Análise de Componentes Principais (PCA)

ENG121 - Reconhecimento de Padrões

Autor: Matheus Castro Silva

RESUMO

O presente trabalho tem como objetivo comparar o desempenho da classificação utilizando a Análise de Componentes Principais (PCA). O classificador escolhido para este trabalho foi o de máquinas de vetor de suporte (SVM), para as 3 bases de dados Wisconsin Breast Cancer, Cancer e USArrests.

Palavras Chaves **- algoritmos, PCA, classificação, reconhecimento de padrões, svm, desempenho, câncer.**

I. INTRODUÇÃO

O presente trabalho busca introduzir a comparação entre as técnicas de classificação, um utilizando o modelo de máquinas de vetor de suporte (SVM) com a Análise de Componentes Principais (PCA) e outro apenas utilizando o modelo de máquinas de vetor de suporte (SVM). Para 3 bases de dados; Wisconsin Breast Cancer, Cancer e USArrests. Realizar uma comparação entre os modelos e por fim identificar cada uma das amostras coletadas é do tipo maligno ou benigno.

II. DEFINIÇÃO DO PROBLEMA

Dado um conjunto da base de dados Wisconsin Breast Cancer identifique se a amostra é benigna ou maligna com base no classificador gerado utilizando SVM+PCA e SVM, por fim realizar a comparação dos resultados obtidos de cada modelo.

III. MODELO DE SVM

Uma máquina de vetores de suporte (SVM, do inglês: support vector machine) é um conceito na ciência da computação para um conjunto de métodos do aprendizado supervisionado que analisam os dados e reconhecem padrões, usado para classificação e análise de regressão. O SVM padrão toma como entrada um conjunto de dados e prediz, para cada entrada dada, qual de duas possíveis classes a entrada faz parte, o que faz do SVM um classificador linear binário não probabilístico. Dados um conjunto de exemplos de treinamento, cada um marcado como pertencente a uma de duas categorias, um algoritmo de treinamento do SVM constrói um modelo que atribui novos exemplos a uma categoria ou outra. Um modelo SVM é uma representação de exemplos como pontos no espaço, mapeados de maneira que os exemplos de cada categoria sejam divididos por um espaço claro que seja tão amplo quanto possível. Os novos exemplos são então mapeados no mesmo espaço e preditos como pertencentes a uma categoria baseados em qual o lado do espaço eles são colocados.

Em outras palavras, o que uma SVM faz é encontrar uma linha de separação, mais comumente chamada de hiperplano entre dados de duas classes. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classe, ver imagem:

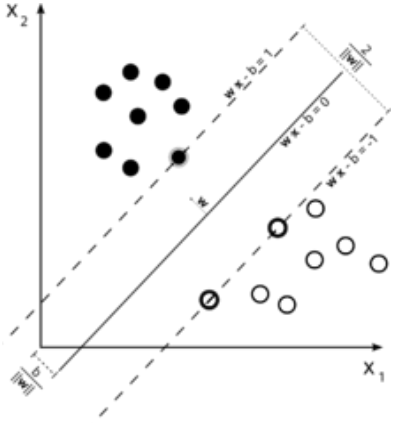

Imagem 1: Ilustração do vetor de suporte entre as duas classes no plano x1 e x2.

IV. TREINAMENTO E TESTE

A base de dados deve ser dividida em dois conjuntos, um de treinamento e outro para os testes, sendo 70% para treinamento e 30% para teste.

V. IMPLEMENTAÇÃO DO MODELO DE SVM

1. Carregar os dados do pacote *mlbench* e substituir os dados faltantes por 0, por exemplo.
2. Dividir de forma aleatória os dados em grupos de treinamento, validação e teste de acordo com uma razão pré definida.
3. Utilizar a função de treinamento que estime os vetores de suporte e classifique os dados de validação, utilizando o pacote *e1071.*

4. Estimar os parâmetros *cost* e *gamma* utilizando os dados de validação de forma gulosa, e ajustar o modelo de treinamento.
5. Calcular o erro quadrático médio (MSE) percentual do classificador.
6. Estimar o MSE, percentual de acerto e o desvio padrão do classificador, referente aos dados de teste.

