Predição de ações judiciais de consumo não registrado: uma abordagem para o problema de classes desbalanceadas

Paulo R. Gomes[1], Dayvson W. Almeida[1], Nelia C. Reis[1], João V. Franca[1], Pedro T. Santos[1], José S. Neto[2], Erika W. Alves[2], Milton S. Oliveira[2]

[1]Núcleo de Computação Aplicada – Universidade Federal do Maranhão (UFMA) – São Luís – MA – Brazil

{paulo.rgomes,dayvsonalmeida,nelia.reis,jvitorfranca,thiagocutrim}@nca.ufma.br

[2]Equatorial Energia – São Luís, MA – Brazil

{jose.sobral, erika.assis, milton.oliveira}@equatorialenergia.com.br

Abstract. *Helping companies to predict consumers who might be brought to justice against them is not a simple task. One of the main problems is the amount of data these companies have, allied to the great imbalance between classes. Thus, this paper aims to predict unregistered energy consumption lawsuits for an electric power company that has a database with more than 2 million customers and a ratio of 1:174 imbalance between classes. For this, we propose an incremental method that uses sampling techniques. The results obtained by this method were 97,07 sensitivity.*

Resumo. *Ajudar empresas a prever os consumidores que podem ser levados à justiça contra as mesmas não é uma tarefa simples. Um dos principais problemas é a quantidade de dados dessas empresas, aliada à grande desproporção entre as classes. Assim, o objetivo deste trabalho é prever processos judiciais de consumo de energia não registrados para uma companhia de energia elétrica que tenha uma base de dados com mais de 2 milhões de clientes e uma razão de 1:174 de desequilíbrio entre as classes. Para isso, propomos um método incremental que utiliza técnicas de amostragem. Os resultados obtidos por esse método foram 97,07 de sensibilidade.*

1. Introdução

Atualmente, muitos pesquisadores vêm estudando problemas em grandes volumes de dados, usando os conceitos de mineração de dados e aprendizado de máquina. Podemos definir a mineração de dados como o processo de explorar grandes quantidades de dados procurando padrões consistentes, como regras de associação ou sequências de tempo, detectando assim as relações entre as variáveis, bem como novos subconjuntos de dados [Tan 2018]. Uma maneira de encontrar padrões e associações consistentes em bancos de dados é usar técnicas de aprendizado de máquina. Tais técnicas têm como estudo explorar a construção de algoritmos que podem aprender por um processo de treinamento e, a partir disso, realizar previsões sobre dados [Mohri et al. 2018].

Quando se trata de grandes volumes de dados, há um problema que sempre assombra os pesquisadores: o desequilíbrio das classes. Um conjunto de dados pode ter várias classes e, na maioria das vezes, elas são desequilibradas. Isso pode ocorrer quando

as instâncias de uma classe superam as instâncias de outras classes. A classe sobrecarregada é chamada de classe majoritária, enquanto a outra é chamada de classe minoritária. No entanto, em muitas situações, a classe que possui menos instâncias é a mais importante [Elrahman and Abraham 2013].

Quando a classe de interesse é relativamente rara e tem um pequeno número de instâncias em comparação com a maioria, o problema de desequilíbrio aumenta ainda mais. Além disso, o custo de classificar erroneamente a classe minoritária é muito alto em comparação com o custo de classificar erroneamente a classe majoritária. De acordo com [Elrahman and Abraham 2013], muitas aplicações do mundo real, tais como diagnóstico médico, detecção de fraude (cartão de crédito, telefonema) e sensoriamento remoto sofrem com este problema.

Um exemplo de desequilíbrio na base de dados é visto nos bancos de dados dos clientes de uma empresa de energia. O mercado de energia em vários países está passando por mudanças, basicamente em decorrência da desregulamentação do mercado e do surgimento de mecanismos judiciais e administrativos de proteção ao consumidor [Ibáñez et al. 2006]. Para ajudar empresas de energia a se relacionarem com clientes e criarem serviços de previsão para ações judiciais, modelos baseados em mineração de dados e aprendizado de máquina são sugeridos.

Um dos problemas mais frequentes em ações judiciais é a cobrança do consumo de energia não registrado (CNR), causados quando há a detecção de furto de energia nas instalações conectadas ao cliente. O problema é comum do Grupo Equatorial de Energia, responsável pela distribuição de energia elétrica nos estados de Pará, Maranhão, Piauí e Alagoas, com mais de 5 milhões de clientes no Brasil, distribuídos um vasto território. Embora o problema de CNR seja mais frequente, corresponde apenas a uma razão de 1: 174 do banco de dados, mostrando um enorme desequilíbrio nas classes CNR e não-CNR.

