**Aula Prática 7**

**Tema: Comparação de Classificadores**

**Murielly Oliveira Nascimento – 11921BSI222**

**Ferramentas Usadas**

As análises foram feitas com o uso da ferramenta Weka. Os algoritmos usados para as análises foram K-vizinhos mais próximos com K=3 e K=6; Árvore de Decisão com o mínimo número de instâncias em cada folha igual a 2 e 4; e BayesNet com algoritmos de busca K2 e TAN, a explicação deste se encontra na próxima seção. Para a divisão da base de dados em treino e teste foi usada a estratégia Cross-Validation com K=10.

As medidas de avaliação usadas foram Precisão, Revocação e Medida F. Todas são baseadas na matriz de confusão:

Tabela

Descrição gerada automaticamente com confiança média

Medida F =

Precisão =

Revocação =

As Bases de Dados escolhidas para testes foram Breast Cancer Data Set e Space Shuttle Autolanding Domain, que podem ser encontradas nos seguintes links:

<https://archive.ics.uci.edu/ml/datasets/breast+cancer>

<https://archive.ics.uci.edu/ml/datasets/Shuttle+Landing+Control>

**O Algoritmo BayesNet**

*A principal vantagem de raciocínio probabilístico sobre raciocínio lógico é o fato de que agentes podem tomar decisões racionais mesmo quando não existe informação suficiente para se provar que uma ação funcionará (*RUSSEL)

Grafos conseguem representar relações causais entre eventos. Uma Rede Bayesiana leva em consideração a dependência dos atributos entre si com o uso de grafos acíclicos. Assim cada nó são instâncias e as arestas dependências condicionais, se dois nós não estão conectados então eles são independentes.

Um possível método para construção das redes bayesianas é como segue:

***Escolha*** *um conjunto de variáveis Xi que descrevam o domínio*

***Escolha*** *uma ordem para as variáveis*

***Enquanto*** *existir variáveis*

***Escolha*** *uma variável Xi e adicione um nó a rede*

***Determine*** *os nós Pais(Xi) dentre os nós que já estejam na rede e que tenham influência direta em Xi*

***Defina*** *a tabela de probabilidades condicionais para Xi.*

De acordo com a professora Inês Dutra da UFRJ, Redes Bayesianas constituem uma forma natural para representação de informações condicionalmente independentes. É uma boa solução para problemas onde conclusões não podem ser obtidas apenas do domínio do problema.

*Como resultado, vários algoritmos especializados em realizar a pesquisa em diferentes tipos de topologias de grafos acíclicos direcionados (DAG) foram desenvolvidos, sendo a maioria deles extensões (usando arcos de aumento) ou modificações de a topologia básica Naive Bayes. Esta abordagem geralmente obtém resultados mais satisfatórios. (ACID, 2005)*

Na ferramenta WEKA é possível escolher os algoritmos de busca a serem usados no classificador NetBayes. Foram usados o K2 e o TAN. O primeiro, de acordo com a professora Carolina Ruiz, busca heuristicamente a mais provável estrutura de rede de crenças dado uma base de dados. O segundo é um acrônimo para Tree Augmented Naive Bayes. É um método de aprendizado bayesiano semi-ingênuo. Ele relaxa a suposição de independência do atributo ingênuo de Bayes, empregando uma estrutura de árvore, na qual cada atributo depende apenas da classe e de um outro atributo. Uma árvore de abrangência ponderada máxima que maximiza a probabilidade dos dados de treinamento é usada para realizar a classificação (ZHENG, 2011)

**Análise da Base Breast Cancer**

**Informações**

A base Breast Cancer é composta por dados coletados da University Medical Centre, Instituto de Oncology, Ljubljana, Yugoslavia. Tem um total de 286 instâncias, sendo 201 da classe câncer de mama *recorrente* e 85 da *não recorrente.* Há a presença de valores ausentes. São 9 atributos do tipo nominal descritos da seguinte forma:

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

* **Idade**: idade do paciente no momento do diagnóstico que cobre os seguintes intervalos (10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.)

Gráfico, Gráfico de barras

Descrição gerada automaticamente

* **Menopausa**: se a paciente está na pré ou pós-menopausa no momento do diagnóstico, descrito da seguinte forma: lt40, ge40, premeno).

Uma imagem contendo Gráfico de barras

Descrição gerada automaticamente

* **Tamanho do tumor**: o maior diâmetro (em mm) do tumor excisado, divido por intervalos (0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59).

Gráfico, Gráfico de barras

Descrição gerada automaticamente

* **Inv-nodes**: o número (intervalo de 0 - 39) de linfonodos axilares que contêm câncer de mama metastático visível no exame histológico (0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39).

Uma imagem contendo Gráfico

Descrição gerada automaticamente

* **Caps de linfonodo**: se o câncer metástase para um linfonodo, embora fora do local original do tumor, ele pode permanecer “contido” pela cápsula do linfonodo. Porém, com o tempo, e com doença mais agressiva, o tumor pode substituir o linfonodo e então penetrar na cápsula, permitindo que invada os tecidos circundantes. (sim ou não)

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

* **Grau de malignidade**: o grau histológico (intervalo 1-3) do tumor. Os tumores de grau 1 consistem predominantemente em células que, embora neoplásicas, retêm muitas de suas características usuais. Os tumores de grau 3 consistem predominantemente em células altamente anormais. (1, 2, 3.)

