Trabalho #2: Classificação não-supervisionada

Instruções

Com uma base de dados pública (pode ser a mesma utilizada no trabalho #1) realizar os procedimentos para a concepção de um sistema de Reconhecimento de Padrões não-supervisionado, incluindo pré-processamento, redução de dimensionalidade, classificação (não supervisionada) e avaliação, conforme visto na aula prática, e considerando as seguintes condições:

1. Pré-processamento: realizar higienização, normalizações, etc necessárias;
2. Redução de dimensionalidade: reduzir a dimensionalidade utilizando a abordagem PCA. Após, selecionar as "n" primeiras Componentes Principais (CPs) que explicam a variabilidade dos dados acumulada até atingir 75%, 90% e 99%, respectivamente; Observe que na prática serão 3 "novas" base de dados, cada uma com o número de CPs (dimensão) correspondente à variância acumulada;
3. Caso a base de dados não forneça os dados particionados em conjuntos de treinamento e teste, separar aleatoriamente o dataset em ⅔ para treinamento e ⅓ para teste;
4. Considerando as bases de dados divididas (⅔ para treinamento e ⅓ para teste), experimentar os algoritmos de classificação não supervisionado abaixo:
    a. K-means (scikit-learn);
    b. Implementar outro método de agrupamento de sua escolha, como por exemplo o Fuzzy-means;
    c. Comparar o resultados das duas abordagens utilizando o critério de avaliação externo (labels) e interno (Silhouette e Critério de Fisher);
5. Fazer uma tabela mostrando o desempenho de classificação de cada classificador não supervisionado (k-means e de sua escolha) para o conjunto de teste, considerando os critérios de avaliação externo (labels) e interno (Silhouette e Critério de Fisher). Apresentar a matriz de confusão para cada algoritmo http://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
   OBS: observe que o treinamento do modelo deve acontecer **apenas** com os dados de treinamento, enquanto o teste **apenas** com os dados de teste.

Variância acumulada do PCA	Variância	Num CPs	Variância	Num CPs	Variância	Num CPs
	75%	1	90%	2	99%	3
K-means (interno)	Acurácia =	0.76	Acurácia =	0.76	Acurácia =	0.76
	TPR =	0.56	TPR =	0.56	TPR =	0.56
	TNR =	0.99	TNR =	0.99	TNR =	0.99
Fuzzy C-Means (interno)	Acurácia =	0.76	Acurácia =	0.79	Acurácia =	0.2
	TPR =	0.56	TPR =	0.61	TPR =	0.37
	TNR =	0.99	TNR =	0.99	TNR =	0.01
K-means (externo)	Silhouette =	0.62	Silhouette =	0.5	Silhouette =	0.47
	Fisher =	4.43	Fisher =	0.83	Fisher =	0.43
Fuzzy C-Means (externo)	Silhouette =	0.62	Silhouette =	0.49	Silhouette =	0.46
	Fisher =	4.43	Fisher =	0.8	Fisher =	0.41