Relatório Projeto 1 - Classificador K-NN

Guilherme F. Plichoski[1]

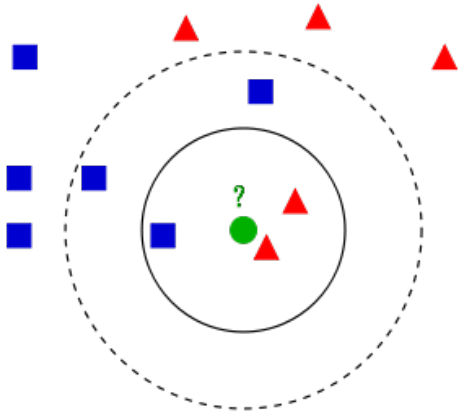
[1]Programa de Pós-Graduação em Computação Aplicada
Universidade Estadual de Santa Catarina (UDESC)
Joinville – SC – Brazil

guilherme.plichoski@edu.udesc.br

1. Classificador K-NN

O algoritmo k-NN (*k-nearest neighbors*) é um método não-paramétrico usado para classificação, onde a entrada consiste nas k amostras do conjunto de treinamento mais próximas da amostra a ser classificada, e a saída consiste na classe predita. Uma amostra é classificada sendo da mesma classe pertencente aos k vizinhos mais próximos. Por exemplo, na Figura 1, se $k = 1$, então a amostra é atribuída a classe do vizinho mais próximo, ou seja, a classe vermelha. Porém, se $k = 5$, a amostra seria atribuída a classe azul [Fukunaga and Narendra 1975].

Figura 1. Ilustração do classificador k-NN.



Dois parâmetros que devem ser determinados para aplicação do k-NN são a métrica de distância utilizada e o valor de k. A métrica de distância mais utilizada é a Distância Euclidiana (L_2-norm) de acordo com a Equação 1.

$$D = \sqrt{(p_1 - q_1)^2 + \ldots + (p_n = q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

Onde, $P = (p_1, \ldots, p_n)$ e $Q = (q_1, \ldots, q_n)$ são dois pontos n-dimensionais. Em relação ao parâmetro k, o valor ótimo está relacionado com a precisão do classificador, podendo variar de acordo com a base de dados. É recomendado o uso de valores ímpares, para evitar ao empate ao atribuir a classe. Dependendo do número de classes presente na base de dados, ainda assim pode haver empate, nestes casos, é recomendado pela literatura

decrementar *k* em 1 até haver desempate. O valor de *k* pode ser definido empiricamente ou através de alguma rotina de otimização. O Algoritmo 1 apresenta o pseudo-codigo para o classificador k-NN.

Algoritmo 1: Pseudo-codigo para o classificador k-NN

```
Preparar conjunto de dados de entrada e saída;
Informar valor de k;
for cada nova amostra do
    Calcular distância para todas as amostras;
    Determinar o conjunto das k's distâncias mais próximas;
    Escolher o rótulo com mais representantes no conjunto dos k vizinhos;
end
Retornar: Conjunto de rótulos de classificação;
```

2. Metodologia

Neste trabalho, duas metodologias foram utilizadas para otimização do classificador k-NN. As abordagens propostas não objetivam somente a otimização do parâmetro *k*, mas também do conjunto de treinamento usado como classificador. Primeiramente, a base de dados deve ser dividida em subconjuntos mutualmente exclusivos de treinamento (Z_1), avaliação (Z_2) e teste (Z_3). O número de amostras de cada subconjunto deve obedecer um percentual predefinido. Além disso, o número de classes dentro de cada subconjunto deve ser uniformemente distribuído. A seguir os métodos utilizados nesse trabalho são especificados.

