Métodos de Classificação
Aprendizagem de Máquina
2023

André Luiz Brun[1]

[1]Colegiado de Ciência da Computação
Campus de Cascavel - UNIOESTE

Resumo. *Este documento consiste na especificação formal do primeiro trabalho da disciplina de Aprendizagem de Máquina (Csc3040) para o ano letivo de 2023. Aqui são apresentadas as atividades a serem desenvolvidas e como cada processo deverá ser realizado. Além disso, o documento contém as informações sobre a formação das equipes, o objeto de trabalho de cada uma e as datas de entrega e apresentação dos relatórios.*

1. Introdução

O objetivo do primeiro trabalho da disciplina consiste em comparar o comportamento, em termos de acurácia, de classificadores baseados em diferentes conceitos sobre uma mesma base de dados. Além disso, pretende-se comparar algumas estratégias de combinação desses classificadores e analisar se a adoção da estratégia de múltiplos classificadores leva a melhores taxas de acerto.

2. Implementação

Nesta seção são descritas como cada etapa do desenvolvimento deve ser realizada segundo os conceitos vistos durante a disciplina.

2.1. Análise descritiva dos dados

Nesta etapa deve-se fazer uma análise descritiva dos dados, apresentando características da base como tamanho, dimensão, origem, número de classes, tipos dos atributos, valores médios, máximos e mínimos dos atributos etc. **Além disso, é necessário fazer uma explanação sobre o problema em questão, falando um pouco da aplicação e do significado dos dados coletados**.

Caso seja de interesse da equipe, pode ser feita uma análise de correlação entre os atributos. Este processo pode ser feito utilizando-se o coeficiente de correlação de Pearson ou mesmo através de uma representação gráfica bidimensional em que cada eixo representa os valores de um dos atributos.

2.2. Divisão do conjunto de dados

O primeiro passo consistirá na divisão da base original em três subconjuntos mutuamente exclusivos: treino, teste e validação (conforme apresentado na Figura 1). A instância que for designada para um conjunto não deve aparecer nos outros.

O conjunto de treino deverá possuir 50% do tamanho do arquivo original. Já as bases de validação e teste, terão 25% da dimensão. No momento de separar a base original

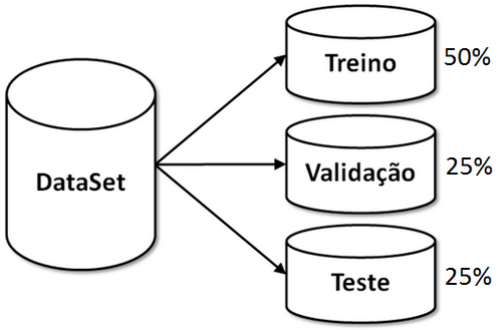

Figura 1. Divisão estratificada do conjunto de entrada

nos três conjuntos (treino, teste e validação), deve-se manter as proporções originais das classes. Por exemplo, se um conjunto possui 200 instâncias da classe A e 100 da classe B, o conjunto de treino terá 100 instâncias da classe A e 50 da classe B.

Importante 1: a escolha das instâncias que formarão cada um dos conjuntos deve ser totalmente aleatória.

Importante 2: lembrem-se de sempre "bagunçar" os conjuntos de dados **antes** de fazer a divisões e de realizar o treinamento. A adoção de aleatoriedade adiciona robustez ao processo.

2.3. Treinamento e Calibração dos Modelos

Depois de formados os conjuntos, o passo seguinte será o treinamento dos modelos de classificação. Nesta tarefa deverão ser implementadas as estratégias dos K Vizinhos mais próximos (KNN), Árvore de Decisão (AD), Naive Bayes (NB), Máquina de Vetor de Suporte (SVM) e Multilayer Perceptron (MLP).

Para se determinar quais os melhores parâmetros dos métodos de classificação, deve-se adotar o conjunto de validação (conforme ilustrado na Figura 2). Por exemplo, digamos que estamos treinando um KNN e queremos decidir qual o melhor K a ser empregado. Deve-se treinar o classificador com o conjunto de treino e então variar o valor de K e analisar quanto o classificador acerta do conjunto de validação. O valor de K que levar à maior acurácia (ou menor taxa de erros) é usado no momento de classificar o conjunto de teste. Os parâmetros que deverão ser definidos para cada classificador são apresentados na Tabela 1.

2.4. Avaliação dos Modelos

Definidos os melhores parâmetros para cada classificador, o passo seguinte será avaliar os seus desempenhos sobre o conjunto de teste (tal processo é ilustrado na Figura 3). Nesta etapa deverá ser guardada a acurácia de cada classificador ao longo das 20 execuções. Ao término desta etapa, terão sido obtidos 100 valores de acurácias (20 para cada classificador).

