Universidade Federal do Rio Grande - FURG
Especialização em Aplicações para Web
Trabalho de Conclusão de Curso

Revisão de ferramentas WEB para Mineração de Dados

Ronaldo Canofre
canofre@inf.ufsm.br

Agenda

- Objetivos e Motivação
- Abordagem Geral
- Revisão Bibliográfica
- Resultados Obtidos
- Considerações Finais

Motivação e Objetivos

- Objetivos
  - Pesquisa e análise de ferramentas web
- Motivação
  - Uso de ferramentas executáveis
  - Portabilidade e aplicações web
  - Facilidade de utilização

Abordagem geral

- Volume de dados gerados
- Quem gera esse volume?
- Eles são importantes?
- Como tirar proveito?

Revisão Bibliográfica 1/2

- Knowledge Discovery in Databases
  - Diversidade de dados
  - Padrões válidos
  - Auxílio em processos e problemas
- Tratamento dos dados
  - Pré e pós processamento

Revisão Bibliográfica 2/2

- **Mineração de Dados**
  - ➡ Dados brutos → Informação útil
- **Técnicas de Mineração**
  - ➡ Agrupamento (Kmeans)
  - ➡ Associação (Apriori)

  } Tarefas descritivas

  - ➡ Classificação (Knn)
  - ➡ Regressão (SVN)

  } Tarefas preditivas

Resultados Obtidos 1/7

- Aplicações em PHP
  - Implementações individuais
  - Pequena variedade
  - KNN, Apriori e Kmeans
- Avaliação inicial

Dados de entrada	Como informar	• Alteração no código
	Tipo de entrada	• Vetor • Arquivo CSV • Métodos das classes
Apresentação da saída		• HTML simples ou não tratado
Configuração do algoritmo		• Edição de variáveis • Métodos das classes
Licença		• GPL – LGPL – MIT – NI

Resultados Obtidos 2/7

- Algoritmo K-means
  - Técnica de Agrupamento
  - K : grupos definidos pelo usuário
  - Definição em torno de uma centroide
  - Condição de parada
  - K <= Instâncias da base

Resultados Obtidos 4/7

- K-means: base de dados selecionadas
  - UCI Machine Learning Repository

Base	Atributos	Tipo	Instancias	Área
seeds	7	Real	210	Agronegócio
Wholesale customers	8	Inteiro	440	Negócios
Turkiye Student Evaluation	33	Inteiro	5820	Educação

Resultados Obtidos 3/7

- K-means: características

Item / Aplicação	Kmeans01	Kmeans02	Kmeans03
K > Instâncias	Permite		
Condição de parada	Estabilidade dos centroides		
Centróide	Randômica		
	Kmeans++	(intervalo dos atributos)	-
Atributos	Sem restrição	Sem restrição	2
Clusters vazios	SIM	NÃO	SIM
Tempo	Base 03	Base 03	Não avaliado
OBS		*set_time_limit / init_set*	Base exemplo

Resultados Obtidos 4/7

- *K-means: resultados obtidos*

Linha	Base de dados	Instâncias por Implementação		
		Kmeans01	Kmeans02	Weka
1	01	[42,64,16,**46**,**42**]	[49, 49, 55, **42**, 15]	[14, **46**, 50, 48, 52]
2	02	[113,63,23,6,235]	[235, 113, 63, 6, 23]	[239, 98, 8, 36, 59]
3	03	[1342, 901, **1140**, 1261, 1176]	[1344,1175, 902, **1140**, 1259]	[760, 731, 1971, 1622, 736]

Resultados Obtidos 4/7

- Algoritmo K-nn
  - Técnica de classificação
  - Medida de proximidade
  - Cálculo de distância entre pontos
  - K : vizinhos mais próximos

Resultados Obtidos 4/7

- K-nn: base de dados selecionadas
  - UCI Machine Learning Repository

Base	Atributos	Tipo	Instancias	Área
blood transfusion	4	Inteiro	748	Negócios
data banknote authentication	4	Real	1372	Computação
íris	4	Real	150	Biologia

Resultados Obtidos 6/7

- **K-nn: características**

Item / Aplicação	Knn01	Knn02	Knn03
Cálculo da Distância	Euclidiana	Euclidiana	Euclidiana Manhattann
Atributos	2	4	Sem restrição
Nova Instância	Sem restrição	Presente na base	Sem restrição
Valor de K	Fixo (4)	Editável	Não define
OBS	Base exemplo	set_time_limit / init_set	Peso nos atributos

Resultados Obtidos 7/7

- K-nn: resultados obtidos

Linha	Base de Dados	Atributos da instância de teste	Classificação	Knn02	Knn03	Weka
1	01	(1,24,6000,77)	Não	Não	Não	Sim
2		(4,6,1500,22)	Sim	Não	Sim	Sim
3	02	(-0.4928,3.060, -1.8356,-2834)	Verdadeira	Verdadeira	Verdadeira	Verdadeira
4		(0.6636,-0.0455, -0.1879,0.2345)	Verdadeira	Verdadeira	Falsa	Verdadeira
5	03	(4.9,2.0,4.0,1.7)	Virginica	Versicolor	Virginica	Versicolor
6		(5.9,3.2,4.8,1.8)	Versicolor	Virginica	Virginica	Virginica
Acurácia média				50%	66%	50%

Conclusão e Trabalhos Futuros

- Conclusões
  - Ferramentas completas em PHP
  - Aplicações individuais
  - Possibilidade de utilização
- Trabalhos Futuros
  - Implementação de ferramentas
  - Comparação de desempenho

- ***Dúvidas e/ou considerações?***