Ficha de Trabalho nº 5

Raciocínio Baseado em Casos - MatLab

Bibliografia

Slides das aulas teóricas.
Documentos do Moodle.

1. Introdução

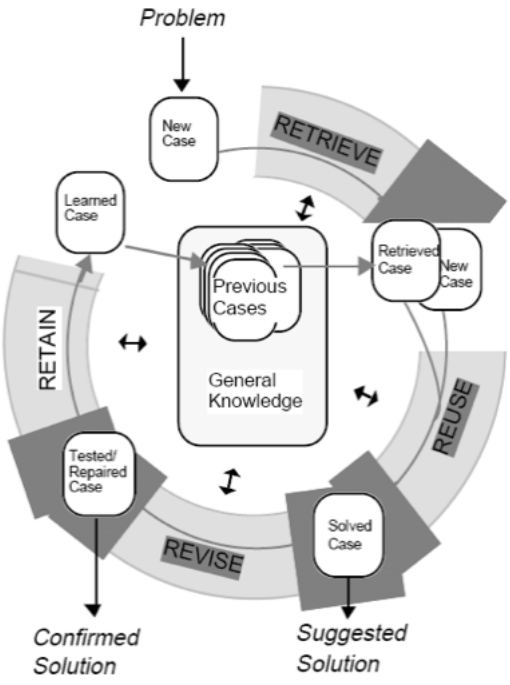
O raciocínio baseado em casos (CBR de Case-Based Reasoning) é uma técnica que procura resolver novos problemas adaptando soluções de problemas anteriores como solução conhecida. Os sistemas baseados em CBR permitem fazer extração do conhecimento a partir de casos ou experiências armazenadas no próprio sistema. Esta extração é feita através da identificação das características e de medidas de similaridade dos casos apresentados a fim de devolver uma melhor solução (resposta).

Resumidamente, um sistema CBR funciona da seguinte forma:

- Armazena experiências anteriores (casos) na memória

Quando surge um novo problema:

- Extrai (*Retrieve*) da memória casos semelhantes sobre situações semelhantes ao novo problema
    - Uma medida de semelhança estima a diferença entre o novo caso e os que estão armazenados
- Reutiliza (*Reuse* + *Revise*) a experiência passada no contexto da nova situação, podendo adaptá-la caso seja necessário
- Armazena (*Retain*) a nova experiência (caso) na memória – aprendizagem



Cada caso descreve uma situação particular e todos os casos são independentes. Um caso consiste num conjunto de pares <atributo, valor>. Quando um novo caso surge usam-se medidas de semelhança para escolher qual o caso da memória que mais se aproxima do atual. Nestas medidas é possível usar pesos para dar mais importância a alguns atributos.

2. Exercício com uma biblioteca de casos de férias na Europa

O ficheiro CSV "TravelCaseBase.csv" fornecido em conjunto com esta ficha consiste numa biblioteca de casos com **1024 viagens** de férias feitas na Europa. Cada caso é descrito pelos seguintes atributos:

- **JourneyCode** – Código identificador único da viagem; assume apenas valores inteiros.
- **HolidayType** – Tipo de férias; assume os seguintes valores: *Active*, *Bathing*, *City*, *Education*, *Language*, *Recreation*, *Skiing* ou *Wandering*.
- **Price** – Preço da viagem; assume valores reais.
- **NumberOfPersons** – Número de pessoas envolvidas na viagem; assume valores reais (uma criança corresponde a 0,5).
- **Region** – Região onde as férias foram feitas; assume qualquer região, país ou cidade válidos.
- **Transportation** – Tipo de transporte utilizado na viagem; assume os seguintes valores: *Car*, *Coach*, *Plane* ou *Train*.
- **Duration** – Número de dias que as férias demoraram; assume valores reais.
- **Season** – Mês do ano em que as férias foram realizadas; assume como valores os nomes dos meses (*January*, *February*, etc.).
- **Accommodation** – Tipo de acomodação utilizada nas férias; assume os seguintes valores: *HolidayFlat*, *OneStar*, *TwoStars*, *ThreeStars*, *FourStars* ou *FiveStars*.
- **Hotel** – Nome do hotel utilizado nas férias; assume qualquer cadeira de caracteres.

Juntamente com esta ficha é fornecida uma implementação parcial, em MatLab, de um sistema de CBR preparado para este *dataset*. Analisando cada uma das 4 fases implementadas neste sistema:

- **Retrieve** – Nesta fase são considerados para efeitos de cálculo de semelhança todos os atributos com exceção do *JourneyCode* e do *Hotel*, que acabam por ser a solução da pesquisa. Para o cálculo de semelhança de cada atributo são aplicadas as seguintes distâncias:
    - HolidayType – Distância assimétrica definida manualmente.
    - Price – Distância euclidiana.
    - NumberOfPersons – Distância linear.
    - Region – Distância de *Haversine*, sendo a latitude e longitude de cada região obtidas automaticamente através da *Google Maps Geolocation API*.
    - Transportation – Distância simétrica definida manualmente.
    - Duration – Distância euclidiana.
    - Season – Distância "circular" entre os meses (por exemplo, a distância entre Novembro e Janeiro é 2).
    - Accommodation – Distância simétrica definida manualmente.

