PUCRS - Escola Politécnica
Inteligência Artificial - Prof. Sílvia

Question & Answer

Trabalho 2 sobre Aprendizagem de Máquina

1. **Objetivo do trabalho:** Experimentar técnicas de classificação tendo como aplicação um sistema de perguntas e respostas em Lingua Portuguesa. Faz parte do escopo do trabalho construir o corpus, processá-lo, classificá-lo e analisar os resultados obtidos.

2. **Corpus:** é um nome dado a uma coleção de documentos (textos, sentenças, etc). O tema desse sistema de perguntas e resposta é o algoritmo Random Forest (Floresta Aleatória). O corpus será construído pela turma seguindo as etapas:

   (a) Levantamento de material bibliográfico sobre o algoritmo Random Forest.

   (b) Definição do formato do arquivo que conterá as perguntas.

   (c) Elaboração de ao menos 20 perguntas com resposta sobre o algoritmo, envolvendo: conceito, propósito, paradigma, tarefa, funcionamento, parâmetros, etc.

   (d) Para cada pergunta, construir ao menos 5 variantes bem definidas (variações linguisticas, paráfrases, ...) das perguntas.

   (e) Análise e união das perguntas feitas por todos os grupos para formar um único corpus.

3. **Etapas do trabalho -** Abaixo as macro-etapas do trabalho:

   (a) Pré-processamento do corpus: definição das classes, normalização morfológica, anotação linguística, extração dos termos, seleção dos termos mais relevantes, estruturação.

   (b) Categorização e análise dos resultados.

   (c) Escrita de um relatório sobre o trabalho realizado.

4. **Descrição da Etapa de Pré-processamento:** O pré-processamento é a etapa mais custosa de qualquer tarefa em Aprendizagem de Máquina (AM). A preparação dos textos é ainda um pouco mais custosa, pois textos são dados desestruturados. Para estruturá-los, ou seja, colocá-los em um formato que viabilize o processamento dos mesmos por um algoritmo de AM, vai incluir:

   (a) **Definição das classes**: Organização das perguntas em classes.

(b) **Normalização Morfológica e Anotação Linguística:** Pode ser feita pelo parser VISL[1] (Figura 1) , Cogroo [2], ou ainda pelo anotador Tree Tagger[3]

(c) O objetivo da **normalização morfológica** é colocar os termos (strings) na mesma "forma". Por exemplo, verbos quando aparecem nos textos estão flexionados: "estudou", "estudaram" e "estuda". A meta, nesse caso, é transformar essas ocorrências em uma forma normal, que no caso dos verbos é o infinito: "estudar". Existem vários tipos de normalização morfológica. A que vamos usar, chama-se "Lematização". A lematização leva os termos para o lema, que, no caso de substantivos, corresponde à palavra no masculino, singular; e, no caso de verbos, no infinitivo.

   i. O objetivo da **anotação linguistica** é prover informações sobre o texto para que possamos, em um segundo momento, escolher os termos mais relevantes de um texto. Anotar um texto é colocar tags (rótulos) em seus termos. Essas tags podem ser morfossintáticas e, até mesmo, semânticas. Nesse trabalho, vamos usar apenas a anotação de Part-Of-Speech (POS). As tags[4] de POS indicam as classes gramaticais das palavras: verbo (V), substantivo (N, PROP) adjetivo (ADJ) e advérbio (ADV).

(d) **Extração do Termos:** Após o processo de anotação, já podemos retirar do texto termos que podem ser úteis na etapa de estruturação. Para cada classe (de pergunta) do corpus, construa uma lista com os tokens mais relevantes. Preserve, em sua implementação, a informação sobre o texto do qual esses termos foram extraídos. Faz parte do seu trabalho identificar as classes gramaticais mais relevantes para este trabalho.

(e) **Seleção dos Termos mais relevantes:** É nessa etapa que precisamos escolher os termos mais relevantes (redução de dimensionalidade) visando a representação dos textos (estruturação). Usando as listas criadas na etapa anterior, crie uma lista geral de termos (sem repetição). Para cada termo dessa lista, contabilize a frequência desse termo no corpus. A seguir, selecione os k primeiros termos mais frequentes. Faz parte do seu trabalho definir o valor de k mais adequado. O resultado dessa seleção é uma lista de termos, conhecida como Bag-of-Words (BoW).