No *método 1*, o algoritmo recebe as bases Z_1, Z_2 e Z_3, uma lista com possíveis valores para *k*, e o número de iterações (*T*) como entrada. A classificação das amostras no subconjunto de avaliação (Z_2) é realizada utilizando o subconjunto de treinamento (Z_1) para cada *k* presente na lista. Ao descobrir o valor de *k* que minimiza o erro, ou seja, maximiza a precisão, substitui-se as amostras classificadas erroneamente em Z_2 por amostras aleatórias da mesma classe em Z_1. Agora, com o valor de *k* fixo, repete-se este processo de troca de amostras por *T* iterações, assim retornando o subconjunto Z_1 que apresentou o menor número de erros. Este mesmo procedimento se repete no *método 2*, contudo ao invés de fixar o parâmetro *k*, a cada iteração *t* cada valor de *k* da lista é testado e considerado o valor que minimiza o erro. Finalmente, para cada método proposto o subconjunto de teste (Z_3) é classificado no subconjunto otimizado Z_1, assim retornando o número de erros constados.

3. Experimentos e Resultados

Para comparar os métodos apresentados neste trabalho, 5 bases de dados diferentes foram utilizadas, as quais serão especificadas a seguir. Como métrica de desempenho, o erro reportado na classificação das amostras no subconjunto Z_3 foi utilizado. Para garantir robustez, cada experimento foi realizado 30 vezes e os valores da média e desvio padrão dos erros foram analisados. Além disso, para avaliar o comportamento dos algoritmos na execução, gráficos da probabilidade média dos erros nos subconjuntos de avaliação e teste para cada iteração serão apresentados. Através destes gráficos é possível verificar

Tabela 1. Parâmetros utilizados nos experimentos.

Percentual do número de amostras			Lista de valores para k	Número de Iterações
Z1	Z2	Z3		
25%	25%	50%	1, 3, 5, 7, 9, 11, 13, 15	30

quando o processo de otimização converge, ou seja, quando não há melhoria em relação a minimização do erro, e também é possível perceber quão o erro da classificação do subconjunto de teste reflete o erro da classificação do subconjunto de avaliação. A tabela 1 apresenta os parâmetros utilizados para os dois métodos.

3.1. Conjunto de dados de plantas íris

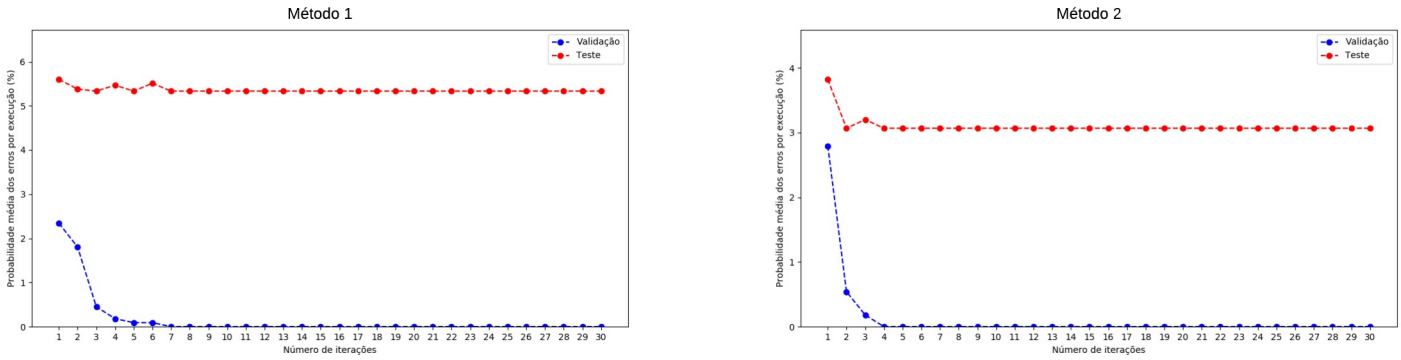
Esta base de dados contém 3 classes i.e setosa, versicolor e virgínia, com 50 amostras de cada e é uma das mais conhecidas em reconhecimento de padrões. Cada classe se refere a um tipo da planta íris. Cada amostra apresenta 4 atributos, sendo o comprimento e a largura da pétala e da sépala [Fisher 1936]. A Tabela 2 apresenta informações detalhadas sobre esta base.