Sabendo disso, o objetivo deste trabalho é desenvolver um método capaz de prever os consumidores CNR com base no problema do desequilíbrio de classes. Para isso, propõe-se um novo método incremental em que o treinamento dos dados é passado em lotes para o classificador e é capaz de tratar o problema da classe desequilibrada e classificar com precisão novos consumidores.

2. Trabalhos relacionados

A literatura mostra muitos métodos para tratar o problema do desbalanceamento. De acordo com [Elrahman and Abraham 2013], existem quatro tipos de técnicas para lidar com o desequilíbrio de classes: (1) métodos de amostragem; (2) aprendizado sensível ao custo; (3) método baseado em reconhecimento (aprendizagem de uma classe); e (4) métodos baseados em conjuntos. Serão abordados alguns métodos que abrangem cada tipo mencionado.

Os métodos de amostragem são compostos por técnicas de subamostragem (*undersampling*) e sobreamostragem (*oversampling*). A subamostragem remove amostras da classe majoritária, com o intuito de equacionar as classes reduzindo a classe majoritária, mantendo os indivíduos mais relevantes, e a sobreamostragem equilibra os dados criando cópias das amostras existentes, ou gerando amostras sintéticas no caso do SMOTE, proposto por [Chawla et al. 2002], adicionando mais amostras à classe minoritária.

A técnica de aprendizagem sensível ao custo, resumidamente, coloca pesos para uma classe, de modo que o modelo que está sendo treinado dá maior significado a essa classe (que em muitos casos pertence à classe minoritária). Outra maneira de fazer isso é adaptar um limite para que, se uma determinada probabilidade na classe minoritária for atendida, essa classe seja considerada positiva. Métodos sensíveis ao custo usam adaptações em classificadores para penalizar a classe majoritária e refletir sobre a classe minoritária. Numerosos métodos usando esta abordagem foram propostos, alguns usando máquina de vetores de suporte, regressão linear.

O método baseado em reconhecimento, basicamente, só tomam conhecimento da classe minoritária, desta forma o classificador tem o trabalho de apenas conhecer uma classe e ser treinado para entender essa classe. Quando um conjunto de dados chega para ser testado, o classificador (que conhece apenas uma classe, a minoria) tem o papel de dizer se essa classe é ou não a classe treinada. O classificador mais conhecido em utilizar essa abordagem é o *SVM One-class* [Manevitz and Yousef 2001].

De acordo com [Elrahman and Abraham 2013], *ensemble* é uma combinação de vários classificadores para melhorar a capacidade de generalização e aumentar a precisão da previsão. As técnicas de combinação mais populares são *boosting* e *bagging*. No *boosting*, cada classificador depende do anterior e foca-se nos erros do anterior. Exemplos que são classificados erroneamente em classificadores anteriores são escolhidos com maior frequência ou mais pesados. Por outro lado, no *bagging*, cada modelo no conjunto vota com mesmo peso. Para promover a variância do modelo, o *bagging* treina cada modelo do Ensemble usando um subconjunto aleatório do conjunto de treinamento.

3. Materiais e Método

Esta seção apresenta a base de dados e a abordagem proposta para resolver o problema do desequilíbrio.

3.1. Base de dados da Companhia de Energia

A base de dados utilizada neste trabalho é uma base privada, da empresa de distribuição de energia denominada Companhia Energética do Maranhão (CEMAR), no Brasil. Consiste em informações de consumo do consumidor, adquiridas dos mais diversos setores da empresa.

Algumas características foram extraídas com base na análise do perfil do consumidor, a fim de criar características relevantes para o modelo. As seguintes características foram extraídas do banco de dados inteiro da etapa anterior: **(1) Informações Gerais:** informações de características individuais de cada cliente, como localização, tipo de cliente (residencial ou comercial); **(2) Consumo de energia:** características que lidam com o perfil de consumo de cada cliente; **(3) Perda de Energia:** características relacionadas à perda de energia de um transformador conectado a um cliente; **(4) Histórico de desligamentos:** características relacionadas a desligamentos; **(5) Faturas:** características que mostram se houve faturas causadas por ter ocorrido CNR com esse cliente; **(6) Lei:** características relacionadas a ações judiciais anteriores tomadas pelo cliente.

É importante observar que os características extraídas diretamente do banco de dados vêm em uma forma grosseira, às vezes categórica e não normalizada. Assim, além da necessidade de transformar essas variáveis, é importante simular intervalos de tempo.

