Projeto de Formatura – 2020

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia de Computação

Tema:

Clusterização no espaço de latência de uma Rede Neural

MOTIVAÇÃO

Em vários cenários do mundo real encontramos diversos tipos de dados, como categóricos, ordinais, contínuos, imagens, séries temporais e outros. Porém, como os algoritmos de clusterização se baseiam na distância entre pontos, eles só conseguem uma boa performance quando trabalham unicamente com dados numéricos.

Além disso, os algoritmos de clusterização não consideram a relação dos atributos com uma variável objetivo, ou seja, não são otimizados para o cenário problema.

Assim, os algoritmos de clusterização não conseguem uma boa performance quando se misturam dados numéricos com não numéricos e os clusteres encontrados não são otimizados para o problema abordado, dificultando encontrar boas soluções para os cenários complexos que temos no mundo real.

OBJETIVOS

Os objetivos são expor a técnica de clusterização no espaço de latência em cenários artificiais, comparar com um algoritmo tradicional chamado k-means e mostrar sua eficácia em problemas complexos e reais.

Para isso, aplicamos a técnica em 3 cenários:

1. Cenário artificial simples e separável
2. Cenário artificial simples e não separável
3. Cenário real de empréstimos P2P

TÉCNICA

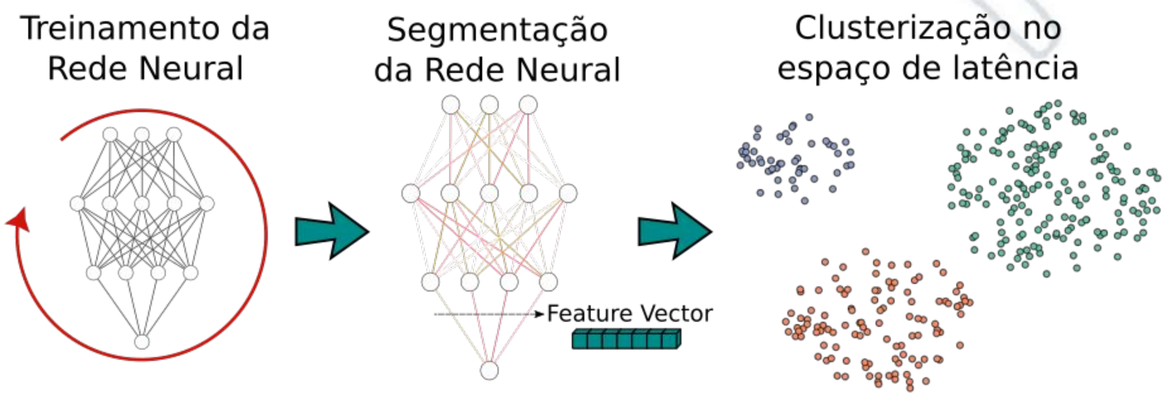

Figura 1: Etapas da clusterização no espaço de latência

A técnica proposta é a de aplicar o algoritmo de clusterização em um espaço de latência gerado por uma rede neural, em contraposição à abordagem tradicional de aplicar o algoritmo diretamente no vetor de entrada.

Para isso, a primeira etapa é treinar uma rede neural para uma variável *target* de interesse do cenário problema. Após o treinamento, segmentamos a rede extraindo o espaço de latência para criar os *feature vectors* das nossas entradas. Por fim, utilizamos o k-means nos *feature vectors* para encontrar os clusteres.

METODOLOGIA

Cada cenário é considerado um projeto próprio e, para cada um, adaptamos a metodologia CRISP-DM. Dividimos o problema em 5 etapas, sendo elas:

1. Entendimento do cenário
2. Exploração dos dados
3. Preparação dos dados
4. Treinamento de modelos
5. Análise dos resultados

RESULTADOS

O K-means aplicado diretamente nas entradas fica enviesado a particionar os dados primeiro pelas variáveis categóricas e só depois pelas numéricas. Por conta disso, o resultado por esse método não é satisfatório uma vez que não ajuda a entender o conjunto de dados no cenário específico, apenas mostra combinações entre as variáveis. Isso pode ser visto na figura 2 (a), que mostra 2 clusteres com variável target mista.

