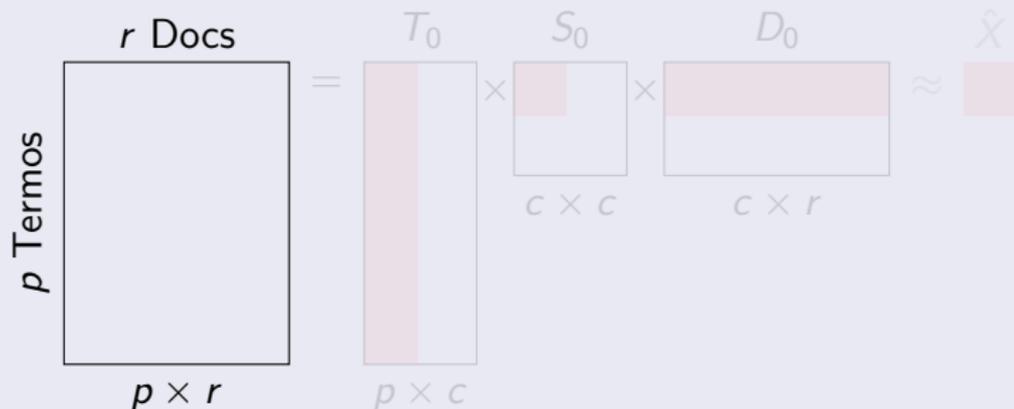
Naive Bayes

$$P(c_k | \vec{d}_j) = \frac{P(c_k)P(\vec{d}_j | c_k)}{P(\vec{d}_j)} \approx \sum_{i=1}^{|\mathcal{T}|} w_{ij} \log \frac{P_{ik}(1 - P_{i\bar{k}})}{P_{i\bar{k}}(1 - P_{ik})}$$

Considera a probabilidade de um documento pertencer a uma determinada classe.

Métodos de Classificação

LSI

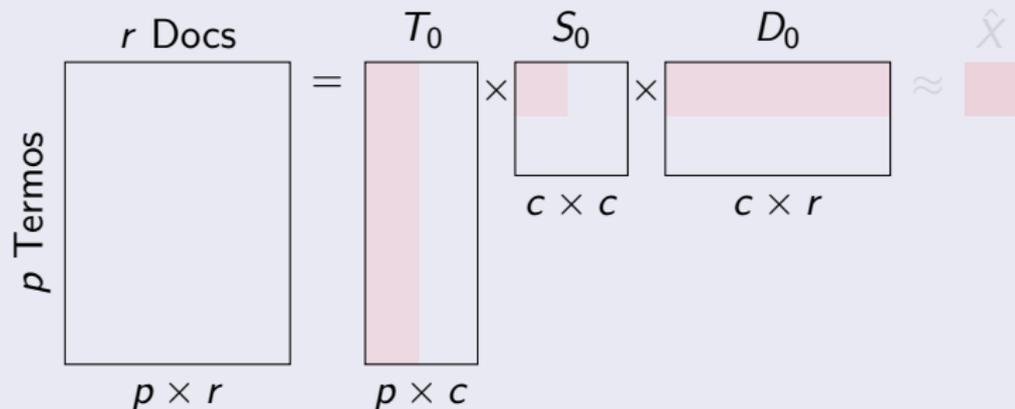


$X = T_0 S_0 D_0$ tal que T_0 e D_0 têm colunas ortonormais e S_0 é diagonal

Usa Singular Value Decomposition para reduzir as dimensões da matriz de termos por documentos.

Métodos de Classificação

LSI

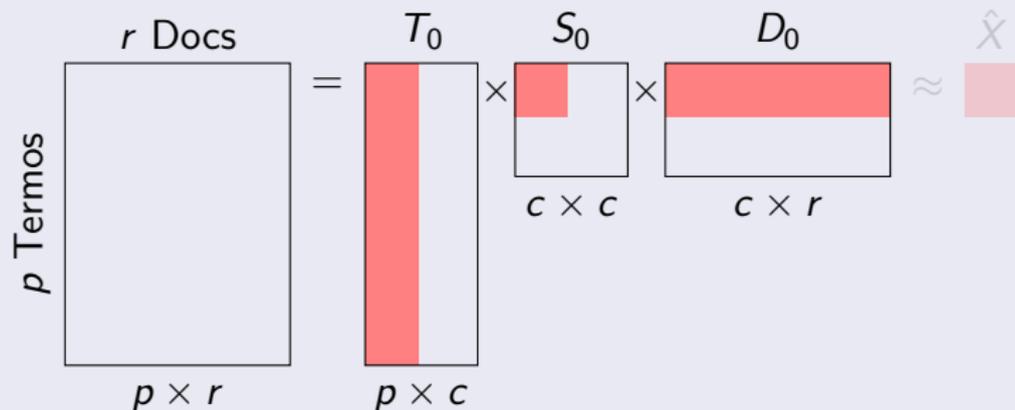


$X = T_0 S_0 D_0$ tal que T_0 e D_0 têm colunas ortonormais e S_0 é diagonal

Usa Singular Value Decomposition para reduzir as dimensões da matriz de termos por documentos.

Métodos de Classificação

LSI

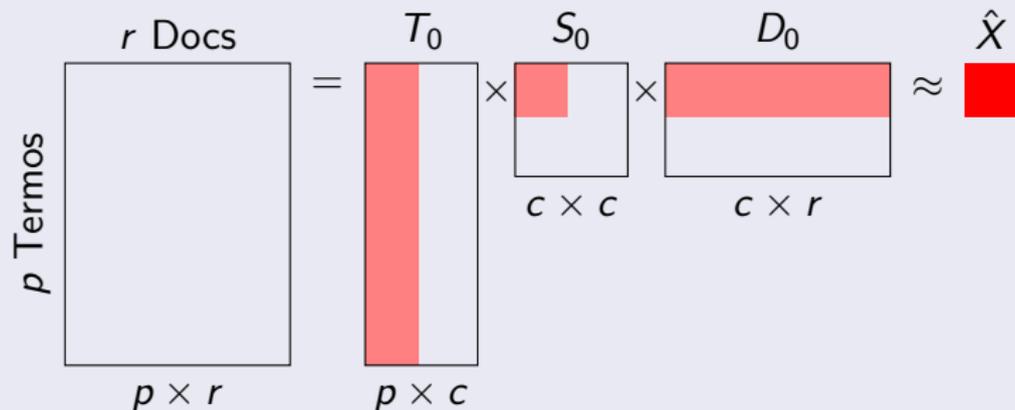


$X = T_0 S_0 D_0$ tal que T_0 e D_0 têm colunas ortonormais e S_0 é diagonal

Usa Singular Value Decomposition para reduzir as dimensões da matriz de termos por documentos.

Métodos de Classificação

LSI

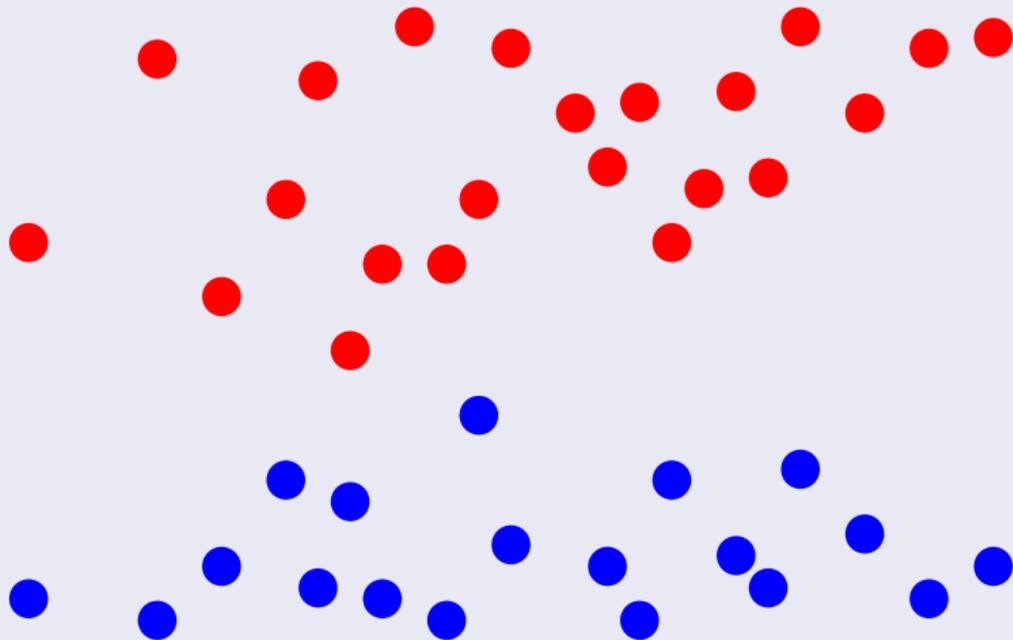


$X = T_0 S_0 D_0$ tal que T_0 e D_0 têm colunas ortonormais e S_0 é diagonal

Usa Singular Value Decomposition para reduzir as dimensões da matriz de termos por documentos.

Métodos de Classificação

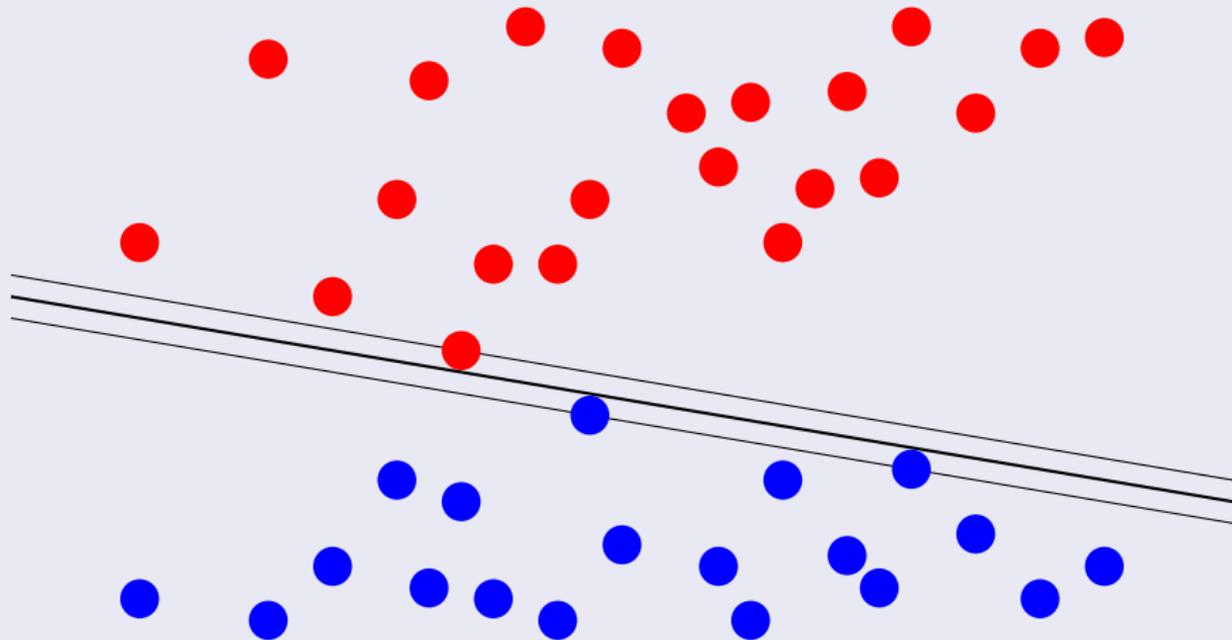
SVM



Determina o hiperplano com maiores margens entre duas classes.

