Tese apresentada à Divisão de Pós-Graduação do Instituto Tecnológico de Aeronáutica como parte dos requisitos para obtenção do título de Mestre em Ciência no Curso de Pós-Graduação em Engenharia Eletrônica e Computação na Área de Sistemas e Controle.

JACKSON PAUL MATSUURA

DISCRETIZAÇÃO PARA APRENDIZAGEM BAYESIANA: APLICAÇÃO NO AUXÍLIO À VALIDAÇÃO DE DADOS EM PROTEÇÃO AO VÔO

Tese aprovada em sua versão final pelos abaixo-assinados:

Prof. Takashi Yoneyama
ORIENTADOR

Prof. Roberto Kawakami Harrop Galvão
CO-ORIENTADOR

Prof. Homero Santiago Maciel
CHEFE DA DIVISÃO DE PÓS-GRADUAÇÃO

Campo Montenegro
São José dos Campos, SP – Brasil
2003

DISCRETIZAÇÃO PARA APRENDIZAGEM BAYESIANA: APLICAÇÃO NO AUXÍLIO À VALIDAÇÃO DE DADOS EM PROTEÇÃO AO VÔO

JACKSON PAUL MATSUURA

Composição da Banca Examinadora:

Prof. Karl Heinz Kienitz	Presidente – ITA
Prof. Takashi Yoneyama	Orientador (ITA)
Prof. Roberto Kawakami Harrop Galvão	Co-orientador (ITA)
Prof. Fábio Gagliardi Cozman	(USP)
Prof. Armando Zeferino Milioni	(ITA)
Prof. Marcelo Gomes da Silva Bruno	(ITA)

ITA

Aos meus pais Mineo (in memoriam) e Ayako.

Agradecimentos:

A Deus por me dar a capacidade e principalmente a oportunidade de realizar esse trabalho.

A minha querida esposa Marlene e aos meus filhos, Jackson Júnior, Jéssica e Juliana pelos momentos em que abdicaram de minha presença, consciente ou inconscientemente, para que esse trabalho pudesse ser completado.

Aos meus orientadores Prof. Dr. Takashi Yoneyama e Prof. Dr. Roberto Kawakami Harrop Galvão, não apenas pela sua orientação acadêmica, mas também por todo apoio e incentivo, prestados antes e durante a realização deste trabalho.

Ao Instituto de Proteção ao Vôo - IPV, pelo fornecimento dos dados utilizados como caso de estudo nesse trabalho, principalmente nas pessoas do Maj. Calheiros e Ten. Marcos Luis.

Aos meus professores e colegas que, dentro de suas limitações de tempo, contribuíram, direta ou indiretamente, em vários níveis com esse trabalho.

Ao “Mestre” Marcelo Santiago do Amaral, que mesmo longe ainda serve de exemplo e inspiração em meus estudos.

Resumo:

A utilização de redes Bayesianas, vem crescendo em diversas áreas e aplicações. As redes Bayesianas podem ser construídas a partir do conhecimento de especialistas ou por algoritmos de aprendizagem que inferem as relações entre as variáveis do domínio a partir de um conjunto de dados de treinamento.

A construção manual de redes Bayesianas vem cada vez mais sendo preterida pelo uso de algoritmos de aprendizagem que, em geral, pressupõem que as variáveis utilizadas na aprendizagem sejam discretas ou, caso sejam contínuas, apresentem uma distribuição Gaussiana, o que normalmente não ocorre na prática.

Portanto para o uso de algoritmos de aprendizagem é necessário que as variáveis contínuas sejam discretizadas segundo algum critério, que no caso mais simples pode ser uma discretização uniforme.

A grande maioria dos métodos de discretização existentes, porém, não são adequados à aprendizagem de redes Bayesianas, pois foram desenvolvidos no contexto de classificação e não de descoberta de conhecimento.

Nesse trabalho é proposto e utilizado um método de discretização de variáveis que leva em conta as distribuições condicionais das mesmas no processo de discretização, objetivando um melhor resultado do processo de aprendizagem.

O método proposto foi utilizado em uma base de dados real de informações de Proteção ao Vôo e a rede Bayesiana construída foi utilizada no auxílio à validação de dados, realizando uma triagem automatizada dos dados.

Foi realizada uma comparação entre o método proposto de discretização e um dos métodos mais comuns.

Os resultados obtidos mostram a efetividade do método de discretização proposto e apontam para um grande potencial dessa nova aplicação da aprendizagem e inferência Bayesiana.

Abstract:

The use of Bayesian networks have increased in diverse areas and applications. Bayesian networks can be built from the knowledge of experts or by learning algorithms that infer the relations between the variables of the domain on the basis of a training data set.

The manual construction of Bayesian networks has been replaced in several applications by the use of learning algorithms, but such algorithms in general assume that the variables used in the learning process are discrete or, in case they are continuous, present a Gaussian distribution, which normally does not occur in real cases.

Therefore for the use of Bayesian learning, variables should be discretized according to some criterion, that in the simplest case can be a uniform discretization.

The great majority of the existing methods of discretization, however, is not adjusted to the learning of Bayesian networks, because they were developed in the context of classification rather than knowledge discovery.

This work proposes and uses a method of discretization of variables that takes into account their conditional distributions in the discretization process, with the purpose of obtaining a better result of the Bayesian learning process.

The proposed method was applied to a real database of Flight Protection information and the resulting Bayesian network was used as an aid to the validation of data, carrying out an automatized screening of the data.

A comparison was carried out between the proposed method of discretization and one of the most common methods employed in the literature.

The obtained results show the effectiveness of the proposed discretization method, pointing to a great potential of this new application of the Bayesian learning and inference.

