UNIVERSIDADE FEDERAL DO CEARÁ (UFC)

Centro de Ciências

Departamento de Computação

Sumário

Machine Learning Probabilístico: Uma Aplicação em Doenças Renais Crônicas

José Alisson Alves Ursulino
Matrícula: 499487
João Paulo Andrade Lima
Matrícula: 510179

31 de março de 2024

1 Resumo

Este trabalho explora a aplicação de técnicas de Machine Learning Probabilístico para a detecção de anomalias em dados associados a doenças renais crônicas. Utilizando modelos como GMM (Mistura de Gaussianas), PPCA (Análise de Componentes Principais Probabilística), FA (Análise de Fatores) e VAE (Autoencoder Variacional), investigamos diferentes abordagens para redução de dimensionalidade e identificação de padrões anômalos. Os resultados indicam que o VAE se destaca como uma ferramenta poderosa na detecção precoce de anomalias em dados complexos de saúde.

2 Contextualização

No contexto da saúde renal, a detecção precoce de anomalias é crucial para intervenções eficazes. Este estudo utiliza modelos probabilísticos de aprendizado de máquina para analisar dados multidimensionais e identificar padrões que podem sinalizar a presença de doenças renais crônicas em estágios iniciais. A abordagem visa melhorar a precisão diagnóstica e proporcionar insights para tratamentos personalizados.

3 Descrição do Conjunto de Dados

3.1 Visão Geral

O conjunto de dados utilizado neste estudo abrange um período de 2 meses na Índia e compreende 25 características, incluindo indicadores vitais como contagem de glóbulos vermelhos, contagem de glóbulos brancos, etc. A variável-alvo, denominada 'classification', categoriza indivíduos como 'ckd' (doença renal crônica) ou 'notckd'. O objetivo principal é empregar técnicas de aprendizado de máquina para prever se um paciente está sofrendo de doença renal crônica.

3.2 Origem do Conjunto de Dados

O conjunto de dados foi obtido no Kaggle, cortesia de Mansoor Iqbal (`https://www.kaggle.com/mansoordaku`). Certas modificações foram feitas no conjunto de dados original para atender aos requisitos deste desafio de pesquisa.

O conjunto de dados não alterado está disponível em: Chronic Kidney Disease (`https://www.kaggle.com/mansoordaku/ckdisease`).

3.3 Informações sobre Atributos

O conjunto de dados abrange 25 atributos, com 11 numéricos e 14 nominais. Esses atributos incluem idade, pressão sanguínea, gravidade específica, albumina, glicose no sangue aleatória e outros indicadores de saúde relevantes.

3.3.1 Atributos Numéricos

- Age: Idade em anos
- Blood Pressure: Pressão sanguínea em mm/Hg
- Specific Gravity: Gravidade específica (nominal) com valores (1.005, 1.010, 1.015, 1.020, 1.025)
- Albumin: Albumina (nominal) com valores (0, 1, 2, 3, 4, 5)

- Sugar: Açúcar (nominal) com valores (0, 1, 2, 3, 4, 5)
- Red Blood Cells: Glóbulos vermelhos (nominal) com valores ('normal', 'abnormal')
- Pus Cell: Células de pus (nominal) com valores ('normal', 'abnormal')
- Pus Cell Clumps: Aglomerados de células de pus (nominal) com valores ('present', 'notpresent')
- Bacteria: Bactéria (nominal) com valores ('present', 'notpresent')
- Blood Glucose Random: Glicose no sangue aleatória (numérico)
- Blood Urea: Ureia no sangue (numérico)
- Serum Creatinine: Creatinina sérica (numérico)
- Sodium: Sódio (numérico)
- Potassium: Potássio (numérico)
- Hemoglobin: Hemoglobina (numérico)
- Packed Cell Volume: Volume globular (numérico)
- White Blood Cell Count: Contagem de glóbulos brancos (numérico)
- Red Blood Cell Count: Contagem de glóbulos vermelhos (numérico)