Métodos de Classificação

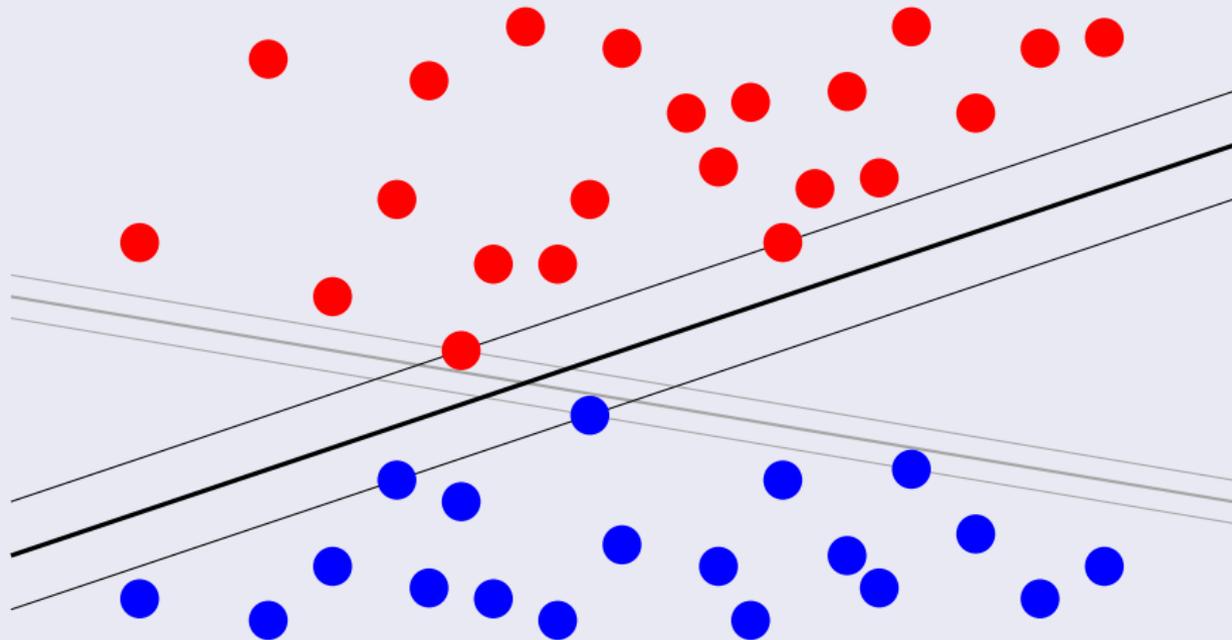
SVM



Determina o hiperplano com maiores margens entre duas classes.

Métodos de Classificação

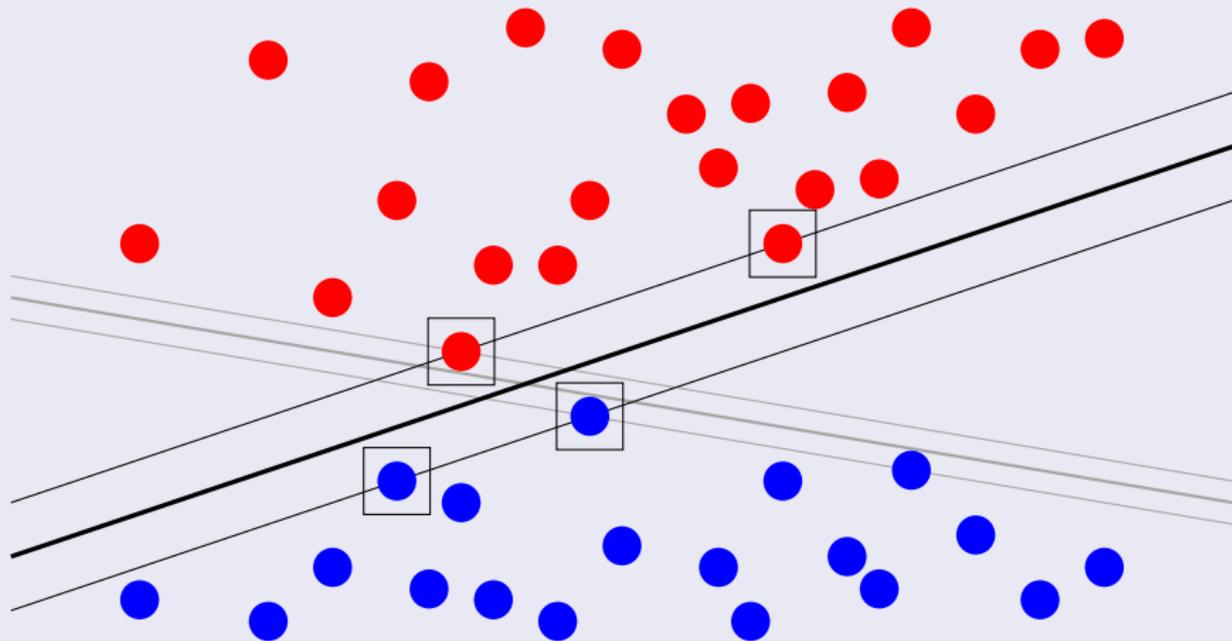
SVM



Determina o hiperplano com maiores margens entre duas classes.

Métodos de Classificação

SVM



Determina o hiperplano com maiores margens entre duas classes.

Métodos de Classificação

SVM

$$\begin{array}{ll} \text{minimizar} & - \sum_{i=1}^n \alpha_i + \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j * K(d_i, d_j) \\ \\ \text{tal que} & \sum_{i=1}^n \alpha_i y_i = 0 \quad \text{e} \quad \forall_i \alpha_i \geq 0 \end{array}$$

Medidas de Avaliação

- Accuracy

$$Accuracy = \frac{\# \text{Documentos correctamente classificados}}{\# \text{Total de documentos}}$$

- MRR

$$MRR(n) = \frac{\sum_{i=1}^{\# \text{Total queries}} ((\frac{1}{rank_i}) \text{ or } 0)}{\# \text{Total queries}}$$

onde $rank_i$ é a posição da primeira resposta correcta para o query i , considerando as primeiras n classes retornadas pelo sistema.

Medidas de Avaliação

- Accuracy

$$Accuracy = \frac{\# \text{Documentos correctamente classificados}}{\# \text{Total de documentos}}$$

- MRR

$$MRR(n) = \frac{\sum_{i=1}^{\# \text{Total queries}} \left(\frac{1}{rank_i} \right) \text{ or } 0}{\# \text{Total queries}}$$

onde $rank_i$ é a posição da primeira resposta correcta para o query i , considerando as primeiras n classes retornadas pelo sistema.

Outline

- 1 Introdução
- 2 Ambiente Experimental**
- 3 Comparação dos Métodos Existentes
- 4 Combinações entre Métodos
- 5 Utilização de Documentos Não Etiquetados
- 6 Contribuições

Conjuntos de Dados

| | Docs Treino | Docs Teste | Total Docs | Menor Classe | Maior Classe |
|--------|----------------|---------------|---------------|-----------------|-----------------|
| Bank37 | 928 | 463 | 1391 | 5 | 346 |
| 20Ng | 11293 | 7528 | 18821 | 628 | 999 |
| R8 | 5485 | 2189 | 7674 | 51 | 3923 |
| R52 | 6532 | 2568 | 9100 | 3 | 3923 |
| Web4 | 2803 | 1396 | 4199 | 504 | 1641 |
| Cade12 | 27322 | 13661 | 40983 | 625 | 8473 |

Números de documentos para os conjuntos de dados: número de documentos de treino, número de documentos de teste, número total de documentos, número de documentos na menor classe, e número de documentos na maior classe.

Conjuntos de Dados

| | Docs Treino | Docs Teste | Total Docs | Menor Classe | Maior Classe |
|--------|----------------|---------------|---------------|-----------------|-----------------|
| Bank37 | 928 | 463 | 1391 | 5 | 346 |
| 20Ng | 11293 | 7528 | 18821 | 628 | 999 |
| R8 | 5485 | 2189 | 7674 | 51 | 3923 |
| R52 | 6532 | 2568 | 9100 | 3 | 3923 |
| Web4 | 2803 | 1396 | 4199 | 504 | 1641 |
| Cade12 | 27322 | 13661 | 40983 | 625 | 8473 |

Números de documentos para os conjuntos de dados: número de documentos de treino, número de documentos de teste, número total de documentos, número de documentos na menor classe, e número de documentos na maior classe.

Conjuntos de Dados

| | Docs Treino | Docs Teste | Total Docs | Menor Classe | Maior Classe |
|--------|----------------|---------------|---------------|-----------------|-----------------|
| Bank37 | 928 | 463 | 1391 | 5 | 346 |
| 20Ng | 11293 | 7528 | 18821 | 628 | 999 |
| R8 | 5485 | 2189 | 7674 | 51 | 3923 |
| R52 | 6532 | 2568 | 9100 | 3 | 3923 |
| Web4 | 2803 | 1396 | 4199 | 504 | 1641 |
| Cade12 | 27322 | 13661 | 40983 | 625 | 8473 |

Números de documentos para os conjuntos de dados: número de documentos de treino, número de documentos de teste, número total de documentos, número de documentos na menor classe, e número de documentos na maior classe.

Conjuntos de Dados

| | Docs Treino | Docs Teste | Total Docs | Menor Classe | Maior Classe |
|--------|----------------|---------------|---------------|-----------------|-----------------|
| Bank37 | 928 | 463 | 1391 | 5 | 346 |
| 20Ng | 11293 | 7528 | 18821 | 628 | 999 |
| R8 | 5485 | 2189 | 7674 | 51 | 3923 |
| R52 | 6532 | 2568 | 9100 | 3 | 3923 |
| Web4 | 2803 | 1396 | 4199 | 504 | 1641 |
| Cade12 | 27322 | 13661 | 40983 | 625 | 8473 |

Números de documentos para os conjuntos de dados: número de documentos de treino, número de documentos de teste, número total de documentos, número de documentos na menor classe, e número de documentos na maior classe.

IREP

- Pré-processa os documentos e passa-os para os métodos.
- Usa as mesmas medidas de avaliação com todos os métodos.
- Permite uma fácil incorporação de novos métodos.
- Permite uma fácil combinação dos métodos existentes.
- Permite uma fácil mudança dos parâmetros para cada método.
- Pode ser chamado repetidamente a partir de uma shell.
- Produz resultados num formato compreensível.

IREP

é uma ferramenta computacional altamente configurável, que pode ser usada para fazer experiências com métodos existentes e facilmente estendida para incorporar novos métodos, medidas de avaliação e conjuntos de dados.