Sumário:

Lista de Figuras:

Lista de Tabelas:

Lista de Abreviaturas e Siglas:

1RD : Discretização 1-Regras (*1-Rules discretization*)

API : Interface de Programação de Aplicação (*Application Programming Interface*)

BN : Rede Bayesiana (*Bayesian Network*)

CMD : Discretização Chimerge

DAG : Grafo Direcionado Acíclico (*Directed Acyclic Graph*)

DBD : Discretização Baseada em Distância (*Distance-based Discretization*)

DMQ : Diferença Média Quadrática

DMQP : Diferença Média Quadrática Ponderada

DTP : Discretização via Tabela de Probabilidades

EFD : Discretização em Freqüências Iguais (*Equal Frequency Discretization*)

EMD : Discretização de Minimização de Entropia (*Entropy Minimization Discretization*)

EWD : Discretização em Intervalos Iguais (*Equal Width Discretization*)

IPV : Instituto de Proteção ao Vôo

K2 : Kutató 2 - denominação de um algoritmo de aprendizagem Bayesiana

LD : Discretização Ociosa (*Lazy Discretization*)

MD : Discretização Multivariável (*Multivariable Discretization*)

MDL : Tamanho Mínimo de Descrição (*Minimum Description Lenght*)

MDLPD : Discretização do Princípio de Tamanho Mínimo de Descrição (*Minimum Description Length Principle Discretization*)

MSBNx : *"Microsoft Research's Bayesian network autoring and evaluation tool"*

PD : Discretização Proporcional (*Proportional Discretization*)

RUD : Discretização não Supervisionada Relativa (*Relative Unsupervided Discretization*)

1. Introdução

1.1. Motivação

A utilização de sistemas baseados em redes Bayesianas, que são uma representação compacta de distribuições de probabilidades conjuntas de um domínio, vem crescendo a cada dia nas mais diversas áreas, como medicina [31], [32], [60], engenharia civil [7], genética [12] e educação [59]. A utilização das redes Bayesianas vem crescendo também em diversos tipos de aplicações, como Sistemas Especialistas [5], Sistemas de Filtragem de Dados [53], Sistemas de Diagnóstico [7], [57], Classificadores [62], Sistemas de Análise [11], [17] e Sistemas de Apoio à Decisão [50].

A construção manual das redes Bayesianas, que implica na definição dos relacionamentos entre as variáveis do domínio e na quantificação de suas probabilidades condicionais, que inicialmente era a única opção, vem cada vez mais sendo preterida pelo uso de algoritmos de aprendizagem, que inferem os relacionamentos e calculam as probabilidades condicionais a partir de um conjunto de dados de treinamento.

Porém, um grande empecilho à aprendizagem Bayesiana, em cenários envolvendo variáveis contínuas, é a falta de um método adequado de discretização, uma vez que a grande maioria dos algoritmos de aprendizagem trata apenas variáveis discretas e que a maioria dos métodos de discretização são voltados para a utilização em classificadores.

Uma discretização adequada é fundamental para que o algoritmo de aprendizagem seja capaz de descobrir corretamente os relacionamentos existentes entre as variáveis do domínio, uma vez que o processo de discretização pode evidenciar ou esconder os padrões apresentados pelo conjunto de aprendizado [3].

O uso dos algoritmos de discretização voltados para classificação em aprendizagem Bayesiana quando o objetivo é a extração de conhecimento pode inviabilizar totalmente a aprendizagem e a utilização da rede assim construída [3]. Dessa forma, para uma aprendizagem Bayesiana eficaz um método de discretização adequado é fundamental [3].

1.2. Escopo

O objetivo principal deste trabalho é a proposta e avaliação de um método de discretização para aprendizagem Bayesiana.

Faz parte do escopo deste trabalho mostrar que a maioria dos métodos de discretização conhecidos não se adequa à aprendizagem Bayesiana e que o método proposto propicia resultados melhores ou no pior dos casos iguais aos métodos atualmente utilizados. Não faz parte do escopo deste trabalho uma demonstração formal da efetividade do método proposto, o que fica como sugestão para trabalhos futuros.

É feita a comparação de resultados obtidos com o método proposto e com um método simples de discretização para comprovar que o primeiro se adequa melhor à aprendizagem Bayesiana.

Para a comprovação da utilidade do método proposto em aplicações reais é proposto e implementado um sistema de validação de dados que utiliza o método de discretização aliado à aprendizagem e inferência Bayesiana. Tal sistema é capaz de realizar uma triagem nos dados, reduzindo muito a quantidade de dados que deve ser analisada por uma pessoa para a identificação de erros. Isso possibilita que grande quantidade de dados seja validada em um menor espaço de tempo e com um menor esforço humano.

Os resultados apresentados pelo sistema de validação implementado são apenas ilustrativos, uma vez que o foco do trabalho está no processo de discretização e na proposição de uma solução para o processo de validação de dados e não na avaliação do desempenho desse processo de validação.

1.3. Contribuições

Entre as contribuições mais relevantes desse trabalho vale ressaltar:

- A proposição de um método de discretização de dados que se adeque às características da aprendizagem Bayesiana, melhorando assim os resultados obtidos com o uso da mesma.

- A comparação entre o método proposto e outros métodos de discretização normalmente utilizados.

- A proposição do uso de aprendizagem e inferência Bayesiana no auxílio à validação de dados, que poderá agilizar muito o processo de validação de dados em várias áreas e nesse caso específico ser usado na validação dos dados de proteção ao vôo do Insituto de Proteção ao Vôo - IPV.

- A proposição de um novo método de comparação do desempenho de redes Bayesianas, baseado na correta identificação de dados incorretos em bases de dados.

1.4. Organização

Para facilitar o entendimento do método de discretização proposto e da proposta de utilização de aprendizagem Bayesiana no auxílio à validação de dados, no Capítulo 2 é feita a apresentação dos fundamentos teóricos usados neste trabalho: redes Bayesianas, aprendizagem Bayesiana e discretização. São também apresentados no Capítulo 2 os métodos de discretização normalmente utilizados e os motivos pelos quais eles não se adequam à aprendizagem Bayesiana.

No Capítulo 3 o método de discretização é proposto e comparado com outros métodos que utilizam probabilidades como métrica para a união de intervalos.

No Capítulo 4 é apresentado o problema de validação de dados, um caso de estudo e a solução baseada no uso de redes Bayesiana.

No Capítulo 5 são descritos os procedimentos utilizados na implementação da aprendizagem e inferência Bayesiana, bem como os testes realizados.

Os resultados obtidos são apresentados e analisados no Capítulo 6 e no Capítulo 7 são apresentadas as conclusões e sugestões para trabalhos futuros.

2. Fundamentação Teórica

2.1. Redes Bayesianas

Um grafo *G* é um par, *G*=(*V*, *E*), onde *V* é um conjunto finito de nós e *E* é um conjunto finito de arcos entre pares de nós de *V*. Caso todos os arcos sejam direcionados o grafo é um grafo direcionado, Figura 2.1(a), caso todos os arcos sejam não direcionados o grafo é um grafo não direcionado, Figura 2.1(b), e caso o grafo tenha arcos direcionados e não direcionados o grafo é misto, Figura 2.1(c) [11].