3.3.2 Atributos Nominais

- Hypertension: Hipertensão (nominal) com valores ('yes', 'no')
- Diabetes Mellitus: Diabetes mellitus (nominal) com valores ('yes', 'no')
- Coronary Artery Disease: Doença arterial coronariana (nominal) com valores ('yes', 'no')
- Appetite: Apetite (nominal) com valores ('good', 'poor')
- Pedal Edema: Edema pedal (nominal) com valores ('yes', 'no')

- Anemia: Anemia (nominal) com valores ('yes', 'no')
- Class: Classe (nominal) com valores ('ckd', 'notckd')

O conjunto de dados está disponível publicamente e foi arquivado pelo Repositório de Aprendizado de Máquina da UCI (Dua und Graff, 2019, `https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease`).

4 Metodologia

Os dados foram divididos em conjuntos de treino (80%) e teste (20%), sendo empregados 5 Folds para realizar a validação cruzada. Na etapa de treinamento, realizamos uma filtragem específica, considerando apenas as instâncias correspondentes a pacientes saudáveis. Essa abordagem foi adotada com o intuito de adaptar o modelo especificamente a esse grupo, visando, assim, a detecção de anomalias nos pacientes dos dados de treino.

Em cada modelo, calculamos um score de anomalia, geralmente utilizando o negativo da verossimilhança, para cada amostra. Inicialmente, efetuamos esses cálculos nos dados de treino, estabelecendo um limiar com base no 1% percentil desses scores. Posteriormente, esse limiar foi aplicado nos dados de teste, permitindo determinar se uma determinada amostra era considerada uma anomalia ou não.

Para avaliar o desempenho de todos os modelos, adotamos as seguintes métricas de avaliação: AUC, acurácia, F1, precisão e recall. Essas métricas proporcionam uma visão abrangente do quão eficaz cada modelo é na detecção de anomalias, considerando diferentes aspectos de seu desempenho.

5 Modelos

5.1 GMM

O GMM é um modelo probabilístico que assume que os dados são originados de várias distribuições gaussianas Rasmussen u. a. (2006). Ele modela a mistura dessas distribuições para descrever a complexidade dos dados. Na detecção de anomalias, o GMM calcula a probabilidade de uma amostra pertencer à mistura. Limiares podem ser definidos com base nessas

probabilidades para identificar anomalias.

Resumo o Algoritmo

1. Inicialização dos parâmetros: médias, covariâncias e pesos das gaussianas.
2. Expectation-Maximization (EM) para otimizar os parâmetros.
3. Cálculo da probabilidade de pertencimento a cada componente para cada amostra.
4. Estabelecimento de limiares para identificar anomalias.

5.2 PPCA

O PPCA é uma extensão probabilística da PCA. Ele assume que os dados são gerados a partir de uma distribuição normal multivariada e incorpora ruído na representação dos dados Bishop und Nasrabadi (2006). Na detecção de anomalias, o PPCA utiliza o erro de reconstrução para calcular os scores de anomalia.

Inicialmente, procedemos com a redução de dimensionalidade dos dados para 2 dimensões (Figura 1), buscando obter uma representação visual aproximada das separações possíveis nos dados. Posteriormente, realizamos a detecção de anomalias considerando de 2 a 10 componentes.

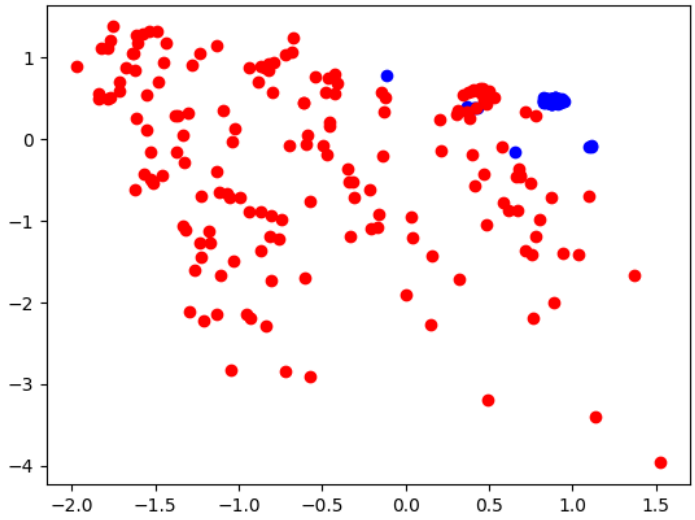
O cálculo do score de anomalia baseou-se no erro de reconstrução dos conjuntos de dados de treino e teste. Para estimar a função de densidade dos dados de treino, empregamos o modelo KernelDensity. Essa escolha foi motivada pela observação de que a distribuição gaussiana multivariada não proporcionava resultados satisfatórios na modelagem da função de densidade, justificando assim a preferência pelo modelo KernelDensity.

