Inteligencia Artificial Data Mining KDD

Inteligencia Artificial

A inteligência artificial (IA) é uma das ciências mais recentes. O trabalho começou logo após a Segunda Guerra Mundial, e o próprio nome foi cunhado em 1956. Seus principais idealizadores foram os cientistas Marvin Minsky, Herbert Simon, Allen Newell, John McCarthy, Warren MuCulloch e Walter Pitts, entre outros.

A construção de máquinas inteligentes fascina a humanidade desde tempos imemoriais. Entretanto, apenas recentemente, com o surgimento do computador moderno, é que a inteligência artificial ganhou meios e massa crítica para se estabelecer como ciência integral, com problemáticas e metodologias próprias. Desde então, seu desenvolvimento tem extrapolado os clássicos programas de xadrez ou de conversão e envolvido áreas como visão computacional, análise e síntese da voz, lógica difusa, redes neurais artificiais e muitas outras.

Inicialmente a IA visava reproduzir o pensamento humano. A Inteligência Artificial abraçou a idéia de reproduzir faculdades humanas como criatividade, auto-aperfeiçoamento e uso da linguagem. Porém, o conceito de inteligência artificial é bastante difícil de se definir. Por essa razão, Inteligência Artificial foi (e continua sendo) uma noção que dispõe de múltiplas interpretações, não raro conflitantes ou circulares.

A questão sobre o que é inteligência artificial pode ser separada em duas partes: "qual a natureza do artificial" e "o que é inteligência". A primeira questão é de resolução relativamente fácil, apontando no entanto para a questão de o que poderá o homem construir.

A segunda questão é consideravelmente mais difícil, levantando a questão da consciência, identidade e mente (incluindo a mente inconsciente) juntamente com a questão de que componentes estão envolvidos no único tipo de inteligência que universalmente se aceita como estando ao alcance do nosso estudo: a inteligência do ser humano. O estudo de animais e de sistemas artificiais que não são modelos triviais, começam a ser considerados como matéria de estudo na área da inteligência.

Uma popular e inicial definição de inteligência artificial, introduzida por John McCarty na famosa conferência de Dartmouth em 1955 é "fazer a máquina comportar-se de tal forma que seja chamada inteligente caso fosse este o comportamento de um ser humano." No entanto, esta definição parece ignorar a possibilidade de existir a IA forte. Outra definição de inteligência artificial é a inteligência que surge de um dispositivo artificial. A maior parte das definições podem ser categorizadas em sistemas que: pensam como um humano; agem como um humano; pensam racionalmente ou agem racionalmente.

Inteligência artificial forte

A investigação em Inteligência artificial forte aborda a criação da forma de inteligência baseada em computador que consiga raciocinar e resolver problemas; uma forma de IA forte é classificada como auto-consciente.

Inteligência artificial fraca

Trata-se da noção de como lidar com problemas não determinísticos.

Uma contribuição prática de Alan Turing foi o que se chamou depois de Teste de Turing (TT), de 1950: em lugar de responder à pergunta "podem-se ter computadores inteligentes?" ele formulou seu teste, que se tornou praticamente o ponto de partida da pesquisa em "Inteligência Artificial".

O teste consiste em se fazer perguntas a uma pessoa e um computador escondidos. Um computador e seus programas passam no TT se, pelas respostas, for impossível a alguém distinguir qual interlocutor é a máquina e qual é a pessoa (ele não especificou o nível intelectual dessa pessoa). No seu artigo original ele fez a previsão de que até 2000 os computadores passariam seu teste. Pois bem, há um concurso anual de programas para o TT, e o resultado dos sistemas ganhadores é tão fraco (o último tem o nome "Ella") que com poucas perguntas logo percebe-se a idiotice das respostas da máquina. É interessante notar que tanto a Máquina de Turing quanto o Teste de Turing talvez derivem da visão que Turing tinha de que o ser humano é uma máquina.

