SAD - Sistema de Avaliação de Desempenho

Especificação de Requisitos

Versão 1

Histórico de Revisões

Versão	Data	Autor	Conteúdo
1	14/04/2014	Daniel Marques	Primeira versão do documento.
2	25/07/2014	Daniel Marques	Revisão do documento.

Documentos de Referência

Título	Versão	Data	Autor
Manual do Weka	3.7.10	31/07/2013	Bouckaert, et. al.

Introdução

Este é o documento de especificação de requisitos do software SAD (Sistema de Avaliação de Desempenho). Este software será desenvolvido durante a disciplina Projeto Final de Programação da PUC-Rio. Este documento deve ser usado como referência para projetar e desenvolver o sistema.

Escopo

O aprendizado supervisionado é um segmento importante da área de aprendizado de máquina. O objetivo deste tipo de técnica é que o algoritmo aprenda uma tarefa a partir de uma lista de exemplos. Essa tarefa pode ser do tipo classificação ou regressão. A primeira consiste em inferir a classe de uma instância. A segunda significa prever o valor de um número real referente à instância.

Existem diversas técnicas e algoritmos de aprendizado de máquina que podem ser utilizados para atingir um nível de acerto satisfatório. Entretanto, empregar essas práticas exige conhecimentos específicos e muitas vezes esforço um grande de desenvolvimento de software. A utilização de valores de referência é uma prática comum para justificar e avaliar o aprendizado de máquina. É possível aplicar métodos simples para chegar a um valor de referência. Por exemplo, em uma tarefa de regressão podemos utilizar a média dos n últimos valores.

Este projeto de programação consiste em um sistema de avaliação de desempenho de métodos. O sistema SAD deve receber como entrada um conjunto de dados. Ele será usado para aplicar métodos simples de predição aos dados e gerará estatísticas de acerto. Estas estatísticas podem revelar a pertinência da aplicação de técnicas de aprendizado de máquina ao problema. Se for o caso, estas também podem ser usadas como valores de referência para os resultados do aprendizado de máquina.

As estratégias de predição empregadas pelo sistema são específicas para conjuntos de dados cujas instâncias tem uma ordem de precedência, e.g. por data ou horário. Existem diversos problemas importantes cujos dados tem essas características, e. g. investimento no mercado financeiro.

Figura 1 - Diagrama simplificado do sistema SAD

Definições, Acrônimos e Abreviaturas

Weka - Coleção de algoritmos de aprendizado de máquina para mineração de dados.

Descrição Geral

Funções do Aplicativo

- Importação dos dados
- Aplicação de estratégias aos dados a fim de fazer a predição de um dos atributos
- Apresentação de estatísticas de resultado
- Exportação das estatísticas ou do resultado bruto para um arquivo
- Executar múltiplos experimentos sequenciais. Cada experimento pode definir um conjunto de dados, estratégia de predição e estatísticas de saída.

Características do Usuário

Os usuários do SAD são estudantes, professores ou pesquisadores da área de aprendizado de máquina. Portanto são usuários especializados em computação e com grande domínio da área.

Postergar Requisitos

- Interfaces gráficas de usuário

- Outros formatos de entrada de dados como o CSV e o XML
- Integrações com o Weka (por exemplo para aplicar os filtros do Weka aos dados ou para usar o conversor de csv para arff)
- Estratégia de predição que faz a combinação linear das colunas de um *dataset* para prever o resultado
- Possibilidade de definição de estratégias pelo usuário

Requisitos Específicos

Interfaces de Usuário

O sistema funcionará em linha de comando. As informações que o usuário deve inserir são especificadas abaixo. Além disso é apresentada a descrição do arquivo de instruções. Este deve especificar todas as opções do sistema que o usuário pretende utilizar em um dado experimento. Por fim, é apresentado o formato do arquivo de dados, que contem os dados doe um experimento. Para maiores detalhes sobre estes arquivos vide manual do usuário.

Informações inseridas pelo usuário

Informação	Descrição
Arquivo de Instruções	Endereço do arquivo que contém as opções dos experimentos a serem realizados.
Opção de debug	Opção que faz com que o sistema exiba mensagens de erro mais detalhadas.
Opção silenciosa	Opção que faz com que o sistema não exiba mensagens.