IREP

- Pré-processa os documentos e passa-os para os métodos.
- Usa as mesmas medidas de avaliação com todos os métodos.
- Permite uma fácil incorporação de novos métodos.
- Permite uma fácil combinação dos métodos existentes.
- Permite uma fácil mudança dos parâmetros para cada método.
- Pode ser chamado repetidamente a partir de uma shell.
- Produz resultados num formato compreensível.

IREP

é uma ferramenta computacional altamente configurável, que pode ser usada para fazer experiências com métodos existentes e facilmente estendida para incorporar novos métodos, medidas de avaliação e conjuntos de dados.

IREP

- Pré-processa os documentos e passa-os para os métodos.
- Usa as mesmas medidas de avaliação com todos os métodos.
- Permite uma fácil incorporação de novos métodos.
- Permite uma fácil combinação dos métodos existentes.
- Permite uma fácil mudança dos parâmetros para cada método.
- Pode ser chamado repetidamente a partir de uma shell.
- Produz resultados num formato compreensível.

IREP

é uma ferramenta computacional altamente configurável, que pode ser usada para fazer experiências com métodos existentes e facilmente estendida para incorporar novos métodos, medidas de avaliação e conjuntos de dados.

IREP

- Pré-processa os documentos e passa-os para os métodos.
- Usa as mesmas medidas de avaliação com todos os métodos.
- Permite uma fácil incorporação de novos métodos.
- Permite uma fácil combinação dos métodos existentes.
- Permite uma fácil mudança dos parâmetros para cada método.
- Pode ser chamado repetidamente a partir de uma shell.
- Produz resultados num formato compreensível.

IREP

é uma ferramenta computacional altamente configurável, que pode ser usada para fazer experiências com métodos existentes e facilmente estendida para incorporar novos métodos, medidas de avaliação e conjuntos de dados.

IREP

- Pré-processa os documentos e passa-os para os métodos.
- Usa as mesmas medidas de avaliação com todos os métodos.
- Permite uma fácil incorporação de novos métodos.
- Permite uma fácil combinação dos métodos existentes.
- Permite uma fácil mudança dos parâmetros para cada método.
- Pode ser chamado repetidamente a partir de uma shell.
- Produz resultados num formato compreensível.

IREP

é uma ferramenta computacional altamente configurável, que pode ser usada para fazer experiências com métodos existentes e facilmente estendida para incorporar novos métodos, medidas de avaliação e conjuntos de dados.

IREP

- Pré-processa os documentos e passa-os para os métodos.
- Usa as mesmas medidas de avaliação com todos os métodos.
- Permite uma fácil incorporação de novos métodos.
- Permite uma fácil combinação dos métodos existentes.
- Permite uma fácil mudança dos parâmetros para cada método.
- Pode ser chamado repetidamente a partir de uma shell.
- Produz resultados num formato compreensível.

IREP

é uma ferramenta computacional altamente configurável, que pode ser usada para fazer experiências com métodos existentes e facilmente estendida para incorporar novos métodos, medidas de avaliação e conjuntos de dados.

IREP

- Pré-processa os documentos e passa-os para os métodos.
- Usa as mesmas medidas de avaliação com todos os métodos.
- Permite uma fácil incorporação de novos métodos.
- Permite uma fácil combinação dos métodos existentes.
- Permite uma fácil mudança dos parâmetros para cada método.
- Pode ser chamado repetidamente a partir de uma shell.
- Produz resultados num formato compreensível.

IREP

é uma ferramenta computacional altamente configurável, que pode ser usada para fazer experiências com métodos existentes e facilmente estendida para incorporar novos métodos, medidas de avaliação e conjuntos de dados.

IREP

- Pré-processa os documentos e passa-os para os métodos.
- Usa as mesmas medidas de avaliação com todos os métodos.
- Permite uma fácil incorporação de novos métodos.
- Permite uma fácil combinação dos métodos existentes.
- Permite uma fácil mudança dos parâmetros para cada método.
- Pode ser chamado repetidamente a partir de uma shell.
- Produz resultados num formato compreensível.

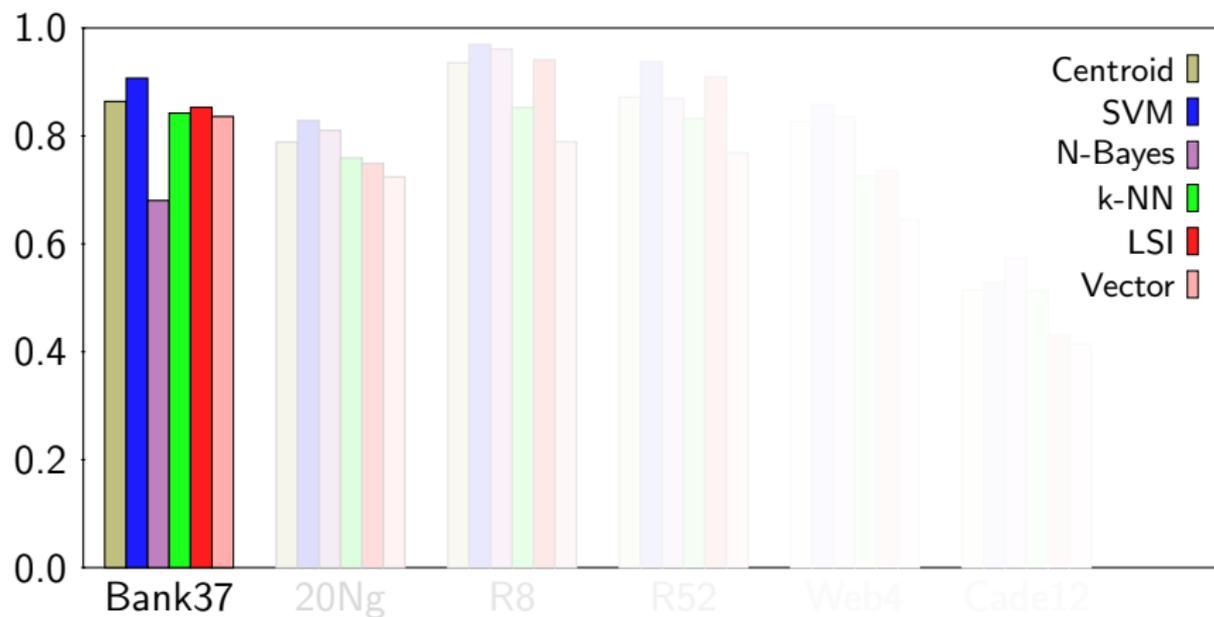
IREP

é uma ferramenta computacional altamente configurável, que pode ser usada para fazer experiências com métodos existentes e facilmente estendida para incorporar novos métodos, medidas de avaliação e conjuntos de dados.

Outline

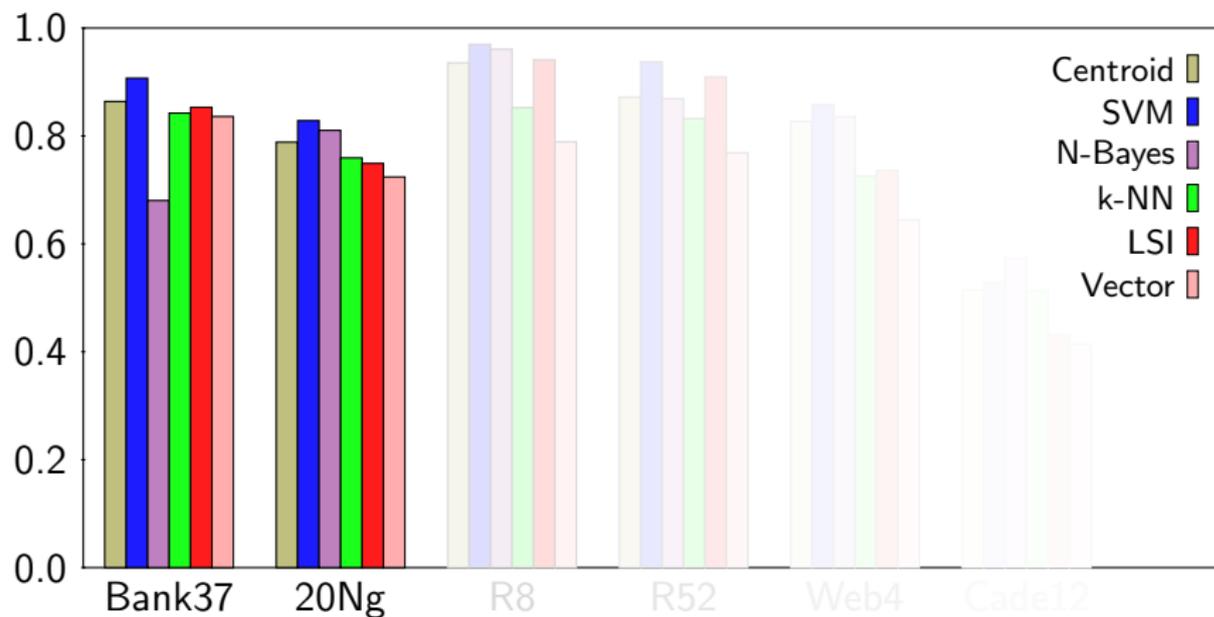
- 1 Introdução
- 2 Ambiente Experimental
- 3 Comparação dos Métodos Existentes**
- 4 Combinações entre Métodos
- 5 Utilização de Documentos Não Etiquetados
- 6 Contribuições

Desempenho dos Métodos Existentes



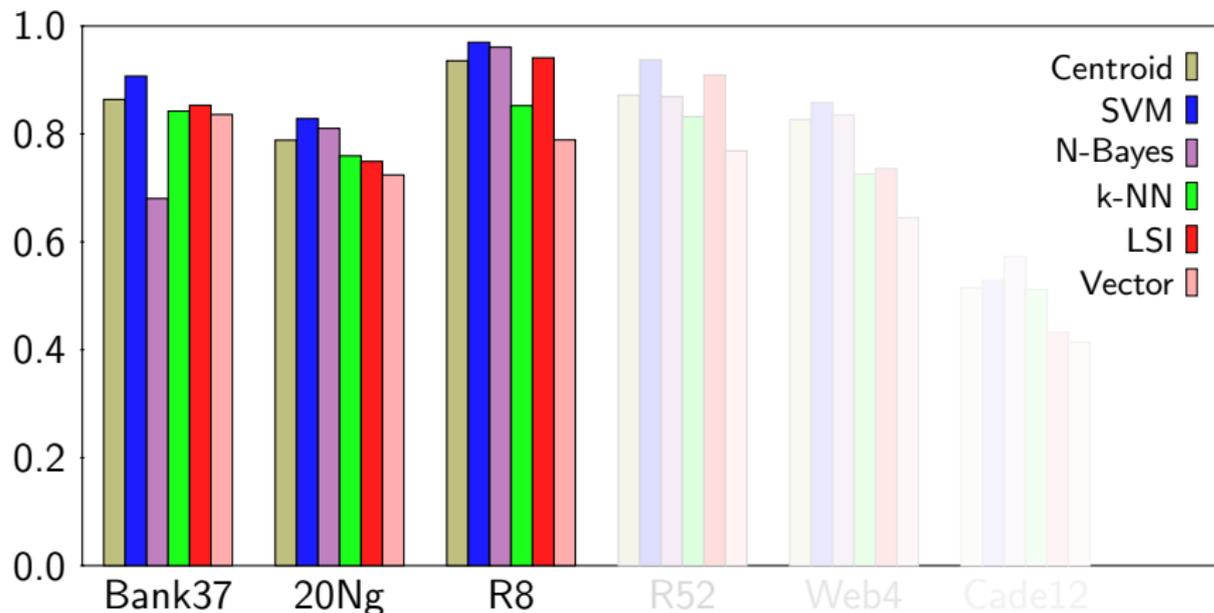
Valores de Accuracy para os seis conjuntos de dados usando cada método de classificação.

