Aprendizado Descritivo

Aula 15 – Aplicações de descoberta de subgrupos

Professor Renato Vimieiro

DCC/ICEx/UFMG

Introdução

- Nessa aula, veremos exemplos de aplicações de descoberta de subgrupos e mineração de modelos excepcionais
- Veremos três exemplos em domínios diferentes:
  - Energia (perfis de consumo energético)
  - Educação (aprendizado em ambientes virtuais)
  - Esportes (prevenção de lesões em atletas profissionais)
- A intenção não é analisar os resultados em si, e, sim, a metodologia usada
  - O objetivo final é como esse tipo de técnica pode ser empregado em cenários mais práticos

Perfis de consumo de energia elétrica

- Nanlin et al. (2014) apresentaram um estudo sobre o perfil de consumo de energia elétrica no Reino Unido
- A ideia central do trabalho é identificar padrões de consumo não usuais com relação ao perfil da residência
- Os autores discutem como a implantação de sistemas de smart grid
- Nesse tipo de rede, os medidores inteligentes medem e transmitem o consumo de energia dos usuários em tempo real
- Assim, ações podem ser tomadas pelas distribuidoras e geradoras de energia para redirecionar a produção para regiões com maior demanda
- O sistema tradicional trabalha com modelos de estimativa de consumo para garantir que a demanda prevista seja atendida satisfatoriamente

Perfis de consumo de energia elétrica

- Alguns países já se encontram mais avançados na substituição dos medidores tradicionais por inteligentes
- Em outros países, como no Brasil, estudos para implantação já foram iniciados, e alguns consumidores já foram integrados ao sistema
- Um dos objetivos desses estudos preliminares é identificar consumidores com padrões não-usuais de consumo
  - Isso permite reajustar a infraestrutura para atender a demandas inesperadas no sistema
- Por exemplo, os gestores do sistema de energia podem definir um padrão de consumo com base em questionários ou no censo de uma região
  - A Cemig aplica um questionário com perguntas sobre equipamentos elétricos/eletrônicos em novas instalações.

Perfis de consumo de energia elétrica

- Os perfis traçados para uma região podem prever um certo número de pessoas por residência, com um certo padrão de consumo de energia
  - Consumidores de bairros mais ricos podem possuir equipamentos que consomem mais energia, ou mesmo uma quantidade maior de equipamentos
  - Enquanto consumidores de bairros mais pobres podem ter menos equipamentos, levando a uma expectativa de menor consumo energético
  - Em outros países, outros fatores também influenciam como o uso ou não de gás em fogões e aquecedores
- Embora sejam úteis para fazer uma avaliação global do consumo, eles podem falhar em regiões mais heterogêneas
  - Alguns bairros podem ter perfis bastante distintos de consumidores: casas de alto e baixo padrão próximas umas das outras, ou ainda residências com poucos moradores ou que estejam fora por longos períodos
- Os dados coletados pelos medidores inteligentes possibilitam fazer esses ajustes mais finos nos perfis e modelos

Perfis de consumo de energia elétrica

- Os autores do trabalho aplicaram descoberta de subgrupos sobre os dados de medidores inteligentes numa base de dados do Reino Unido
- Os dados consistem na medição do consumo (em kWh) em intervalos de 30min, iniciando às 0h e terminado às 23h30
  - Foram feitas 48 medições por dia
- Foram coletados dados de cerca 5000 consumidores durante 1,5 ano
- Além dos dados de consumo, foram coletados dados socio-demográficos sobre os moradores e residências