Tabela 2. Informações detalhadas sobre o conjunto de dados de plantas íris.

Características da Base:	Multivariada
Características dos atributos:	Real
Número de amostras:	150
Número de atributos:	4
Valores faltantes?	Não
Data de referência:	01/07/1988

A seguir, a Figura 2 apresenta os gráficos da probabilidade média dos erros nos subconjuntos de avaliação e teste para os métodos 1 e 2, respectivamente.

Figura 2. Gráficos da probabilidade média dos erros para os métodos 1 e 2, respectivamente.



É possível perceber que ambos os métodos convergem em torno da 5ª iteração. Em relação ao subconjunto de avaliação, podemos observar que os dois métodos tem um comportamento bem similar. Contudo para o subconjunto de teste, o método 2 obteve uma

probabilidade média dos erros menor que no método 1. Os valores dos erros absolutos para os 30 experimentos foram 2.30±1.15 para o método 2 e 4.00±1.31 para o método 1.

3.2. Conjunto de dados de localização de proteínas de E.coli

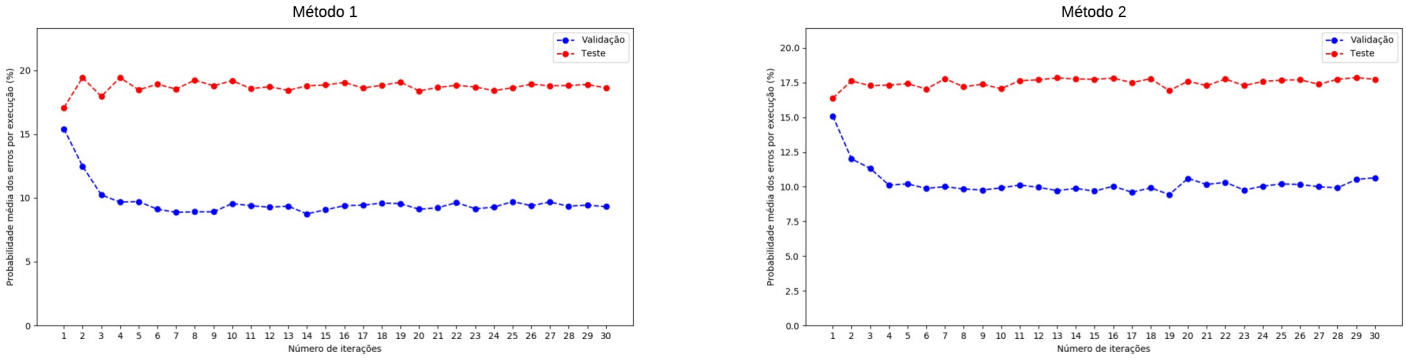
Este conjunto se refere ao problema de classificar proteínas de E.coli de acordo com a localização celular correspondente. É composto por 8 classes i.e citoplasma, membrana interna sem sequência de sinal, periplasma, membrana interna com sequência de sinal inacessível, membrana externa, lipoproteína da membrana externa, lipoproteína da membrana interna e membrana interna com sequência de sinal acessível. Foram utilizados 7 atributos para a classificação i.e. métodos McGeoch e von Heijne para reconhecimento de sequencia de sinal, sinal peptidase II de von Heijne, presença de carga no terminal N das lipoproteínas, pontuação da análise discriminante do amino ácido, pontuação do programa de previsão da região de abrangência da membrana ALOM e a pontuação do programa ALOM após a exclusão do sinal putativo de clivagem [Horton and Nakai 1996]. A Tabela 3 apresenta informações detalhadas sobre esta base.

Tabela 3. Informações detalhadas sobre o conjunto de dados de proteínas de E.coli.

Características da Base:	Multivariada
Características dos atributos:	Real
Número de amostras:	336
Número de atributos:	7
Valores faltantes?	Não
Data de referência:	01/09/1996

A seguir, a Figura 4 apresenta os gráficos da probabilidade média dos erros nos subconjuntos de avaliação e teste para os métodos 1 e 2, respectivamente.

