Aprendizado Descritivo

Aula 10 – Algoritmos exaustivos para descoberta de subgrupos

Professor Renato Vimieiro

DCC/ICEx/UFMG

Introdução

- Como discutimos anteriormente, descoberta de subgrupos se tornou uma das principais tarefas de mineração de padrões discriminativos
- A ideia central dessa tarefa é encontrar padrões caracterizando subgrupos de transações em que a propriedade de interesse possui uma distribuição não usual com respeito à população
- Existem diferentes abordagens conforme a propriedade de interesse
  - O mais comum é identificar grupos com distribuições não usuais de alvos categóricos (classes); especialmente em problemas com duas classes.
  - Mas existem também métodos para alvos numéricos
- Hoje veremos um método exaustivo para identificação de subgrupos de interesse que pode ser usado em ambos os tipos de alvo

SD-Map

- O algoritmo SD-Map foi proposto por Atzmueller e Puppe em 2006.
- O SD-Map é um método exaustivo baseado no FP-Growth
- Ao contrário de outras abordagens anteriores baseadas em algoritmos de mineração de itemsets frequentes, esse é um dos poucos que ainda possui implementações em ferramentas atuais para descoberta de subgrupos (PySubgroup)
- Em sua versão inicial, o algoritmo foi desenvolvido especificamente para o caso de rótulos binários
  - Posteriormente, foi estendido para acomodar outros atributos alvo
- A modificação inicial proposta pelos autores é bastante simples, porém, antes de apresentá-la, vamos discutir a formalização da tarefa adotada por eles

Conceitos básicos

- Os autores assumem que a base de dados de entrada é composta somente por atributos categóricos
    - Atributos numéricos precisam ser discretizados na etapa de pré-processamento
- Assim, eles definem que a base possui um conjunto de n atributos $\mathcal{A}$
- Cada atributo $a \in \mathcal{A}$ possui um domínio de valores possíveis $dom(a)$
- Um objeto na base é descrito por valores associados a cada um dos n atributos
    - Um objeto $o = (a_1 = v_1, a_2 = v_2, \dots, a_n = v_n)$, em que $v_i \in dom(a_i)$ para todo a_i

Conceitos básicos

- Como dito, no caso específico do SD-Map, os autores consideram a variável alvo como um rótulo binário (classe positiva e negativa)
- Assim, a função $t: O \rightarrow \{+, -\}$ indica o rótulo (o valor da variável alvo) para um objeto o
- A linguagem de descrição dos subgrupos no SD-Map é composta por conjunto de predicados construídos com os atributos de $\mathcal{A}$
  - Uma descrição $d = \{e_1, e_2, \ldots, e_k\}$, em que $e_i = (a_i, V_i), a_i \in \mathcal{A}, V_i \subseteq dom(a_i)$ e para todo i $\neq$ j, $e_i = (a_i, V_i), e_j = \left(a_j, V_j\right), a_i \neq a_j$
  - Os predicados são seletores de valor para os diferentes atributos
  - Uma descrição deve impor uma única restrição de valores, no máximo, para cada atributo

Conceitos básicos

- A linguagem de descrição dos subgrupos é, portanto, o conjunto de todas as descrições possíveis de se construir com as regras anteriores
  - Ela é denotada por Σ
- A linguagem Σ define o espaço de busca do algoritmo
- Os autores assumem que ela é pré-definida pelo usuário
  - Ou seja, ela é considerada um parâmetro do algoritmo e não é computada por ele
  - Alternativamente, pode-se assumir uma linguagem padrão e usá-la na descoberta dos padrões
  - Eles assumem dois casos padrão:
    - seletores do tipo $a = v \equiv a \in \{v\}$
    - seletores com restrição de conjuntos de valores como anteriormente

Conceitos básicos

- Cada seletor/predicado atua como uma função indicadora que, para todo objeto o, informa se o satisfaz ou não a função
  - Um objeto satisfaz um seletor $e_i = (a_i, V_i)$ se o valor do atributo a_i para o objeto o pertence ao conjunto V_i
- Uma descrição, portanto, deve ser lida como a conjunção dos seus seletores
  - Um objeto o é descrito por $d \in \Sigma$ se ele satisfaz cada um dos $e_i \in d$
  - $d \equiv e_1 \wedge e_2 \wedge \cdots \wedge e_k$
- Em outras palavras, cada seletor é visto como um item, e a descrição um itemset na mineração de padrões frequentes