Desempenho dos Métodos Existentes



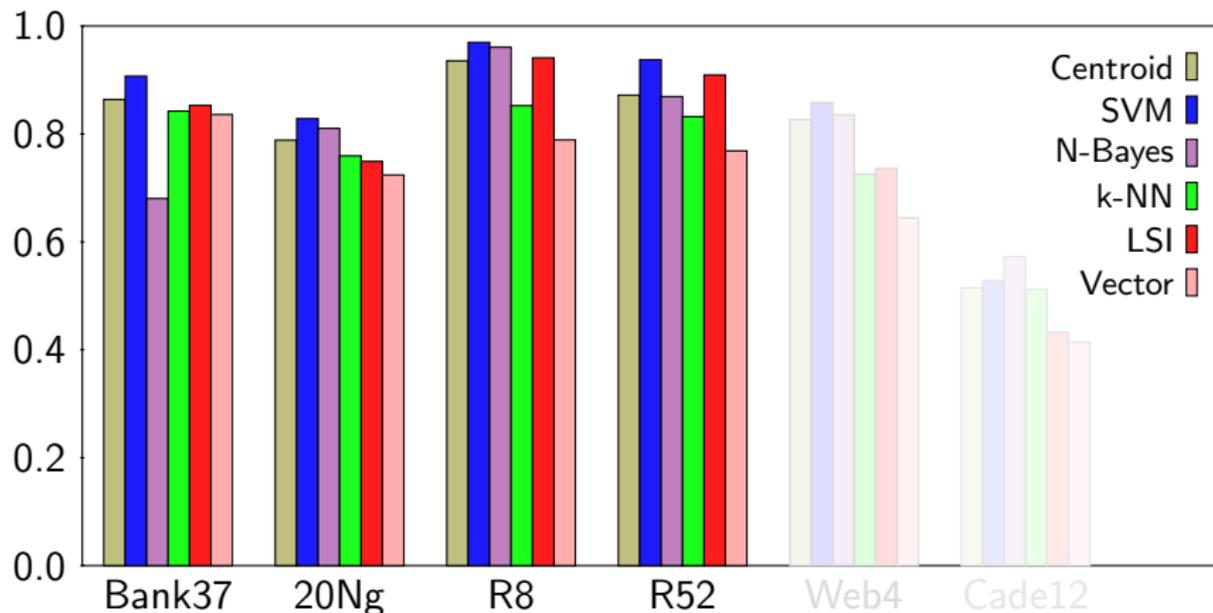
Valores de Accuracy para os seis conjuntos de dados usando cada método de classificação.

Desempenho dos Métodos Existentes



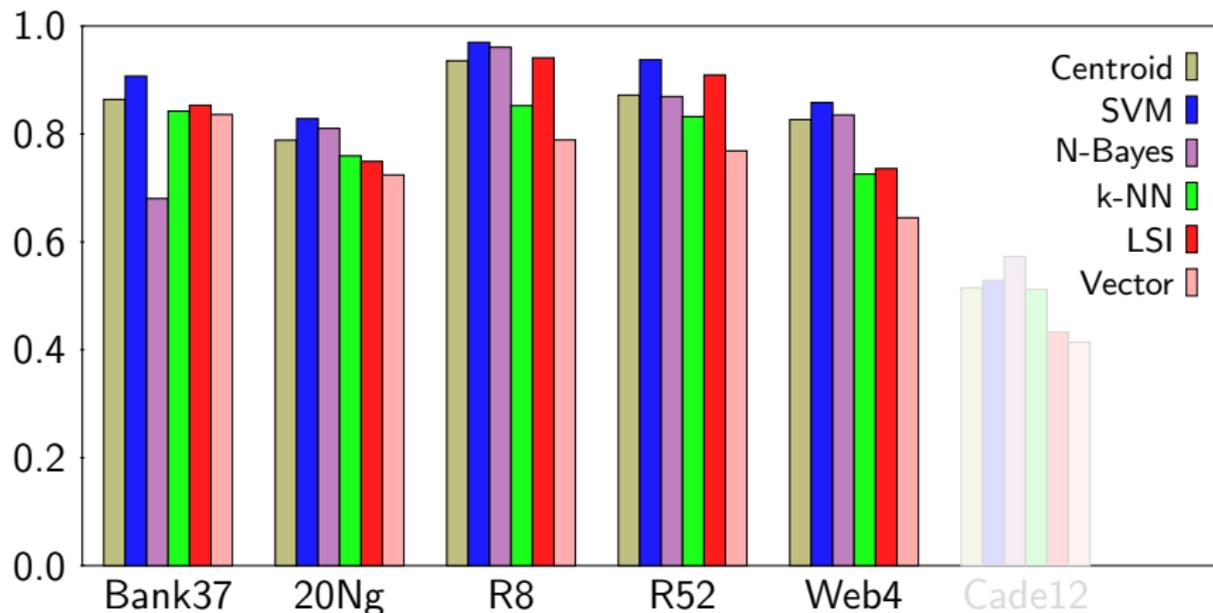
Valores de Accuracy para os seis conjuntos de dados usando cada método de classificação.

Desempenho dos Métodos Existentes



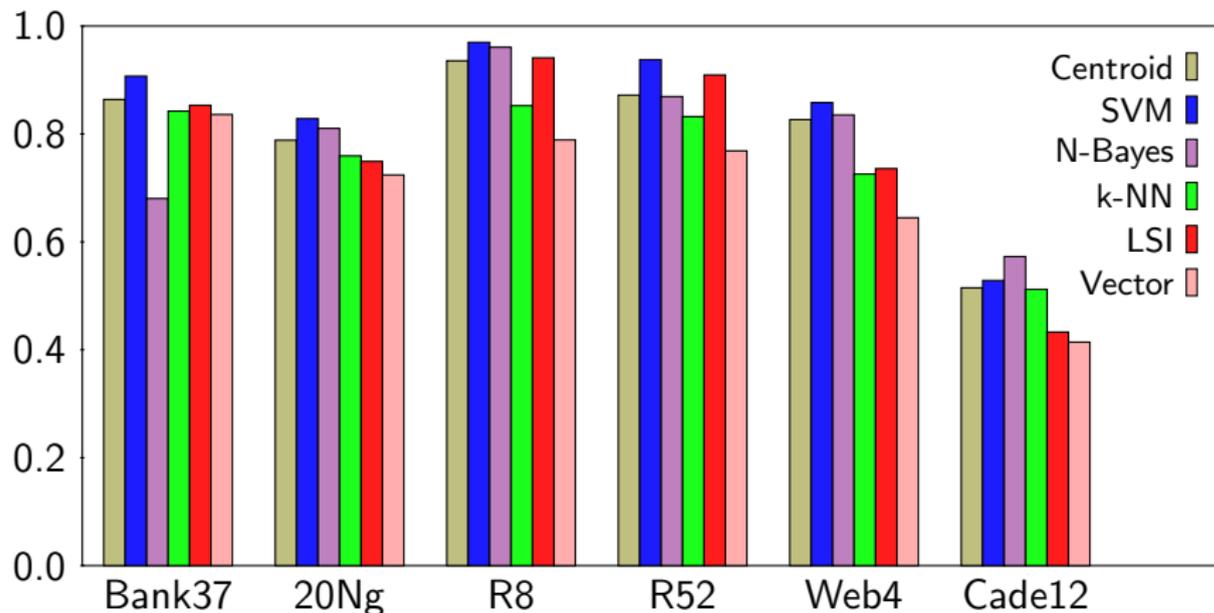
Valores de Accuracy para os seis conjuntos de dados usando cada método de classificação.

Desempenho dos Métodos Existentes



Valores de Accuracy para os seis conjuntos de dados usando cada método de classificação.

Desempenho dos Métodos Existentes



Valores de Accuracy para os seis conjuntos de dados usando cada método de classificação.

Desempenho dos Métodos Existentes

- Dos vários métodos baseados em centroides, C-NormSum é o melhor.
- A qualidade dos resultados obtidos com C-NormSum é quase tão boa como a obtida com SVM, e melhor do que com Vector e k-NN.
- C-NormSum apresenta uma boa relação entre o tempo gasto em treino e teste e a qualidade dos resultados obtidos.
- Usar *tfidf* para calcular os pesos dos termos dos documentos é geralmente melhor do que usar *td*.

Desempenho dos Métodos Existentes

- Dos vários métodos baseados em centroides, C-NormSum é o melhor.
- A qualidade dos resultados obtidos com C-NormSum é quase tão boa como a obtida com SVM, e melhor do que com Vector e k-NN.
- C-NormSum apresenta uma boa relação entre o tempo gasto em treino e teste e a qualidade dos resultados obtidos.
- Usar *tfidf* para calcular os pesos dos termos dos documentos é geralmente melhor do que usar *td*.

Desempenho dos Métodos Existentes

- Dos vários métodos baseados em centroides, C-NormSum é o melhor.
- A qualidade dos resultados obtidos com C-NormSum é quase tão boa como a obtida com SVM, e melhor do que com Vector e k-NN.
- C-NormSum apresenta uma boa relação entre o tempo gasto em treino e teste e a qualidade dos resultados obtidos.
- Usar *tfidf* para calcular os pesos dos termos dos documentos é geralmente melhor do que usar *td*.

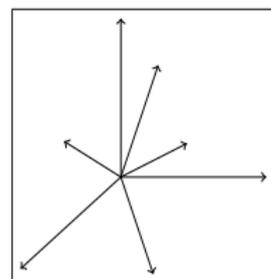
Desempenho dos Métodos Existentes

- Dos vários métodos baseados em centroides, C-NormSum é o melhor.
- A qualidade dos resultados obtidos com C-NormSum é quase tão boa como a obtida com SVM, e melhor do que com Vector e k-NN.
- C-NormSum apresenta uma boa relação entre o tempo gasto em treino e teste e a qualidade dos resultados obtidos.
- Usar *tfidf* para calcular os pesos dos termos dos documentos é geralmente melhor do que usar *td*.

