



Seleção de Atributos com Otimização Multiobjetivo para melhoria da Previsão de *Churn*

Mariana Dias Pennone

mapennone@gmail.com

Profa. Dra. Priscila Cristina Berbert Rampazzo

Universidade Estadual de Campinas/Faculdade de Ciências Aplicadas

RESUMO

A Seleção de Atributos (SA) tem o objetivo de se reduzir a dimensionalidade dos dados em uma Base de Dados e, assim, melhorar a capacidade de análise destes pelo computador. Neste estudo, a abordagem utilizada para se obter um subconjunto de atributos é a *wrapper*, na qual a interação entre um algoritmo de aprendizado e um algoritmo de otimização será responsável pela seleção e combinação dos atributos mais relevantes para a previsão de *churn* em uma determinada empresa. Como múltiplos critérios estão sendo avaliados durante o processo de otimização da seleção, o problema está sendo abordado por através de um algoritmo de Otimização Multiobjetivo (OM). Considerando que a solução de um problema de OM é constituída por um conjunto de pontos, a aplicação de Algoritmos Evolutivos (AE), especialmente os Algoritmos Genéticos (AG), a este problema parece imediata, pois estes consideram simultaneamente um conjunto de indivíduos para representação das soluções. Além de melhorar o processo de previsão do *churn*, este trabalho busca explicar suas causas.

PALAVRAS CHAVE. Seleção de Atributos, *Machine Learning*, *Churn*, Algoritmo Genético.

1. Introdução

A Seleção de Atributos (SA) é uma área de pesquisa em Aprendizado de Máquina com o objetivo de se reduzir a dimensionalidade dos dados em uma Base de Dados e, assim, melhorar o desempenho computacional exigido para análises e previsões. Com a redução dos atributos, há um aumento do conhecimento acerca das características mais importantes de um problema e a eliminação de dados irrelevantes ou redundantes. Para que esta redução de dimensões ocorra, um subconjunto de atributos relevantes é selecionado na fase de pré-processamento dos dados. Este subconjunto pode ser obtido a partir de diferentes abordagens e, para este projeto, será utilizada a abordagem *wrapper*, onde um algoritmo de aprendizado é usado para guiar o processo de seleção de atributos [Spolaôr et al., 2017]. Pensando na Seleção de Atributos como um processo de busca combinatorial, o objetivo deste projeto é tratá-la como um problema de Otimização Matemática que será resolvido através de um Método Metaheurístico. Como os subconjuntos de atributos serão avaliados considerando múltiplos critérios, a Seleção de Atributos será modelada como um problema de Otimização Multiobjetivo.

Uma das possíveis aplicações destas metodologias é a previsão de *churn* de clientes de prestadoras de serviços. *Churn* é definido como a propensão de clientes cessarem os contratos de negócios com estas empresas em um dado período [Xiao et al., 2015]. A proposição de políticas de retenção de clientes é importante em um mercado competitivo.



2. Otimização Multiobjetivo

Matematicamente, a OM define um problema de minimização¹ como:

Minimizar $[f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_k(\mathbf{x})]$, sujeito a $\mathbf{x} \in \mathbf{X}$.

\mathbf{F} é uma função vetorial representando os k objetivos. Cada componente de \mathbf{F} será denotado por f_i , $i = 1, \dots, k$. \mathbf{X} é um subconjunto de decisões factíveis sobre o qual \mathbf{F} está definida. \mathbf{x} é o vetor de variáveis de decisão. O espaço de busca \mathbf{X} é parcialmente ordenado no sentido de que duas soluções arbitrárias são relacionadas de duas possíveis maneiras: ou uma domina a outra ou nenhuma delas domina [Abraham e Jain, 2005]. Este conceito pode ser formalizado da seguinte forma. Sejam \mathbf{a} e $\mathbf{b} \in \mathbf{X}$, dizemos que \mathbf{a} domina \mathbf{b} se e somente se: $\forall i \in \{1, 2, \dots, k\}$, $f_i(\mathbf{a}) \leq f_i(\mathbf{b})$ e $\exists j \in \{1, 2, \dots, k\}$, $f_j(\mathbf{a}) < f_j(\mathbf{b})$.