SD-Map

- Como foi dito, o SD-Map é uma adaptação do FP-Growth para descoberta de subgrupos
- O SD-Map adapta o filtro por suporte, e elimina subgrupos não frequentes num rótulo de interesse
- Considerando que descrições de subgrupos podem ser vistas como itemsets, seletores como itens, e desejamos usar o FP-Growth, o primeiro passo do algoritmo é transformar a base de dados categórica em uma base de dados binária em que os itens são os seletores
- Se os seletores forem do tipo $a = v$, então o algoritmo FP-Growth pode ser usado praticamente em sua forma original
    - Os nós da árvore precisam acomodar a contagem de objetos em cada classe
    - Ao final, os padrões são filtrados para retornar os k melhores conforme a medida de qualidade

SD-Map

- Se os seletores contiverem mais que um valor, então modificações adicionais precisam ser feitas
- Veja que um seletor $e_i = (a_i, V_i)$ em que $|V_i| = q$ é equivalente à expressão $e_{i1} = (a_i, V_{i1}) \vee e_{i2} = (a_i, V_{i2}) \vee \cdots \vee e_{iq} = \left(a_i, V_{iq}\right)$
    - Os seletores são disjunções de seletores de valores únicos
- Dessa forma, se todos os possíveis seletores forem considerados, haverá sobreposição entre alguns, podendo levar a FP-trees muito maiores e padrões redundantes

SD-Map

- Uma abordagem ingênua para resolver o problema poderia ser:
  - Descartar nós que representem seletores usados para condicionar a FP-Tree; se o seletor condicionante foi usado antes, o nó é um subconjunto dele
  - Quando for gerar os padrões através da combinação dos nós (último passo do FP-Growth quando a árvore é um caminho), considerar apenas combinações de seletores sobre conjuntos de valores disjuntos
- Alternativamente, pode-se introduzir o conceito de negação de seletores somente durante a fase de descoberta de padrões
  - Por De Morgan, um seletor contendo disjunções pode ser representado por um equivalente envolvendo conjunções de negações
- Se $dom(a_i) = \{v_1, v_2, v_3, v_4\}$ e tivermos um seletor $e_i = (a_i, \{v_1, v_2\})$, então o mesmo é equivalente a $\neg(a_i, \{v_3\}) \wedge \neg(a_i, \{v_4\}) \equiv (a_i, \neg\{v_3\}) \wedge (a_i, \neg\{v_4\})$
  - A semântica da árvore continua sendo a mesma, de conjunção de itens
  - Pode-se incluir um seletor negativo para valores faltantes, levando a uma maior precisão das medidas de qualidade

Extensões do SD-Map

- Apesar dos autores demonstrarem que o SD-Map alcançou um bom desempenho comparado a outras abordagens exaustivas, pode-se observar que não há nenhuma poda no espaço de busca com respeito à medida de qualidade empregada
  - Ou seja, a única modificação com respeito a mineração de itemsets frequentes tradicional é a possibilidade de computar a função de qualidade junto com a descoberta dos padrões.
  - A filtragem é feita em um pós-processamento
- Posteriormente, observou-se que uma abordagem branch-and-bound poderia ser usada, podando ramos da FP-Tree com estimativas de qualidade menores que um limiar definido pelos k melhores padrões encontrados até o momento

Estimativas otimistas

- Grosskreutz et al. definiram o conceito de estimativas otimistas justas (*tight optimistic estimates*) para funções de qualidade de subgrupos
- Uma função oe é uma estimativa otimista justa para uma função de qualidade q se $\forall_{s,s' \in \Sigma} s \preccurlyeq s' \rightarrow oe(s) \geq q(s')$
- Em outras palavras, toda especialização de um subgrupo s deve ter qualidade limitada pela estimativa de qualidade de s
- Dessa forma, podemos acrescentar a estimativa de qualidade aos nós da árvore e usá-la durante a mineração dos padrões para efetuar podas

Estimativas otimistas

- Considere a medida de qualidade Piatetsky-Shapiro definida por:
  - $PS(X) = n[P(+|X) - P(+)]$, onde n é o tamanho (absoluto) do subgrupo
- É fácil perceber que PS é compatível com o WRAcc
  - A única diferença está no fato dela usar o tamanho do subgrupo ao invés do seu suporte
- Teorema: A função $oe(X) = nP(+|X)[1 - P(+)]$ é uma estimativa otimista justa para PS
- Prova (intuição):

