Pontifícia Universidade Católica de Minas Gerais
Instituto de Ciências Exatas e Informática
Plan. de Capacidade e Avaliação de Sistemas Computacionais | Prof. Felipe Soares

Trabalho Prático 1
Caracterização de Cargas de Trabalho

Neste trabalho prático, vamos treinar os conceitos relacionados à caracterização de cargas de trabalho estudados até o momento. Para isso utilizaremos duas abordagens diferentes para a geração do modelo de carga de trabalho:

- regressões e
- agrupamentos.

A abordagem aplicando técnicas de regressão deve ser realizada utilizando **algum editor de planilha ou programação**. Já para a aplicação da técnica de agrupamento, utilizaremos o *software* **Weka 3.8** que é bastante utilizado para mineração de dados e aprendizagem de máquina (*machine learning*).

Iremos caracterizar uma base de dados composta por processos submetidos a um supercomputador.

O arquivo correspondente a essa carga de trabalho encontra-se no Canvas (logSupercomputador.txt). Esse arquivo é composto por três colunas:

- A primeira coluna apresenta o instante de tempo em que o processo foi submetido ao supercomputador (coluna *SubTime*);
- A segunda indica o número de processadores utilizados para a execução do processo (coluna *NumProcs*);
- A terceira coluna representa o tempo de execução desse processo (*ExeTime*), em segundos.

Parte 1 - Caracterização de Cargas de Trabalho utilizando Distribuições de Probabilidade:

Inicialmente, você deve abrir o arquivo "logSupercomputador.txt" no **editor de planilha**, que fará a conversão dos dados desse arquivo, alocando cada campo (*SubTime*, *NumProcs* e *ExeTime*) na coluna correspondente.

Em seguida, você deve criar uma quarta coluna que armazenará o intervalo de tempo entre chegadas de processos sucessivos (*InterArrivalTime*). Você deve calcular o tempo entre chegadas de cada dupla de processos sucessivos e armazenar o valor calculado nessa quarta coluna.

O próximo passo é calcular a média, o desvio padrão e o coeficiente de variação do tempo de execução dos processos, do número de processadores e do tempo entre chegadas de processos sucessivos.

Pergunta 1 - A média representa adequadamente essa base de dados? Justifique sua resposta.

Agora, você deve gerar um histograma para o tempo de execução dos processos. Para isso, você deve seguir os seguintes passos: inicialmente, você deve definir os blocos de dados em que os valores obtidos para essa métrica (tempo de execução dos processos) serão categorizados. Por exemplo, esses blocos poderiam ser: i) 0 a 100; ii) 100 a 200; iii) 200 a 500; iv) 500 a 1.000; v) 1.000 a 10.000; etc. Não necessariamente esses são os melhores blocos de dados para essa métrica. Esses valores foram indicados apenas para exemplificação.

Então deve-se criar uma coluna para cada um dos blocos de dados definidos.

Em seguida, os valores da métrica devem ser categorizados no bloco de dados correspondente, de acordo com sua característica. Por exemplo, o primeiro processo indicado no *log* analisado gastou 4.743 segundos executando. Assim, ele deve ser categorizado no bloco de dados v) 1.000 a 10.000 (utilize funções lógicas para realizar essa categorização).

O próximo passo é calcular a média, o desvio padrão e o coeficiente de variação do tempo de execução dos processos, para cada um dos blocos de dados definidos. Caso o coeficiente de variação de um determinado bloco apresente valor superior a 0,5, esse bloco de dados deve ser subdividido em blocos menores, pois esse valor para o coeficiente de variação indica que esse bloco não apresenta dados homogêneos.

Pergunta 2 - A média obtida para os valores de cada bloco de dados representa mais adequadamente a base de dados analisada? Compare os valores obtidos neste passo com a média obtida anteriormente, para a base de dados inteira.

Em seguida, compute o número de processos em cada bloco de dados e a correspondente porcentagem de processos em cada bloco em relação ao número total de processos.

Por fim, gere um histograma (gráfico de barras) para o tempo de execução dos processos utilizando a porcentagem calculada.

Adicione ao histograma uma linha de tendência. Para isso, clique com o botão direito do *mouse* nas barras do histograma gerado e selecione a opção “Adicionar linha de tendência...”. Ao adicioná-la, teste todas as distribuições de probabilidade disponíveis e solicite a exibição do coeficiente de correlação (R^2). Observe que, quanto mais próximo de 1 for o coeficiente de correlação, mais adequada é a curva aos dados. A distribuição de probabilidade que melhor descreve os dados apresentados nada mais é do que a própria curva que se encaixou melhor nos dados do histograma. Qual é a distribuição de probabilidade que melhor representa os tempos de execução dos processos apresentados no *log* analisado?

Repita esses passos e gere também histogramas para o número de processadores e o tempo entre chegadas de processos sucessivos.