Gráfico

Descrição gerada automaticamente com confiança baixa

* **Mama**: o câncer de mama pode obviamente ocorrer em qualquer uma das mamas. (direita ou esquerda)

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente com confiança média

* **Quadrante da mama**: a mama pode ser dividida em quatro quadrantes, tendo como ponto central o mamilo (esquerda cima, esquerda baixo, direita cima, direita baixo, central).

Gráfico, Gráfico de barras

Descrição gerada automaticamente

* **Irradiação**: a radioterapia é um tratamento que utiliza raios-x de alta energia para destruir as células cancerígenas (sim ou não).

Uma imagem contendo Gráfico

Descrição gerada automaticamente

* **Classe**: A classificação é dada por recorrente e não recorrente.

Gráfico

Descrição gerada automaticamente

**Pré-processamento**

O atributo classe é o primeiro na base de dados. Para a ferramenta Weka reconhecê-lo como a *classe* usei o filtro Reorder para colocá-la como o último atributo. Os dados da base são categóricos e precisam ser convertidos em numéricos para análises. O filtro usado para a conversão foi OrdinalToNumeric. Por fim os valores ausentes foram substituídos pelas médias dos respectivos atributos usando o filtro ReplaceMissingValues.

Interface gráfica do usuário, Texto, Aplicativo, Email

Descrição gerada automaticamente

* Classe

Gráfico, Gráfico de barras

Descrição gerada automaticamente com confiança média

* Irradiação

Gráfico

Descrição gerada automaticamente

* Idade

Gráfico, Gráfico de barras

Descrição gerada automaticamente

* Menopausa

Gráfico, Gráfico de barras

Descrição gerada automaticamente

* Tamanho do Tumor

Gráfico

Descrição gerada automaticamente

* Inv-Nodes

Uma imagem contendo Texto

Descrição gerada automaticamente

* Caps de linfonodo

Gráfico

Descrição gerada automaticamente

* Grau de malignidade

Gráfico, Gráfico de barras

Descrição gerada automaticamente

* Mama

Gráfico

Descrição gerada automaticamente

* Quadrante da mama

Gráfico

Descrição gerada automaticamente

**Avaliação**

Para todos os casos estou considerando a classe Recorrentes como positiva, por ter o menor número de instâncias e a classe Não Recorrente como negativa. A ferramenta Weka faz o cálculo de considerando cada uma das classes, porém para as análises feitas nesse trabalho usei a lógica acima.

A matriz de confusão do Algoritmo K-vizinhos mais próximos é como segue:

* Com K = 3

Texto

Descrição gerada automaticamente

* Com K = 6

Texto

Descrição gerada automaticamente

|  |  |  |
| --- | --- | --- |
|  | K = 3 | K = 6 |
| Precisão | 0,422 | 0,563 |
| Revocação | 0,318 | 0,212 |
| Medida F | 0,362 | 0,308 |

A matriz de confusão do Algoritmo Árvore de Decisão e a árvore gerada são como seguem:

* Com número de instâncias por folha igual a 2

Texto

Descrição gerada automaticamente

Diagrama

Descrição gerada automaticamente

* Com número de instâncias por folha igual a 4

Texto

Descrição gerada automaticamente

Diagrama

Descrição gerada automaticamente

|  |  |  |
| --- | --- | --- |
|  | N = 2 | N = 4 |
| Precisão | 0,512 | 0,500 |
| Revocação | 0,247 | 0,271 |
| Medida F | 0,333 | 0,351 |

A matriz de confusão e o grafo gerado pelo algoritmo BayesNet são como seguem:

* Com algoritmo de busca K2

Texto

Descrição gerada automaticamente

Diagrama

Descrição gerada automaticamente

* Com algoritmo de busca TAN

Texto, Carta

Descrição gerada automaticamente

Imagem em preto e branco

Descrição gerada automaticamente com confiança média

|  |  |  |
| --- | --- | --- |
|  | K2 | TAN |
| Precisão | 0,544 | 0,692 |
| Revocação | 0,365 | 0,318 |
| Medida F | 0,437 | 0,435 |

**Análise da Base de Dados Shuttle Landing Control**

**Informações**

A Base de Dados Shuttle Landing Control foi criada artificialmente por Roger Burke e sua equipe na NASA para avaliar sobre quais condições o pouso automático seria melhor do que o manual. Trata-se de uma base de dados pequena com 15 instâncias, valores ausentes e 6 atributos do tipo categórico descritos a seguir:

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

* **Estabilidade**: estável e muito estável (1 e 2)

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

* **Erro**: tamanho do erro de pequeno a grande (1 a 4)

Interface gráfica do usuário

Descrição gerada automaticamente com confiança média

* **Sinal**: positivo ou negativo (1 e 2)

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

* **Vento**: Cabeça ou cauda (1 e 2)

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

* **Magnitude**: Força do vento (1), leve (2), média (3), forte(4)

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

* **Visibilidade**: Sim ou Não

Interface gráfica do usuário

Descrição gerada automaticamente com confiança média

* **Classe**: dividida em manual (1) e automático (2).