Figura 3. Gráficos da probabilidade média dos erros para os métodos 1 e 2, respectivamente.



Neste experimento, não podemos afirmar que o algoritmo converge até a 30ª iteração, contudo, a variação apartir da 10ª iteração é mínima. Com isto, podemos inferir uma tendência da estabilização do erro. Em relação a precisão, os dois métodos atingiram um resultado muito similar, em torno 17.5% de probabilidade média dos erros. Os valores absolutos para as os métodos 1 e 2 foram 31.37±5.47 e 28.43±4.04, respectivamente.

3.3. Conjunto de amostras de medidas geométricas do trigo

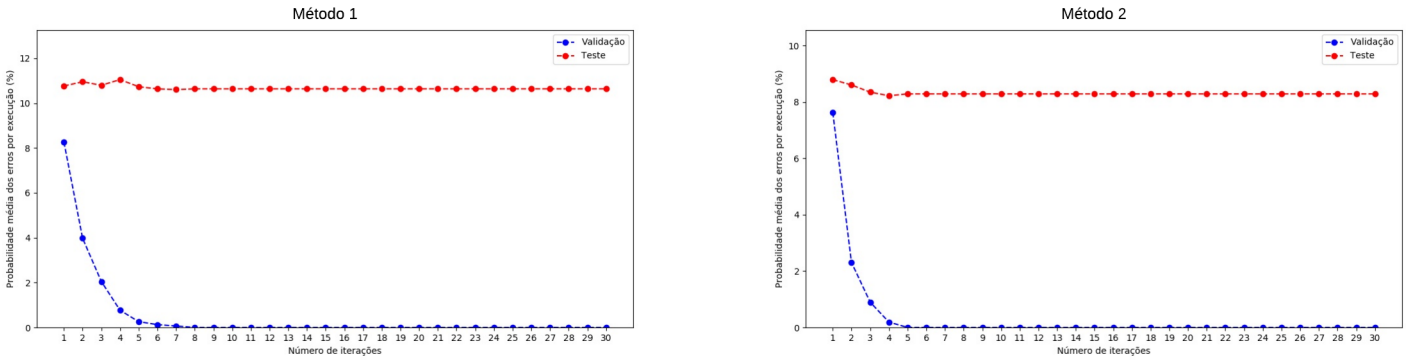
Este conjunto se refere a medidas geométricas de três variantes do trigo i.e. kama, rosa e canadense. Os 7 atributos de cada amostra foram medidos utilizando técnicas de raio-X, e as imagens foram capturadas em placas KODAK de raio-X. Foram coletadas informações de área e perímetro, medidas de compacidade, comprimento e largura do núcleo, coeficiente de assimetria e comprimento do sulco do núcleo [Charytanowicz et al. 2010]. Para cada espécie de trigo foram coletadas 70 amostras. A Tabela 4 apresenta informações detalhadas sobre esta base.

Tabela 4. Informações detalhadas sobre o conjunto de dados de amostras de trigo.

Características da Base:	Multivariada
Características dos atributos:	Real
Número de amostras:	210
Número de atributos:	7
Valores faltantes?	Não verificado
Data de referência:	29/09/2012

A seguir, a Figura 4 apresenta os gráficos da probabilidade média dos erros nos subconjuntos de avaliação e teste para os métodos 1 e 2, respectivamente.

Figura 4. Gráficos da probabilidade média dos erros para os métodos 1 e 2, respectivamente.



Em relação aos resultados, os dois métodos obtiveram um comportamento muito similar, convergindo em torno da 6 iteração. Porém, o método 2 obteve um ganho de aproximadamente 2% na probabilidade média dos erros. Os valores absolutos dos erros para as 30 iterações foram de 11.17 ± 2.59 e 8.70 ± 1.78 para os métodos 1 e 2, respectivamente.

