Mineração de Dados Complexos

INF-0611 - Dados Complexos e Recuperação de Informação

Trabalho Final da Disciplina

Trabalho em Dupla

Contexto e Base de Dados

Tal como você aprenderá ao longo das demais disciplinas do curso, um dos desafios da área de Mineração de Dados Complexos é transformar um conjunto de dados denso e pouco compreensível (à primeira vista) em uma história que faça sentido em um contexto específico. Assim como em narrativas, uma sequência cronológica de fatos é importante para o entendimento do enredo, assim como é importante relatar como os dados foram coletados, manuseados e representados até que as descobertas (informação) possam se tornar viáveis.

Dentro do contexto da biologia, uma tarefa de grande importância é o reconhecimento de espécies de árvores. Dada uma coleção de imagens de árvores, podem ser usadas técnicas de visão computacional, processamento de imagens e aprendizado de máquinas, para desenvolver modelos que ajudem na classificação e/ou recuperação automática de imagens de árvores. Em muitos casos, imagens de árvores inteiras podem ser difíceis de tratar, pois requerem um grande trabalho de pre-processamento para poder delimitar o contorno da árvore na imagem, já que uma espécie pode estar junta com outras no seu habitat natural. No entanto, imagens das folhas das árvores, em muitos casos, caracterizam de melhor forma a uma determinada espécie de árvore e demandam um custo menor de pre-procesamento.

Tem-se criado várias bases públicas de imagens de folhas de árvores (com a respectiva anotação da espécie) para ajudar na projeção e desenvolvimento de algoritmos para classificação e/ou recuperação automática de espécies de árvores. A base de imagens originais que será usada no projeto final pode ser encontrada em:

http://www.cvl.isy.liu.se/en/research/datasets/swedish-leaf/

A base consta de 1125 imagens de folhas de árvores, que pertencem a 11 espécies diferentes de árvores (75 imagens por cada espécie) que podem ser encontrados na Suécia. A Figura 1 mostra uma imagem de exemplo de cada espécie.

Figura 1. Exemplo de uma folha de cada espécie.

Uma abordagem amplamente utilizada para representar as imagens de folhas é através de descritores de contornos. A Figura 2 mostra dois exemplos da saída de um desses descritores para duas espécies diferentes.

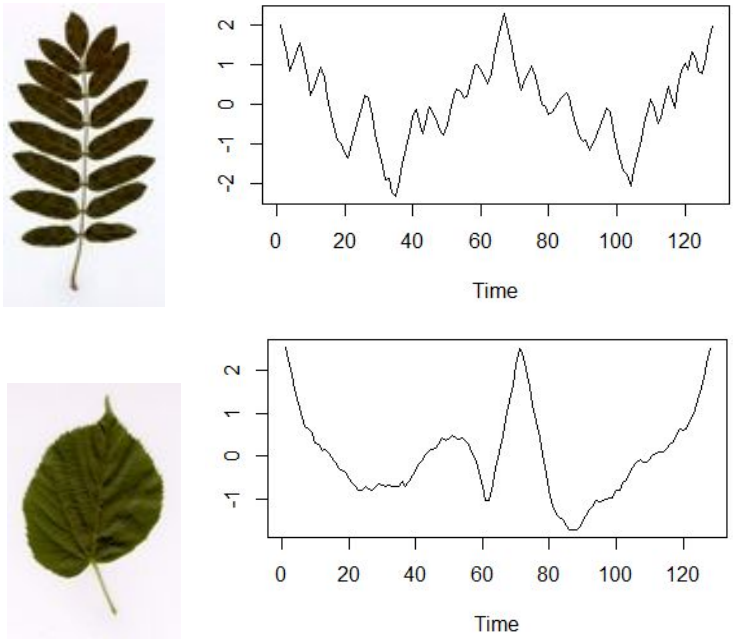

Figura 2. Exemplo de um descritor de contorno para duas espécies diferentes..