VI. IMPLEMENTAÇÃO DA TÉCNICA DE PCA

A Análise de Componentes Principais (ACP) ou Principal Component Analysis (PCA) é um procedimento matemático que utiliza uma transformação ortogonal (ortogonalização de vetores) para converter um conjunto de observações de variáveis possivelmente correlacionadas num conjunto de valores de variáveis linearmente não correlacionadas chamadas de componentes principais. O número de componentes principais é menor ou igual ao número de variáveis originais. Esta transformação é definida de forma que o primeiro componente principal tem a maior variância possível (ou seja, é responsável pelo máximo de variabilidade nos dados), e cada componente seguinte, por sua vez, tem a máxima variância sob a restrição de ser ortogonal a (i.e., não correlacionado com) os componentes anteriores. Os componentes principais são garantidamente independentes apenas se os dados forem normalmente distribuídos (conjuntamente). O PCA é sensível à escala relativa das variáveis originais. Dependendo da área de aplicação, o PCA é também conhecido como transformada de Karhunen-Loève (KLT) discreta, transformada de Hotelling ou decomposição ortogonal própria (POD).

O PCA foi inventado em 1901 por Karl Pearson. Agora, é mais comumente usado como uma ferramenta de Análise Exploratória de Dados e para fazer modelos preditivos. PCA pode ser feito por decomposição em autovalores (Valores Próprios) de uma matriz covariância, geralmente depois de centralizar (e normalizar ou usar pontuações-Z) a matriz de dados para cada atributo. Os resultados de PCA são geralmente discutidos em termos pontuações (scores) de componentes, também chamados de pontuações de fatores (os valores de variável transformados correspondem a um ponto de dado particular), e carregamentos (loadings), i.e., o peso pelo qual cada variável normalizada original deve ser multiplicada para se obter a pontuação de componente.

O PCA é a mais simples das verdadeiras análises multivariadas por autovetores. Com frequência, sua operação pode ser tomada como sendo reveladora da estrutura interna dos dados, de uma forma que melhor explica a variância nos dados. Se visualizarmos um conjunto de dados multivariados em um espaço de alta dimensão, com 1 eixo por variável, o PCA pode ser usado para fornecer uma visualização em dimensões mais baixas dos mesmos dados, uma verdadeira "sombra" do objeto original quando visto de seu ponto mais informativo. Isto é feito usando-se apenas os primeiros componentes principais, de forma que a dimensionalidade dos dados transformados é reduzida. A figura abaixo mostra as possíveis direções para estes vetores *u*.

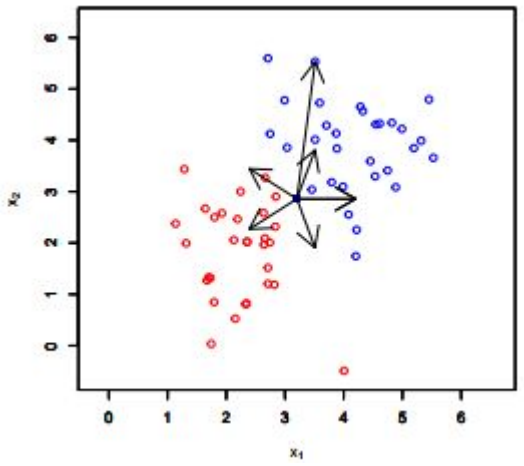

Imagem 2: Vetores *u* indicando direções arbitrárias para as possíveis retas que passam pelo ponto médio.

O PCA é fortemente ligado à análise de fatores (Factorial Analysis); de fato, alguns pacotes estatísticos propositadamente confluem as técnicas. A verdadeira análise de fatores faz suposições diferentes sobre a estrutura subjacente dos dados e encontra os autovetores de uma matriz levemente diferente.