Perfis de consumo de energia elétrica

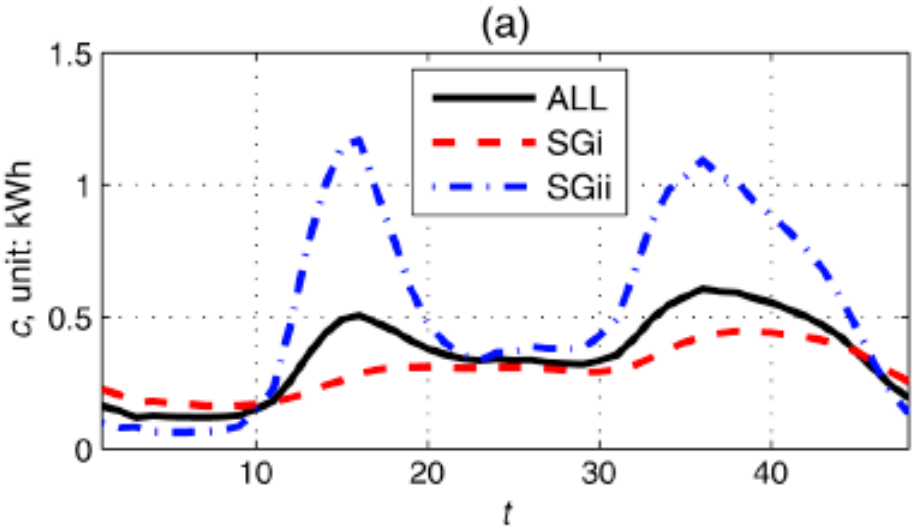
Socio-demographic variables	Description	Number of categories	Example(s)
GSP group	Grid Supply Point Group in U.K. which are regional electricity distribution networks	Total 14 3 in dataset[a]	Southern; South Wales; North Scotland
Age	Age of head of household	6	Age 26-35
Decision maker type	Type of person deciding household matters	13	Young Couple
Family lifestage	The combined stage of life and family status including children	14	Young family with children
Household composition	People living together and their relationships to one another	13	Male homesharers
Household income band	Total household income per year	10	£30,000 to £39,999
Mains gas flag	Whether a household is connected to the Main gas network; if Yes, it's assumed that the household uses gas	2	Connected to gas; not connected to gas
Mosaic public sector group	Classification on citizen's location, demographics, lifestyles and behaviors	15	Young, well-educated city dwellers; Wealthy people living in the most sought after neighborhoods
Mosaic public sector type	Subcategories of Mosaic Public Sector Group	69	Young professional families settling in better quality older terraces
Number of bedrooms	Number of Bedrooms of the property	5	5 + bedrooms
Property age	When the property was built	6	1871–1919
Property type 2011	Type of property in 2011	5	Purpose built flats; Farm
Property value fine	Estimated property value	25	£500,001 to £600,000
Tenure 2011	Property ownership in 2011	3	Privately rented

