Levantamento de Conceitos e Métodos de Seleção de Features Baseados em Teoria da Informação
Teoria da Informação

Gabriel R. Martins[1], **André L. M. Dutra**[2]

[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

{garoma20, andre}@dcc.ufmg.br

Abstract. *Feature selection involves selecting a subset of features from a dataset to train and use in machine learning models, aiming to reduce data dimensionality. As such, techniques that maximize relevance, minimize redundancy, and maximize conditional redundancy stand out. Information-theoretic approaches use mutual information to calculate these properties, combining them into an objective function maximized in a feature space search. Among existing approaches, JMI, which linearly combines average relevance, redundancy, and conditional redundancy, stands out as the most promising and stable.*

Resumo. *Feature selection consiste em selecionar um subconjunto de features de um dataset para o treino e uso em modelos de aprendizado de máquina, visando reduzir a dimensionalidade dos dados. Nesse contexto, destacam-se técnicas que maximizam a relevância, minimizam a redundância e maximizam a redundância condicional das features. Abordagens baseadas em teoria da informação usam o conceito de informação mútua para calcular essas propriedades, combinando-as em uma função objetivo maximizada em uma busca em espaço de features. Dentre as abordagens existentes, a JMI, que combina linearmente os valores médios de relevância, redundância e redundância condicional, se destaca como a mais promissora e estável.*

1. Feature Selection

Feature Selection é uma técnica comum no contexto de aprendizado de máquina que consiste na seleção de um subconjunto dos atributos de uma base de dados utilizando alguma métrica para selecionar os atributos mais relevantes para um problema [Cai et al. 2018]. Com o grande avanço da tecnologia moderna houve um crescimento na quantidade de dados gerados, com isso, surgiu a necessidade de uma ferramenta que pudesse filtrar dados de interesse dentre todos os dados coletados. Essa técnica que surgiu como solução para esse problema é *feature selection*, e, atualmente, ela é utilizada principalmente no contexto de aprendizado de máquina, no qual o pré-processamento dos dados é uma tarefa relevante que pode afetar a qualidade do modelo a ser criado. Dentre as vantagens do uso dessa técnica estão a melhora na taxa de aprendizado do modelo, a diminuição do tempo gasto para o treinamento e a simplificação dos resultados obtidos. Isso é possível porque com uma quantidade de dados menores o tempo gasto para seu processamento é menor naturalmente, porém não há necessariamente perda de informação ao reduzir o número de atributos dado que os mais relevantes tendem a se manter. Além disso, com

menos atributos o resultado do modelo se torna mais compreensível, porque é possível distinguir mais claramente qual o impacto de cada atributo na solução do problema bem como mapear esse impacto no significado real para cada atributo.

1.1. Modelos de *Feature Selection*

Existem diversas técnicas para realizar *feature selection*, porém, elas tendem a seguir um dentre 3 padrões: modelos do tipo *filter*, modelos de tipo *wrapper* ou modelos do tipo *embedded*. Os modelos do tipo *filter* são modelos mais simples e rápidos onde, em geral, se calcula um *score* para cada atributo utilizando de alguma métrica e os atributos com maior *score* são selecionados como subconjunto relevante para o modelo. É uma maneira rápida de selecionar um subconjunto relevante, porém pode ignorar a correlação entre os atributos de um modelo, podendo perder informações úteis. Já os modelos tipo *wrapper* são mais custosos, eles funcionam testando os vários subconjuntos possíveis de atributos no modelo, realizando o treinamento para cada subconjunto testado e avaliando seus resultados. Portanto, é um processo mais caro se comparado aos modelos *filter*, porém tende a ter resultados melhores para a criação de subconjuntos, porque testa o desempenho deles diretamente no modelo. Vale ressaltar que a quantidade de subconjuntos de atributos pode ser muito grande tornando esta etapa muito custosa, sendo assim, normalmente utiliza-se de algum algoritmo de busca em espaço de estados com o intuito de encontrar o subconjunto com melhor desempenho sem ter que testar todos os subconjuntos possíveis. Enfim, os modelos do tipo *embedded* são um meio termo entre os dois modelos apresentados anteriormente. Neste caso, a seleção do subconjunto mais relevante para o modelo é uma tarefa atribuída diretamente ao modelo, ou seja, o próprio modelo decide a cada iteração quais sãos os atributos mais relevantes para a resolução do problema, podendo alterar ao longo das iterações com intuito de selecionar ao fim o subconjunto que apresentou maior relevância naquele contexto.

1.2. Feature importance

Neste contexto de *feature selection* uma importante tarefa é como definir quais atributos são mais relevantes para resolução de um problema de aprendizado de máquina. Surge assim a definição de *Feature Importance* que é exatamente uma métrica que mede a contribuição de cada um dos atributos para a resolução de um problema, métrica essa que pode ser calculada de diversas formas distintas. Como resultado do uso de uma métrica como essa, os modelos tendem a se tornar mais compreensíveis e explicáveis[Rengasamy et al. 2022], o que é importante em aplicações multidisciplinares, como por exemplo aplicações médicas, onde diversas áreas são responsáveis por interpretar e aceitar o resultado de um modelo.