Em outras palavras, “ \mathbf{a} não é pior que \mathbf{b} em nenhum dos objetivos e é melhor em pelo menos um”. Neste caso, a solução esperada é composta por um conjunto de pontos de equilíbrio, ou seja, soluções consideradas iguais entre si e superiores em relação às demais. Este conjunto é referido como Conjunto Pareto-ótimo. Formalmente, dizemos que um indivíduo \mathbf{a} é Pareto-ótimo se e somente se \mathbf{a} é não-dominado em relação a \mathbf{X} , i.e, nenhum vetor do espaço de busca domina \mathbf{a} . É importante observar que soluções do Conjunto Pareto-ótimo não podem ser melhoradas em um objetivo sem causar degradação em pelo menos um dos outros. Finalmente, os valores de $\mathbf{F}(\mathbf{x})$ cujo \mathbf{x} está no Conjunto Pareto-ótimo formam a Fronteira de Pareto.

3. Algoritmos Genéticos

Algoritmos Genéticos (AGs) são uma classe de Algoritmos Evolutivos [Bäck et al., 2000], nos quais uma população inicial de soluções para o problema é gerada aleatoriamente e cada indivíduo (solução individual) é avaliado por uma medida de aptidão; em problemas de otimização esta medida pode ser representada pela(s) função(ões)-objetivo. Indivíduos mais aptos são selecionados, sendo capazes de produzir descendentes. Cada descendente é gerado a partir dos operadores genéticos de *crossover* e *mutação*.

O NSGA-II (*Nondominated Sorting Genetic Algorithm*) [Deb et al., 2000], AG proposto na literatura para lidar com problemas com múltiplos objetivos, é caracterizado pela forma como os indivíduos da população são avaliados. A avaliação é feita a partir de dois parâmetros diferentes, *ranking* e *crowding distance*, calculados a partir dos atributos selecionados que compõem um indivíduo.

Inicialmente, o processo de avaliação irá separar os indivíduos em diferentes fronteiras, de acordo com os critérios de dominância (*ranking*). Desta forma, indivíduos com melhores desempenhos nas funções-objetivo são classificados em fronteiras melhores. Vale ressaltar que indivíduos que compõem a mesma fronteira são indiferentes. Em seguida, é calculada a distância de aglomeração (*crowding distance*) [Deb et al., 2000] de cada indivíduo, que considera em seu cálculo a distância de cada indivíduo aos indivíduos mais próximos, a partir de suas funções-objetivo. Esta distância é utilizada como critério de desempate para indivíduos de uma mesma fronteira, já que privilegia pontos extremos ou pontos mais espalhados pelas diferentes fronteiras. Após a população de filhos ser gerada pelos métodos de *crossover* e *mutação*, todos os indivíduos presentes na geração passam pelos operadores de *ranking* e *crowding distance* e são classificados em diferentes fronteiras de acordo com o desempenho de suas funções-objetivo. Somente os indivíduos alocados nas melhores fronteiras seguem para a geração seguinte.

¹Definições para o problema de maximização são análogas. Lembrando que maximizar $f(\mathbf{x})$ equivale a -minimizar $-f(\mathbf{x})$



4. Base de Dados e Modelo de Aprendizado de Máquina

Empresas de serviços telefônicos, provedores de serviços de Internet, TV paga, seguradoras, dentre outros, em geral usam análise *churn* como uma de suas principais métricas de negócios, já que o custo para obter um novo cliente é aproximadamente 5 a 6 vezes maior do que reter um cliente antigo [Xiao et al., 2015]. Dessa forma, a previsão de *churn* de clientes é uma das etapas principais para maximizar o valor dos clientes para a empresa, identificando clientes mais vulneráveis à rotatividade para a aplicação de programas de marketing para retenção.

A base escolhida para as análises de seleção de atributos faz parte do *IBM Sample Data Sets* e está disponível no Kaggle [kag, 2018]. O *data set* simula os dados de um empresa de telecomunicações e inclui informações sobre: (i) Clientes que deixaram o serviço no último mês (*churn*); (ii) Informações sobre o cliente (gênero, senioridade, casado ou não, apresenta dependentes ou não); (iii) Serviços que cada cliente assinou (telefone, múltiplas linhas, internet, segurança online, *backup online*, proteção de dispositivo, suporte técnico e streaming de TV e filmes); (iv) Informações da conta do cliente (há quanto tempo ele é cliente, contrato, forma de pagamento, faturamento sem papel, cobranças mensais e cobranças totais).

