PUC-MG - Pontifícia Universidade Católica de Minas Gerais
Curso : Ciência da Computação
Disciplina : Inteligência Artificial
Professora : Cristiane Neri Nobre
Turno: Tarde
Alunos : Filipe Arthur Ferreira Silva, Henrique Augusto Rodrigues

Lista de exercícios 04

Questão 01

O algoritmo CART (Classification and Regression Trees), é um algoritmo baseado em árvore que funciona recursivamente examinando várias maneiras de particionar ou dividir dados de forma binária (ou seja, cada nó possui até dois filhos) localmente em segmentos menores com base em diferentes valores e combinações de preditores, selecionando no final a combinação com melhor desempenho. O algoritmo utiliza o Índice GINI para calcular a impureza da base de dados analisada, considerando uma divisão binária para cada atributo, onde cada um possui os valores distintos que ocorrem na base de dados.

Questão 02

Aprender com dados desequilibrados está entre os problemas cruciais enfrentados pela comunidade de aprendizado de máquina.
As distribuições de classes desequilibradas afetam o processo de treinamento dos classificadores, levando a um viés desfavorável em relação à (s) classe (s) majoritária (s).
Tal situação não pode ser aceita na maioria das aplicações do mundo real (por exemplo, medicina ou detecção de intrusão) e, portanto, algoritmos para combater o problema de desequilíbrio de classes têm sido um foco de intensa pesquisa por mais de duas décadas.
Os aplicativos contemporâneos ampliaram nossa visão do problema de dados desequilibrados, confirmando que classes desproporcionais não são a única fonte de problemas de aprendizagem.

Uma proporção de desequilíbrio de classe distorcida é frequentemente acompanhada por fatores adicionais, como instâncias difíceis e limítrofes, pequenas disjunções, pequeno tamanho de amostra ou a natureza flutuante de dados de streaming.
Esses desafios continuamente emergentes mantêm o campo em expansão, exigindo soluções novas e eficazes que podem analisar, compreender e enfrentar essas dificuldades de nível de dados.

No entanto, apesar de seus recursos poderosos, as arquiteturas profundas ainda são muito vulneráveis a distribuições de dados desequilibradas e são afetadas

por novos desafios, como representações de dados complexas, a relação entre dados desequilibrados e embeddings extraídos e o aprendizado de um número extremamente grande de classes.
Essas distribuições de classes distorcidas representam um desafio para os modelos de aprendizado de máquina, pois os classificadores padrão são orientados por uma função de perda de 0-1 que assume uma penalidade uniforme para ambas as classes.
Tanto a subamostragem quanto a sobreamostragem podem ser realizadas de maneira aleatória, o que tem baixa complexidade, mas leva a um comportamento potencialmente instável (por exemplo, removendo instâncias importantes ou aumentando as ruidosas).
Propomos DeepSMOTE - um algoritmo de sobreamostragem novo e inovador dedicado a aprimorar modelos de aprendizado profundo e combater o viés de aprendizado causado por classes desequilibradas.

Métodos de sobreamostragem baseados em métricas, como SMOTE, também podem ser caros do ponto de vista computacional porque exigem acesso ao conjunto de dados completo durante o treinamento e a inferência.
Acessar o conjunto de dados completo, especialmente ao lidar com dados de imagem ou fala, pode ser desafiador ao usar sistemas de aprendizado profundo que também requerem grandes quantidades de memória para armazenar gradientes.
Para que um método de sobreamostragem ser aplicado com sucesso a modelos de aprendizagem profunda, acreditamos que deve atender a três critérios essenciais: 1) Deve operar de forma ponta a ponta, aceitando entrada bruta, como imagens (ou seja, semelhante a VAEs, WAEs e GANs).

usar um discriminador / gerador em um GAN, que é fundamentalmente semelhante a um codificador / decodificador porque o discriminador efetivamente codifica a entrada (sem a camada final totalmente conectada) e o gerador (decodificador) gera a saída.
Todas as classes são usadas durante o treinamento para que o codificador / decodificador possa aprender a reconstruir imagens de classes majoritárias e minoritárias a partir dos dados desequilibrados.