1.3. Maldição da Dimensionalidade

Maldição da dimensionalidade é a expressão dada ao fenômeno que relaciona o aumento no número de atributos de uma base de dados com o baixo desempenho de um modelo de aprendizado de máquina [Altman 2018]. Isso acontece porque quanto maior o número de atributos para um dado maior a quantidade de instâncias únicas possíveis para aquele dado, considerando um número fixo de instâncias em uma base de dados, se o número de atributos aumenta, os dados tendem a ficar esparsos e representar uma pequena parcela da

quantidade possível de instâncias. Como consequência disso, modelos com muitos atributos tendem a exigir uma grande quantidade de dados para evitar *overfitting* do modelo, fazendo com que *feature selection* seja uma técnica necessária nesse contexto.

2. Propriedades Objetivo em Feature Selection

Embora uma variedade de abordagens, baseadas ou não em teoria da informação, existam para o problema de *feature selection*, observa-se em grande parte das técnicas que as features finais de uma seleção possuem, em geral, uma série de propriedades que as tornam mais ou menos desejáveis em relação às demais features [Li et al. 2017]. Desse modo, todo algoritmo de seleção de features se resume a, diretamente ou não, selecionar o conjunto de features que maximiza a combinação destas propriedades. Esta seção se reserva a descrever estas propriedades, e de que maneira elas se relacionam ao formato dos dados de um dataset.

Como esta seção aborda propriedades gerais de features, o conceito de ”correlação”será abordado de maneira generalizada para abranger diferentes técnicas utilizando diferentes métricas de correlação. Portanto, nesta seção uma métrica de correlação se refere a qualquer função $f_c : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{R}$ que mapeie um par de distribuições quaisquer (discretas ou contínuas) a um valor real quantificando o grau de correlação entre elas, possuindo todas as propriedades de uma função de similaridade.

2.1. Relevância

Dado um dataset composto de n features $X_1, X_2, \cdots, X_n$ e um target Y, a relevância de uma feature X_i em relação ao target é definida como $rel(X_i, Y) = f_c(X_i, Y)$. No contexto de feature selection, deseja-se encontrar o conjunto S que maximize a soma das relevâncias de cada feature. A relevância ocorre quando uma feature é fortemente correlacionada ao target, carregando maior capacidade de previsão do target em um contexto de classificação.

2.2. Redundância

Dado um dataset composto de n features $X_1, X_2, \cdots, X_n$ e um conjunto de features selecionadas S, a redundância de uma feature X_i em relação a S é definida como $red(X_i, S) = \sum_{X_j \in S} f_c(X_i, X_j)$. No contexto de feature selection, deseja-se encontrar o conjunto S que minimize a soma das redundâncias entre as features. A redundância ocorre quando duas features fortemente correlacionadas carregam a mesma informação, carregando a mesma capacidade de inferência sobre o target. Isso aumenta a dimensionalidade dos dados sem agregar valor preditivo, o que é negativo considerando a maldição da dimensionalidade, além de que a informação compartilhada recebe peso duplo no modelo, o que pode desbalanceá-lo.

2.3. Redundância Condicional

Dado um dataset composto de n features $X_1, X_2, \cdots, X_n$, um target Y e um conjunto de features selecionadas S, a redundância de uma feature X_i em relação a S condicional ao target Y é definida como $red_{cond}(X_i, S|Y) = \sum_{X_j \in S} f_c(X_i, X_j|Y)$. A redundância condicional possui em si uma série de peculiaridades. Primeiramente, ela requer que a métrica de correlação admita o cálculo de correlação condicional entre variáveis, o que

nem sempre é o caso. Segundo que, no contexto de feature selection, deseja-se encontrar o conjunto S que maximize o somatório das redundâncias de cada feature, o que soa contraintuitivo em comparação às conclusões tomadas no item anterior. A intuição por trás da redundância condicional se dá no fato de que, dado que o valor do target é conhecido, é esperado que features fortemente correlacionadas a ele se correlacionem, mesmo que esta correlação não ocorra num geral, pois ambas estão causalmente relacionadas ao target. Este é o princípio por trás do Paradoxo de Simpson, que descreve variáveis que apresentam uma determinada correlação ao terem seus dados avaliados em grupos (condicionalmente), mas sua correlação se dissipa ou se inverte quando seus dados são avaliados no todo [Blyth 1972].

3. Teoria da Informação Aplicada a Feature Selection

Revisões literárias sobre técnicas de feature selection [Li et al. 2017], especialmente as baseadas em teoria da informação [Brown et al. 2012], mostram que a maioria das abordagens relevantes para o problema envolvem uma busca no espaço de estados de conjuntos de features, otimizando uma função-objetivo que combina termos que quantificam as propriedades mencionadas anteriormente. Este paradgima é utilizado porque o problema de selecionar o conjunto ótimo S de features é NP-difícil.