Há quem diga que essa visão está absolutamente errada, do ponto de vista linguístico, já que associamos à "máquina" um artefato inventado e eventualmente construído. Dizem eles: "Nenhum ser humano foi inventado ou construído". Afirma-se ainda que a comparação, feita por Turing, entre o homem e a máquina é sinônimo de sua "ingenuidade social", pois as máquinas são infinitamente mais simples do que o homem, apesar de, paradoxalmente, se afirmar que a vida é complexa. No entanto, esta linha de raciocínio é questionável, afinal de contas, os computadores modernos podem ser considerados "complexos" quando comparados ao COLOSSUS (computador cujo o desenvolvimento foi liderado por Turing, em 1943), ou a qualquer máquina do início do século XX. O fato é que se considerarmos que o homem é uma "máquina", onde todos os componentes têm uma razão para existir, necessariamente somos levados a imaginar que alguém o construiu. Esta "incredulidade" vem do fato de a ciência "ainda" não ter encontrado respostas para esta pergunta.

A inteligência artificial fraca centra a sua investigação na criação de inteligência artificial que não é capaz de verdadeiramente raciocinar e resolver problemas. Uma tal máquina com esta característica de inteligência agiria como se fosse inteligente, mas não tem autoconsciência ou noção de si. O teste clássico para aferição da inteligência em máquinas é o Teste de Turing.

Há diversos campos dentro da IA fraca, e um deles é a Linguagem Natural, que trata de estudar e tentar reproduzir os processos de desenvolvimento que resultaram no funcionamento normal da língua. Muitos destes campos utilizam softwares específicos e linguagens de programação criadas para suas finalidades. Um exemplo bastante conhecido é o programa A.L.I.C.E. (Artificial Linguistic Internet Computer Entity, ou Entidade Computadorizada de Linguagem Artificial para Internet), um software que simula uma conversa humana. Programado em Java e desenvolvido com regras heurísticas para os caracteres de conversação, seu desenvolvimento resultou na AIML (Artificial Intelligence Markup Language), uma linguagem específica para tais programas e seus vários clones, chamados de Alicebots.

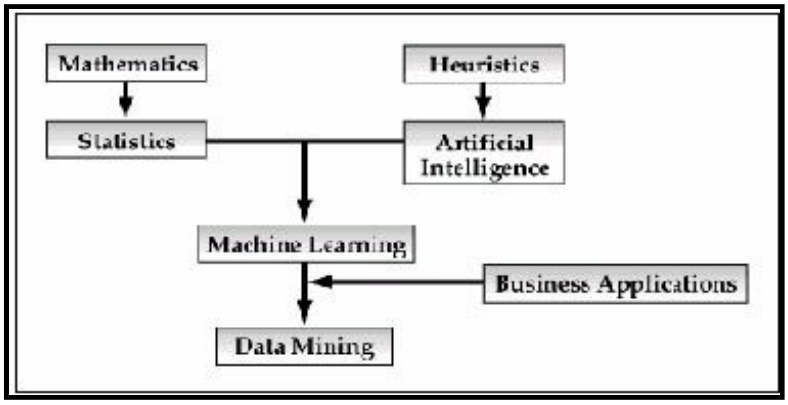
Muito do trabalho neste campo tem sido feito com simulações em computador de inteligência baseado num conjunto predefinido de regras. Poucos têm sido os progressos na IA forte. Mas dependendo da definição de IA utilizada, pode-se dizer que avanços consideráveis na IA fraca já foram alcançados.

Data Mining

Definição

Data mining (ou mineração de dados) é o processo de extrair informação válida, previamente desconhecida e de máxima abrangência a partir de grandes bases de dados, usando-as para efetuar decisões cruciais.

Data mining vai muito além da simples consulta a um banco de dados, no sentido de que permite aos usuários explorar e inferir informação útil a partir dos dados, descobrindo relacionamentos escondidos no banco de dados. Pode ser considerada uma forma de descobrimento de conhecimento em bancos de dados (KDD - Knowledge Discovery in Databases), área de pesquisa de bastante evidência no momento, envolvendo Inteligência Artificial e Banco de Dados.