Para a identificação e previsão de *churn*, o modelo de previsão escolhido foi a Regressão Logística. A regressão logística é utilizada quando a variável dependente (Y) é de natureza binária e as variáveis independentes podem ser categóricas ou não. Em vez de modelar essa resposta Y diretamente, a regressão logística modela a probabilidade de Y pertencer a uma categoria específica [James et al., 2013]. Uma das medidas de desempenho utilizadas para a otimização no AG foi a métrica de precisão, calculada pela divisão entre as classificações positivas corretas e todas as classificações realizadas pelo modelo.

5. Resultados

Antes de aplicar o método de seleção de atributos usando o Algoritmo Genético implementado, foi realizada a previsão com Regressão Logística utilizando todos os atributos do *data-frame* selecionado. Para a previsão, foram utilizados os módulos já implementados do *Scikit-learn* em Python. A matriz de confusão encontra-se na Figura 1. A partir da matriz confusão obtemos uma acurácia de previsão de 80,26%. Porém, a precisão da previsão de *churn* é de apenas 69%. Desta forma, é a melhoria desta métrica que será buscado no AG, juntamente com a busca de minimização do número de atributos utilizados.

		Previsto	
		Não Churn	Churn
Classe Real	Não Churn	1152	116
	Churn	231	259

Figura 1: Matriz confusão dos resultados obtidos.

O algoritmo foi implementado em Linguagem de Programação *Python*. Como dito anteriormente, com duas funções-objetivo: (1) minimização do número de atributos (colunas da base de dados) utilizadas para a previsão; (2) maximização da precisão na previsão de *churn*.

Para os resultados obtidos, foram realizados testes variando o tamanho da população e a quantidade de gerações entre os valores 100 e 500. Para este resumo, serão apresentados somente os resultados obtidos com o teste de 500 indivíduos e 500 gerações. O primeiro gráfico (Figura 2) mostra a composição das três primeiras fronteiras não-dominadas resultante e o segundo gráfico (Figura 3) diz respeito à frequência relativa (em porcentagem) que cada atributo apareceu nos indivíduos selecionados na primeira fronteira não-dominada.

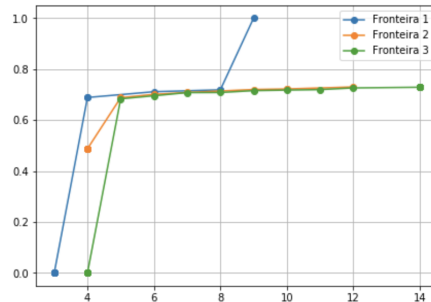


Figura 2: Três primeiras fronteiras não-dominadas para o teste com 500 indivíduos e 500 gerações.

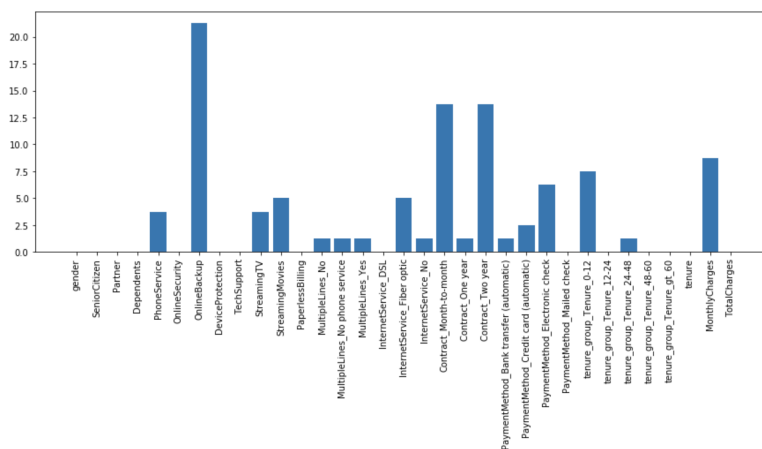


Figura 3: Frequência relativa do aparecimento dos atributos na fronteira não-dominada para o teste com 500 indivíduos e 500 gerações.

A partir de um total de 33 colunas de atributos na base de dados original (após tratamento dos dados), podemos notar que o AG obteve bons resultados de melhoria na precisão da previsão de *churn*. Há casos de precisão de previsão de *churn* melhor que de 69%, valor obtido com a previsão utilizando todos os atributos disponíveis.