VII. ANÁLISE DA SEPARABILIDADE COM PCA

Utilizando o conjunto de amostras referente às duas classes abaixo, foi feito uma análise da separabilidade em 2 e 3 dimensões, utilizando a técnica de PCA com as componentes mais relevantes para as bases de dados utilizadas. As imagens a seguir mostram os resultados obtidos:

a. BreastCancer

Utilizando a técnica do PCA, as características mais relevantes segundo o PCA é mostrado na imagem abaixo:

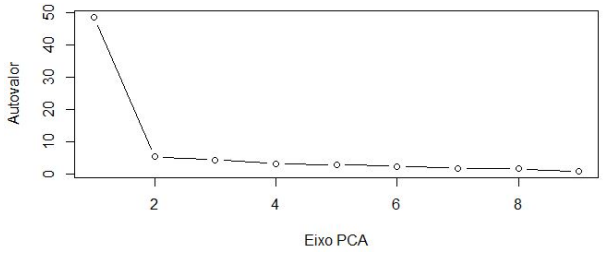

Imagem 3: Ilustração da relevância de cada caraterística do dataset.

As amostras resultantes do algoritmo PCA no plano resultante está abaixo, azul representa o tipo de câncer benigno e as em vermelho o maligno. As imagens 4 e 5 contém a visualização de separação das amostras.

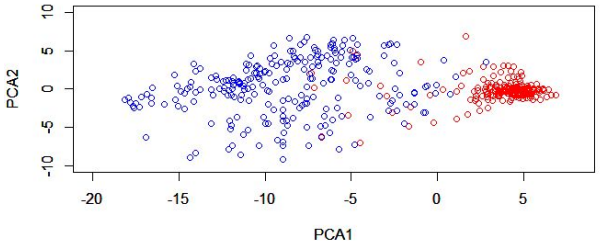

Imagem 4: Ilustração da separação dos dados utilizando apenas as 2 componentes mais relevantes.

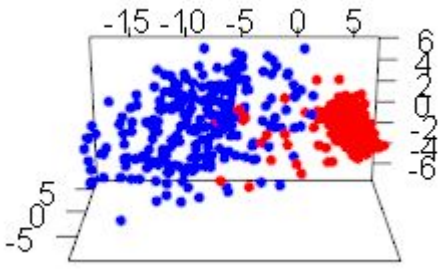

Imagem 5: Ilustração da separação dos dados utilizando apenas as 3 componentes mais relevantes.

1. Porcentagem de Acerto BreastCancer SVM com PCA:
   0.96774 ou 96.77%
2. Porcentagem de Acerto BreastCancer SVM sem PCA:
   0.93548 ou 93.55%

b. Cancer

Utilizando a técnica do PCA, as características mais relevantes segundo o PCA é mostrado na imagem abaixo:

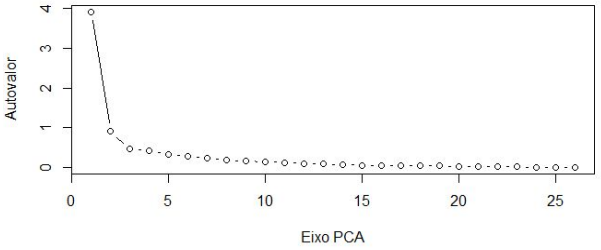

Imagem 6: Ilustração da relevância de cada caraterística do dataset.

As amostras resultantes do algoritmo PCA no plano resultante está abaixo, azul representa o tipo de câncer benigno e as em vermelho o maligno. As imagens 7 e 8 contém a visualização de separação das amostras.

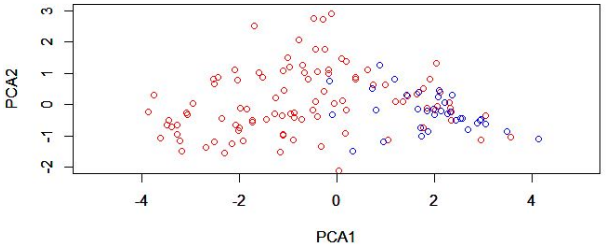

Imagem 7: Ilustração da separação dos dados utilizando apenas as 2 componentes mais relevantes.

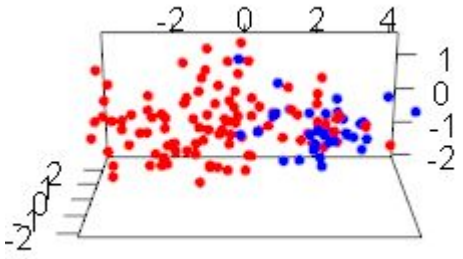

Imagem 8: Ilustração da separação dos dados utilizando apenas as 3 componentes mais relevantes.