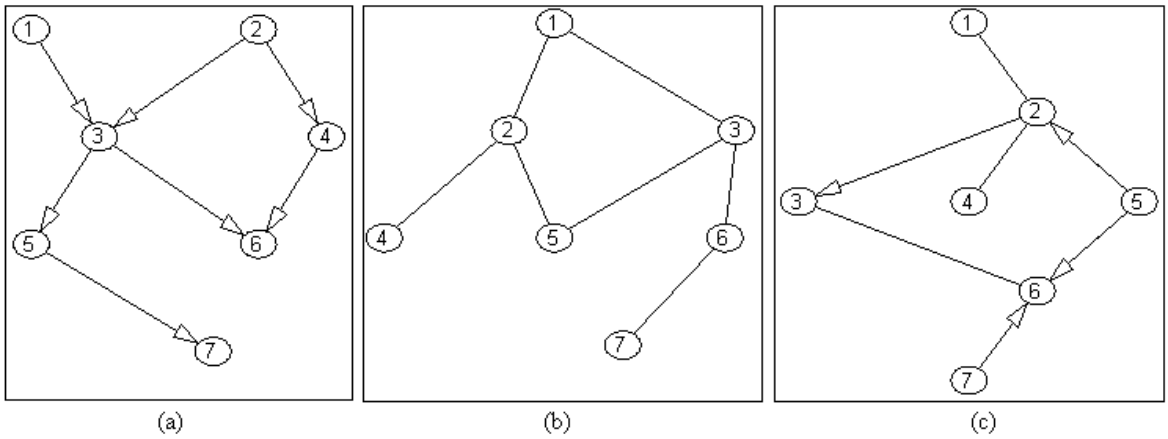

Figura 2.1: Exemplos de grafos: (a) grafo direcionado; (b) grafo não direcionado; (c) grafo misto.

Um grafo direcionado onde é possível, a partir de um nó e percorrendo os arcos nas direções indicadas, voltar ao mesmo nó, é um grafo direcionado cíclico, Figura 2.2(b). Caso não seja possível, a partir de nenhum nó, percorrendo os arcos nas direções indicadas, voltar ao mesmo nó, o grafo direcionado é acíclico, Figura 2.2(a).

As redes Bayesianas – (*Bayesian Network* - BN), também conhecidas como redes de crença Bayesianas, grafos causais, redes causais, redes de crença e modelos recursivos[4], são grafos direcionados acíclicos onde os nós representam variáveis aleatórias de um domínio específico e os arcos indicam a existência de dependência probabilística entre os nós ligados. Essas dependências são expressas por probabilidades condicionais [41], [48], [49].

Um exemplo de BN é apresentado na Figura 2.3, onde as variáveis do domínio são *CHUVA*, *VELOCIDADE BAIXA*, *PISTA ESCORREGADIA* e *ACIDENTE*. Cada uma das variáveis pode assumir apenas dois valores, *sim* ou *não*.

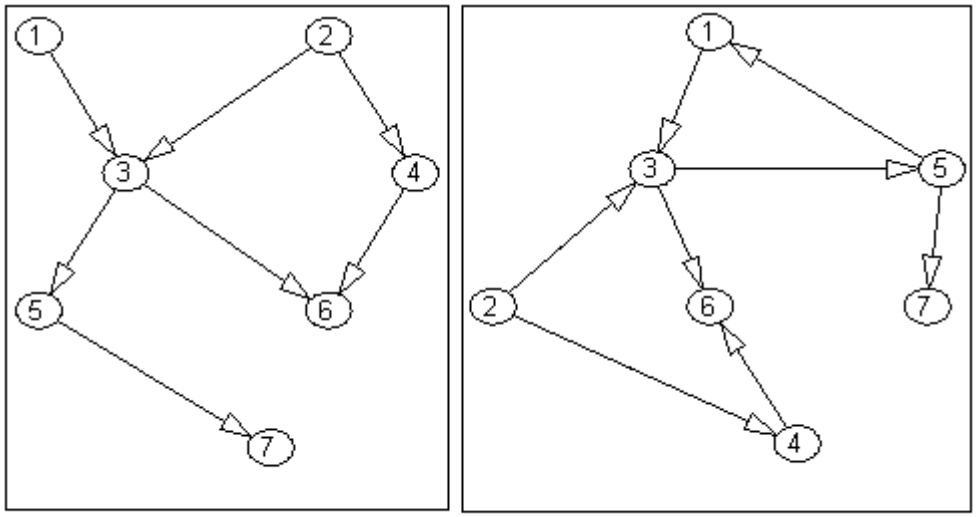

Figura 2.2: Exemplos de grafos: (a) grafo direcionado acíclico; (b) grafo direcionado cíclico.

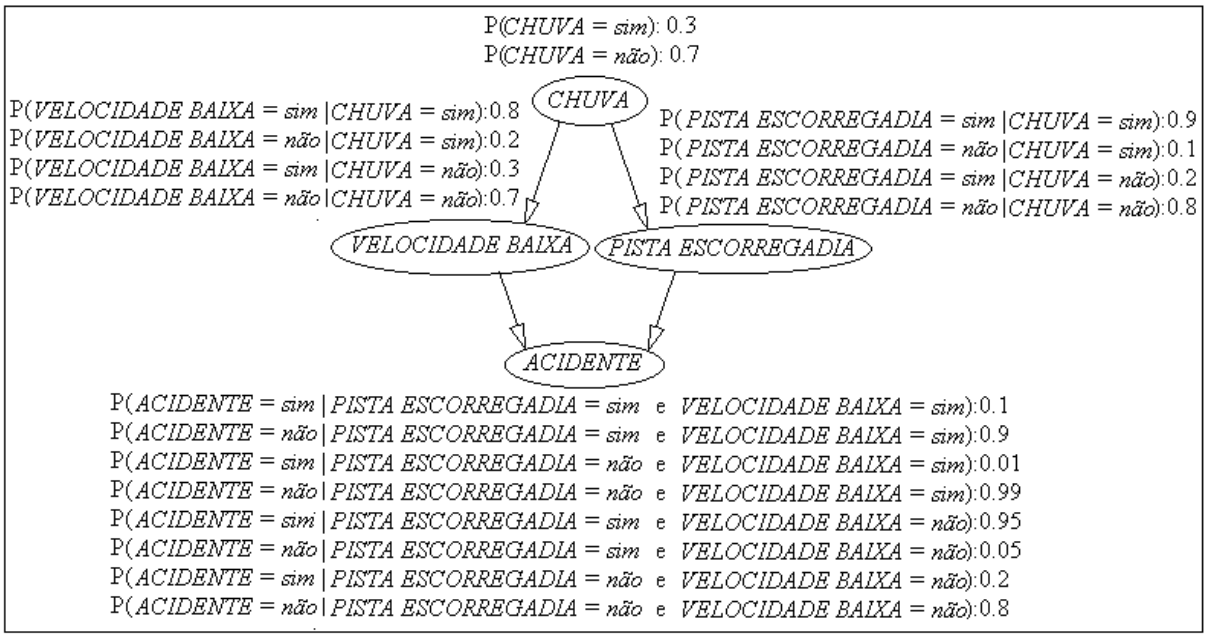

Figura 2.3: Exemplo de rede Bayesiana.