3.4. Conjunto de dados de análise química do vinho

Este conjunto contém amostras de três tipos de vinhos coletadas na mesma região da Itália, porém de três cultivadores diferentes. Os atributos de cada amostra foram coletados através de análises químicas realizadas com estes vinhos, sendo a quantidade de álcool, ácido málico, quantidade de cinza, alcalinidade das cinzas, quantidade de

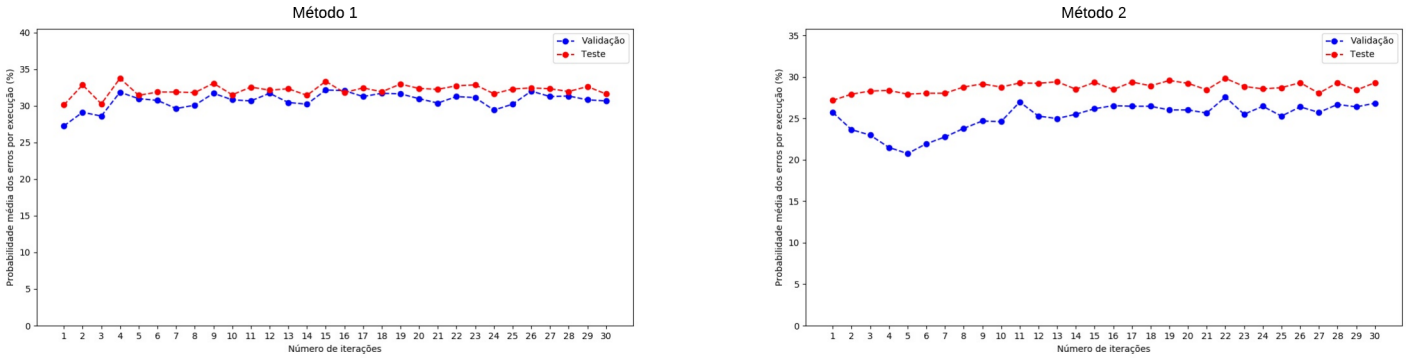
magnésio, fenóis totais, flavonóides, fenóis não flavonóides, proantocianidinas, intensidade da cor, matiz, OD280/OD315 diluição do vinho e prolina [Aeberhard et al. 1993]. A Tabela 5 apresenta informações detalhadas sobre esta base.

Tabela 5. Informações detalhadas sobre o conjunto de dados de amostras de trigo.

Características da Base:	Multivariada
Características dos atributos:	Inteiro e Real
Número de amostras:	178
Número de atributos:	13
Valores faltantes?	Não
Data de referência:	01/07/1991

A seguir, a Figura 5 apresenta os gráficos da probabilidade média dos erros nos subconjuntos de avaliação e teste para os métodos 1 e 2, respectivamente.

Figura 5. Gráficos da probabilidade média dos erros para os métodos 1 e 2, respectivamente.



Em relação aos outros experimentos, este obteve um comportamento mais irregular. Contudo, a variação da probabilidade média do erro foi muito baixa para os dois métodos por toda otimização. Neste caso, também o método 2 teve um ganho de aproximadamente 2% com o método 2. Os valores absolutos médio dos erros foram 28.90$\pm$2.37 para o método 1 e 25.80$\pm$2.61 para o método 2.

3.5. Conjunto de dados de localização de proteínas de Leveduras

Este conjunto se refere ao problema de classificar proteínas de Leveduras de acordo com a localização celular correspondente. É composto por 10 classes i.e citosólico ou citoesqulético, nuclear, mitocondrial, três tipos de proteína de membrana, uma com sinal não clivado, outra com sinal clivado e outra sem sinal N-terminal, extracelular, vacuolar, peroxisomal e lúmen do retículo endoplasmático. Foram utilizados 8 atributos para a classificação i.e. métodos McGeoch e von Heijne para reconhecimento de sequencia de sinal, pontuação do programa de previsão da região de abrangência da membrana ALOM, pontuação de análise discriminante do conteúdo de aminoácidos da região N-terminal (20 resíduos de comprimento) de mitocôndrias e não-mitocôndrias, presença