Processo

Data Mining ou Mineração de Dados consiste em um processo analítico projetado para explorar grandes quantidades de dados (tipicamente relacionados a negócios, mercado ou pesquisas científicas), na busca de padrões consistentes e/ou relacionamentos sistemáticos entre variáveis e, então, validá-los aplicando os padrões detectados a novos subconjuntos de dados.

Técnicas distintas como redes neurais, indução de árvores de decisão, sistemas baseados em regras e programas estatísticos, tanto isoladamente quanto em combinação, podem ser aplicadas ao problema. Em geral, o processo de busca é interativo, de forma que os analistas revêem o resultado, formando um novo conjunto de questões para refinar a busca em um determinado aspecto das descobertas e, realimentam o sistema com novos parâmetros. Ao final do processo, o sistema de data mining gera um relatório das descobertas, que passa então a ser interpretado pelos analistas de mineração. De posse da interpretação das informações, torna-se possível a obtenção de algum tipo de conhecimento.

Etapas

As operações necessárias para se efetuar uma análise por Data Mining pode ser dividida em 4 fases: Análise do Problema, Preparação dos dados, Modelagem e, Análise e Validação dos Resultados.

a) Análise do problema: o processo de análise deve se iniciar declarando-se um objetivo a ser requerido, ou seja, deve-se definir que tipo de

conhecimento está se lidando através do problema exposto. A importância desta análise concentra-se na possibilidade de se selecionar os dados necessários e da definição das técnicas a serem utilizadas na análise.

b) Preparação dos Dados: Esta fase, segundo Edelstein (1998) consiste em 5 subfases: coletânea de dados, Avaliação, consolidação e limpeza, seleção dos dados e transformação. De uma forma resumida, esta fase consiste na aquisição, limpeza e enriquecimento dos dados, que segundo Manilla (1994), pode tomar até 80% do tempo necessário para todo o processo, devido às bem conhecidas dificuldades de integração de bases de dados heterogêneas.

- ***Coletânea dos dados:*** **Os dados podem vir de diversas fontes, desde as capturadas nas operações internas e de transações, até as fontes externas como dados demográficos, informações de cartão de crédito e outros, vinculando a seleção ao problema a ser analisado.**
- ***Avaliação:*** **Um exame dos dados identifica possíveis características que afetarão a qualidade do modelo. Os dados necessários podem residir em uma ou muita bases de dados. Os dados de fonte podem residir nas bases de transações ou nos armazéns dos dados. Ainda outros, podem residir em uma base que pertença a uma outra companhia tal como um departamento de crédito. Quando os dados vêm de múltiplas fontes, devem ser consolidados em uma única base de dados. Nesse tempo pode-se ter problemas na consolidação como inconsistência dos dados, codificações diferentes, e valores inconsistentes para o mesmo artigo de dados.**
- ***Consolidação e limpeza:*** **Nesta etapa é feita a construção da base de dados a ser trabalhada. Assim, consolida-se os dados e repara-se os problemas identificados no exame dos dados. Desta forma, deve-se buscar obter as informações necessárias para corrigir os erros, ou remover os registros com campos vazios e/ou incompletos ou mesmo colocar um valor comum nos campos vazios.**
- ***Seleção dos dados:*** **Com os dados recolhidos para a construção do modelo, é necessário selecionar os dados específicos para o modelo. Para um modelo de predição, isto significa geralmente selecionar as variáveis independentes (ou colunas), as variáveis dependentes, e os casos, sendo este último utilizado para treinar o modelo.**
- ***Transformação:*** **Após a seleção dos dados, algumas transformações adicionais dos dados podem ser necessárias. A ferramenta escolhida pode também ditar como se deve representar os dados. Muitas árvores da decisão usadas para a classificação requerem dados contínuos, tais como a renda, estejam agrupados em escalas como elevado, baixo, ou médio.**

c) Modelagem: Baseando-se na definição dos problemas e na base de dados construída, tem-se nesta fase a definição das tarefas serem utilizadas e quais técnicas estarão associadas a tarefa definida, visto que uma tarefa pode ser executada por diversos algoritmos. Assim cria-se um modelo que deverá ser analisado.