As técnicas adotadas para tratar essas variáveis foram modeladas com o *Feature Tools* [1] é descrita abaixo:

1. Normalização de variáveis numéricas. A normalização consiste no escalonamento e conversão de valores das diferentes variáveis selecionadas para um único intervalo de valores;
2. Codificação de variáveis categóricas para *One-hot code* [Buckman et al. 2018] ou para números inteiros. Para habilitar o processamento por algoritmos de aprendizado de máquina, o processo de codificação transforma essas variáveis em valores numéricos;
3. Geração de estatísticas como média e desvio padrão de variáveis numéricas e contadores para variáveis categóricas. Tais estatísticas podem facilitar a visualização e identificação de padrões nos dados. Para separar informações antigas e recentes do consumidor para o mecanismo de classificação, foram geradas estatísticas para diferentes períodos da história do consumidor: de 01/01/2016 a 28/02/2019 (histórico completo), 01/02/2019 até 28/02/2019 (Fevereiro), 01/01/2019 a 31/01/2019 (Janeiro), 30/09/2018 até 01/04/2019.

Assim, os dados brutos foram processados para agregar informações temporais usando estatísticas. Dessa forma, novas características foram geradas com base nos dados originais extraídos do banco de dados.

A base de dados foi dividida em dois grupos, um para treinamento e outro para teste. Em ambas as bases de dados, foram considerados indivíduos de CNR, aqueles que tiveram processos, o resto eram não-CNR.

3.2. Método

Esta seção explica o método proposto para tratar o problema de desequilíbrio na base de dados da empresa de energia.

3.2.1. XGBoost incremental

O método proposto adiciona as vantagens do uso de técnicas de amostragem. Como mencionado, a proporção do banco de dados do cliente CNR é de 1:174, um claro problema de desequilíbrio de classes. Um dos métodos mais convencionais e mais usados na literatura são as técnicas de amostragem. No entanto, neste caso, além de um problema de desequilíbrio de classe, temos um outro problema, o número de indivíduos no banco de dados, que atinge cerca de dois milhões e meio de indivíduos.

A solução aplicada para esse problema foi um tipo de sub-amostragem, já que assim o classificador utilizado poderá aprender a classe minoritária e fugir de um possível *overfitting*, adotada durante o processo de escolha da abordagem, que consiste primariamente de uma nova divisão na base de treinamento, baseada na classe, sendo assim criada uma base contendo os indivíduos CNR outra contendo os indivíduos não-CNR da base de treinamento.

[1] Featuretools - Uma ferramenta Python de software livre para engenharia automatizada de recursos, disponível em: www.featuretools.com

A partir de então, o método usa o algoritmo *eXtreme Gradient Boosting* (XGBoost). O algoritmo XGBoost é um modelo de aprendizado de máquina escalonável para o aprimoramento de árvore baseado em Árvores de Decisão de Intensidade de Gradiente, que tem mostrado resultados muito bons em classificação [Chen and Guestrin 2016]. O XGBoost difere dos modelos clássicos de otimização de árvore para lidar com dados esparsos e possui um procedimento de esboço quantil ponderado justificado teoricamente, que permite manipular pesos de instância na aprendizagem da árvore. O XGBoost estima a classe alvo por uma série de árvores de decisão e define o peso quantificado para cada nó folha [Chen and Guestrin 2016].

Como o banco de dados é altamente desequilibrado e muito grande (tanto em indivíduos quanto em características), não é viável carregar todos os indivíduos para treinar o modelo. Assim, a estratégia de usar lotes é proposta como uma solução. A forma de treinamento em lotes, em vez de enviar o arquivo de treinamento completo para o classificador, apresenta o classificador com uma quantidade resumida de indivíduos durante cada etapa de treinamento. Ou seja, se tivermos um grupo de treinamento com 1.000.000 de pessoas e escolhermos criar 5 lotes, cada etapa do treinamento de 200.000 pessoas será transferida para o treinamento. Esta é uma maneira de lidar com o problema de bases muito grandes para o treinamento, mas ainda não há garantia de tratamento do desequilíbrio de classes.