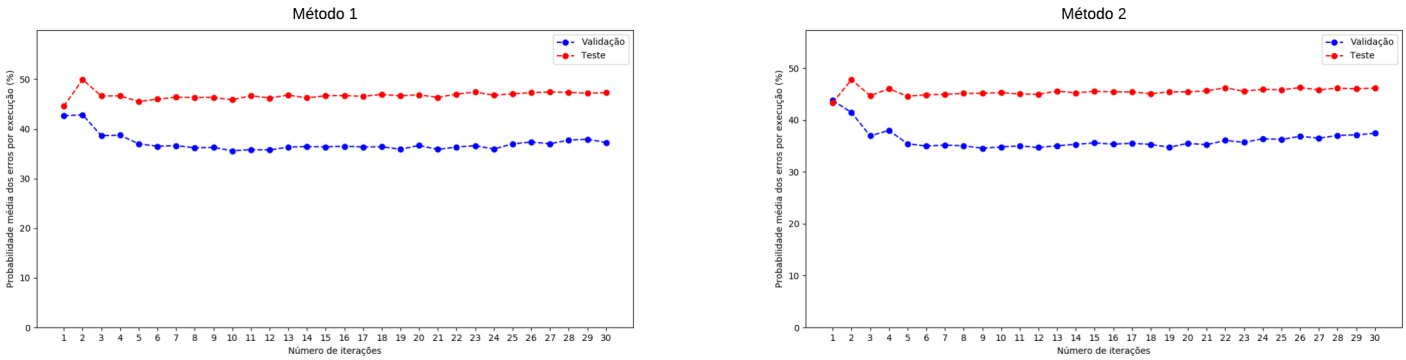
de ”HDEL”, sinal de direcionamento peroxissomal no terminal C, pontuação de análise discriminante do conteúdo de aminoácidos vacuolares e extracelulares e pontuação da análise discriminante dos sinais de localização nuclear de proteínas nucleares e não nucleares [Horton and Nakai 1996]. A Tabela 6 apresenta informações detalhadas sobre esta base.

Tabela 6. Informações detalhadas sobre o conjunto de dados de localização de proteínas de Leveduras.

Características da Base:	Multivariada
Características dos atributos:	Real
Número de amostras:	1484
Número de atributos:	8
Valores faltantes?	Não
Data de referência:	01/09/1996

A seguir, a Figura 6 apresenta os gráficos da probabilidade média dos erros nos subconjuntos de avaliação e teste para os métodos 1 e 2, respectivamente.

Figura 6. Gráficos da probabilidade média dos erros para os métodos 1 e 2, respectivamente.