Arquivo de Instruções

O arquivo de instruções deve ser escrito em um formato pré-definido. Neste arquivo é possível definir diversos experimentos subsequentes a serem executados pelo sistema. Cada experimento pode definir opções diferentes (indicadas a seguir). As opções são listadas abaixo:

Opção	Descrição
Arquivo com os dados	O sistema utiliza o arquivo cujo caminho (com nome do arquivo) foi especificado neste campo.

Método de predição	O usuário pode escolher um método de predição por experimento. Cada método de predição pode ter diversas sub-opções. Quando um determinado método for escolhido, obrigatoriamente o usuário deve inserir as informações necessárias (opções do método).
Opções do métodos de predição	Cada método pode ter opções diferentes.

Formato sugerido do arquivo de Instruções

O arquivo de instruções deve obedecer ao formato a seguir:

```
Entry{
        data = "<Caminho para o arquivo de dados 1>",
        strategy = "<Nome da estratégia de predição 1>",
        option = “<valor opção>“,
}
Entry{
        data = "<Caminho para o arquivo de dados 2>",
        strategy = "<Nome da estratégia de predição 2>",
        option = “<valor opção>“,
}

…

Entry{
        data = "<Caminho para o arquivo de dados k>",
        strategy = "<Nome da estratégia de predição k>",
        option = “<valor opção>“,
}
```

Formato sugerido do arquivo de dados

O arquivo deve obedecer ao formato a seguir:

```
Entry{"<Valor 1>"}
Entry{"<Valor 2>"}
...
Entry{"<Valor k>"}
```

Requisitos Não Funcionais

ID	Título	Descrição
RF001	Tipos de dados	Os tipos de dados compreendidos pelo sistema devem ser Numérico (inteiro ou real) e nominal (lista de valores para classificação).
RF002	Ordem de precedência das instâncias	As instâncias que compões os dados de entrada para o sistema devem ter uma ordenação específica, e.g. por data ou horário. As estratégias de predição devem assumir que na instância *n+1* os dados das *n* instâncias anteriores estão disponíveis.

Requisitos Funcionais

ID	Título	Descrição
RF001	Métodos de predição	• Média aritmética do atributo (em todo o conjunto de treino). Baseada nos valores reais e não nos previstos. Somente para tarefas de regressão • Média aritmética dos últimos n valores. Requer uma ordem de precedência dos registros. É necessário que a entrada esteja ordenada de forma correta. Deve ser baseado nos valores reais dos atributos e não nos previstos. Somente para tarefas de regressão. • Repetir o valor anterior • Classe ou valor mais numeroso (em todo o conjunto de treino). Baseado nos valores reais e não nos previstos. • Classe ou valor mais numeroso entre os n últimos. Requer uma ordem de precedência dos registros. É necessário que a entrada esteja ordenada de forma correta. Deve ser baseado

		nos valores reais dos atributos e não nos previstos.
RF002	Estatísticas de saída	Estatísticas análogas às do Weka (quando aplicáveis): ● Número de instâncias corretamente classificadas ● Número de instâncias incorretamente classificadas ● Número total de instâncias ● *Mean absolute error*, onde: $$MAE = \frac{\sum_{i=1}^{n} \lvert fi - yi \rvert}{n}$$ ● *Root mean squared error*, onde: $$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (f - y)^2}{n}}$$
RF003	Arquivos de entrada de dados	O arquivo com os dados deve conter os valores de saída para a instâncias. Cada instância tem um único valor de saída que pode ser numérico ou nominal. Cada linha deve conter apenas um valor. A ordem dos valores deve representar a ordem real das instâncias. Vide seção Formato do Arquivo de Dados.
RF004	Múltiplos experimentos	Deve ser possível especificar múltiplos experimentos sequenciais a para o sistema. Este executa os experimentos um apos o outro sem intervenção do usuário.
RF005	Apresentação dos resultados	Deve ser possível ver os resultados consolidadeos na tela tela (linha de comando) ou armazenado-los em arquivo.
RF006	Exportação dos resultados brutos	Deve ser possível exportar os resultados (valor real, valore predito) de todas as instâncias para um arquivo.

RF007	Verbosidade	Deve ser possível escolher o nível de verbosidade da execução do programa. O usuário deve poder escolher desde mensagens detalhadas até nenhuma.