Outline

- 1 Introdução
- 2 Ambiente Experimental
- 3 Comparação dos Métodos Existentes
- 4 Combinções entre Métodos**
- 5 Utilização de Documentos Não Etiquetados
- 6 Contribuições

Métodos de Classificação Existentes



Espaço de termos
 p dimensional

Similaridade do Coseno

→ Vector

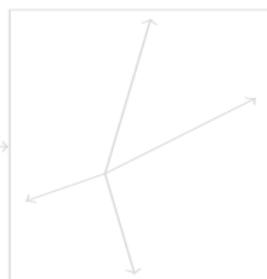
k-NN + Similaridade do Coseno

→ k-NN

Kernel + Estratégia de Votação

→ SVM

SVD

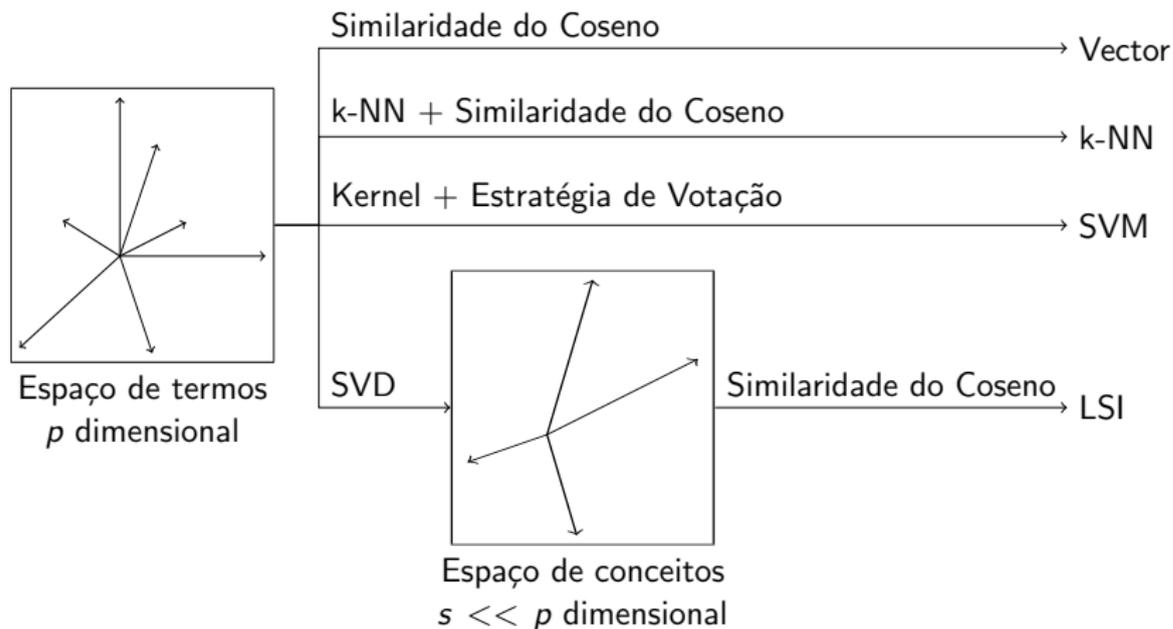


Espaço de conceitos
 $s \ll p$ dimensional

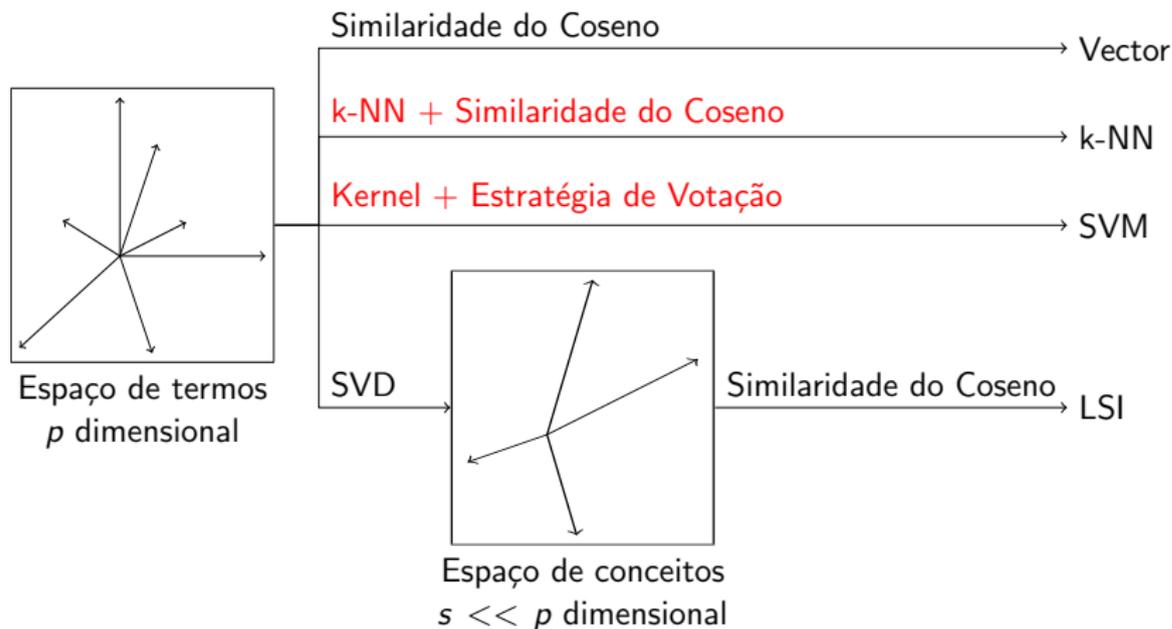
Similaridade do Coseno

→ LSI

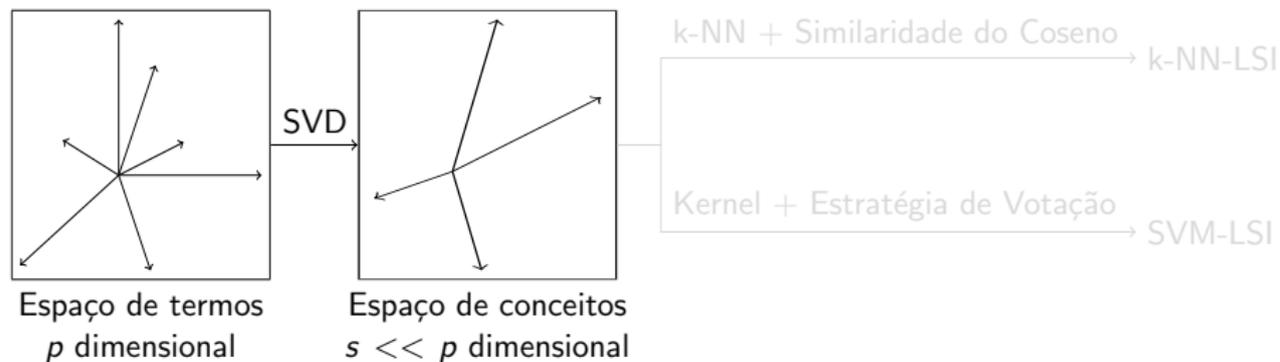
Métodos de Classificação Existentes



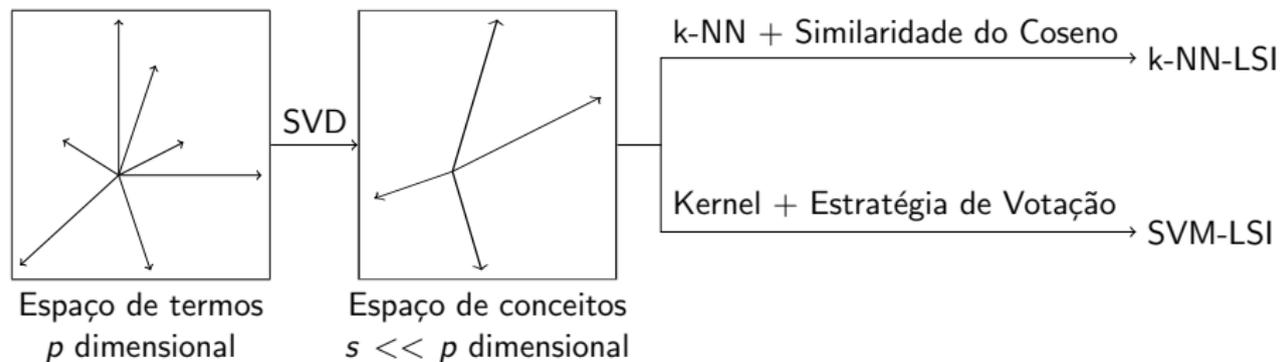
Métodos de Classificação Existentes