Interface gráfica do usuário

Descrição gerada automaticamente

**Pré-processamento**

Assim como na base Breast-Cancer é preciso reordenar os atributos, transformá-los em numéricos e tratar os valores ausentes. Novamente os filtros Reorder, OrdinalToNumeric e ReplaceMissingValues foram usados.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

* **Visibilidade**

Gráfico

Descrição gerada automaticamente

* **Estabilidade**

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

* **Erro**

Gráfico, Gráfico de barras

Descrição gerada automaticamente

* **Sinal**

Gráfico

Descrição gerada automaticamente com confiança média

* **Vento**

Uma imagem contendo Gráfico

Descrição gerada automaticamente

* **Magnitude**

Gráfico

Descrição gerada automaticamente

* **Classe**

Uma imagem contendo Interface gráfica do usuário

Descrição gerada automaticamente

**Avaliação**

Nesse caso, estou considerando a classe Manual, 6 instâncias, como positivo e Automático, 9 instâncias, como negativo.

A matriz de confusão do Algoritmo K-vizinhos mais próximos é como segue:

* Com K = 3

Texto

Descrição gerada automaticamente

* Com K = 6

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

|  |  |  |
| --- | --- | --- |
|  | K = 3 | K = 6 |
| Precisão | - | - |
| Revocação | 0 | 0 |
| Medida F | - | - |

A matriz de confusão e a árvore gerada com o Algoritmo Árvore de Decisão são como seguem:

* Com número de instâncias por folha igual a 2.

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Diagrama

Descrição gerada automaticamente

* Com número de instâncias por folha igual a 4

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Diagrama

Descrição gerada automaticamente

|  |  |  |
| --- | --- | --- |
|  | N = 2 | N = 4 |
| Precisão | 0,400 | 0,667 |
| Revocação | 0,333 | 1 |
| Medida F | 0,364 | 0,800 |

A matriz de confusão e o grafo gerados pelo algoritmo BayesNet são como seguem:

* Com algoritmo de busca K2

Texto

Descrição gerada automaticamente com confiança média

Diagrama

Descrição gerada automaticamente

* Com o algoritmo de busca TAN

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Diagrama

Descrição gerada automaticamente

|  |  |  |
| --- | --- | --- |
|  | K2 | TAN |
| Precisão | 0,250 | 0,250 |
| Revocação | 0,167 | 0,167 |
| Medida F | 0,200 | 0,200 |

**Comparação dos Algoritmos**

Neste trabalho foram usadas duas bases de dados de tamanhos diferentes. A primeira Breast Cancer tinha 286 instâncias e a Shuttle Landing Control 15 instâncias. Para avaliar qual algoritmo teve o melhor desempenho analisamos os resultados que cada um teve nas duas bases de dados. Como duas configurações dos algoritmos foram usadas estou considerando essas para as análises finais: K=6 para o Algoritmo K-Vizinhos mais próximos, N=4 para a Árvore de Decisão, e o algoritmo de busca TAN para NetBayes.

* Precisão

|  |  |  |
| --- | --- | --- |
|  | Breast-Cancer | Shuttle Landing Control |
| K vizinhos mais próximos | 0,563 | - |
| Árvore de Decisão | 0,500 | 0,667 |
| NetBayes | 0,692 | 0,250 |

* Revocação

|  |  |  |
| --- | --- | --- |
|  | Breast-Cancer | Shuttle Landing Control |
| K vizinhos mais próximos | 0,212 | 0 |
| Árvore de Decisão | 0,271 | 1 |
| NetBayes | 0,318 | 0,167 |

* Medida F

|  |  |  |
| --- | --- | --- |
|  | Breast-Cancer | Shuttle Landing Control |
| K vizinhos mais próximos | 0,308 | - |
| Árvore de Decisão | 0,351 | 0,800 |
| NetBayes | 0,435 | 0,200 |

O Algoritmo com maior precisão, revocação e Medida F, na base de dados Breast Cancer é o NetBayes e na Shuttle Landing Control é a Árvore de Decisão. Podemos inferir que o algoritmo NetBayes possui uma performance melhor em bases de dados grandes, enquanto a Árvore de Decisão tem uma performance melhor em bases de dados pequenas.

**Referências Bibliográficas**

Russel, J. Stuart & Norvig, Peter. “Artifical Intelligence: A modern Approach”. Prentice Hall.

ACID, S. “Learning Bayesian Network Classifiers: Searching in a Space of Partially Directed Acyclic Graphs”. Springer Science 2005

RUIZ, C. “Illustration of the K2 Algorithm for Learning Bayes Net Structures”. Department of Computer Science, WPI.

Zheng, F., Webb, G.I. (2011). Tree Augmented Naive Bayes. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8\_850