Ao criar lotes para treinamento, pode ocorrer que, em vários lotes, haja somente indivíduos não-CNR ou uma quantidade não equilibrada de indivíduos não-CNR que especializam o classificador apenas nessa classe. Então, para lidar com o problema do desbalanceamento, propomos um modelo incremental em lotes, que antes garante que em cada lote há indivíduos de ambas as classes, pegando indivíduos dos dois arquivos - contendo os indivíduos da classe não CNR e CNR separadamente - gerados a partir da base de treinamento. Em seguida, lotes com tamanho variando em potências de 2, variando de 8 a 32 mil indivíduos por lote são criados. Para compor os indivíduos em cada lote, são mantidas proporções entre CNR e não-CNR, para cada lote as proporções são utilizadas: 1:1, 1:2, 1:3, 1:5 e 1:10. Ou seja, diferentes arquiteturas são criadas para executar o treinamento. Essas proporções para cada estágio de treinamento permitem que o classificador encontre um relacionamento equilibrado entre os indivíduos das duas classes.

No processo de treinamento, uma parte da base de dados de treinamento é usada para validar cada arquitetura (base de validação). Em cada iteração, um modelo é treinado. Assim, na primeira iteração, o primeiro tamanho de lote com a primeira proporção (8 indivíduos, proporção 1: 1) é treinado e esse modelo aplicado à base de dados de validação e as métricas são coletadas. As próximas iterações aumentam a proporção do lote até atingir 1:10, portanto, o tamanho do lote também é aumentado. Em todas as iterações, o modelo é aplicado à base de dados de validação para calcular as métricas de validação até que a melhor arquitetura seja encontrada. A Figura 1 a seguir demonstra o procedimento do método:

Ao final, é possível encontrar o tamanho ideal dos lotes e a proporção entre as classes em cada lote, que irá treinar um modelo robusto capaz de realizar o treinamento com toda a base de dados, independentemente do tamanho pelo uso de lotes, e tratar o problema de desequilíbrio de classes usando uma proporção no treinamento do modelo.

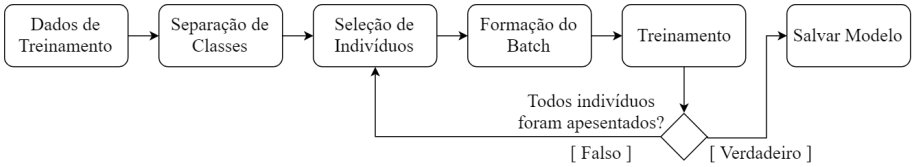

Figura 1. Procedimento do método

Por fim, métricas de validação são calculadas para validar a robustez do modelo. As métricas utilizadas são: Sensibilidade, é a razão entre verdadeiros positivos: a capacidade do sistema de prever com precisão a condição dos casos que realmente o possuem; Especificidade, é a proporção de verdadeiros negativos: a capacidade do sistema de prever corretamente a ausência da condição para casos que não a possuem; e Acurácia, é a proporção de previsões sem levar em conta o que é positivo e o que é negativo.

4. Resultados

Esta seção apresenta e discute os resultados obtidos com o método proposto para a previsão de ações judiciais de consumo de energia não registrado. Para avaliação, a base de dados adquirida foi dividida em dois conjuntos: treinamento e teste.

Para o treinamento, foram usados 60% do banco de dados de treinamento total e 40% para o teste, mantendo a proporção real do banco de dados para indivíduos CNR e não-CNR. Nos materiais (Seção 3.1), 1169 características são definidas por assunto, das quais 255 são características nativas retiradas da própria aquisição de dados, e 914 são características geradas por variações temporais e transformação de categórico para *One-hot* . Assim, mostramo-se que o método proposto é robusto, não apenas no tratamento do problema de desbalanceamento, mas também no número de características. A proporção entre clientes com CNR e sem CNR é descrita na Tabela 1.

Tabela 1. Proporção entre clientes CNR e não-CNR nas bases de treino e teste

Database	CNR	não-CNR	Total
Treino	8.560	1.476.042	1.484.602
Teste	5.714	998.442	1.004.156
Total	14.274	2.474.484	2.488.758

4.1. XGBoost Incremental

Como explicado na seção 3.2.1, um método baseado em XGBoost incremental é proposto para tratar o problema em questão. Este método agrega os avanços apresentados nas técnicas de amostragem, além de envolver agregação de árvores de decisão aprimoradas em tempo de treinamento.

O método trata do problema do desequilíbrio de classes e do problema do número de indivíduos da base de dados usando uma abordagem incremental para encontrar o tamanho de lote ideal e a proporção de indivíduos relativa às classes nos lotes. Para encontrar esses dois parâmetros no treinamento, 10% da base de dados de treinamento foi utilizada para validação. Depois de realizar este treinamento, o método encontrou o melhor tamanho de lote do tamanho de 8192 com uma proporção de 1: 1 de indivíduos CNR e não-CNR.