A saída de um descritor de contorno pode ser tratada como uma série temporal, já que também existe uma relação temporal (de ordem) implícita entre as posições do vetor de características gerado pelos descritores de contorno. Por tanto, as técnicas utilizadas para tratar com séries temporais, também podem ser aplicadas sobre os vetores de características gerados por tais descritores. Por exemplo, a Figura 3 mostra o *Recurrence Plot* gerado para as mesmas séries mostradas acima.

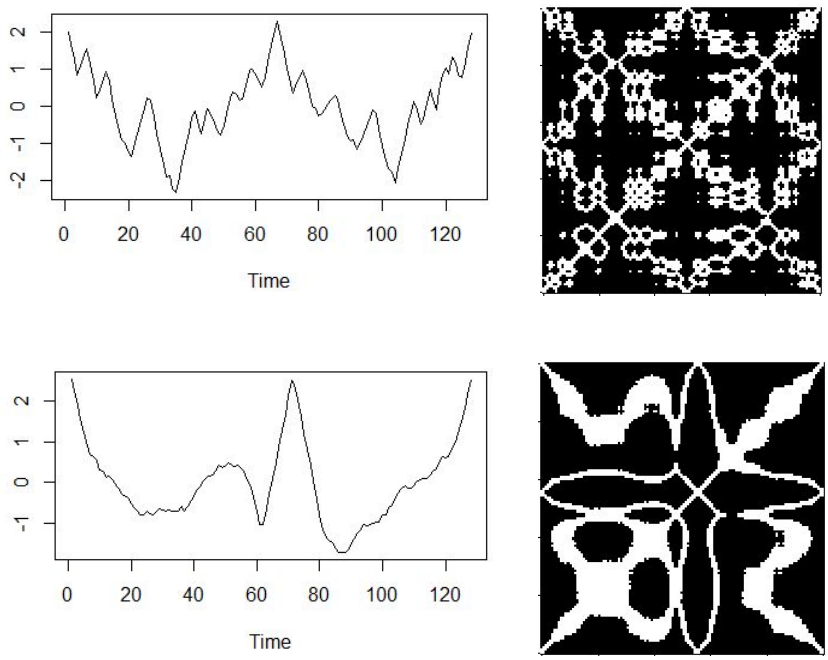

Figura 3. *Recurrence Plot* de duas especies de arvores diferentes.

A tarefa (história) que propomos a você está inserida no contexto de Recuperação da Informação. A ideia é que você aplique o que foi visto em sala de aula para criar sua solução que possibilite a **Busca e Recuperação de Informação** relevante da base.

Enunciado

Objetivo

Projetar e implementar uma solução de busca de informação em que, dada como entrada uma série que representa a saída de um descritor de contorno referente a uma imagem (chamaremos de *query*), recuperar as imagens cujas séries são relevantes para aquela *query*. Nesse contexto, será considerado relevante aquelas imagens recuperadas que pertençam à mesma espécie que a *query*.

Sua solução deverá recuperar as 100 (cem) séries mais relevantes para a consulta estipulada. O escopo deste trabalho é na camada de processamento de um sistema de recuperação de informação.

Roteiro

Na Figura 4 ilustramos o fluxo de trabalho esperado para a resolução deste trabalho.

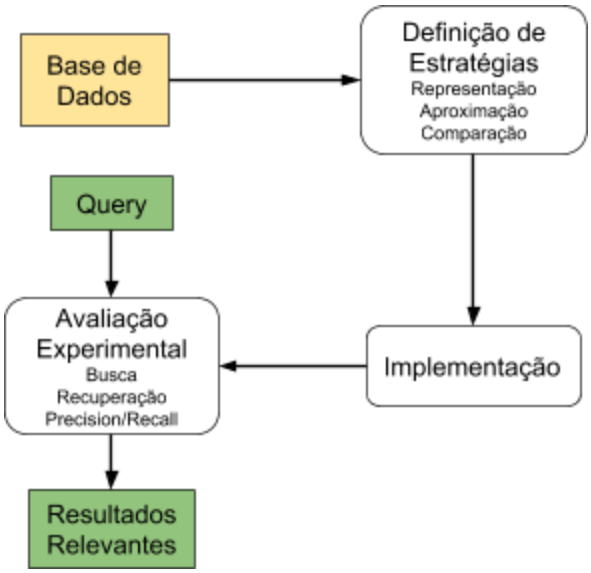

Figura 4. Fluxo de trabalho.