A semelhança total é calculada através de uma distância linear pesada e os valores de cada atributo são normalizados através dos valores máximos existentes para cada um no *dataset* (com exceção da distância entre regiões, onde é utilizada um máximo conhecido para a normalização). Desta fase são apenas devolvidos os casos que apresentem uma semelhança superior a um *threshold* previamente definido (90% por omissão).

- **Reuse** – Nesta fase é aplicada uma regressão linear múltipla aos casos obtidos do *retrieve*, regressão essa que utiliza como termos independentes o número de pessoas e a duração das viagens e como termo dependente o preço da viagem. A aplicação desta técnica permite ajustar o preço inicialmente inserido pelo utilizador para o novo caso em função dos casos semelhantes que já existem.
- **Revise** – Nesta fase é pedido ao utilizador que identifique dentro dos casos devolvidos do *retrieve* qual é o mais semelhante ao novo caso, identificação essa que é feita através do atributo *JourneyCode*. Para além disso, é pedido ao utilizador que indique se pretende atualizar o preço do novo caso para o valor que foi estimado na fase de *reuse*, ou se pretende manter o valor definido inicialmente.
- **Retain** – Nesta fase é pedido ao utilizador que indique se pretende que o novo caso seja acrescentado à biblioteca existente. Se assim for, o caso é acrescentado ao fim do ficheiro CSV com os valores definidos inicialmente (eventualmente com o preço atualizado), com um *JourneyCode* novo e com o hotel do caso identificado como o mais semelhante.

Cada uma das fases encontra-se implementada numa função de MatLab independente, existindo uma função geral chamada *cbr* que é responsável pela execução do processo de forma ordeira. A definição dos dados do novo caso é feita manualmente numa estrutura chamada *new_case,* que se encontra definida no início da função *cbr*.

a) Analise o ficheiro **retrieve.m** disponibilizado no Moodle.

- Defina os valores para a similaridade assimétrica do atributo *HolidayType* na função *get_holiday_type_similarities*.
- Defina os valores para a similaridade simétrica do atributo *Accommodation* na função *get_accommodation_similarities*.
- Implemente a fórmula genérica da distância linear na função *calculate_linear_distance*.
- Implemente a fórmula genérica da distância euclidiana na função *calculate_euclidean_distance*.
- Implemente a fórmula genérica para as distâncias locais (simétricas ou assimétricas) na função *calculate_local_distance*.
- Sabendo que a circunferência da Terra são 40075 quilómetros e que, por esse motivo, essa é aproximadamente a distância máxima entre quaisquer dois pontos geográficos, implemente a normalização da distância de *Haversine* na última linha da função *calculate_haversine_distance*.

- Preencha os pesos de cada atributo no vetor *weighting_factors*, assumindo que os atributos *Price* e *Season* são os mais relevantes no cálculo da semelhança.
- Utilizando a fórmula da distância linear pesada, implemente o cálculo da semelhança final e guarde o valor na variável *final_similarity*.

b) Analise o ficheiro **reuse.m** disponibilizado no Moodle.

- Implemente a fórmula que permite chegar ao novo valor estimado do preço, sabendo que o modelo matemático se trata de uma regressão linear múltipla.

c) Analise o ficheiro **revise.m** disponibilizado no Moodle.

- A maioria dos sistemas de CBR permitem que numa fase de *revise* o utilizador possa atualizar/corrigir os valores inseridos inicialmente para o novo caso em função da observação feita aos casos semelhantes que existem. A implementação atual apenas permite a atualização do valor do atributo *Price* para o novo valor estimado na fase de *reuse*. Faça as alterações necessárias para que o utilizador possa nesta fase alterar qualquer um dos atributos inseridos inicialmente para o novo caso.

d) A implementação atual do sistema de CBR assume que todos os atributos utilizados para o cálculo de semelhança são fornecidos inicialmente. Esta situação não corresponde à realidade, já que muitas das vezes não é possível definir valores para todos os atributos. Faça as alterações necessários no sistema para que seja possível deixar atributos por definir nos novos casos, considerando as seguintes regras:

- Para efeitos do cálculo de semelhança devem apenas ser utilizados os atributos que tenham valores definidos.
- Caso o utilizador decida acrescentar o novo caso à biblioteca na fase de *retain*, os atributos que não se encontrem definidos devem ser obtidos do caso identificado como mais semelhante na fase de *revise* (à semelhança do que já acontece com o atributo *Hotel*).

e) Execute o sistema e faça variar os seguintes parâmetros para perceber o impacto que têm nos resultados obtidos:

- *Threshold* mínimo de semelhança.
- Peso de cada atributo na fase de *retrieve*.
- As várias distâncias simétricas e assimétricas definidas na fase de *retrieve*.

f) **[Proposta de Trabalho]** Atualmente a fase de *reuse* contempla apenas 3 atributos e propõe novos valores adaptados para apenas um deles. Proponha e implemente novas abordagens que tirem partido de mais atributos e que permitam fazer mais adaptações ao novo caso em função da informação obtida dos casos semelhantes.

g) **[Proposta de Trabalho]** Implemente uma interface gráfica em MatLab para o sistema em causa, permitindo assim que os dados dos novos casos e as restantes informações sejam pedidas e fornecidas através dessa interface.