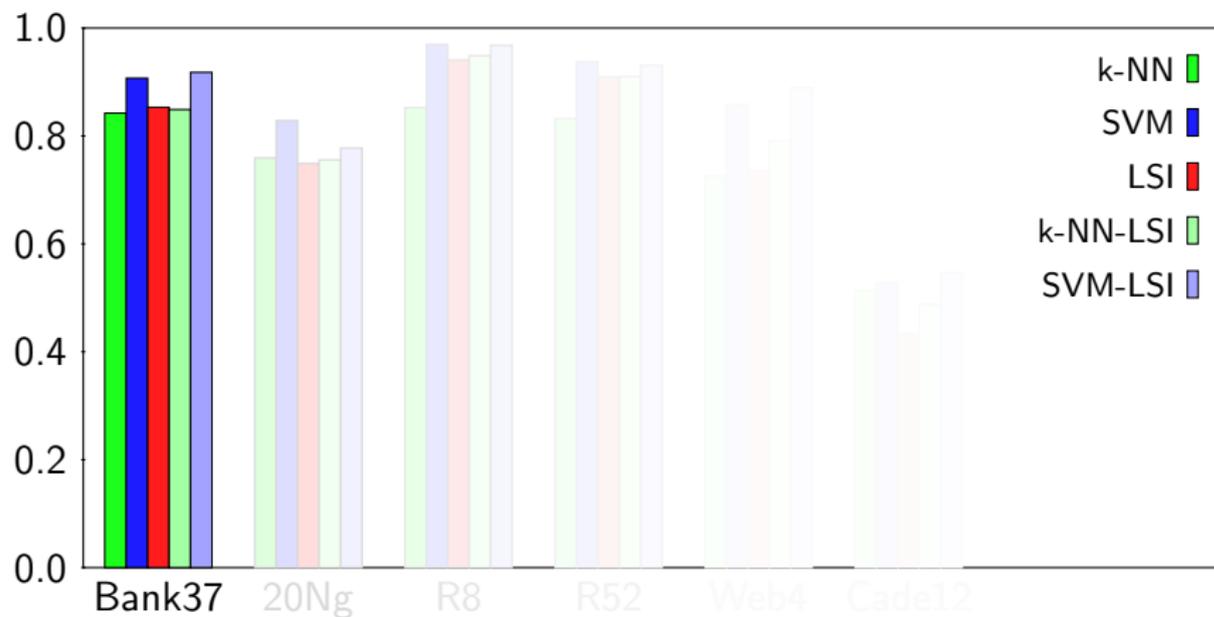
Combinções com LSI



Combinções com LSI

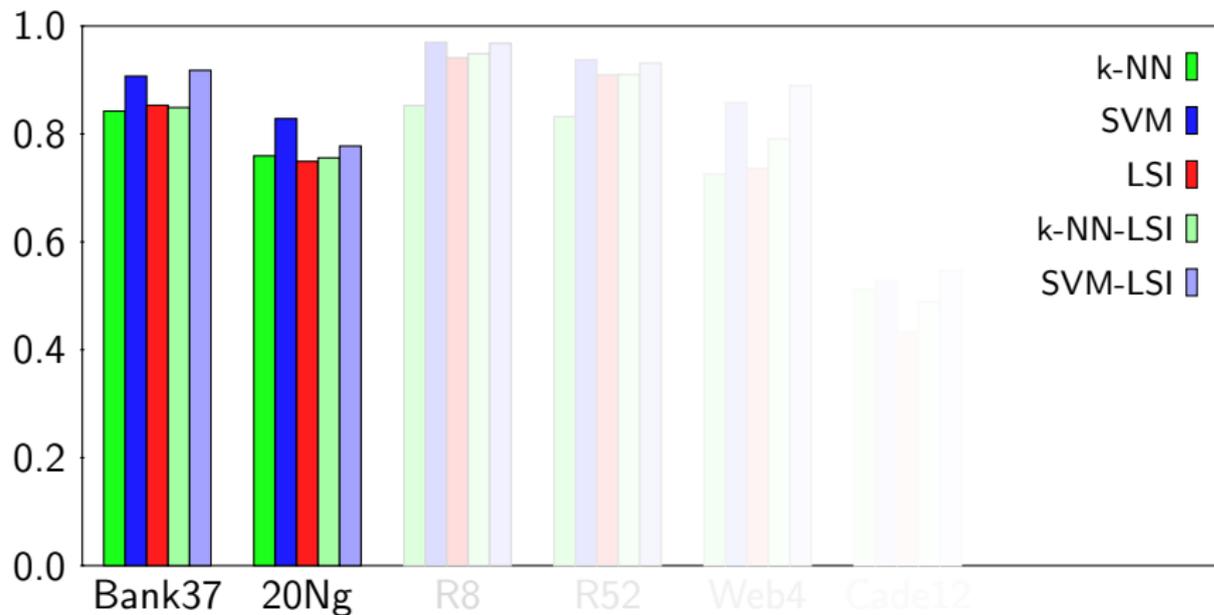


Desempenho das Combinações entre Métodos



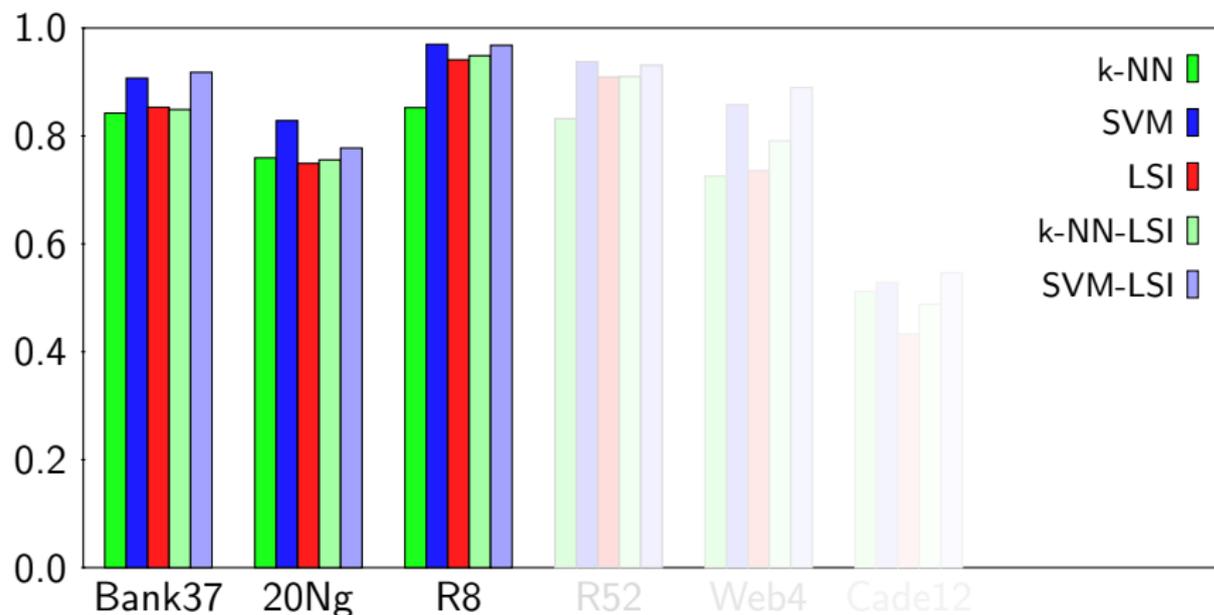
Valores de Accuracy para os seis conjuntos de dados usando cada método.

Desempenho das Combinações entre Métodos



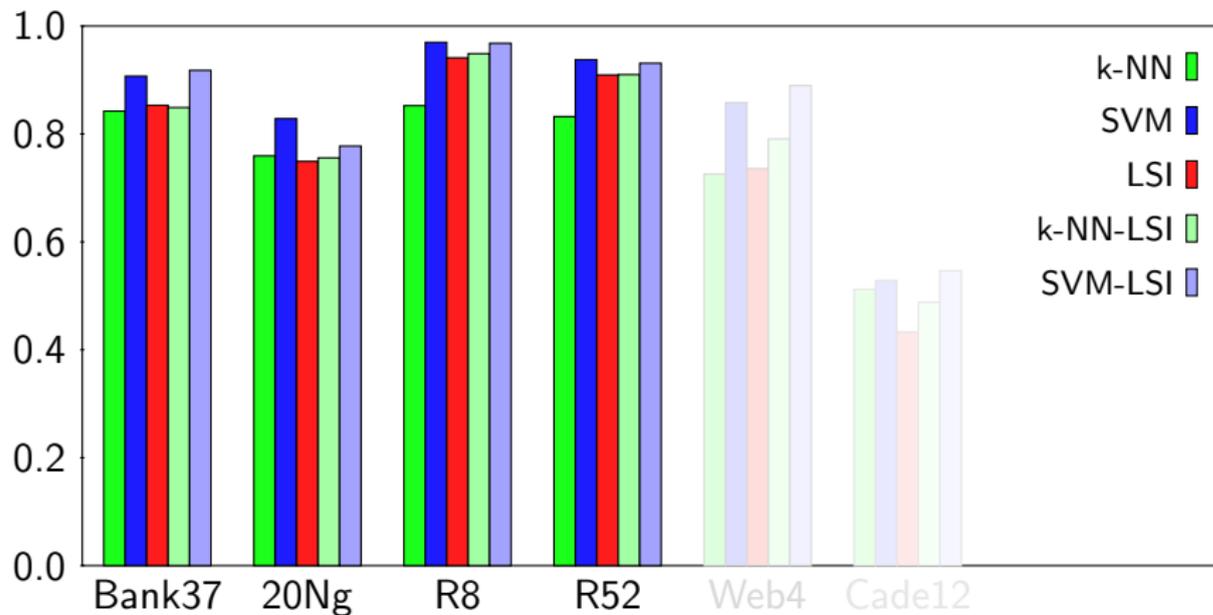
Valores de Accuracy para os seis conjuntos de dados usando cada método.

Desempenho das Combinações entre Métodos



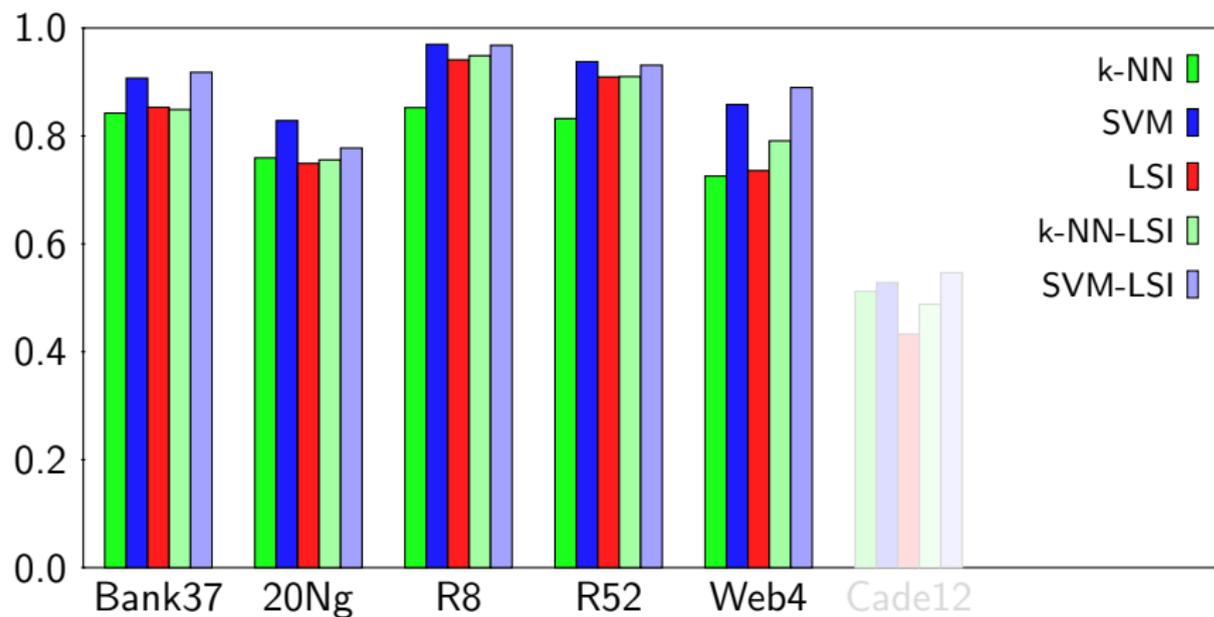
Valores de Accuracy para os seis conjuntos de dados usando cada método.

Desempenho das Combinações entre Métodos



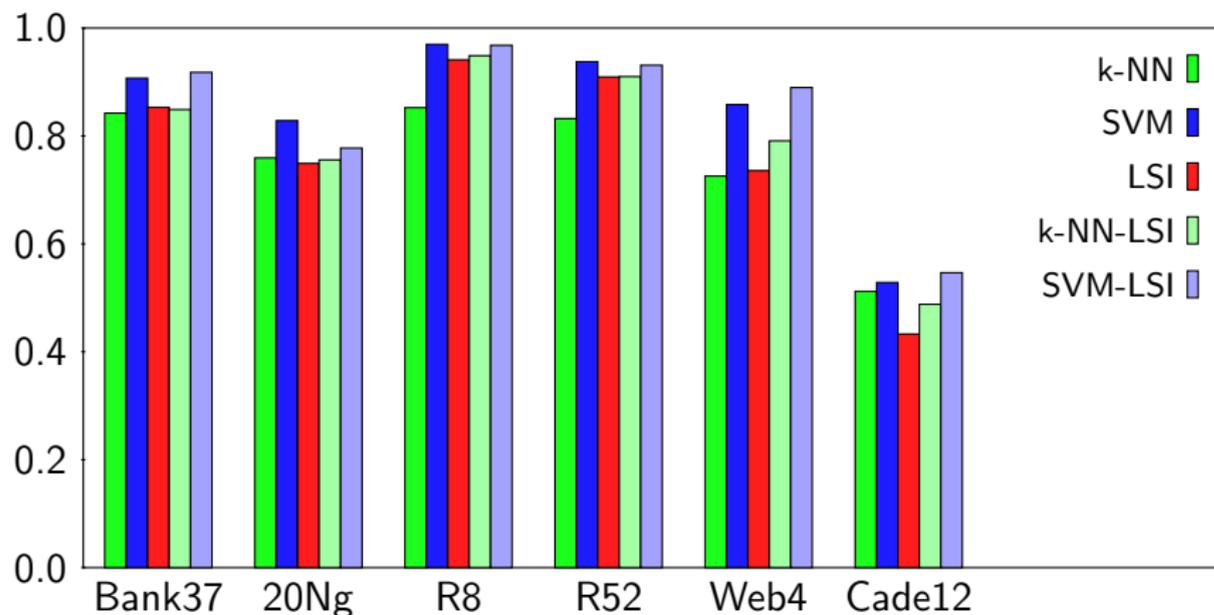
Valores de Accuracy para os seis conjuntos de dados usando cada método.