O conjunto original de 1125 imagens foi dividido em dois sub-conjuntos de imagens: 500 imagens para a base de dados (onde serão realizadas as buscas), e 625 imagens para usá-las como consultas. Serão fornecidos no moodle os vetores de características gerados pelo descritor de contorno para cada imagem da base de dados e as consultas, os quais tem 128 dimensões. O arquivo "**SwedishLeaf_TRAIN.csv**" contém os vetores de características das imagens da base de dados, e o arquivo "**SwedishLeaf_TEST.csv**" contém os vetores de características das imagens de consulta. Em ambos casos, cada linha representa uma imagem, sendo que a primeira coluna contém a espécie da árvore (um inteiro entre 1 e 15), e as 128 colunas restantes, o vetor de características gerado pelo descritor de contorno.

Na etapa de **Definição de Estratégias** de representação, aproximação e comparação dos dados, a fim de pavimentar a busca e a sua futura recuperação. Nesta etapa, você deverá tomar algumas decisões, tais como:

- Como representarei os meus dados?

- Qual será o meu descritor? Usarei um descritor específico para extrair informações de interesse (vetor de características)? Que atributos representam melhor meus dados?
- Adotarei quais métricas de similaridade/distância para comparar duas séries de dados?
- Como modelar uma série temporal? Como modelar a série de dados de outra forma? Com qual resolução dos dados trabalharei?
- Preciso fazer alguma transformação em meus dados? Precisarei usar o SAX para reduzir a dimensionalidade da série e obter uma representação simbólica? Dynamic Time Warping (DTW) funcionaria melhor? É necessário normalização? Usar o *Recurrence Plot* para transformar as series em imagens e aplicar CBIR (usando LBP tal vez)?

Requisito: você deve projetar **duas estratégias**, as duas explicitamente fazendo uso de **séries temporais**. Para cada estratégia considere duas medidas de distância/similaridade diferentes.

Definida as estratégias, chegamos à etapa de **Implementação**. Neste momento, você deverá implementar, em Linguagem R, a solução para o objetivo definido neste trabalho.

Finalmente, para a etapa de **Avaliação** da sua solução, devemos definir o que é relevante e o que não é, a fim de avaliar o resultado da sua busca/recuperação por meio das métricas precisão (precision) e revocação (recall). Como mencionado acima, uma imagem da base de dados será considerada relevante se pertence à mesma espécie que a imagem de consulta.

A fim de avaliar a eficiência do seu método, você deve realizar experimentos com as imagens encontradas no arquivo **"SwedishLeaf_TEST.csv"**. Você deve relatar a **revocação (recall) e precisão (precision) médias** das 625 *queries* usando P@K, sendo K = {5, 10, 15, 20, 25, 30, …, 100}, ou seja, analisará a precisão média dos K primeiros itens da lista retornada pela sua solução/método e sua respectiva revocação média.

Mostrar:

- Gráfico comparando as 4 curvas de precisão-revocação (dois métodos propostos, cada um usando duas funções de distância/similaridade distintas);
- Discussão dos resultados.
- Outras análises que julgar relevante.

Protocolo de Submissão dos Resultados

Você deve submeter um arquivo compactado (formato zip), contendo os seguintes documentos/arquivos:

- Código R referente às implementações realizadas, devidamente comentado;
- Relatório técnico (em PDF) contendo, nome completo dos integrantes da dupla e descrição, dentre outros assuntos, sobre eventuais procedimentos de pré-processamento realizados, suas decisões de projeto, os algoritmos efetivamente implementados,

eventuais dificuldades enfrentadas, os resultados, as suas análises e interpretação dos resultados.