A tabela 2 mostra os resultados das métricas de validação do método incremental proposto e apenas XGBoost sem tratar o desbalanceamento de classe (sem o uso de lotes).

Tabela 2. Resultados do método XGBoost incremental comparado com XGBoost.

Método	ACC	SEN	SPEC	TP	FN	FP	TN
XGBoost	99,50	20,58	99,96	1176	4538	388	998054
XGBoost incremental	90,94	97,07	90,91	5547	167	90710	907732

Como pode ser visto na Tabela 2, ao usar lotes, o treinamento do XGBoost com 4096 de cada classe em cada iteração resultou em um bom equilíbrio entre classes no treinamento, permitindo que o classificador tenha bons resultados, alcançando 97.07% de sensibilidade, mostrando que pode prever com eficiência clientes CNR e 90.91% de especificidade. Comparado ao uso bruto do XGBoost, que teve um incrível desequilíbrio de 1:174, fez com que o classificador aprendesse mais sobre a classe não-CNR com apenas 20,58% de sensibilidade, mesmo tendo 99,50% de acurácia.

Para mostrar como o método proposto é robusto e sensível, não apenas em lidar com o problema do desbalanceamento, mas em classificar corretamente os casos de CNR, o método proposto é comparado com outros métodos considerados de última geração, tais como: LSTM (Long Short-term Memory) [Hochreiter and Schmidhuber 1997]; Regressão Logística (LR) [Hosmer Jr et al. 2013]; e Random Forest (RF) [Breiman 2001]. Tabela 3 mostra a comparação do método proposto e outros métodos considerados de última geração.

Tabela 3. Comparação entre o método proposto e outros considerados ótimos.

Métodos	ACC	SEN	SPEC
RF	99,43	1,27	99,99
LR	99,42	7,94	99,95
LSTM	95,72	84,28	95,79
XGBoost incremental	90.94	97.07	90.91

A tabela 3 mostra que os métodos RF e LR não foram capazes de aprender a classe minoritária, mostrando sensibilidade abaixo de 8% em ambos os casos. Em contraste, eles alcançaram uma especificidade de mais de 99%, isto é provavelmente devido ao desequilíbrio de classes, porque ao treinar o classificador, eles só aprendem a classe não-CNR, precisamente na proporção de 1:174. O método LSTM apresentou métricas interessantes, mas o método proposto se mostrou mais sensível e robusto na tarefa de tratar o problema do desequilíbrio de classes e na correta classificação dos consumidores com CNR.

5. Conclusão

Este trabalho apresentou um método para prever ações judiciais de consumo de energia não registradas em empresas de energia, propondo uma abordagem que lide com problema de desbalanceamento entre classes. O método proposto é um algoritmo de treinamento incremental, onde várias árvores são criadas e as mesmas são agregadas ao conjunto principal. O método é capaz de tratar o problema da classe desbalanceada e classificar com precisão novos consumidores de CNR e não-CNR.

Para avaliar o trabalho, utilizou-se uma base de dados da Companhia Energética do Maranhão, uma empresa de eletricidade que cobre todos os municípios de um estado do Brasil, e para analisar processos em todos os clientes, há um desequilíbrio natural de 1:173 casos de CNR e não-CNR.

O método se mostrou muito eficiente, uma vez que atingiu o objetivo de lidar com grandes volumes de dados, onde os mesmos se encontram completamente desbalanceados. Como trabalhos futuros, sugere-se que métodos de aprendizagem sensível ao custo sejam agregados para geração de lotes, ainda, testar o método em outras companhias elétricas e assuntos de processos diferentes.

Agradecimentos

Trabalho apoiado pelo projeto Sipaju, financiado pela Equatorial Energia e pela Agência Nacional de Energia Elétrica (ANEEL) através do P&D No PD-00037-0031/2018.

Referências

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. (2018). Thermometer encoding: One hot way to resist adversarial examples.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Elrahman, S. M. A. and Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1(2013):332–340.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Ibáñez, V. A., Hartmann, P., and Calvo, P. Z. (2006). Antecedents of customer loyalty in residential energy markets: Service quality, satisfaction, trust and switching costs. *The Service Industries Journal*, 26(6):633–650.

Manevitz, L. M. and Yousef, M. (2001). One-class svms for document classification. *Journal of machine Learning research*, 2(Dec):139–154.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Tan, P.-N. (2018). *Introduction to data mining*. Pearson Education India.