Como há poucos exemplos de classes minoritárias, os exemplos de classes majoritárias são usados para treinar o modelo para aprender os padrões básicos de reconstrução inerentes aos dados.
no entanto, ao contrário das imagens usadas durante a reconstrução fase de perda de treinamento, as imagens amostradas são todas da mesma classe.

Ao alterar a ordem das imagens reconstruídas, que são todas da mesma classe, introduzimos efetivamente a variação no processo de codificação / decodificação.

A perda de penalidade é baseada na diferença do erro quadrático médio (MSE) entre D0 e D1, D1 e D2, etc., como se uma imagem fosse superamostrado por SMOTE (ou seja, como se uma imagem fosse gerada com base na diferença entre uma imagem e o vizinho da imagem).

Portanto, evitamos a necessidade de um discriminador porque usamos dados de treinamento para treinar o gerador, simplesmente alterando a ordem das imagens codificadas / decodificadas.
simular a metodologia SMOTE durante DeepSMOTE treinamento selecionando uma amostra de classe e calculando uma distância entre a instância e seus vizinhos (no espaço de incorporação ou recurso), exceto que a distância (MSE) durante o treinamento é usada como uma penalidade implícita na perda de reconstrução.
Conforme observado por Arjovsky et al., muitos modelos de aprendizagem profunda generativos incorporam efetivamente um termo de penalidade, ou ruído, em sua função de perda, para transmitir diversidade na distribuição do modelo.

O uso do termo de penalidade e a fidelidade do SMOTE na interpolação de amostras sintéticas durante a fase de inferência nos permite evitar o uso de um discriminador, que normalmente é usado pelos modelos GAN e WAE.
principal diferença entre nas fases de treinamento e geração do DeepSMOTE, durante a fase de geração de dados, o SMOTE é substituído pela etapa de permutação da ordem.
ao passo que, durante o treinamento, a variação é introduzida pela permutação da ordem dos exemplos de treinamento que são codificados e então decodificados e também por meio da perda de penalidade.

Por ser capaz de trabalhar em imagens brutas e extrair recursos delas, o DeepSMOTE pode gerar instâncias artificiais mais significativas do que abordagens baseadas em pixels, mesmo usando regras relativamente mais simples para geração de instâncias.
Nosso método é fácil de ajustar e usar em qualquer dado, tanto como uma solução de caixa preta quanto como um trampolim para o desenvolvimento de novas e robustas arquiteturas de aprendizado profundo.

O DeepSMOTE pode ser visto como uma solução em nível de dados para o desequilíbrio de classe, pois cria instâncias artificiais que equilibram o conjunto de treinamento, que podem então ser usados para treinar qualquer classificador profundo sem sofrer viés.
DeepSMOTE exclusivamente cumpriu três características cruciais de um algoritmo de reamostragem bem-sucedido neste domínio: a capacidade de operar em imagens brutas, a criação de embeddings de baixa dimensão eficientes e a geração de imagens artificiais de alta qualidade.

Extensos estudos experimentais mostram que DeepSMOTE não só supera algoritmos de sobreamostragem baseados em pixel e GAN de última geração, mas também oferece robustez incomparável para taxas de desequilíbrio variáveis com alta estabilidade de modelo, enquanto gera imagens artificiais de excelente qualidade.

Nossos próximos esforços se concentrarão em aprimorar o DeepSMOTE com informações sobre o nível de classe e dificuldades no nível da instância, o que permitirá lidar melhor com regiões desafiadoras do espaço de recursos.
Planejamos aprimorar nossa função de perda dedicada com penalidades em nível de instância para focar o treinamento do codificador / decodificador em instâncias

que exibem características limítrofes / sobrepostas, enquanto descartando instâncias externas e com ruído.
Tal função de perda insensível à distorção composta fará uma ponte os mundos entre as abordagens de nível de dados e de algoritmo para aprender com dados desequilibrados.

Além disso, queremos tornar o DeepSMOTE adequado para cenários de aprendizagem contínua e ao longo da vida, onde há uma necessidade de lidar com proporções de classes dinâmicas e gerar novas instâncias artificiais.
Imaginamos que DeepSMOTE pode não só ajudar a conter o desequilíbrio de classes online, mas também ajudar a aumentar a robustez de modelos de aprendizagem ao longo da vida ao esquecimento catastrófico.