Para a variável *CHUVA*, o valor *sim* indica que no local sendo analisado, em um determinado momento, estava chovendo. O valor *não* indica que não estava chovendo.

Para a variável *VELOCIDADE BAIXA*, o valor *sim* indica que um veículo passando pelo local sendo analisado estava abaixo de um certo limite de velocidade, por exemplo, 30km/h. O valor *não* indica que um veículo passando pelo local analisado estava acima do limite de velocidade definido.

Para a variável *PISTA ESCORREGADIA*, o valor *sim* indica que a superfície da pista no local da análise estava escorregadia. O valor *não* indica que as condições da pista eram normais.

Para a variável *ACIDENTE,* o valor sim indica que um veículo ao passar pelo local se envolveu em um acidente. O valor não indica que um veículo passou pelo local sem se envolver em nenhum acidente.

Os arcos do grafo direcionado acíclico (*Directed Acyclic Graph* - DAG), que representam as relações existentes entre as variáveis, indicam que a variável *CHUVA* se relaciona com as variáveis *VELOCIDADE BAIXA* e *PISTA ESCORREGADIA* e que as variáveis *VELOCIDADE BAIXA* e *PISTA ESCORREGADIA* se relacionam com a variável *ACIDENTE*.

As probabilidades condicionais de cada nó quantificam essas relações. As probabilidades condicionais do nó *PISTA ESCORREGADIA,* por exemplo, explicitam as probabilidades da superfície da pista estar escorregadia ou normal uma vez estabelecido se está chovendo ou não. Se o valor da variável *CHUVA* for *sim* então a probabilidade do valor da variável *PISTA ESCORREGADIA* ser *sim* é de 0.9. Se o valor da variável *CHUVA* for *não* então a probabilidade do valor da variável *PISTA ESCORREGADIA* ser *sim* é de 0.2.

Por serem uma ferramenta prática poderosa de representação do conhecimento e inferência sob condições de incerteza [8], [17], o uso das BNs vem crescendo constantemente [41], principalmente nas áreas de economia, processamento de padrões [11], reconhecimento da fala, algoritmos de reconstrução de imagens e comparação de modelos. Elas têm sido aplicadas com sucesso [19] em sistemas especialistas [5], [32], [50], [60], instrumentos de diagnóstico [7], [57], sistemas de tomada de decisão [54] e classificadores [27], [53], [56], [62].

As BNs são um subconjunto das redes probabilísticas, cuja principal característica é a habilidade para explorar a estrutura de um grafo e reduzir o cálculo da probabilidade de um evento, dada a evidência disponível, a uma série de cálculos locais.

Para isso são usados somente valores obtidos de um nó e seus vizinhos em uma estrutura de grafo, evitando que a função de distribuição de probabilidades conjunta global tenha que ser calculada. A representação gráfica também explicita relações de dependências e constitui uma ferramenta poderosa na aquisição de conhecimentos e no processo de verificação [57]. Ou seja, as redes probabilísticas são uma representação compacta da função de distribuição de probabilidade conjunta global de todas as variáveis aleatórias do domínio modelado [57].

Outro exemplo de redes probabilísticas são as redes de Markov [48], que são representadas por grafos não direcionados, mas sua interpretação e principalmente sua

construção são bem mais complexas do que as das BNs, por isso apesar de também terem o mesmo poder de representatividade compacta da função de distribuição de probabilidades conjunta elas não são tão populares.

Extensões das BNs que além dos nós representando as variáveis do domínio em questão contenham nós representando decisões ou vantagens são chamadas de redes de decisão ou diagramas de influência [43]; o foco desse trabalho no entanto são as BNs segundo sua definição, onde todos os nós representam variáveis do domínio.

Uma BN pode ser dividida em duas partes, um DAG e um conjunto de probabilidades condicionais. O DAG representa qualitativamente as dependências entre as variáveis e as probabilidades condicionais quantificam essa dependência [24], [38].

Um caso particular de BN muito utilizada em tarefas de classificação é a rede *naive-Bayes*. Em uma rede *naive-Bayes* um nó representa a variável a ser classificada e existem arcos da variável classificada para todos os outros nós, chamados de atributos da classe. Os atributos não possuem arcos entre si, pois assume-se que eles sejam independentes dado a ocorrência de uma classe. O DAG de uma rede *naive-Bayes,* portanto, apresenta um nó raiz e apenas arcos saindo desse nó e indo para cada um dos outros nós, conforme ilustrado na figura 2.4.

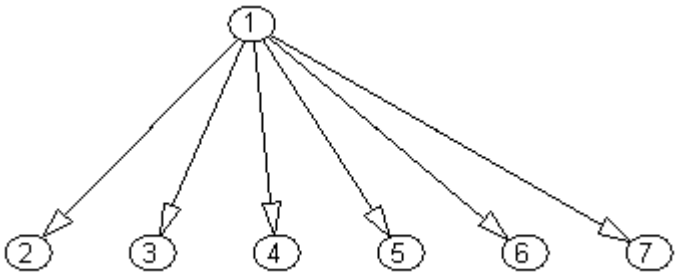

Figura 2.4: Exemplo de DAG de uma rede *naive-Bayes*.

Por ser uma versão simplificada de uma BN e assumir a independência entre os nós, uma vez definida a classe do nó raiz, fica claro que seu poder de representatividade da distribuição de probabilidade conjunta global é limitado. Contudo para tarefas de classificação onde apenas a probabilidade de ocorrência da classe precisa ser calculada e não a probabilidade conjunta global essa simplificação tem se mostrado muito útil.

Tendo sido esclarecido o poder de representatividade das BNs, será exposto a seguir seu poder de raciocínio ou inferência. Para tanto, será necessária antes a definição de alguns

termos que serão usados durante esse trabalho, de modo a facilitar a interpretação das informações aqui apresentadas.