Para que o processo tenha base para análise estatística, deverão ser executadas 20 repetições. Os valores a serem comparados deverão ser os valores médios das 20 execuções. Um exemplo de representação dos resultados é ilustrado na Tabela 2 onde cada coluna corresponde ao desempenho de um método de classificação monolítico ao

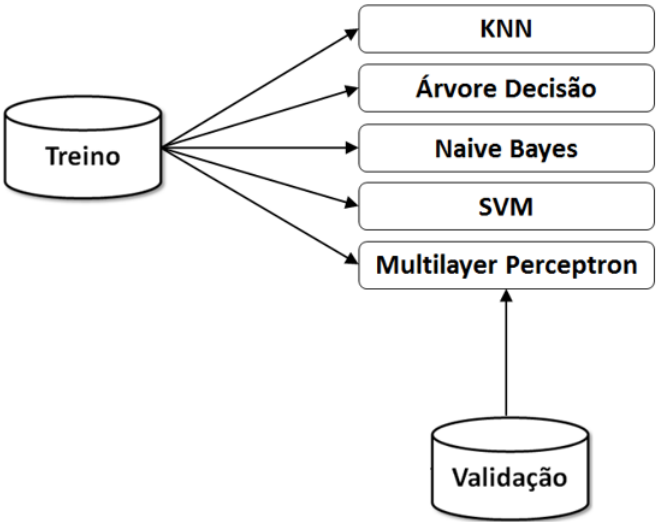

Figura 2. Adoção do conjunto de validação na estimação dos parâmetros

Tabela 1. Conjunto de parâmetros a serem calibrados através do Grid-search

Classificador	Parâmetros
KNN	K distance
AD	criterion max_depth min_samples_split min_samples_leaf
SVM	kernel C
NB	-
MLP	hidden_layer_sizes activation max_iter learning_rate

longo das 20 execuções do experimento. Na última linha são apresentados a acurácia média e o desvio padrão do longo das execuções.

2.5. Análise Comparativa

A última etapa consiste na comparação das acurácias dos métodos para descobrir qual deles obteve o melhor desempenho. Para tanto, deve-se executar dois testes estatísticos. O primeiro servirá para detectar se há diferença entre o desempenho dos algoritmos (independente de qual foi melhor ou pior). O segundo teste estatístico serve para comparar, dois a dois, os classificadores com o objetivo de avaliar se eles têm desempenhos significativamente diferentes e quem é o melhor.

Para avaliar se há pelo menos um classificador com desempenho diferente dos demais utilizem o teste de Kruskal-Wallis com 5% de significância. Caso haja pelo menos um classificador com comportamento diferente deve-se aplicar o teste de Mann-Whitney

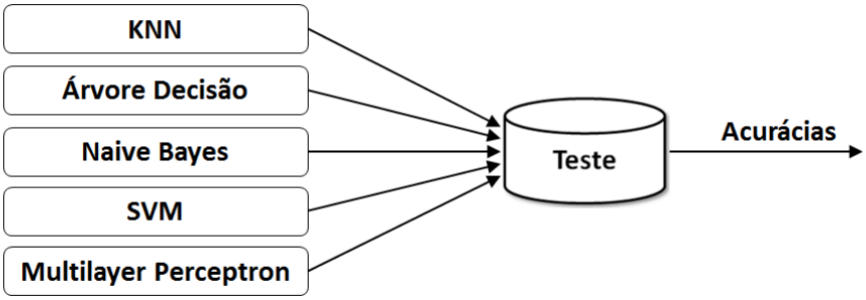

Figura 3. Adoção do conjunto de validação na estimação dos parâmetros

Tabela 2. Exemplo de estrutura para análise dos resultados dos sistemas monolíticos

Repetição	KNN	AD	NB	SVM	MLP
1	Acc	Acc	Acc	Acc	Acc
2	Acc	Acc	Acc	Acc	Acc
...	...	...	...	...	...
20	Acc	Acc	Acc	Acc	Acc
	Média (DP)	Média (DP)	Média (DP)	Média (DP)	Média (DP)

(bicaudal), também com 5% de significância, para identificar quais classificadores apresentaram comportamento discrepante.

Os testes podem ser realizados via código em python, usando a biblioteca scipy (conforme exemplo visto em sala) ou pelos endereços Kruskal-Wallis e Mann-Whitney.

2.6. Sistemas de Múltiplos Classificadores

Além das estratégias de classificação apresentadas anteriormente deve-se implementar três abordagens de combinação de classificadores:

- Regra do Soma
- Voto Majoritário
- Borda Count

Estas abordagens devem combinar a opinião dos cinco classificadores desenvolvidos na primeira etapa do trabalho. Ao término do processo de execução será obtido uma estrutura similar à apresentada na Tabela 3. Essa representação seque os mesmos princípios daquela com os desempenhos dos modelos monolíticos (Tabela 2).