Desempenho das Combinações entre Métodos



Valores de Accuracy para os seis conjuntos de dados usando cada método.

Desempenho das Combinações entre Métodos



Valores de Accuracy para os seis conjuntos de dados usando cada método.

Desempenho das Combinações entre Métodos

- k-NN-LSI implica uma pequena alteração relativamente a LSI e apresenta melhores resultados do que k-NN e LSI.
- SVM-LSI é melhor do que SVM na média dos vários conjuntos de dados.

Desempenho das Combinações entre Métodos

- k-NN-LSI implica uma pequena alteração relativamente a LSI e apresenta melhores resultados do que k-NN e LSI.
- SVM-LSI é melhor do que SVM na média dos vários conjuntos de dados.

Outline

- 1 Introdução
- 2 Ambiente Experimental
- 3 Comparação dos Métodos Existentes
- 4 Combinações entre Métodos
- 5 Utilização de Documentos Não Etiquetados**
- 6 Contribuições

Utilização de Documentos Não Etiquetados

Quando usar documentos não etiquetados:

- Quando existem pequenas quantidades de documentos etiquetados.
- Quando existem muitos documentos não etiquetados.
- Quando é difícil ou “caro” classificar mais documentos.

Como incorporar a informação dos documentos não etiquetados:

- Usando EM.
- Incrementalmente.

Porquê usar um método baseado em centroides:

- Porque é rápido.
- Porque tem uma boa Accuracy.

Utilização de Documentos Não Etiquetados

Quando usar documentos não etiquetados:

- Quando existem pequenas quantidades de documentos etiquetados.
- Quando existem muitos documentos não etiquetados.
- Quando é difícil ou “caro” classificar mais documentos.

Como incorporar a informação dos documentos não etiquetados:

- Usando EM.
- Incrementalmente.

Porquê usar um método baseado em centroides:

- Porque é rápido.
- Porque tem uma boa Accuracy.

Utilização de Documentos Não Etiquetados

Quando usar documentos não etiquetados:

- Quando existem pequenas quantidades de documentos etiquetados.
- Quando existem muitos documentos não etiquetados.
- Quando é difícil ou “caro” classificar mais documentos.

Como incorporar a informação dos documentos não etiquetados:

- Usando EM.
- Incrementalmente.

Porquê usar um método baseado em centroides:

- Porque é rápido.
- Porque tem uma boa Accuracy.

Incorporar a Informação Usando EM

Se todo o conjunto de dados está disponível desde o início, como numa biblioteca.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Estimação: Para cada documento não etiquetado $d_j \in U$, classificá-lo de acordo com os centroides disponíveis.

Maximização: Para cada classe c_j , actualizar o seu centroide $\vec{c}_{j_{new}}$, considerando os documentos etiquetados e as etiquetas para os documentos não etiquetados obtidas no passo anterior.

Iterar: Até que os centroides não mudem em duas iterações consecutivas.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Usando EM

Se todo o conjunto de dados está disponível desde o início, como numa biblioteca.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Estimação: Para cada documento não etiquetado $d_j \in U$, classificá-lo de acordo com os centroides disponíveis.

Maximização: Para cada classe c_j , actualizar o seu centroide $\vec{c}_{j_{new}}$, considerando os documentos etiquetados e as etiquetas para os documentos não etiquetados obtidas no passo anterior.

Iterar: Até que os centroides não mudem em duas iterações consecutivas.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Usando EM

Se todo o conjunto de dados está disponível desde o início, como numa biblioteca.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Estimação: Para cada documento não etiquetado $d_j \in U$, classificá-lo de acordo com os centroides disponíveis.

Maximização: Para cada classe c_j , actualizar o seu centroide $\vec{c}_{j_{new}}$, considerando os documentos etiquetados e as etiquetas para os documentos não etiquetados obtidas no passo anterior.

Iterar: Até que os centroides não mudem em duas iterações consecutivas.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Usando EM

Se todo o conjunto de dados está disponível desde o início, como numa biblioteca.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Estimação: Para cada documento não etiquetado $d_j \in U$, classificá-lo de acordo com os centroides disponíveis.

Maximização: Para cada classe c_j , actualizar o seu centroide $\vec{c}_{j_{new}}$, considerando os documentos etiquetados e as etiquetas para os documentos não etiquetados obtidas no passo anterior.

Iterar: Até que os centroides não mudem em duas iterações consecutivas.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Usando EM

Se todo o conjunto de dados está disponível desde o início, como numa biblioteca.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Estimação: Para cada documento não etiquetado $d_j \in U$, classificá-lo de acordo com os centroides disponíveis.

Maximização: Para cada classe c_j , actualizar o seu centroide $\vec{c}_{j_{new}}$, considerando os documentos etiquetados e as etiquetas para os documentos não etiquetados obtidas no passo anterior.

Iterar: Até que os centroides não mudem em duas iterações consecutivas.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Usando EM

Se todo o conjunto de dados está disponível desde o início, como numa biblioteca.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Estimação: Para cada documento não etiquetado $d_j \in U$, classificá-lo de acordo com os centroides disponíveis.

Maximização: Para cada classe c_j , actualizar o seu centroide $\vec{c}_{j_{new}}$, considerando os documentos etiquetados e as etiquetas para os documentos não etiquetados obtidas no passo anterior.

Iterar: Até que os centroides não mudem em duas iterações consecutivas.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Usando EM

Se todo o conjunto de dados está disponível desde o início, como numa biblioteca.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Estimação: Para cada documento não etiquetado $d_j \in U$, classificá-lo de acordo com os centroides disponíveis.

Maximização: Para cada classe c_j , actualizar o seu centroide $\vec{c}_{j_{new}}$, considerando os documentos etiquetados e as etiquetas para os documentos não etiquetados obtidas no passo anterior.

Iterar: Até que os centroides não mudem em duas iterações consecutivas.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Incrementalmente

Se o conjunto de dados muda ao longo do tempo, como uma linha de notícias ou na internet.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Iterar: Para cada documento não etiquetado $d_j \in U$:

- Classificar d_j de acordo com a sua semelhança a cada um dos centroides.
- Actualizar os centroides com o novo documento d_j classificado no passo anterior.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Incrementalmente

Se o conjunto de dados muda ao longo do tempo, como uma linha de notícias ou na internet.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Iterar: Para cada documento não etiquetado $d_j \in U$:

- 1. Classificar d_j de acordo com a sua semelhança a cada um dos centroides.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Incrementalmente

Se o conjunto de dados muda ao longo do tempo, como uma linha de notícias ou na internet.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Iterar: Para cada documento não etiquetado $d_j \in U$:

- Classificar d_j de acordo com a sua semelhança a cada um dos centroides.
- Actualizar os centroides com o novo documento d_j classificado no passo anterior.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Incrementalmente

Se o conjunto de dados muda ao longo do tempo, como uma linha de notícias ou na internet.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Iterar: Para cada documento não etiquetado $d_j \in U$:

- Classificar d_j de acordo com a sua semelhança a cada um dos centroides.
- Actualizar os centroides com o novo documento d_j classificado no passo anterior.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Incrementalmente

Se o conjunto de dados muda ao longo do tempo, como uma linha de notícias ou na internet.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Iterar: Para cada documento não etiquetado $d_j \in U$:

- Classificar d_j de acordo com a sua semelhança a cada um dos centroides.
- Actualizar os centroides com o novo documento d_j classificado no passo anterior.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Incrementalmente

Se o conjunto de dados muda ao longo do tempo, como uma linha de notícias ou na internet.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Iterar: Para cada documento não etiquetado $d_j \in U$:

- Classificar d_j de acordo com a sua semelhança a cada um dos centroides.
- Actualizar os centroides com o novo documento d_j classificado no passo anterior.

Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Incorporar a Informação Incrementalmente

Se o conjunto de dados muda ao longo do tempo, como uma linha de notícias ou na internet.

Entradas: Um conjunto de documentos etiquetados, L , e um conjunto de documentos não etiquetados U .

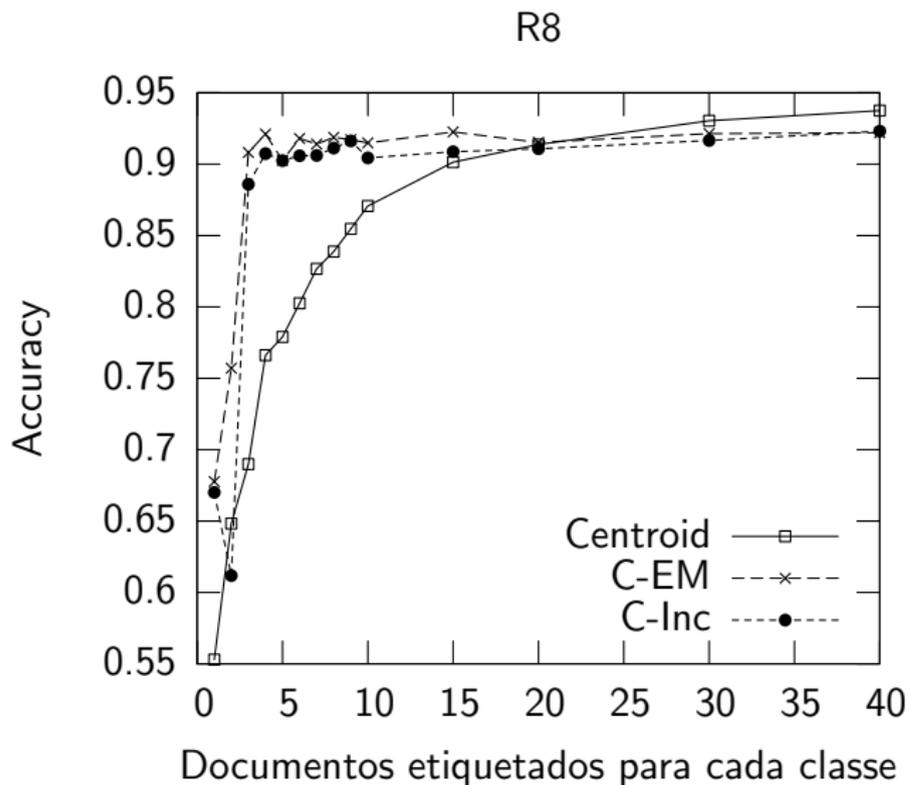
Inicialização: Para cada classe c_j que apareça em L , determinar o centroide da classe \vec{c}_j , usando uma das fórmulas para os centroides e considerando apenas os documentos etiquetados.