- Variável – atributo – característica – campo: elemento representado por um nó da BN. Normalmente referenciado como variável, mas em caso de classificadores passa a ser chamado de atributo ou característica da variável de classe e quando a BN é construída ou usada em bancos de dados os nós representam os campos ou atributos de uma tabela [8];
- Instância – ocorrência: determinação do valor de uma variável, $E_1=e_1$, o valor da variável E_1 é e_1;
- Registro – tupla: determinação simultânea dos valores de todas as variáveis da BN. No caso de um banco de dados corresponde a uma linha da tabela;
- Evidência: conjunto de ocorrências $E = \{ E_1=e_1, E_2=e_2 , ... , E_n=e_n\}$;
- Probabilidade a priori: probabilidade do valor de uma variável ou dos valores de um subconjunto de variáveis assumir um dado valor ou subconjunto de valores, estimada a partir das probabilidades conhecidas sem considerar qualquer evidência [60];
- Probabilidade a posteriori: probabilidade de uma variável ou subconjunto de variáveis estimada a partir das probabilidades conhecidas e de uma evidência [60];
- Antecessores e sucessores de um nó *A* em um DAG: caso haja um arco entre os nós *A* e *B* partindo do nó *A* em direção ao nó *B*, então *B* é sucessor de *A*. Se a direção do arco for de *B* para *A* então B é antecessor de *A*;
- Nó folha: nó que não tem sucessores.

Tomando como base o BN da figura 2.3 as variáveis ou campos seriam: *CHUVA*, *VELOCIDADE BAIXA*, *PISTA ESCORREGADIA* e *ACIDENTE*. Uma instância ou ocorrência seria a determinação de que em uma dada situação estava chovendo, *CHUVA = sim*, ou de que em uma dada situação a pista não estava escorregadia, *PISTA ESCORREDIA = não*. Um exemplo de registro seria a determinação de que em uma dada situação estava chovendo, a pista estava escorregadia, um carro estava em baixa velocidade e não ocorreu acidente, {*CHUVA = sim*, *PISTA ESCORREGADIA = sim*, *VELOCIDADE BAIXA = sim*, *ACIDENTE = não*}. Um exemplo de evidência seria que em uma dada situação a pista estivesse escorregadia e um carro estivesse em baixa velocidade, {*PISTA ESCORREGADIA = sim*, *VELOCIDADE BAIXA = sim*}.

Uma probabilidade a priori seria a probabilidade de chover de 0.3 e uma probabilidade a posteriori seria a probabilidade de ocorrer um acidente, dada a evidência de que a pista está escorregadia e de que o carro não está a uma velocidade baixa, que seria de 0.95.

O raciocínio ou inferência Bayesiana ou ainda a propagação de crença [29], [R12] é justamente o processo de atualização das probabilidades a posteriori baseado nas probabilidades condicionais e na evidência fornecida [4], [30]. Nas BNs a evidência pode ser definida para qualquer subconjunto de nós e a probabilidade a posteriori pode ser calculada para qualquer outro subconjunto de nós [20].

Como uma BN representa um modelo probabilístico completo do domínio, e como é possível a partir dela calcular a distribuição de probabilidade conjunta para todas as variáveis envolvidas, fica claro que a rede contém informação suficiente para calcular probabilidades de quaisquer variáveis E_1, E_2, ..., E_n [48].

Em particular, podemos calcular a probabilidade da variável E_i assumir o valor e_i dado que $\{E_k = e_k\}$, k=1..n, $k \neq i$, ou seja, a probabilidade a posteriori de E_i.

Essa propagação de crença pode ser calculada pelo teorema de Bayes e pelo teorema da probabilidade total [37], [47], [22].

Usando novamente a BN da figura 2.3 vamos ilustrar o processo de inferência Bayesiana.

Dada a evidência de que está chovendo, $\{CHUVA = sim\}$, qual seria a probabilidade de ocorrer um acidente, P(*ACIDENTE = sim*) ?

Como *CHUVA = sim* a probabilidade de um veículo estar a uma velocidade baixa será: P(*VELOCIDADE BAIXA = sim*) = P(*VELOCIDADE BAIXA = sim* | *CHUVA = sim*) = 0.8.

Da mesma forma concluímos que P(*VELOCIDADE BAIXA = não*) = 0.2, P(*PISTA ESCORREGADIA = sim*) = 0.9 e que P(*PISTA ESCORREGADIA = não*) = 0.1.

Com esses valores de probabilidade a posteriori podemos agora atualizar a probabilidade de *ACIDENTE usando o teorema da probabilidade total*:

P(*ACIDENTE = sim*) = 0.1*0.9*0.8+0.01*0.1*0.8+0.95*0.9*0.2+0.2*0.1*0.2 = 0.2478.

P(*ACIDENTE = não*) = 0.9*0.9*0.8+0.99*0.1*0.8+0.05*0.9*0.2+0.8*0.1*0.2 = 0.7522.

Caso a evidência fosse da variável *ACIDENTE*, por exemplo, para atualizar a probabilidade das variáveis *VELOCIDADE BAIXA* e *PISTA ESCORREGADIA* bastaria usar o teorema de Bayes para calcular as probabilidades condicionais P(*VELOCIDADE BAIXA = sim* | *ACIDENTE = sim*) e P(*PISTA ESCORREGADIA = sim* | *ACIDENTE = sim*) a partir das probabilidades da BN e usar novamente o teorema da probabilidade total.

Vários algoritmos de propagação, visando diminuir o tempo necessário para a atualização das probabilidades a posteriori vem sendo propostos e estudados [14], [24], [28], [29], [34],

[35], [46], [59], [60] mas não se pode afirmar que um deles seja o melhor para todas as BNs, alguns se destacam apenas para alguns casos específicos. Mas mesmo sem o uso de um algoritmo de propagação de crença ótimo, sistemas que possuem BNs como base de conhecimento tem se mostrado muito eficientes e tem sido muito difundidos ganhando importância inclusive em áreas comerciais [29].

Uma vez exposta a conceituação das BNs, sua potencialidade e sua utilização para inferência passaremos a explorar sua construção.

2.2. Aprendizagem Bayesiana

Originalmente o conceito das BNs foi desenvolvido supondo-se uma dependência de especialistas humanos para a definição do DAG, ou seja, da estrutura ou topologia da BN e para a estimação das probabilidades condicionais [49], mas elas podem ser construídas tanto a partir do conhecimento de especialistas humanos quanto a partir de bases de dados, com a utilização de algoritmos de aprendizagem Bayesiana [31].

A construção manual de uma BN pode ser um processo bastante trabalhoso e caro para grandes aplicações [19] e em domínios complexos sua especificação além de consumir bastante tempo está propensa a erros [36].

Por esse motivo os esforços dirigidos para o desenvolvimento de métodos que possam construir BNs diretamente de um banco de dados, ao invés do discernimento de especialistas humanos, vem crescendo constantemente [49].