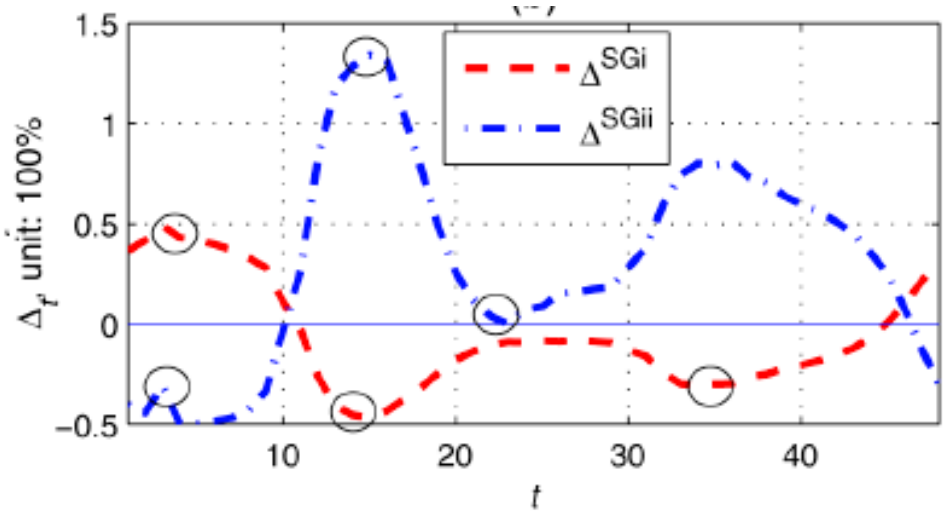
Perfis de consumo de energia elétrica

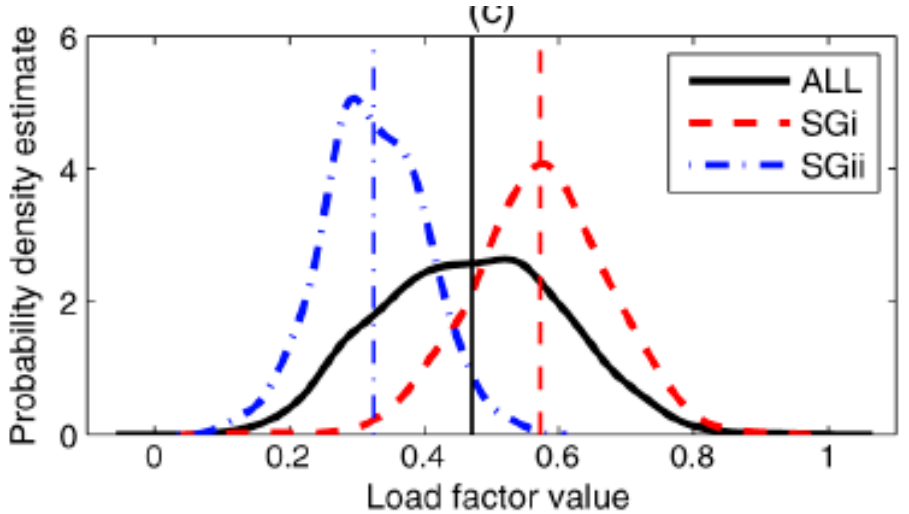
- Os seletores usados para encontrar os subgrupos foram:
  - Média do consumo em cada ponto de medição (numérico): $c_{t_1} \ldots c_{t_{48}}$
  - Variáveis sócio-demográficas (numérico, categórico)
- Um dos alvos usados no estudo é o fator de carga médio diário
  - $\bar{l} = \frac{\sum c_t}{\max c_t \times 48} = \frac{mean\ c_t}{\max c_t}$
- Residências com fator de carga alto possuem um padrão de consumo mais constante, enquanto residências com fator de carga baixo indicam residências com picos de consumo
- Uma das hipóteses investigadas no trabalho foi avaliar a possibilidade o consumo nos pontos de medição para determinar o tipo de consumo
- Medida de qualidade CWRAcc
  - $CWRAcc = \frac{n}{N} \int |f(x) - f_{sg}(x)| dx$, $f(x)$ é a função de densidade de probabilidade na população e $f_{sg}(x)$ é a fdp no subgrupo

Perfis de consumo de energia elétrica

	Condition	Nr. of samples	Group's mean load factor, $\bar{l}$	$\delta^{\bar{l}}$
ALL	All households	4779	0.47	–
SGi	$c_{t_4} \geq 0.11$ AND $c_{t_{14}} \leq 0.51$ AND $c_{t_{35}} \leq 0.74$	981	0.57	21%
SGii	$c_{t_3} \leq 0.14$ AND $c_{t_{15}} \geq 0.68$ AND $c_{t_{22}} \leq 0.58$	561	0.33	−30%







Figuras retiradas de Nanjin et al. (2014)

Aprendizado em ambientes virtuais

- Romero et al. (2009) investigaram o uso de descoberta de subgrupos para avaliar a contribuição do Moodle no processo de aprendizado dos alunos da Universidade de Córdoba, Espanha
- Eles usaram dados de 5 disciplinas com maior uso da plataforma para avaliar a contribuição no processo de aprendizado
- A nota foi discretizada em conceitos
- A tabela mostra as variáveis usadas