Parte 2 - Caracterização de Cargas de Trabalho aplicando Técnicas de Agrupamento (*Clustering*):

Você deve abrir o *software* **Weka 3.8** e clicar no botão “*Explorer*”. Na tela que se abrir, você deve clicar no botão “*Open file...*” para abrir o arquivo “logSupercomputador.dat.arff” também disponibilizado no Canvas juntamente com este enunciado. Esse arquivo é composto por quatro colunas (*SubTime*, *NumProcs*, *ExeTime* e *InterArrivalTime*).

Remova o atributo *SubTime* da base de dados, pois não faz sentido caracterizar bases de dados por meio desse atributo. Ele já foi utilizado para a determinação dos valores do atributo *InterArrivalTime*, que será utilizado na caracterização dessa carga de trabalho.

Continuando o pré-processamento dos dados, é necessário tratar valores ausentes. Para isso, clique no botão "*Choose*" do painel "*Filter*", encontre e selecione o filtro "*ReplaceMissingValues*". Em seguida, clique no botão "*Apply*", também presente no painel "*Filter*", para que o filtro selecionado seja efetivamente aplicado à base de dados. Lembre-se de selecionar o valor "*No class*", no *combobox* que aparece logo acima do histograma, antes de aplicar o filtro selecionado. Opcionalmente, existem vários outros filtros que podem ser aplicados a essa base de dados.

Na aba "*Cluster*", clique no botão "*Choose*" do painel "*Clusterer*" e selecione a técnica de agrupamento "*SimpleKMeans*". Com um duplo clique no nome da técnica, você poderá alterar alguns dos parâmetros utilizados durante sua execução. Altere o valor do *combobox* "*displayStdDevs*" para "*True*" para que seja exibido, ao final da execução do algoritmo de agrupamento, os desvios padrão encontrados para os centróides dos *clusters*. Varie também o número de *clusters* (modificando o valor do campo "*numClusters*") de 2 a 12. Para cada uma dessas configurações para o número de *clusters* encontrados, execute a técnica de agrupamento, clicando no botão "*Start*" e avalie os agrupamentos gerados e exibidos no painel "*Clusterer output*". Para cada execução do algoritmo de agrupamento, guarde a média e o desvio padrão dos centróides de cada *cluster* gerado e calcule também seu coeficiente de variação (CV). Além disso, guarde também a frequência de cada grupo gerado (número de integrantes do *cluster*).

Pergunta 3 - Qual conjunto de *clusters* representa melhor essa base de dados? Justifique sua resposta.

Para cada execução do algoritmo de agrupamento, você pode visualizar os *clusters* gerados clicando na aba "*Visualize*". Para entender melhor os gráficos apresentados, clique em um dos gráficos e, em seguida, na nova tela que se abrir, fixe o eixo "Y" e varie o "X". Além disso, aumente um pouco o valor do campo "*Jitter*" para visualizar melhor a diversidade de pontos nos *clusters*.

O que deve ser entregue:

Este trabalho poderá ser realizado em grupos de até 6 pessoas.

Deve ser entregue um relatório (**usando LaTeX - utilizar "SBC Conferences Template"**) como resultado da realização deste trabalho prático (máximo 10 páginas). Anexa a esse relatório deve ser apresentada a planilha eletrônica onde os cálculos da primeira parte deste trabalho prático foram realizados. Esse relatório deve conter o seguinte conteúdo:

- Introdução sobre o trabalho realizado, explicando o contexto do trabalho, a base de dados, os cálculos realizados e o resumo do que foi feito para chegar aos resultados.

  Estrutura:
  - Resumo
  - Introdução
  - Metodologia
    - Materiais
    - Método
  - Resultados
  - Conclusão

- Média, desvio padrão e coeficiente de variação (CV) do tempo de execução dos processos, do número de processadores e do tempo entre chegadas de processos sucessivos. Essas médias representam adequadamente essa base de dados? Justifique sua resposta.

- Histograma para o tempo de execução dos processos. Esse histograma deve apresentar também a linha de tendência mais adequada para descrevê-lo. Indique também o coeficiente de correlação (R^2) entre o histograma e a linha de tendência escolhida. Qual é a distribuição de probabilidade que melhor representa os tempos de execução dos processos apresentados no *log* analisado? Justifique sua resposta.

  Indique também a média, o desvio padrão e o coeficiente de variação do tempo de execução dos processos, para cada um dos blocos de dados definidos no histograma. A média obtida para os valores de cada bloco de dados representa mais adequadamente a base de dados analisada? Compare os valores obtidos nesse passo com a média obtida anteriormente, para a base de dados inteira. Justifique sua resposta.

  Indique também o número de processos em cada bloco de dados e a correspondente porcentagem de processos em cada bloco em relação ao número total de processos.

- As mesmas análises e resultados do item anterior para o histograma do número de processadores e do tempo entre chegadas de processos sucessivos.

- Média, desvio padrão e coeficiente de variação (CV) dos centróides de cada *cluster* gerado em cada execução do algoritmo de agrupamento (cada vez que o valor do campo "*numClusters*" é alterado temos uma nova execução do algoritmo de agrupamento). Além disso, indique também a frequência de cada grupo gerado (número de integrantes do *cluster*). Qual conjunto de *clusters* representa melhor essa base de dados? Justifique sua resposta.