O aprender pode ser interpretado como o processo de aquisição de uma representação interna efetiva para as restrições existentes no mundo, fatos e regras [48] ou, em outras palavras, a aquisição de conceitos e de conhecimentos estruturados [44].

O problema de aprendizagem Bayesiana pode ser enunciado como: dado um conjunto de treinamento $D=\{ u_1, u_2, ..., u_n \}$, de tuplas de U, onde U é o conjunto das variáveis do domínio e cada u_i corresponde aos valores das ocorrências de cada uma das variáveis de U, encontrar a rede B que melhor se adequa ao conjunto D [19]. A formalização do conceito de adequação e a maneira utilizada para encontrar B é que vão diferenciar os métodos de aprendizagem Bayesiana.

De maneira geral o aprendizado pode ocorrer com ou sem supervisão. Quando a aquisição dos conceitos e do conhecimento estruturado é orientada por um supervisor ou uma função de supervisão que classifica os resultados apresentados em certos ou errados, tem-se o aprendizado supervisionado. Quando o aprendizado ocorre sem esse tipo de orientação tem-se

o aprendizado não supervisionado. Dependendo da intensidade de envolvimento do supervisor o aprendizado pode ser por descoberta, por exemplos ou por programação. O aprendizado por descoberta é aquele em que a interferência do supervisor sobre o processo de aprendizado é mínima. O aprendizado por exemplos é aquele em que o supervisor fornece amostras representativas do universo de conhecimentos que devem ser generalizadas. O aprendizado por programação é aquele em que o supervisor incorpora diretamente os seus conhecimentos ao processo [44].

Usando esse critério, dividimos os métodos de aprendizagem Bayesiana em supervisionados e não-supervisionados. Um método de aprendizagem Bayesiana será supervisionado se existir uma função de avaliação do desempenho da rede, no caso de classificadores Bayesianos, isso ocorre na forma da variável a ser classificada. É um aprendizado por exemplos, onde cada tupla é um exemplo de classificação e a função de avaliação normalmente medirá a capacidade da BN classificar corretamente as variáveis, dada a evidência dos atributos. Caso não haja essa função de avaliação de desempenho, ou seja, caso a BN não tenha o objetivo de classificar uma variável, o método será não-supervisionado. Os métodos desse tipo são também chamados de aprendizagem por regressão ou predição contínua [52].

O fato de um método de aprendizagem Bayesiana ser não-supervisionado não elimina, porém, a necessidade de avaliar a aprendizagem.

Como enunciado anteriormente o problema de aprendizagem Bayesiana objetiva encontrar a rede B que melhor se adequa à D. Podemos então dividir os métodos ou algoritmos de aprendizagem Bayesiana não-supervisionados em duas categorias. A primeira usa métodos de busca heurística [44], [51] para construir possíveis BNs e então as avalia usando um método de pontuação, que estima a adequação da rede em análise com relação ao conjunto de dados fornecidos [31], [19] e não com relação à correta classificação de uma variável, como no caso de aprendizagem supervisionada. O processo de construção de possíveis BNs continua até que a pontuação das novas redes não seja significantemente melhor que as das anteriores ou até que um critério de parada seja atingido.

A segunda categoria constrói BNs pela análise de relações de dependência entre os nós. As possíveis relações de dependência entre os vários nós são avaliadas pelo uso de testes de independência condicional e são criados os arcos para as dependências mais relevantes. Ambas têm suas vantagens e desvantagens; geralmente a primeira categoria apresenta resultados mais rápidos, mas devido à sua natureza heurística pode não encontrar a melhor

solução. A segunda categoria converge para a melhor solução, mas apenas em situações onde as distribuições de probabilidade satisfazem certas hipóteses [8].

A primeira categoria é preferível em casos reais, onde não se sabe a priori se as distribuições de probabilidade satisfazem as hipóteses necessárias para a convergência da segunda categoria [31].

Falando agora especificamente dos algoritmos de aprendizagem baseados em busca heurística, podemos dividir a aprendizagem em duas tarefas: o aprendizado dos parâmetros numéricos ou das probabilidades condicionais para um dado DAG e o aprendizado do próprio DAG. Essas duas tarefas são dependentes, pois as probabilidades são calculadas para um DAG específico e a avaliação da adequação de um DAG normalmente faz uso da probabilidade conjunta ou das probabilidades condicionais. Desse modo para cada possível DAG é necessário que as probabilidades condicionais sejam calculadas. O processo de avaliação de cada DAG demandará, portanto, um esforço proporcional à quantidade de dados usada na aprendizagem.

Uma vez que as buscas heurísticas não garantem a descoberta da melhor solução possível, poderia parecer intuitivo que uma busca exaustiva por todos os DAG possíveis deveria ser executada, mas na prática essa busca exaustiva só pode ser realizada quando a BN a ser aprendida tiver um número pequeno de nós.

A quantidade, f(n), de possíveis DAGs em função do número, n, de nós da rede pode ser calculada pela seguinte função recursiva [13]:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i) \qquad (2.1)$$

Caso os nós já estejam ordenados, ou seja, caso a possível direção dos nós entre cada par de nós já seja conhecida a quantidade de possíveis DAGs, g(n), poderá ser calculada pela seguinte fórmula:

$$g(n) = 2^{\binom{n}{2}} = 2^{n(n-1)/2} \qquad (2.2)$$

A quantidade de possíveis DAGs para domínios com 1 a 10 variáveis, para qualquer ordem dos nós e para uma única ordem dos nós foi calculada e é apresentada nas Tabelas 2.1 e 2.2.

Tabela 2.1: Quantidade de possíveis DAGs para redes com até 10 nós.

Número de nós	Quantidade de possíveis DAGs
1	1
2	3
3	25
4	543
5	29 281
6	3 781 503
7	1.1×10^{9}
8	7.8×10^{11}
9	1.2×10^{15}
10	4.2×10^{18}

Tabela 2.2: Quantidade de possíveis DAGs para redes com até 10 nós e ordem dos nós definida.

Número de nós	Quantidade de possíveis DAGs
1	1
2	2
3	8
4	64
5	1 024
6	32 768
7	2 097 152
8	3.0×10^{8}
9	6.9×10^{10}
10	3.5×10^{13}

Fica claro que devido à necessidade do cálculo das probabilidades condicionais para a avaliação de cada possível DAG e que devido à quantidade de possíveis DAGs, mesmo para uma ordem fixa dos nós, não é possível fazer uma busca exaustiva para redes com muitos nós, e por isso o uso de buscas heurísticas se faz necessário.