Neste contexto, técnicas baseadas em teoria da informação usam a informação mútua de Shannon como métrica de correlação entre variáveis. A relevância é calculada como a informação mútua entre a feature e o target $I(X_i; Y)$, a redundância entre pares de features como sua informação mútua $I(X_i; X_j)$, e a redundância condicional como a informação mútua condicional $I(X_i; X_j|Y)$. O uso da informação mútua em aprendizado segue de uma derivação do uso da entropia de Shannon, cujo uso como função de perda já é consolidado em diversos modelos estado da arte [Bridle 1990]. Como a entropia do target é a função a ser minimizada, a feature que minimiza condicionalmente essa entropia, e que portanto tem maior relevância e informação mútua ao target, é a que mais contribui para a previsão do target. Assim, a informação mútua é uma métrica de correlação adequada para o cálculo da relevância, o que se generaliza para as demais propriedades.

3.1. Técnicas de Feature Selection Baseadas em Informação Mútua

Após realizar uma revisão da literatura sobre métodos de feature selection baseados em teoria da informação, Brown et al. criaram um framework que generaliza todas as funções-objetivo baseadas em informação mútua da literatura para a busca em espaço de features: $J_{CMI}(X_k) = I(X_k; y) + \sum_{X_j \in S} g[I(X_j; X_k), I(X_j; X_k|Y)]$[Brown et al. 2012].

Onde a função g mede o ganho sobre a redundância e redundância condicional de cada feature em relação à feature X_k. Quanto maior o valor de $J_{CMI}(X_k)$, melhor é o conjunto de features $S \cup \{X_k\}$. De acordo com o artigo, todas as técnicas na literatura seguem essa fórmula, combinando relevância com uma soma de ganhos sobre a redundância geral e condicional, variando apenas a função de ganho. Por exemplo, a técnica mais simples, MIM [Lewis 1992], que maximiza apenas a relevância, usa $g(a, b) = 0$. A técnica mais utilizada, mRMR [Peng et al. 2005], usa $g(a, b) = -a/|S|$, calculando a relevância menos a redundância média. A técnica com melhor desempenho geral e equilíbrio entre performance e estabilidade, JMI [Yang and Moody 1999], define $g(a, b) = (b - a)/|S|$,

somando a relevância e a média das redundâncias condicionais e subtraindo a média das redundâncias, funcionando como uma versão aprimorada da mRMR.

3.2. Seleção de Features Contínuas Baseada em Informação Mútua

Embora técnicas de feature selection sejam o estado da arte para seleção de features, elas funcionam apenas com variáveis discretas. Para features contínuas, a principal abordagem baseada em teoria da informação é particionar (binning) os dados, transformando variáveis contínuas em discretas. Isso permite, inclusive, a análise de datasets mistos sem segregar variáveis. O Maximal Information Coefficient (MIC), criado por Reshef et al. [Reshef et al. 2011], é uma heurística que encontra o número de bins que maximiza a informação mútua entre pares de variáveis contínuas. O MIC e suas variações são a principal abordagem para seleção de features contínuas, sendo usados também em outros contextos, como a importância de features.

4. Busca em Espaço de Estados

O problema de busca em espaço de estados consiste em 3 atributos básicos: um conjunto de estados, um estado inicial e um conjunto de estados alvo. A ideia é produzir um algoritmo capaz de encontrar um estado alvo a partir do estado inicial passando por estados do conjunto de estados.Esse problema é np-difícil, porém existem duas maneiras de abordá-lo, com algoritmos de busca sem informação ou com informação. No contexto de *feature selection*, a ideia é buscar um subcojunto de atributos ótimo para o modelo dentre todos os possíveis, para isso, a heurística gulosa normalmente é utilizada. Essa heurística é o um dos métodos de busca informada mais simples na qual, a cada passo na busca, se escolhe o atributo que maximiza uma métrica específica. Em modelos de *feature selection* baseados em teoria de informação, utiliza-se, normalmente, uma das fórmulas para cálculo de *score* dos atributos como equação base para funcionamento dessa heurística.

Referências

Altman, N., K. M. (2018). The curse(s) of dimensionality. *Nat Methods*, 15:399–400.

Blyth, C. R. (1972). On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366.

Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Soulié, F. F. and Hérault, J., editors, *Neurocomputing*, pages 227–236, Berlin, Heidelberg. Springer Berlin Heidelberg.

Brown, G., Pocock, A., Zhao, M.-J., and Luján, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, 13(1):27–66.

Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79.

Lewis, D. D. (1992). Feature selection and feature extraction for text categorization. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2017). Feature selection: A data perspective. *ACM Comput. Surv.*, 50(6).

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.

Rengasamy, D., Mase, J. M., Kumar, A., Rothwell, B., Torres, M. T., Alexander, M. R., Winkler, D. A., and Figueredo, G. P. (2022). Feature importance in machine learning models: A fuzzy information fusion approach. *Neurocomputing*, 511:163–174.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524.

Yang, H. and Moody, J. (1999). Data visualization and feature selection: New algorithms for nongaussian data. *Advances in neural information processing systems*, 12.