Name	Description	Type
course	Identification of the course	Discrete
nAssigment	Number of assignments completed	Continuous
nAssigmentP	Number of assignments passed	Continuous
nAssigmentF	Number of assignments failed	Continuous
nQuizz	Number of quizzes completed	Continuous
nQuizzP	Number of quizzes passed	Continuous
nQuizzF	Number of quizzes failed	Continuous
nMessagesC	Number of messages sent to the chat	Continuous
nMessagesT	Number of messages sent to the teacher	Continuous
nMessagesF	Number of messages sent to the forum	Continuous
nRead	Number of forum messages read	Continuous

Figura retirada de Carmona e Elizondo (2011)

Aprendizado em ambientes virtuais

IF course = C110 AND nAssignment = High AND nPosts = High THEN mark = Good (Accuracy: 0.9285, Significance: 6.5348, Coverage:0.1575).

This rule shows that in the ProjectManagement (C110) course, the students who have completed a high number of assignments and sent a lot of messages to the forum, have also obtained good marks. The teacher must continue to promote these types of activity in this course because of their effectiveness for the students in the final mark obtained.

IF course = C29 AND nMessagesT = Very low THEN mark = Fail (Accuracy: 0.8560, Significance: 59.1774, Coverage: 0.2520).

In the AppliedComputerScienceBasis (C29) course, most of the students who have sent a very low number of messages to the teacher have failed. Using this information, the teacher can direct more attention to these students because they have a higher probability of failing.

Figuras retiradas de Carmona e Elizondo (2011)

Aprendizado em ambientes virtuais

IF course = C110 OR C88 AND nPosts = High OR Very High AND nQuizP = Medium OR High OR Very High THEN mark = Good (Accuracy: 0.7382, Significance: 43.4771, Coverage: 0.2431).

This rule shows that if the students of the course Project- Management (C110) or Comput-erScienceBasis (C88) have sent a high or very high number of messages to the forum, and they have also obtained a medium, high or very high score in the quizzes, then they obtain good marks.

IF course = C29 OR C110 OR C111 AND nAssignmentF = Very High OR High OR Medium AND nQuizF = Very High OR High OR Medium AND nMessagesT = Very low OR Low THEN mark = Fail (Accuracy: 0.8667, Significance: 61.8034, Coverage: 0.4726).

This rule shows that if the students of the course ProgrammingForEngineers (C29) or ProjectManagement (C110) or ComputerScienceBasis (C88) have failed in a very high, high or medium number of assignments, have failed in a very high, high or medium number of quizzes, and have sent a very low or low number of messages to the teacher, then they have obtained a fail in their final marks.

Figuras retiradas de Carmona e Elizondo (2011)

Prevenção de lesões em atletas profissionais

- de Leeuw et al. (2022) apresentaram uma aplicação de descoberta de subgrupos para prevenção de lesões em atletas profissionais de vôlei masculino
- Eles avaliaram a rotina de treinamentos de 10 atletas profissionais que participaram de competições internacionais durante 24 semanas
- O risco de lesão foi avaliado através do questionário Oslo Sports Trauma Research Center (OSTRC)
- Foram avaliados quatro fatores
  - Dificuldade de participação em treinos e competições
  - Redução no volume de treinamento
  - Desempenho afetado
  - Apresentação de sintomas ou incômodo

Prevenção de lesões em atletas profissionais

- Esse questionário é amplamente usado para avaliar o risco de lesões
- O resultado desse questionário foi normalizado numa escala de 0-100
- Além disso, foram anotados diariamente a rotina de treinamento dos atletas
  - Tipo de exercício: fortalecimento muscular superior, inferior ou completa
  - Peso usado
  - Número de repetições
- A percepção do bem-estar foi avaliada com base em questionário que os atletas respondiam diariamente antes do café da manhã
- Finalmente, no caso de treinos técnicos e táticos, sensores usados pelos atletas contavam o número e mediam a altura de saltos executados durante o treino