Resumo do Algoritmo

1. Redução de dimensionalidade para representação latente.
2. Cálculo do erro de reconstrução.
3. Estabelecimento de limiares com base nos scores de anomalia.

Para esse modelo, o melhor número de componentes para a métrica AUC foi de 7 componentes.

Figura 1: Plot do PPCA para 2 componentes



	Auc	Accuracy	F1	Precision	Recall
PPCA 2 Components	0.97004	0.92857	0.93528	0.96519	0.91329
PPCA 3 Components	0.97716	0.91429	0.92877	0.95709	0.90384
PPCA 4 Components	0.98115	0.91786	0.93253	0.95709	0.91074
PPCA 5 Components	0.98259	0.91071	0.92475	0.96156	0.89074
PPCA 6 Components	0.98289	0.90357	0.91923	0.96858	0.87562
PPCA 7 Components	0.98369	0.90000	0.91547	0.96858	0.86872
PPCA 8 Components	0.98343	0.90357	0.91922	0.96884	0.87539
PPCA 9 Components	0.98249	0.88571	0.90377	0.97350	0.84511
PPCA 10 Components	0.98104	0.88571	0.90470	0.97468	0.84914

Tabela 1: Tabela com os resultados da detecção de anomalias para o PPCA

5.3 FA

A FA é uma técnica de modelagem que assume que as observações são causadas por um número menor de fatores latentes. Na detecção de anomalias, a FA procura representar os dados através desses fatores, permi-

tindo identificar padrões incomuns Murphy (2012).

Resumo do Algoritmo

1. Redução de dimensionalidade para representação por fatores.
2. Cálculo do erro de reconstrução.
3. Estabelecimento de limiares com base nos scores de anomalia.

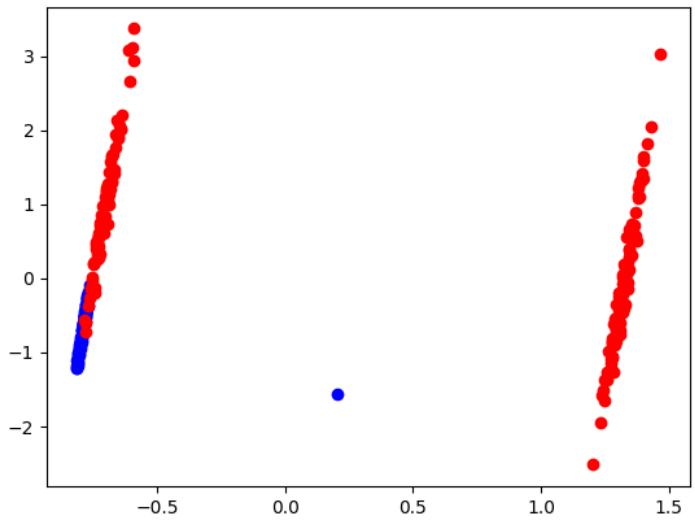
No contexto da Análise de Fatores (FA), a abordagem metodológica adotada apresentou semelhanças com a implementada para o PPCA. Inicialmente, conduzimos uma redução para duas dimensões com o objetivo de facilitar a visualização dos dados, conforme ilustrado na Figura 2. Em seguida, prosseguimos com a detecção de anomalias, considerando um intervalo de 2 a 10 fatores. O cálculo do score de anomalia seguiu os mesmos procedimentos previamente delineados.