3. Porcentagem de Acerto BreastCancer SVM com PCA:
   0.80769 ou 80.77%
4. Porcentagem de Acerto BreastCancer SVM sem PCA:
   0.71153 ou 71.15%

c. USArrests

Utilizando a técnica do PCA, as características mais relevantes segundo o PCA é mostrado na imagem abaixo:

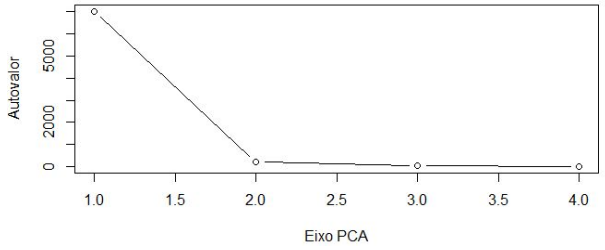

Imagem 9: Ilustração da relevância de cada caraterística do dataset.

As imagens 10 e 11 contém a visualização das amostras com 2 e 3 componentes do PCA.

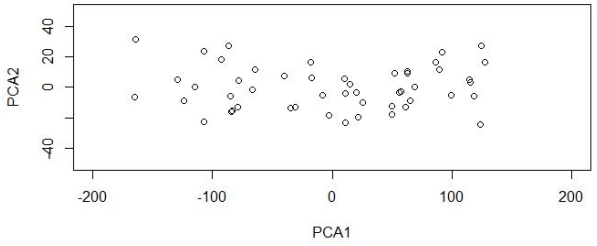

Imagem 10: Ilustração da separação dos dados utilizando apenas as 2 componentes mais relevantes.

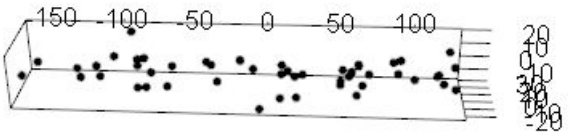

Imagem 11: Ilustração da separação dos dados utilizando apenas as 3 componentes mais relevantes.

Como dito anteriormente, a Análise de Componentes Principais tenta maximizar a variância dos dados, em menos dimensões. Para quantificar isso, calculou-se a variância explicada.

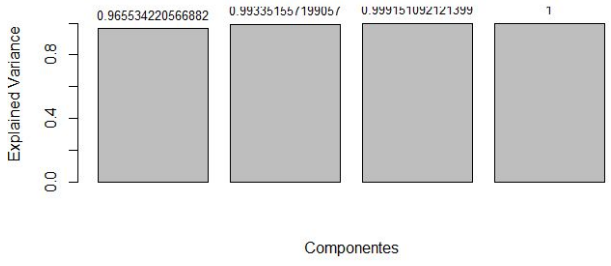

Imagem 12:Análise de componentes principais.

VIII. CONCLUSÕES

O método de Máquina de vetores de suporte (SVM) utilizando todas as características é bastante satisfatório, e quase sempre se mostra eficiente na separação das classes. Mas quando utiliza-se o SVM combinado com a técnica de Análise de componentes principais (PCA), o algoritmo se mostra mais robusto e tende a obter resultados razoavelmente melhores que somente o SVM.

O desempenho do modelo de SVM depende de dois parâmetros, cost e gamma e quanto melhor ajustados forem estes parâmetros, melhor será o resultado obtido do modelo.

Como foi visto, ambos os métodos satisfizeram as expectativas do trabalho. Os dois apresentam pontos positivos e negativos um em relação ao outro e a escolha entre estes modelos depende do problema a ser resolvido.

IX. REFERÊNCIAS

[1] R tutorial, http://www.r-tutor.com/r-introduction, acessado 30/08/2017.

[2] Antônio de Pádua Braga - Notas de Aulas de Redes Neurais Artificiais e de Reconhecimento de Padrões, acessado 22/09/2017.

[3] Análise de componentes principais (PCA), https://pt.wikipedia.org/wiki/An%C3%A1lise_de_componentes_principais, acessado 23/10/2017.