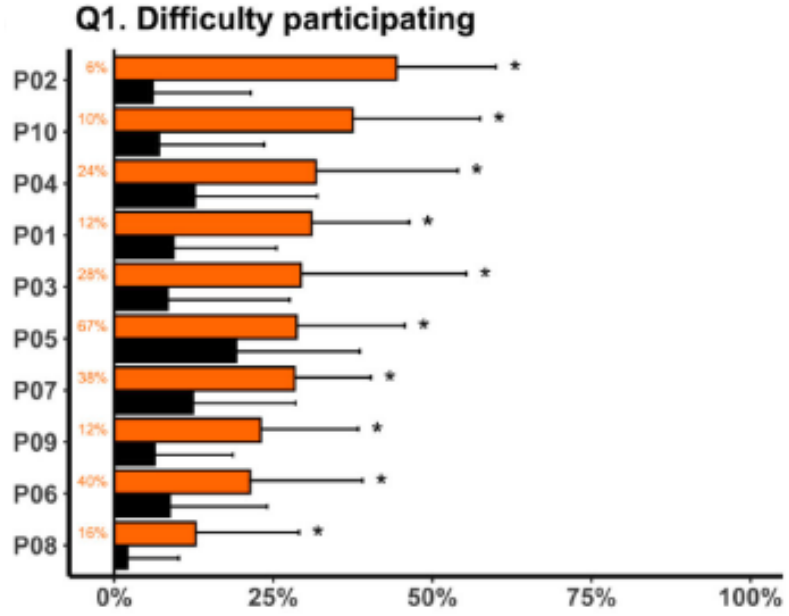
Prevenção de lesões em atletas profissionais

- Eles usaram seletores que agregavam o valor coletado nas rotinas de treinamento para descobrir subgrupos com distribuições não usuais para as variáveis do OSTRC
  - As funções de agregação foram: média, desvio-padrão, primeiro quartil, terceiro quartil e soma
  - Os tempos de agregação foram: 7, 14 e 28 dias para avaliar efeitos de curto, médio e longo prazo
- A medida de qualidade é baseado no coeficiente R^2 em modelos de regressão
  - Avaliam se o subgrupo é capaz de explicar a variância

Prevenção de lesões em atletas profissionais

- Atleta P02 reportou desconforto ~45% mais vezes quando realizou mais de 196 saltos nas últimas duas semanas
  - O atleta era líbero

Player	Window	Condition	Subgroup size
P01	14 days	Stand. dev. number of jumps ≥ 86.8[a]	28%
P02	14 days	Total number of jumps ≥ 196[a]	6%
P03	28 days	Stand. dev. daily number of jumps ≤ 22.2[a]	28%
P04	14 days	Average jump height ≥ 54.7 cm	24%
P05	28 days	Third quartile number of jumps ≥ 65.75	42%
P06	28 days	First quartile of daily mood scores ≥ 8	40%
P07	28 days	Stand. dev. daily number of high jumps ≥ 1.70	38%
P08	14 days	Average daily sleep duration ≤ 7.11 h	16%
P09	14 days	Average jump height ≤ 48.3 cm	12%
P10	14 days	Stand. dev. weight percentage upper body exercises ≥ 0.08	10%



Figuras retiradas de de Leeuw et al. (2022)

Leitura

- de Leeuw, A. W., van der Zwaard, S., van Baar, R., & Knobbe, A. (2021). Personalized machine learning approach to injury monitoring in elite volleyball players. *European Journal of Sport Science*, *22*(4), 511–520. https://doi.org/10.1080/17461391.2021.1887369
- Carmona, C. J., & Elizondo, D. (2011). *Subgroup Discovery: Real-World Applications*. Techincal Report.
- Romero, C., González, P., Ventura, S., Del Jesús, M. J., & Herrera, F. (2009). Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data. *Expert Systems with Applications*, *36*(2), 1632-1644.
- Jin, N., Flach, P., Wilcox, T., Sellman, R., Thumim, J., & Knobbe, A. (2014). Subgroup discovery in smart electricity meter data. *IEEE Transactions on Industrial Informatics*, *10*(2), 1327-1336.

Aprendizado Descritivo

Aula 15 – Aplicações de descoberta de subgrupos

Professor Renato Vimieiro

DCC/ICEx/UFMG