(f) **Estruturação:** Nessa fase, vamos usar uma representação vetorial para estruturar os textos. A BoW funcionará com os atributos (campos) do texto. A representação vetorial mais simples é a binária, que

[1] Disponivel para consulta on-line em https://visl.sdu.dk/visl/pt/parsing/automatic/parse.php

[2] Disponivel para download em http://cogroo.sourceforge.net/(Disponivel em http://cogroo.sourceforge.net/

[3] Disponivel para download em http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ ou para consulta on-line em https://gramatica.usc.es/~gamallo/php/tagger/TaggerPT.php

[4] As tags exemplificadas são do VISL. Variam conforme o anotador usado.

indica se um termo da BoW está ou não no texto. Por exemplo, supondo que a BoW é formada pelo vetor [P1,P2,P3,P4] e existem as seguintes perguntas já pré-processados (cada pergunta com sua lista de termos): T1 = {P4, P5, P6} sobre Conceito; T2 = {P1,P3, P7} sobre Tarefa; T3 = {P8,P4, P5} sobre Conceito; T4 = {P1,P8, P9} sobre Tarefa e T5 = {P1,P4, P9} sobre Conceito. O arquivo arff para o Weka, (ferramenta que usaremos nas etapas seguintes), com os textos estruturados, ficaria assim:

```
@relation Arquivo
@attribute P1 integer
@attribute P2 integer
@attribute P3 integer
@attribute P4 integer
@attribute classe {Conceito,Tarefa}
@data
0, 0, 0, 1, Conceito
1, 0, 1, 0, Tarefa
0, 0, 0, 1, Conceito
1, 0, 0, 0, Tarefa
1, 0, 0, 1, Conceito
```

5. **Descrição da Etapa de Categorização:** A categorização (ou classificação) de textos é o processo de automaticamente atribuir uma ou mais categorias predefinidas a documentos textuais. Nessa etapa testaremos algoritmos de classificação sobre o corpus pré-processado. O objetivo dessa etapa é testar com diferentes algoritmos (ao menos 3, incluindo Redes Bayesianas e MultiLayer Perceptron) e comentar aquele que melhor classificou os textos. Usar, nessa etapa, 80% dos textos (de cada classe) para treino e os restantes para teste. Nos 80% utilize validação cruzada para definir o modelo.

6. **Descrição do Relatório:** Deve ser entregue um relatório descrevendo: o objetivo do trabalho, a construção do corpus, pré-processamento (descrever o pré-processamento realizado, configurações da BoW); para cada tarefa, mencionar os algoritmos testados e detalhar a análise dos resultados (tomar como base as medidas usuais), bem como incluir comentários sobre o desenvolvimento do mesmo e a sua conclusão.

7. **Desenvolvimento e Entrega:** O trabalho poderá ser desenvolvido ao longo das aulas práticas da disciplina. Entrega Corpus: 10/06/2019 em um fórum próprio para isso no moodle. Entrega final: 27/06/2019 também via moodle (fontes e versão final do relatório)..

8. **Forma de avaliação**:

   (a) Etapa de construção do corpus (participação na tarefa e qualidade das perguntas e respostas elaboradas): 2,0 pts (etapa fundamental, todos devem participar)

   (b) Etapas de pre-processamento (da normalização à estruturação): 4,0 pts (dependente da avaliação presencial do dia 27/06/2019)

(c) Etapa de Aprendizagem - tarefa de Categorização (Weka): 2,0 pts (dependente da avaliação presencial do dia 27/06/2019)

(d) Análise dos resultados (relatório): 2,0 pts

(e) Pontos Extras:

- Implementação do ciclo completo do sistema de perguntas e respostas. Permitindo a entrada de novas perguntas, execução do algoritmo para identificação da classe da perguna e apresentação da resposta correspondente: 2,0 pts