Uma observação de destaque durante a aplicação da Análise de Fatores (FA) foi a notável discrepância observada no gráfico bidimensional em comparação com o PCA e PPCA. No caso da FA, o gráfico conseguiu efetivamente segregar os dados em dois grupos distintos, não necessariamente correlacionados com as categorias de pacientes saudáveis e não saudáveis. Essa característica evidencia a capacidade da FA em identificar padrões complexos nos dados, ampliando a compreensão da variabilidade presente no conjunto de dados analisado.

	Auc	Accuracy	F1	Precision	Recall
FA 2 Factors	0.98494	0.93214	0.94739	0.90233	1.00000
FA 3 Factors	0.98445	0.92143	0.93754	0.88385	1.00000
FA 4 Factors	0.98280	0.88571	0.91765	0.84951	1.00000
FA 5 Factors	0.98190	0.85714	0.90063	0.82025	1.00000
FA 6 Factors	0.98241	0.70357	0.80375	0.67744	1.00000
FA 7 Factors	0.98087	0.73214	0.81891	0.69979	1.00000
FA 8 Factors	0.97906	0.75000	0.82876	0.71258	1.00000
FA 9 Factors	0.97799	0.76071	0.83480	0.72222	1.00000
FA 10 Factors	0.97695	0.77500	0.84274	0.73296	1.00000

Tabela 2: Tabela com os resultados da detecção de anomalias para o FA

Figura 2: Plot do FA para 2 fatores



5.4 VAE

O VAE é um tipo de autoencoder que mapeia os dados para uma distribuição latente. Ele introduz a variabilidade nesse processo, tornando-o mais robusto Murphy (2022). Na detecção de anomalias, o VAE utiliza o negativo da verossimilhança marginal como score de anomalia.

Resumo do Algoritmo

1. Encoder mapeia os dados para uma distribuição latente.
2. Amostragem na distribuição latente.
3. Decoder reconstrói os dados.
4. Cálculo do negativo da verossimilhança marginal.
5. Estabelecimento de limiares com base nos scores de anomalia.

5.4.1 Estrutura do modelo

- Encoder com 2 camadas, uma com (input_dim, hidden_dim) e outra com (hidden_dim, latent_dim)
- Função de ativação ReLU no final de ambas.
- Decoder com 2 camadas, uma com (latent_dim, hidden_dim) e outra com (hidden_dim, input_dim).
- Função de ativação ReLU no final da primeira.

O modelo em questão possui três hiperparâmetros distintos, a saber: a dimensão latente, a dimensão oculta e a taxa de aprendizado. Dessa maneira, o processo de otimização dos hiperparâmetros abarcará esses três elementos essenciais, cada um desempenhando um papel fundamental na configuração do modelo.

A condução da otimização dos hiperparâmetros foi efetuada por meio da biblioteca Ax. Nesse contexto, os três hiperparâmetros mencionados anteriormente foram alvos do processo de otimização. Optamos por realizar 25 tentativas de otimização, uma quantidade que se mostrou satisfatória para atingir resultados eficazes e robustos. Essa abordagem assegurou uma exploração abrangente do espaço de hiperparâmetros, contribuindo para a obtenção de configurações ótimas que maximizam o desempenho do modelo.

	hidden_dim	latent_dim	lr
best_parameters	22	10	0.00529

Tabela 3: Tabela com os melhores hiperparâmetros obtidos na otimização

5.4.2 Detecção de anomalias

No que se refere à detecção de anomalia, o procedimento seguiu uma abordagem semelhante aos modelos anteriores. A distinção principal reside no cálculo do score de anomalia específico para o Modelo de Autoencoder Variacional (VAE). Nesse caso, adotamos o negativo da verossimilhança marginal, utilizando um número de amostras igual a 50.

Vale destacar que o VAE se destacou como o modelo com os melhores resultados em comparação aos demais, apresentando desempenho superior em todas as métricas avaliadas. Essa superioridade evidencia a eficácia do VAE na tarefa de detecção de anomalias, destacando-o como uma escolha robusta para este cenário específico.