Um subgrupo atingiria o máximo de qualidade se ele cobrisse somente objetos positivos. Logo, a melhor estimativa de qualidade tem de levar em conta esse fator. Um subgrupo 'puro' pode ser obtido a partir de X eliminando os falsos positivos (e alguns verdadeiros positivos). O tamanho máximo que esse subgrupo pode alcançar é $nP(+|X)$. ■

SD-Map*

- Atzmueller e Lemmerich (2009) propuseram uma modificação ao SD-Map para acomodar podas por estimativa de qualidade
  - O algoritmo foi chamado de SD-Map*
- Para computar a estimativa de qualidade, precisamos armazenar nos nós somente o tamanho do subgrupo (frequência) e o número de objetos da classe positiva
  - Os demais valores são constantes para um conjunto de dados
- Os autores adicionaram modificações ao algoritmo original

SD-Map*

- <u>Poda 1</u>: durante a construção das árvores condicionais, ramos condicionados com estimativas abaixo do pior subgrupo encontrado até o momento podem ser descartados
- <u>Poda 2</u>: durante a construção das árvores condicionais, nós (seletores) com estimativas abaixo do mínimo podem ser eliminados (tal como era feito com o suporte)
- <u>Reordenação</u>: os seletores podem ser reordenados dinamicamente para que os mais promissores sejam avaliados primeiro quando as árvores forem construídas
- Os valores armazenados nos nós variam conforme a medida de qualidade e estimativa usada, devendo ser ajustada de acordo. A estrutura da árvore permanece a mesma

Atributos alvo numéricos

- Os autores do SD-Map* apresentaram, ainda, a possibilidade de usá-lo para encontrar subgrupos com atributos alvo numéricos
- Existem diversas medidas de qualidade para alvos numéricos, dentre elas uma adaptação da PS
  - $PS_c(X) = n[\mu_X - \mu]$ sendo μ_X a média do alvo no subgrupo, e μ a média na população

Atributos alvo numéricos

- Teorema: A função $oe(X) = \sum_{o \in X, t(o) > 0} t(o) - \mu$ é uma estimativa otimista justa para PS_c
- Prova:

$$PS_c(X) = n\left[\frac{\sum_{o \in X} t(o)}{n} - \frac{n \cdot \mu}{n}\right]$$

$$= \sum_{o \in X} t(o) - n \cdot \mu = \sum_{o \in X} (t(o) - \mu)$$

Como a função otimista considera somente objetos com valores alvo positivos, a qualidade de qualquer subgrupo será no máximo esse valor

■

Atributos alvo numéricos

- Veja que o caso binário pode ser tratado como um caso particular em que positivo => t(o)=1, e negativo => t(o)=0
- Os mesmos autores publicaram em 2016 uma versão estendida do artigo onde eles apresentam estimativas para diversas outras medidas de qualidade
- Eles também apresentam um segundo algoritmo baseado em bitsets para situações em que a FP-Tree possuía desempenho ruim

Leitura

- Atzmueller, M., Puppe, F. (2006). SD-Map – A Fast Algorithm for Exhaustive Subgroup Discovery. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds) Knowledge Discovery in Databases: PKDD 2006. PKDD 2006. Lecture Notes in Computer Science(), vol 4213. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11871637_6
- Atzmueller, M., Lemmerich, F. (2009). Fast Subgroup Discovery for Continuous Target Concepts. In: Rauch, J., Raś, Z.W., Berka, P., Elomaa, T. (eds) Foundations of Intelligent Systems. ISMIS 2009. Lecture Notes in Computer Science(), vol 5722. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04125-9_7
- Grosskreutz, H., Rüping, S., Wrobel, S. (2008). Tight Optimistic Estimates for Fast Subgroup Discovery. In: Daelemans, W., Goethals, B., Morik, K. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008. Lecture Notes in Computer Science(), vol 5211. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-87479-9_47
- Lemmerich, F., Atzmueller, M. & Puppe, F. Fast exhaustive subgroup discovery with numerical target concepts. Data Min Knowl Disc 30, 711–762 (2016). https://doi.org/10.1007/s10618-015-0436-8

Aprendizado Descritivo

Aula 10 – Algoritmos exaustivos para descoberta de subgrupos

Professor Renato Vimieiro

DCC/ICEx/UFMG