O gráfico da Figura 3 apresenta a frequência de utilização dos atributos na fronteira não-dominada do teste de 500 indivíduos e 500 fronteiras com o objetivo de se analisar quais atributos explicam melhor os casos de *churn*. Comparando-se esses resultados com as importâncias de cada atributo calculadas na regressão logística (Figura 4), não é possível chegar a uma conclusão bem definida de quais atributos explicam melhor o processo de *churn* a partir do AG. Podemos perceber que os atributos mais selecionados pelo AG foram Online Backup, contrato de mês a mês e contrato de dois anos. Apesar de essas características apresentarem importâncias relevantes no gráfico da Figura 4, nem todos os atributos de maior importância foram captados.

6. Conclusões

O processo de Seleção de Atributos, realizado através de um Algoritmo Genético Multiobjetivo, pode ser capaz de melhorar o processo de previsão e principalmente, pode conseguir indicar quais atributos explicam o *churn* previsto. Tão importante quanto o processo de previsão, é explicar as causas ou principais responsáveis por este resultado. Além disso, a Seleção de Atributos favorece a capacidade de generalização dos modelos aprendidos, evitando o *overfitting*, e permite reduzir os

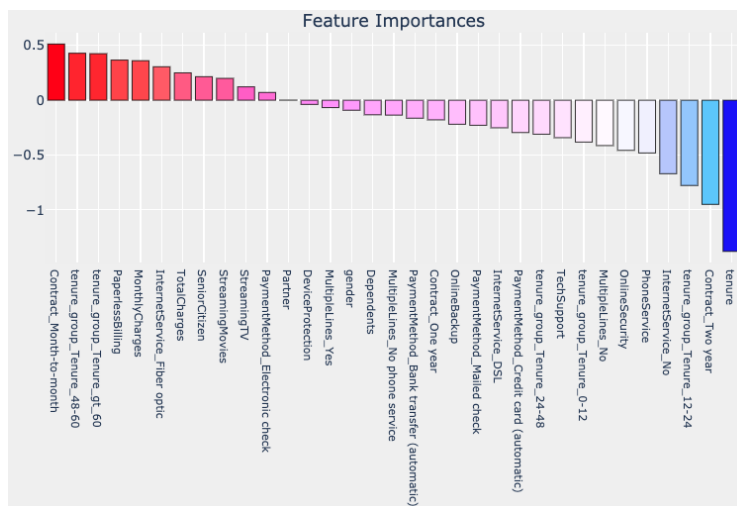


Figura 4: Importância de cada atributo no modelo de regressão logística para a previsão de *churn*.

custos envolvidos no armazenamento dos dados e no treinamento do modelo [Guyon e Elisseeff, 2003]. A pesquisa realizada obteve bons resultados na melhoria da precisão da previsão de *churn*, cabendo ao decisor escolher em qual ponto das fronteiras não-dominadas obtidas ele deseja atuar (*trade-off* entre volume de dados *versus* precisão da previsão).

Referências

- (2018). Kaggle, Machine Learning and Data Science Community. <https://www.kaggle.com/blastchar/telco-customer-churn>.
- Abraham, A. e Jain, L. (2005). *Evolutionary Multiobjective Optimization: Theoretical Advances and Applications*, chapter Evolutionary Multiobjective Optimization, p. 1–6. Springer-Verlag, Londres, 1 edition.
- Bäck, T., Fogel, D. B., e Michalewicz, Z. (2000). *Evolutionary Computation 1: Basic Algorithms and Operators*. Institute of Physics Publishing, Bristol, UK, 1 edition.
- Deb, K., Agrawal, S., Pratab, A., e Meyarivan, T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In *Proceedings of VI Conference Parallel Problem Solving from Nature*, volume 1917, p. 849–858, Paris, France. Springer-Verlag.
- Guyon, I. e Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.
- James, G., Witten, D., e Tibshirani, T. H. R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York.
- Spolaôr, N., Lorena, A. C., e Lee, H. D. (2017). Feature selection via pareto multi-objective genetic algorithms. *Applied Artificial Intelligence*, 31(9-10):764–791.
- Xiao, J., Xiao, Y., Huang, A., Liu, D., e Wang, S. (2015). Feature-selection-based dynamic transfer ensemble model for customer churn prediction. *Knowledge and Information Systems*, 43(1): 29–51.