Tabela 3. Exemplo de estrutura para análise dos resultados dos SMCs

Repetição	Soma	Voto Majoritário	Borda Count
1	Acc	Acc	Acc
2	Acc	Acc	Acc
...	...	...	...
20	Acc	Acc	Acc
	Média (DP)	Média (DP)	Média (DP)

A avaliação dos modelos de combinação de classificadores deverão seguir os mesmos princípios especificados para os modelos monolíticos. Deve-se comparar a média de

acurácia das vinte execuções de forma a identificar, através do teste de Kruskal-Walllis se há diferença significativa entre as estratégias de múltiplos classificadores. Caso haja pelo menos uma discrepante (rejeitando-se H_0), devem então ser aplicado o teste de Mann-Whitney par-a-par.

Importante: uma vez que as estratégias de combinação utilizam o percentual de confiança do voto de cada classificador, recomendo que, no momento de avaliar os classificadores (sobre o conjunto de testes), sejam salvos esses percentuais de confiança de cada classificador para cada instância. Dessa forma ganha-se tempo na execução do trabalho.

2.7. Comparação entre abordagem monolítica e SMC

O último passo do processo de análise consiste na comparação, em termos de acurácia, do melhor modelo monolítico e da melhor estratégia baseada em sistemas de múltiplos classificadores. Uma vez que a comparação dar-se-á apenas entre dois modelos, será utilizado o teste estatístico de Mann-Whitney.

2.8. Como fazer?

A linguagem adotada é de escolha da dupla. Entretanto, é fortemente indicado o uso de Python ou Java.

Não é necessário implementar os métodos de classificação. Neste caso, pode-se e é indicado, que sejam utilizadas implementações prontas dos métodos, ficando a carga da dupla apenas a implementação do framework e análise dos parâmetros e resultados.

Da mesma forma, para o carregamento, aleatorização, divisão e sorteio dos conjuntos de treino, teste e validação podem ser utilizadas funções próprias das linguagens.

Caso a linguagem tenha implementada as soluções de combinação elencadas acima (soma, voto majoritário e borda count), a equipe poderá fazer uso delas nos experimentos, sem ter que reimplementá-las.

3. Equipes

Na Tabela 4 são apresentadas as composições de cada equipe bem como o problema sobre qual cada uma trabalhará. Além disso, são apresentados os endereços eletrônicos onde as bases de dados podem ser obtidas.

4. O que deve ser entregue

4.1. Relatório

Deve ser elaborado um relatório técnico em formato pdf contendo:

- Detalhamento de quais foram os parâmetros empregados em cada método de classificação e em qual faixa de valores cada parâmetro foi variado. Por exemplo, no KNN, seria possível variar o valor de k entre 1 e 20.
- Análise estatística indicando se há diferença significativa no desempenho dos métodos, ou seja, se há algum classificador que seja diferente dos demais.
- A análise deve mostrar quais classificadores tiveram desempenho similar e quais foram mais acurados durante o processo.

O formato do relatório deve ser a formatação presente neste texto. As regras para tal podem ser obtidas no link download. No arquivo disponível pode-se utilizar a formatação em arquivo .doc ou em latex.

Tabela 4. Formação das equipes e conjunto de dados para o trabalho

ID	Equipe	Base	Fonte
1	Felipi Lima Matozinho João Luiz Reolon	Vehicle	link
2	Gabriel Norato Claro Maria Eduarda Crema Carlos	German	link
3	Jaqueline Cavaller Faino Davi Marchetti Giacomel	ILPD	link
4	Bruno Stafuzza Maion Rodrigo da Rosa	WDVG	link
5	Heloisa Aparecida Alves Vinicius Muller de Freitas	CTG	link
6	Gustavo Pauli da Luz Guilherme de Oliveira Correia	ImageSegmentation	link
7	Rafael Gotz Gabriel Alves Mazzuco Rodrigo Brickmann Rocha	Phoneme	link

4.2. Código-fonte

Além do relatório citado, cada equipe deverá enviar os códigos fonte construídos para a execução dos experimentos. Ambos arquivos podem ser compactados e enviados como arquivo único.

5. Para quando?

O trabalho deverá ser submetido no link disponibilizado na turma de disciplina dentro do ambiente Microsoft Teams até as 23:59 do dia 23/10/2023.

As apresentações serão realizadas na aula do dia 24/10/2023.

Cada grupo terá 15 minutos para apresentar o trabalho realizado, focando na descrição do problema, nos desempenhos obtidos e no resultado da análise estatística.