Por outro lado, a técnica de aplicar o K-means no espaço de latência conseguiu particionar os dados muito bem, pois permitiu entender o conjunto de dados específico para o problema. Como resultado, conseguimos identificar clusteres diferentes com variáveis *target* únicas, mas a grande vantagem foi isolar os subconjuntos não separáveis. Isso pode ser visto na figura 2 (b), na qual três clusteres têm variável *target* única e outro com *output* misto, que é exatamente o subgrupo do conjunto de dados não separável.

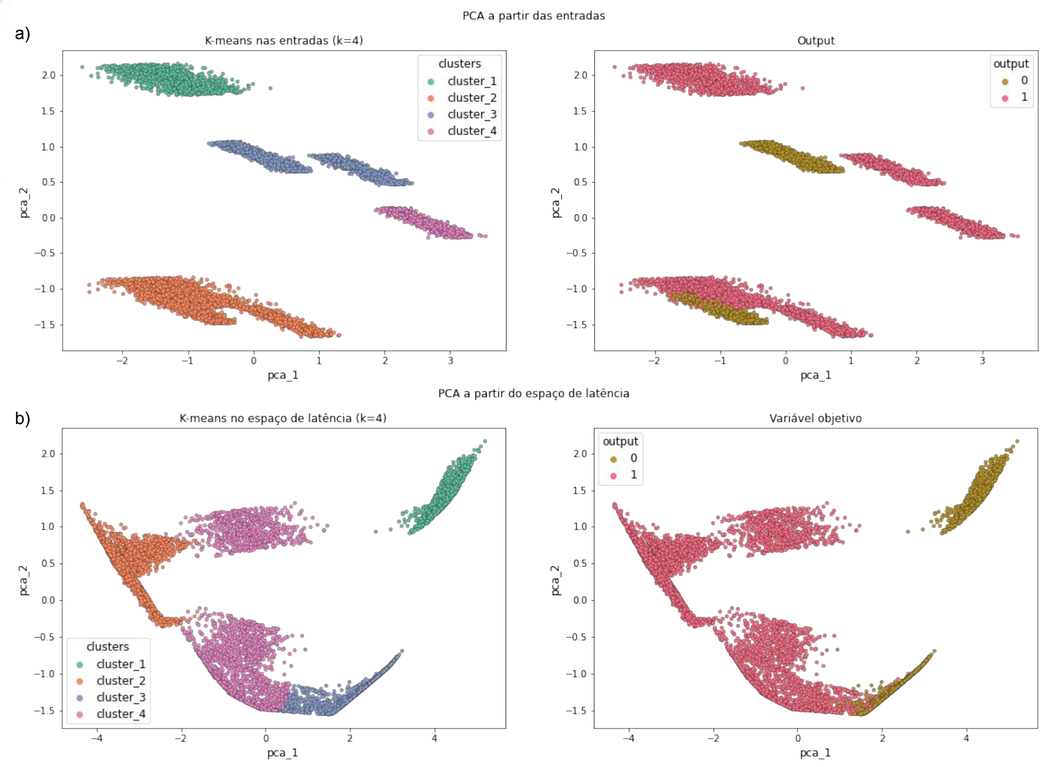

Figura 2: Visualização dos clusteres por PCA do cenário 2

Em particular, no cenário de empréstimos, a clusterização pelo espaço de latência encontrou grupos de empréstimos com uma inadimplência diferente do esperado pelo modelo de risco da fintech.

A partir desses grupos, identificamos oportunidades de melhoria no modelo. Assim, é possível ajustar a taxa de juros para ser mais justa ao tomador de crédito, e melhorar o gerenciamento de risco do investidor.

Integrantes: Alexandre X. Ciuffatelli
Pedro Miguel M. Goncalves

Professor(a) Orientador(a): Prof. Dr. Jorge Luis Risco Becerra