Iterar: Para cada documento não etiquetado $d_j \in U$:

- Classificar d_j de acordo com a sua semelhança a cada um dos centroides.
- Actualizar os centroides com o novo documento d_j classificado no passo anterior.

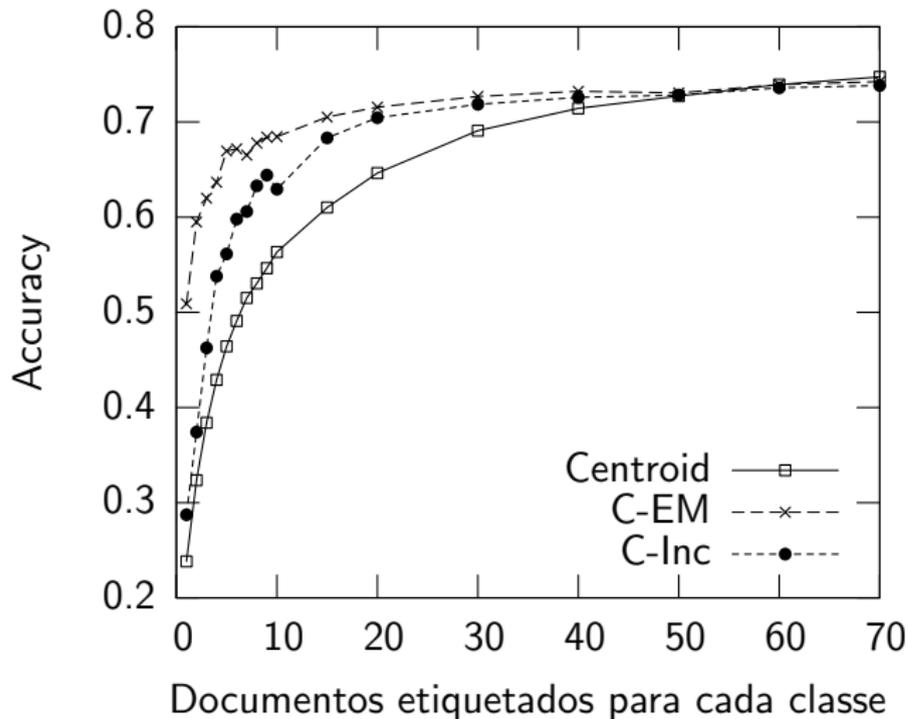
Saídas: Para cada classe c_j , o centroide \vec{c}_j .

Desempenho da Utilização de Docs Não Etiquetados



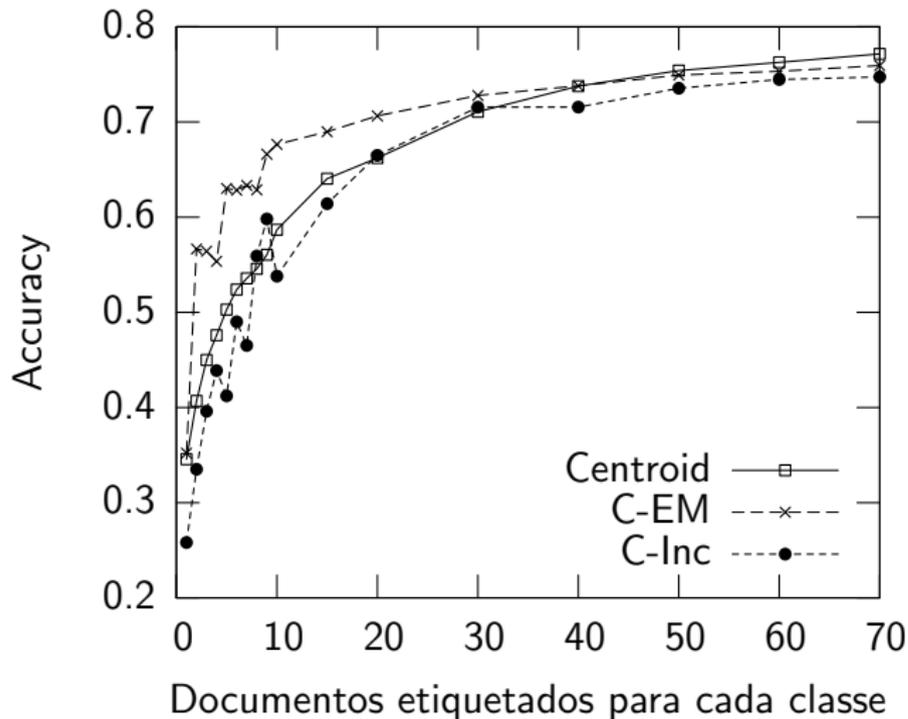
Desempenho da Utilização de Docs Não Etiquetados

20Ng



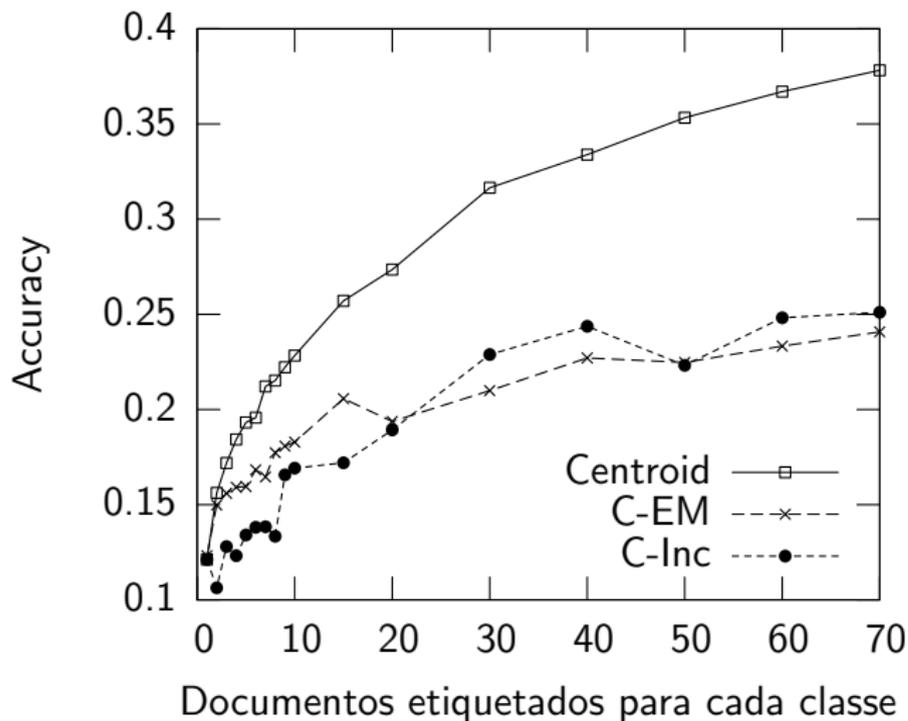
Desempenho da Utilização de Docs Não Etiquetados

Web4



Desempenho da Utilização de Docs Não Etiquetados

Cade12



Desempenho da Utilização de Docs Não Etiquetados

- A incorporação dos documentos não etiquetados usando C-EM é em geral melhor do que incrementalmente, em especial quando há poucos documentos etiquetados para cada classe.
- Se o modelo inicial dos dados for suficientemente preciso, usar documentos não etiquetados melhora os resultados.
- Se o modelo inicial dos dados não for suficientemente preciso, usar documentos não etiquetados piora os resultados.

Desempenho da Utilização de Docs Não Etiquetados

- A incorporação dos documentos não etiquetados usando C-EM é em geral melhor do que incrementalmente, em especial quando há poucos documentos etiquetados para cada classe.
- Se o modelo inicial dos dados for suficientemente preciso, usar documentos não etiquetados melhora os resultados.
- Se o modelo inicial dos dados não for suficientemente preciso, usar documentos não etiquetados piora os resultados.

Desempenho da Utilização de Docs Não Etiquetados

- A incorporação dos documentos não etiquetados usando C-EM é em geral melhor do que incrementalmente, em especial quando há poucos documentos etiquetados para cada classe.
- Se o modelo inicial dos dados for suficientemente preciso, usar documentos não etiquetados melhora os resultados.
- Se o modelo inicial dos dados não for suficientemente preciso, usar documentos não etiquetados piora os resultados.

Outline

- 1 Introdução
- 2 Ambiente Experimental
- 3 Comparação dos Métodos Existentes
- 4 Combinações entre Métodos
- 5 Utilização de Documentos Não Etiquetados
- 6 Contribuições**

Contribuições Principais

- Desenvolvimento de uma ferramenta altamente configurável para fazer experiências com vários métodos de classificação.
- Comparação exaustiva de 13 métodos de classificação usando 4 colecções de dados standard e uma criada para este trabalho.
- Proposta de dois novos métodos de classificação que correspondem à combinação de métodos existentes.
- Proposta de dois algoritmos para incorporar documentos não etiquetados num método baseado em centroides usando EM e incrementalmente.
- Estudo empírico de quando é que se devem usar documentos não etiquetados com um método baseado em centroides.

Contribuições Principais

- Desenvolvimento de uma ferramenta altamente configurável para fazer experiências com vários métodos de classificação.
- Comparação exaustiva de 13 métodos de classificação usando 4 colecções de dados standard e uma criada para este trabalho.
- Proposta de dois novos métodos de classificação que correspondem à combinação de métodos existentes.
- Proposta de dois algoritmos para incorporar documentos não etiquetados num método baseado em centroides usando EM e incrementalmente.
- Estudo empírico de quando é que se devem usar documentos não etiquetados com um método baseado em centroides.

Contribuições Principais

- Desenvolvimento de uma ferramenta altamente configurável para fazer experiências com vários métodos de classificação.
- Comparação exaustiva de 13 métodos de classificação usando 4 colecções de dados standard e uma criada para este trabalho.
- Proposta de dois novos métodos de classificação que correspondem à combinação de métodos existentes.
- Proposta de dois algoritmos para incorporar documentos não etiquetados num método baseado em centroides usando EM e incrementalmente.
- Estudo empírico de quando é que se devem usar documentos não etiquetados com um método baseado em centroides.

Contribuições Principais

- Desenvolvimento de uma ferramenta altamente configurável para fazer experiências com vários métodos de classificação.
- Comparação exaustiva de 13 métodos de classificação usando 4 colecções de dados standard e uma criada para este trabalho.
- Proposta de dois novos métodos de classificação que correspondem à combinação de métodos existentes.
- Proposta de dois algoritmos para incorporar documentos não etiquetados num método baseado em centroides usando EM e incrementalmente.
- Estudo empírico de quando é que se devem usar documentos não etiquetados com um método baseado em centroides.

Contribuições Principais

- Desenvolvimento de uma ferramenta altamente configurável para fazer experiências com vários métodos de classificação.
- Comparação exaustiva de 13 métodos de classificação usando 4 colecções de dados standard e uma criada para este trabalho.
- Proposta de dois novos métodos de classificação que correspondem à combinação de métodos existentes.
- Proposta de dois algoritmos para incorporar documentos não etiquetados num método baseado em centroides usando EM e incrementalmente.
- Estudo empírico de quando é que se devem usar documentos não etiquetados com um método baseado em centroides.

Obrigada.