	Auc	Accuracy	F1	Precision	Recall
VAE	1.00000	0.98929	0.99173	0.98381	1.00000

Tabela 4: Tabela com os resultados do VAE

6 Conclusão

No tocante aos modelos de redução de dimensionalidade que adotaram a estratégia do erro de reconstrução, observou-se que o PCA (utilizado unicamente para efeitos comparativos) com 2 componentes apresentou os resultados mais favoráveis em termos da métrica F1. Em contrapartida, para a métrica AUC, destacou-se o PCA com 4 componentes, evidenciando um desempenho superior. Pode-se inferir que o modelo PCA demonstrou vantagem em relação aos modelos de Análise de Fatores (FA) e PPCA neste contexto específico, conforme ilustrado nas tabelas abaixo:

	Auc	Accuracy	F1	Precision	Recall
PCA 2 Components	0.98229	0.94643	0.95682	0.91770	1.00000
FA 2 Factors	0.98494	0.93214	0.94739	0.90233	1.00000
PPCA 2 Components	0.97004	0.92857	0.93528	0.96519	0.91329

Tabela 5: Tabela com o melhor número de componentes/fatores para cada modelo, com base no F1

Ao compararmos todos os modelos, destaca-se que o Modelo de Autoencoder Variacional (VAE) foi, de longe, o mais proeminente. Tal resultado era esperado, considerando a natureza robusta e poderosa desse modelo. A utilização de uma ferramenta de otimização de hiperparâmetros contribuiu para aprimorar ainda mais a precisão do VAE.

	Auc	Accuracy	F1	Precision	Recall
PCA 4 Components	0.98855	0.90357	0.92648	0.87083	1.00000
FA 2 Factors	0.98494	0.93214	0.94739	0.90233	1.00000
PPCA 7 Components	0.98369	0.90000	0.91547	0.96858	0.86872

Tabela 6: Tabela com o melhor número de componentes/fatores para cada modelo, com base no AUC

Na condução da pesquisa, poderíamos ter considerado a inclusão das funções de ativação como hiperparâmetros a serem otimizados, além de explorar um número mais elevado de tentativas para a otimização bayesiana. Contudo, julgamos que os resultados obtidos com as configurações atuais já se mostraram suficientemente elucidativos.

A tabela apresentada abaixo sintetiza os resultados do projeto, excluindo o PCA, uma vez que a nossa intenção é comparar exclusivamente os modelos probabilísticos. Este resumo consolidado fornece uma visão abrangente do desempenho relativo de cada modelo, destacando o VAE como a escolha mais eficaz para a detecção de anomalias no contexto investigado.

	Auc	Accuracy	F1	Precision	Recall
VAE	1.00000	0.98929	0.99173	0.98381	1.00000
FA 2 Factors	0.98494	0.93214	0.94739	0.90233	1.00000
GMM	0.99867	0.95357	0.93922	1.00000	0.88851
FA 3 Factors	0.98445	0.92143	0.93754	0.88385	1.00000
PPCA 2 Components	0.97004	0.92857	0.93528	0.96519	0.91329

Tabela 7: Tabela com os 5 melhores modelos probabilísticos com base na métrica F1

Os resultados destacam o VAE como o modelo mais eficaz na detecção de anomalias em dados de doenças renais crônicas. A abordagem probabilística oferecida pelos modelos explorados revela-se promissora para melhorar a compreensão e identificação precoce de padrões associados a condições de saúde complexas.

Referências

[Bishop und Nasrabadi 2006] BISHOP, Christopher M. ; NASRABADI, Nasser M.: *Pattern recognition and machine learning*. Springer, 2006

[Dua und Graff 2019] DUA, D. ; GRAFF, C.: *UCI Machine Learning Repository*. `http://archive.ics.uci.edu/ml`. 2019

[Murphy 2012] MURPHY, Kevin P.: *Machine learning: a probabilistic perspective*. MIT press, 2012

[Murphy 2022] MURPHY, Kevin P.: *Probabilistic machine learning: an introduction*. MIT press, 2022

[Rasmussen u. a. 2006] RASMUSSEN, Carl E. ; WILLIAMS, Christopher K. u. a.: *Gaussian processes for machine learning*. Bd. 1. Springer, 2006