Pelo fato da busca heurística achar um valor ótimo local, a inicialização da busca é de fundamental importância para a solução obtida. Como quanto maior a estrutura ou quanto maior o número de arcos em uma BN, maior é o espaço em memória necessário para armazená-la e mais lenta é a propagação de informações [48], os algoritmos de aprendizagem Bayesiana baseados em busca normalmente realizam a busca adicionando nós a uma estrutura inicialmente sem nós.

Existem várias propostas de algoritmos de aprendizagem Bayesiana e alguns estudos sobre elas [1], [9], [10], [24], [25], [26], [23], [29], [36], [58], [13], [28], mas nada totalmente

conclusivo. Um desses algoritmos de aprendizado será um pouco mais detalhado e utilizado nesse trabalho.

Conforme afirmado anteriormente, os métodos de busca heurística normalmente usam a probabilidade conjunta ou as probabilidades condicionais para avaliar o quão uma rede é adequada ao conjunto de treinamento. Falta ainda expormos como as probabilidades condicionais são calculadas, uma vez que a partir destas a probabilidade conjunta pode ser facilmente obtida.

Mas antes de expor como essas probabilidades condicionais são calculadas é preciso fazer uma diferenciação entre os tipos de variáveis.

Podemos dividir as variáveis em variáveis qualitativas ou categóricas e variáveis quantitativas ou numéricas.

Variáveis qualitativas são variáveis que assumem valores entre uma série de possíveis categorias finitas. Variáveis qualitativas podem em alguns casos ser ordenadas, mas não é possível aplicar operações aritméticas sobre elas [65].

Variáveis quantitativas podem ser ordenadas e submetidas a operações aritméticas. As variáveis quantitativas podem, por sua vez, ser classificadas em dois grupos, discretas e contínuas. Uma variável quantitativa discreta assume apenas valores enumeráveis dentro de sua faixa de variação. Uma variável quantitativa contínua pode assumir qualquer valor real dentro de seu limite de variação.

Para nosso estudo as variáveis qualitativas e as variáveis quantitativas discretas com um número finito de possíveis valores podem ser agrupadas em uma mesma categoria à qual nos referiremos como variáveis discretas. As variáveis quantitativas discretas que podem assumir infinitos valores podem ser agrupadas às variáveis quantitativas contínuas, as quais chamaremos por abuso de linguagem de variáveis contínuas.

Os nós de uma BN podem representar tanto variáveis discretas, que assumem valores de um conjunto finito, quanto variáveis contínuas, que podem assumir infinitos valores. De acordo com as variáveis representadas as BN podem ser classificadas em três categorias: BNs discretas, BNs contínuas e BNs mistas [38].

A diferença entre as categorias está na representação das probabilidades condicionais. Enquanto que em BNs discretas as probabilidades condicionais são representadas por tabelas de probabilidade condicional, em BNs contínuas elas são expressas por funções de densidade ou distribuição de probabilidade condicional. E em BNs mistas as probabilidades condicionais serão representadas por uma mistura de tabelas e funções.

A tabela de probabilidade condicional de uma variável discreta condicionada por outras variáveis discretas pode ser estimada a partir das freqüências de ocorrência dessas variáveis no conjunto de treinamento pelo uso direto da definição da probabilidade condicional [37].

Fica claro, porém, que essa estimativa não pode ser usada para variáveis contínuas, uma vez que a probabilidade condicional deve ser representada por uma função e não por um conjunto de valores finitos. Para representar corretamente as probabilidades condicionais de uma variável contínua é preciso estimar tanto a família da função distribuição de probabilidade quanto seus parâmetros. Infelizmente essas funções de distribuição geralmente não são conhecidas para dados reais [65]. Uma convenção é assumir que a variável tenha uma distribuição normal [45]. Contudo, estamos fazendo fortes hipóteses sobre a natureza dos dados. Se essas hipóteses são garantidas então o modelo produzido pode ser uma boa aproximação da realidade [20], mas caso contrário nada se pode afirmar.

Uma alternativa a esse problema é a utilização de métodos de discretização [19]. A discretização é uma técnica muito comum para lidar com valores contínuos e consiste em unir valores adjacentes em intervalos se sua distinção contribui pouco para a estrutura do problema [63]. Através da discretização é criada uma variável discreta para representar a variável contínua. Cada valor da variável discreta corresponde a um intervalo de valores da variável contínua. A variável discreta resultante desse processo é usada no lugar da variável contínua [65].

2.3. Discretização

A discretização pode ser definida como o processo de transformação de uma variável contínua em uma variável discreta [6], [62] e [65].

O uso de variáveis discretas tornam tanto o processo de aprendizagem como o processo de inferência Bayesiana mais simples e eficientes, diminuindo a necessidade de poder computacional de processamento e armazenagem e o tempo necessário para a construção de uma BN e para a realização de inferências com a rede construída. O uso de variáveis discretas pode ainda, em diversos casos, resultar em uma BN mais adequada ao domínio do problema.

Mesmo que pudéssemos garantir as hipóteses sobre a natureza dos dados, feitas para a estimação de funções distribuição de probabilidade, ou seja, mesmo que as BNs contínuas ou mistas fossem uma boa aproximação da realidade existem outros fatores que favorecem o uso de técnicas de discretização:

- Vários algoritmos de aprendizagem e inferência não podem trabalhar com valores contínuos [30], [40];
- Não existem muitos algoritmos de inferência eficientes para modelos contínuos e mistos [19];
- Para usuários e especialistas características discretizadas são mais fáceis de entender, usar e explicar. [39]
- Os algoritmos de aprendizagem que podem tratar variáveis contínuas normalmente demandam menos recursos quanto tratam somente variáveis discretas, desse modo o aprendizado com variáveis discretas é mais rápido [17], [33], [39], [52], [62], [65];
- Apesar de capturar apenas características grosseiras da distribuição das variáveis contínuas, a discretização resulta em modelos que podem ser usados eficientemente para inferência probabilística e tomada de decisão [50];
- Para tarefas específicas como classificação, é suficiente estimar a probabilidade de um dado pertencer a um certo intervalo [20].
- A aplicação de algoritmos de indução desenvolvidos para valores discretos a valores contínuos pode gerar muitos relacionamentos com pouco poder preditivo [63];
- Os classificadores resultantes de um processo de discretização são freqüentemente significativamente menos complexos e algumas vezes mais precisos que os classificadores obtidos com dados contínuos [18].