d) Análise e Validação dos Resultados: Não importa o quão exato um modelo promete ser, não há nenhuma garantia de que este refletirá o mundo real. Um modelo válido não é necessariamente um modelo correto. Uma das razões principais para este problema é que sempre há suposições implícitas no modelo. Variáveis tais como a taxa de inflação podem ser parte de um modelo para predizer o propensão de um indivíduo a comprar, mas se ocorrer um salto da inflação de 3% para 17% isto afetará certamente o comportamento deste. Conseqüentemente, é importante testar um modelo no mundo real.

Data Mining na Prática: Algoritmo K-Means

A idéia do algoritmo K-Means (também chamado de K-Médias) é fornecer uma classificação de informações de acordo com os próprios dados. Esta classificação, como será vista a seguir, é baseada em análise e comparações entre os valores numéricos dos dados. Desta maneira, o algoritmo automaticamente vai fornecer uma classificação automática sem a necessidade de nenhuma supervisão humana, ou seja, sem nenhuma pré-classificação existente. Por causa desta característica, o K-Means é considerado como um algoritmo de mineração de dados não supervisionado.

Para entender como o algoritmo funciona, vamos imaginar que temos uma tabela com linhas e colunas que contêm os dados a serem classificados. Nesta tabela, cada coluna é chamada de dimensão e cada linha contém informações para cada dimensão, que também são chamadas de ocorrências ou pontos. Geralmente, trabalha-se com dados contínuos neste algoritmo, mas nada impede que dados discretos sejam utilizados, deste que eles sejam mapeados para valores numéricos correspondentes.

O algoritmo vai analisar todos os dados desta tabela e criar classificações. Isto é, o algoritmo vai indicar uma classe (cluster) e vai dizer quais linhas pertencem a esta classe. O usuário deve fornecer ao algoritmo a quantidade de classes que ele deseja. Este número de classes que deve ser passada para o algoritmo é chamado de k e é daí que vem a primeira letra do algoritmo: K-Means.

Para gerar as classes e classificar as ocorrências, o algoritmo faz uma comparação entre cada valor de cada linha por meio da distância. Geralmente utiliza-se a distância euclidiana para calcular o quão 'longe' uma ocorrência está da outra. A maneira de calcular esta distância vai depender da quantidade de atributos da tabela fornecida. Após o cálculo das distâncias o algoritmo calcula centróides para cada uma das classes. Conforme o algoritmo vai iterando, o valor de cada centróide é refinado pela média dos valores de cada atributo de cada ocorrência que pertence a este centróide. Com isso, o algoritmo gera k centróides e coloca as ocorrências da tabela de acordo com sua distância dos centróides.

Para simplificar a explicação de como o algoritmo funciona vou apresentar o algoritmo K-Means em cinco passos:

PASSO 01: Fornecer valores para os centróides.
Neste passo os k centróides devem receber valores iniciais. No início do algoritmo geralmente escolhe-se os k primeiros pontos da tabela. Também é importante colocar todos os pontos em um centróide qualquer para que o algoritmo possa iniciar seu processamento.

PASSO 02: Gerar uma matriz de distância entre cada ponto e os centróides.
Neste passo, a distância entre cada ponto e os centróides é calculada. A parte mais 'pesada' de cálculos ocorre neste passo pois se temos N pontos e k centróides teremos que calcular N x k distâncias neste passo.

