Aprendizado Supervisionado e Não Supervisionado

Tipos de problema

Luciana Nascimento Santana Prachedes

Mineração de Dados - 2022/1

UNIVERSIDADE
FEDERAL DE JUIZ DE FORA

Aprendizado Supervisionado - Regressão

Exemplo: Data Mining of Agricultural Yield Data: A Comparison of Regression Models [1]

O trabalho fala sobre o uso de técnicas de mineração de dados, especificamente regressão, na previsão da colheita em plantações. Os dados coletados e armazenados de plantações são oriundos de agricultura de precisão, que diz respeito ao uso de tecnologias de GPS e pequenos sensores para medir e avaliar as condições das áreas de atividades agronômicas dada a variabilidade do solo e clima. O objetivo é fazer a predição da colheita baseado nesses dados. O trabalho trata esse objetivo como tarefa multi-dimensional de regressão do ponto de vista da mineração de dados. São apresentadas e avaliadas quatro técnicas e é recomendado a de melhor desempenho (em relação a alta acurácia e alta generalidade do modelo dentro do contexto proposto).

Aprendizado Supervisionado - Regressão

Descrição dos dados: Coletados de 2003 a 2006, em três campos na Alemanha. Cada campo tem sete atributos: quantidade de fertilizante (N1, N2, N3), vegetação (REIP32, REIP49), condutividade elétrica do solo (EM38) e a colheita para um certo ano - que é o atributo alvo.

Aqui, a colheita é medida em toneladas métricas por hectare (t/ha).

Dado o conjunto de treinamento com esses atributos, com técnicas de regressão (MLP, RBF, Regression Tree, e SVM) os autores encontram uma função aproximada e avaliam a qualidade da função com RMSE (Root Mean Square Error) e MAE (Mean Absolute Error).

Foi usada a razão de 9:1 para treinamento e teste dos modelos.

A técnica com melhores resultados foi a Support Vector Machine (SVM).

Aprendizado Supervisionado - Classificação

Exemplo: Analysing Soil Data using Data Mining Classification Techniques [2]

O trabalho visa prever o tipo de um solo baseado em técnicas de mineração de dados, especificamente classificação. São usadas três técnicas de classificação diferentes - JRip, J48 e Naive Bayes - e a de melhor desempenho, baseado na acurácia do modelo, foi a JRip.

Os dados foram coletados de um laboratório de testagem de solo e possuem os seguintes atributos: Village Name, Soil Type or Color, Soil Texture, PH, EC (Electrical Conductivity), Lime Status, Phosphorous. São abordadas duas classes de solo: vermelho e preto.

JRip classificou corretamente mais instâncias do que os outros modelos.

Aprendizado Não Supervisionado - Agrupamento

Exemplo: Customer Data Clustering Using Data Mining Technique [3]

O trabalho faz uso de técnicas de mineração, especificamente agrupamento, sobre um conjunto de dados de clientes de uma loja para segmentá-los em grupos visando localizar os clientes de alto lucro, alto valor e baixo risco.

Os atributos utilizados foram: recência, lucro total do cliente, receita total do cliente, e departamento de receita principal. Foram formados quatro grupos e os resultados mostraram fatos interessantes, tal como o grupo 1 ser o grupo mais lucrativo, pois representava cerca de 35% da receita, mas apenas 6% dos clientes. Esse conhecimento é útil para criação de estratégias de negócio e marketing.

Aprendizado Não Supervisionado - Regras de Associação

Exemplo: Mining association rules for the quality improvement of the production process [4]

O trabalho utiliza técnicas de mineração de dados, especificamente regras de associação, para extrair conhecimento sobre operações e gestão da informação. O exemplo de aplicação detalha um experimento industrial no qual o processo de fabricação de um fornecedor de produtos de perfuração é analisado. Disfunções na gestão de operações e perda de tempo de produção são problemas que impactam o desempenho e qualidade dos sistemas industriais, bem como o seu custo de produção. Por isso, é interessante analisar esses casos.

Aprendizado Não Supervisionado - Regras de Associação

Os dados são oriundos de uma empresa real e os atributos presentes são: tamanho de peça sendo produzido (pequeno, médio e grande), diferentes disfunções no processo de produção (sete tipos) e atrasos de diferentes tempos (1h, 4h e 10h).

Uma das regras de associação encontradas: Medium, Dys1 -> Delay4h.

Os resultados mostraram que a disfunção Dys1 representa 56,3% dos custos dos atrasos e, portanto, é a principal causa desses custos.

Diferenças

- Aprendizado Supervisionado
  - Existência de um alvo
    - Regressão: valor contínuo procurado, alvo numérico
    - Classificação: discreto, classifica/categoriza em classes-alvo
- Aprendizado Não Supervisionado
  - Não existe alvo conhecido
    - Agrupamento: não há grupos pré-definidos, busca por similaridades
    - Regras de associação: implicações, probabilidade de co-ocorrência de itens do conjunto de dados

Referências

1. Ruß, G. (2009). Data Mining of Agricultural Yield Data: A Comparison of Regression Models. Lecture Notes in Computer Science, 24–37. doi:10.1007/978-3-642-03067-3_3
2. Rajeswari, V., & Arunesh, K. (2016). Analysing Soil Data using Data Mining Classification Techniques. Indian Journal of Science and Technology, 9(19). doi:10.17485/ijst/2016/v9i19/93873
3. Rajagopal, D. (2011). Customer data clustering using data mining technique. arXiv preprint arXiv:1112.2663.
4. Kamsu-Foguem, B., Rigal, F., & Mauget, F. (2013). Mining association rules for the quality improvement of the production process. Expert Systems with Applications, 40(4), 1034–1045. doi:10.1016/j.eswa.2012.08.039