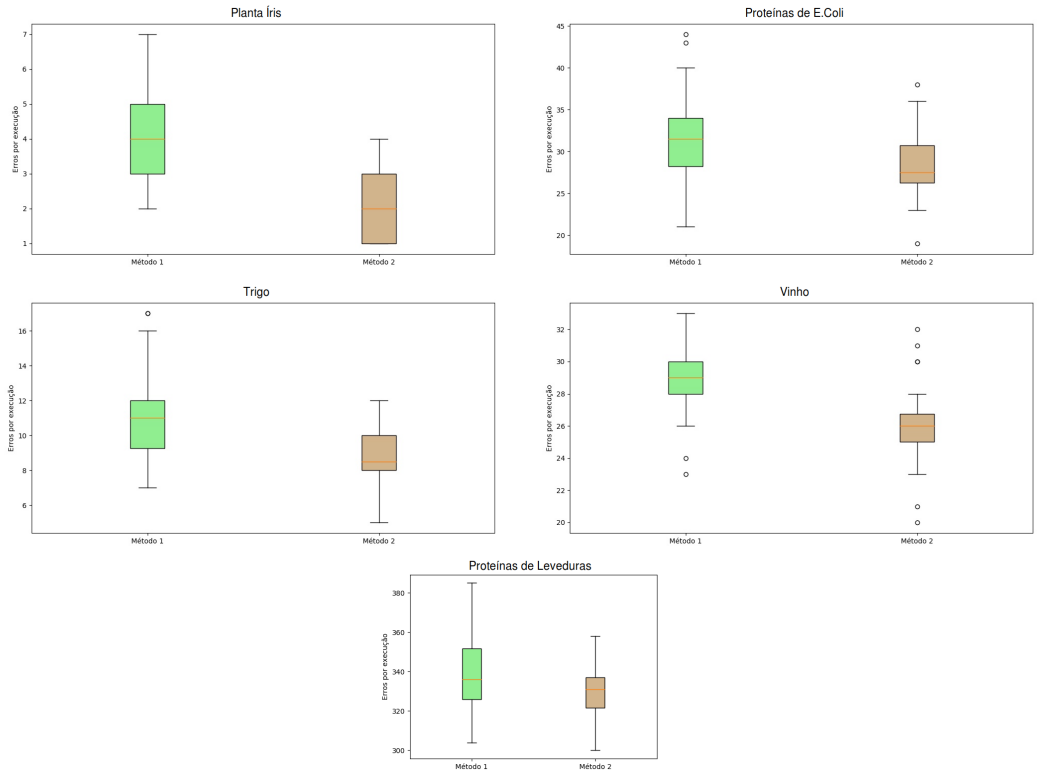


Neste caso, os dois métodos apresentam um comportamento bem similar, assim pode-se perceber a convergência do algoritmo em torno da 10ª iteração para os dois métodos. A probabilidade média do erro para os experimentos nos dois métodos foi em torno de 45%, sendo a mais baixa em comparação com as outras bases de dados. Isso se deve a alta complexidade do subconjunto contendo 1484 amostras. Os valores absolutos dos erros para as 30 iterações foram de 347.30 ± 15.07 para o método 1 e 332.63 ± 13.94 para o método 2.

3.6. Comparação entre os métodos

Para avaliar se os dois métodos tem diferença estatística, a Figura 3.6 apresenta os diagramas de caixa dos dois métodos para cada base de dados apresentada neste trabalho. O gráfico de diagrama de caixa apresenta a dispersão dos erros por meio de quartis.

Como pode-se observar, os diagramas de caixa apresentam uma interseção em todos os casos, assim, é difícil de se avaliar se existe uma diferença real entre os dois



métodos. Para uma conclusão precisa, utiliza-se um teste de hipótese que usa conceitos estatísticos para rejeitar ou não uma hipótese nula. Assumindo que as amostras (experimentos) provém de uma população normal, neste trabalho foi utilizado o *teste t* [Lilja 2005]. A Tabela 7 mostra o *p-value* resultante do teste.

Tabela 7. *P-value* para cada base de dados.

Base de Dados	***p-value***
Planta Íris	0.126
Proteínas de E.coli	0.021
Trigo	7.657e-05
Vinho	1.08e-05
Proteínas de Leveduras	0.0002

Com um nível de confiança de 95%, podemos rejeitar a hipótese de que há diferença significativa entre os dois métodos para o conjunto de dados de plantas Íris ($p\text{-}value >= 0.05$), já para os outros experimentos não pode-se descartar essa hipótese ($p\text{-}value <= 0.05$). Assim, como em todos os casos o método 2 obteve melhor probabilidade média dos erros, este método torna-se mais atrativo em relação ao método 1.

Referências

Aeberhard, S., Coomans, D., and Vel, O. D. (1993). Improvements to the classification performance of rda. *Journal of chemometrics*, 7(2):99–115.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P. A., Łukasik, S., and Żak, S. (2010). Complete gradient clustering algorithm for features analysis of x-ray images. In *Information technologies in biomedicine*, pages 15–24. Springer.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of human genetics*, 7(2):179–188.

Fukunaga, K. and Narendra, P. M. (1975). A branch and bound algorithm for computing k-nearest neighbors. *IEEE transactions on computers*, 100(7):750–753.

Horton, P. and Nakai, K. (1996). A probabilistic classification system for predicting the cellular localization sites of proteins. In *Ismb*, volume 4, pages 109–115.

Lilja, D. J. (2005). *Measuring computer performance: a practitioner's guide*. Cambridge university press.