PASSO 03: Colocar cada ponto nas classes de acordo com a sua distância do centróide da classe.
Aqui, os pontos são classificados de acordo com sua distância dos centróides de cada classe. A classificação funciona assim: o centróide que está mais perto deste ponto vai 'incorporá-lo', ou seja, o ponto vai pertencer à classe representada pelo centróide que está mais perto do ponto. É importante dizer

que o algoritmo termina se nenhum ponto 'mudar' de classe, ou seja, se nenhum ponto for 'incorporado' a uma classe diferente da que ele estava antes deste passo.

PASSO 04: Calcular os novos centróides para cada classe.
Neste momento, os valores das coordenadas dos centróides são refinados. Para cada classe que possui mais de um ponto o novo valor dos centróides é calculado fazendo-se a média de cada atributo de todos os pontos que pertencem a esta classe.

PASSO 05: Repetir até a convergência.
O algoritmo volta para o PASSO 02 repetindo iterativamente o refinamento do cálculo das coordenadas dos centróides.
Desta maneira teremos uma classificação que coloca cada ponto em apenas uma classe. Desta maneira dizemos que este algoritmo faz uma classificação hard (hard clustering) uma vez que cada ponto só pode ser classificado em uma classe. Outros algoritmos trabalham com o conceito de classificação soft onde existe uma métrica que diz o quão 'dentro' de cada classe o ponto está.

Topo

KDD

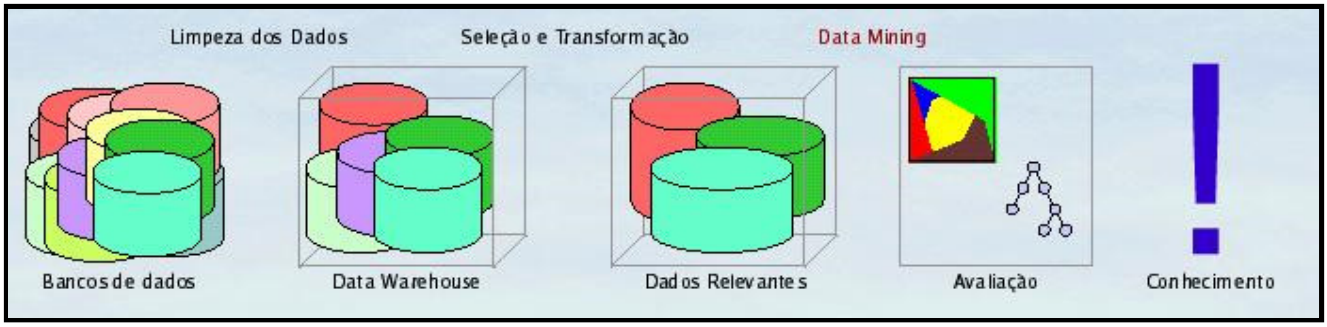
KDD consiste, fundamentalmente, na estruturação do banco de dados; na seleção, preparação e pré-processamento dos dados; na transformação, adequação e redução da dimensionalidade dos dados; no processo de Data Mining; e nas análises, assimilações, interpretações e uso do conhecimento extraído do banco de dados, através do processo de Data Mining.

É um processo geral de descoberta de conhecimentos úteis, previamente desconhecidos a partir de grandes bancos de dados. Possui várias etapas:

- **Interdependentes;**
- **Podem ser repetidas;**
- **Nem sempre tem distinções claras entre si.**

Processo

- **Seleção: escolha de dados para processamento**
- **Pré-processamento: enriquecimento dos dados, desnormalização de banco de dados.**
- **Transformações: filtragem, normalizações, reprojeções, adequações a algoritmos.**
- **Data Mining: extração de padrões, classificação, agrupamentos.**
- **Interpretação: visualização, validação.**



Topo

Referências Bibliográficas

www.pcs.usp.br/~pcs5000/PCS5000 DM SI.pdf
Acesso em:09/09/2006(10:32)

www.cce.puc-rio.br/informatica/dataminingcentro.htm
Acesso em:09/09/2006(10:53)

www.intelliwise.com/reports/i2002.htm
Acesso em:09/09/2006(11:15)

Wikipedia
Acesso em:15/09/2006(00:27)

Topo