É interessante notar que várias das aplicações atuais de redes Bayesianas envolvem discretização feita por especialistas humanos [19]. A discretização produz uma generalização inerente, pelo agrupamento de dados em alguns intervalos, representando-os de forma mais geral. Além disso, a discretização tem alguma plausibilidade psicológica uma vez que em muitos casos as pessoas aparentemente realizam um passo de pré-processamento similar, representando dados naturalmente contínuos, como temperatura, tempo e velocidade em valores nominais, como por, exemplo, temperatura alta, temperatura normal, temperatura baixa; tempo bom, tempo ruim, velocidade alta, velocidade baixa [61].

Em termos práticos quando estamos trabalhando com aprendizagem Bayesiana a partir de bases de dados, as variáveis apresentadas são todas discretas, uma vez que para cada tupla presente no banco cada variável assume um único valor. Desse modo as variáveis de aprendizagem poderão assumir apenas um número finito de valores.

Ao utilizarmos a rede, porém, podem ser apresentados valores distintos dos da base de dados. Dessa forma as variáveis originalmente contínuas, que por serem armazenadas em uma

base de dados se tornaram discretas, não podem simplesmente ser encaradas como se discretas fossem.

Outro fator que deve ser levado em conta é que muitos dos problemas que normalmente ocorrem com variáveis contínuas, como aprendizagem mais lenta, ausência de generalização, dificuldade de entendimento por parte de usuários e especialistas, necessidade de mais recursos para o aprendizado, geração de muitos relacionamentos com pouco poder preditivo, também ocorrem se a quantidade de possíveis valores ou a quantidade de intervalos de uma variável discreta for muito grande.

Além disso, há outros problemas associados ao uso de variáveis discretas com uma grande quantidade de possíveis valores, como geração de redes de complexidade elevada, necessidade de mais espaço para armazenamento da rede e propagação de crença mais lenta. Por isso a discretização deve resultar preferencialmente em variáveis discretas com poucos intervalos.

A discretização em si pode também ser encarada como uma forma de descoberta de conhecimento, onde valores críticos em um domínio contínuo podem ser revelados [33]. Os intervalos de discretização, portanto, não devem esconder padrões de relacionamento entre as variáveis. Eles devem ser escolhidos com cuidado ou descobertas potenciais podem ser perdidas [3].

Existem vários métodos de discretização [65], a maioria deles foi desenvolvida para sistemas de classificação, outros têm uma utilização mais genérica.

Os dois métodos de discretização mais simples e muito utilizados são a Discretização em Intervalos Iguais (*Equal Width Discretization* - EWD) [6], [16], [40], [61], [65] e a Discretização em Freqüências Iguais (*Equal Frequency Discretization* - EFD) [6], [39], [61], [65]. Na EWD são criados k intervalos de igual tamanho onde k é um número inteiro maior que zero definido externamente. Na EFD são criados k intervalos de modo que cada intervalo contenha aproximadamente o mesmo número de ocorrência de valores. Esses métodos de discretização não levam em conta qualquer informação sobre possíveis relacionamentos entre as variáveis e, portanto podem fazer com que os padrões de relacionamento sejam perdidos. Sua simplicidade de implementação, porém, os torna muito populares, principalmente na comparação de desempenho com outros métodos de discretização.

A maioria dos métodos de discretização, como EWD e EFD produzem intervalos disjuntos, onde cada valor só pode pertencer a um único intervalo. Existem, porém, métodos que discretizam as variáveis em intervalos não disjuntos, como a Discretização para Aprendizagem Fuzzy (*Fuzzy Learning Discretization*), a Discretização Fuzzy (*Fuzzy*

Discretization) e a Discretização não Disjunta (*Non-Disjoint Discretization*) [65]. Os algoritmos de aprendizagem Bayesiana, no entanto, só conseguem lidar com intervalos disjuntos [65] e por isso esses métodos de discretização não são úteis em nosso contexto.

Outro método de discretização de aplicação específica é a Discretização Qualitativa Dinâmica (*Dynamic Qualitative Discretization*), utilizada em séries temporais, uma vez que os intervalos de discretização podem mudar de acordo com os valores apresentados [65].

Um método de discretização bastante interessante, mas pouco usado na prática é a Discretização Ociosa (*Lazy Discretization* - LD). A LD não discretiza os dados para o aprendizado. Os dados são discretizados quando uma evidência é apresentada. A discretização é feita com base nos valores apresentados como evidência e os intervalos são definidos de modo que os valores apresentados fiquem no centro de um intervalo. Como o aprendizado só pode ser realizado após a discretização e como para evidências diferentes intervalos diferentes são definidos, sua utilização acaba se tornando extremamente lenta na prática [65].

Como citado anteriormente a grande maioria dos métodos de discretização para o uso com BNs foi elaborada com o objetivo de melhorar o desempenho das BNs como classificadores e portanto utilizam informações relacionadas à variável a ser classificada na definição dos intervalos de discretização.

Alguns desses métodos se baseiam diretamente no erro de classificação para a definição dos intervalos, como a Discretização baseada em Erro (*Error-based Discretization*) [33] , a Discretização Dinâmica (*Dynamic Discretization*) [65] e a Discretização de Melhoria Interativa (*Interative-improvement Discretization*) [65]. Já a Discretização Sensível a Custo (*Cost-sensitive Discretization*) [65] permite que a definição dos intervalos seja feita baseada numa função custo calculada a partir dos erros de classificação.

Outros métodos, apesar de se basearem na variável de classe, não usam o erro de classificação para definir os intervalos de discretização, mas sim outros tipos de medida como entropia, quantidade de informação, ou uma métrica própria, mas sempre com relação à variável de classe. Alguns desses métodos também utilizam técnicas de agrupamento (*clustering*) [44], [51] para auxiliar na definição dos intervalos. Exemplos desses métodos são a Discretização de Minimização de Entropia (*Entropy Minimization Discretization* - EMD) [3], [6], [16], [33], Discretização Chimerge – CMD [6], [39], [61], Discretização baseada em Distância (*Distance-based Discretization* – DBD) [6], [39], Discretização 1-Regras (*1-Rules discretization* – 1RD) [3], [16], [39] e Discretização ConMerge [63].

Esses métodos objetivam o aumento da informação mútua entre cada variável e a variável de classe, bem como a maximização da precisão preditiva.

Em geral e particularmente quando uma rede não será usada apenas para classificação, é preferível que a discretização de cada variável aumente ou preserve sua informação mútua com respeito a todas as variáveis diretamente relacionadas [19].

Na descoberta de conhecimento normalmente os dados são analisados de uma maneira exploratória, onde a ênfase não está na precisão preditiva, mas na descoberta de padrões previamente escondidos. Deste modo o critério de escolha de intervalos de discretização deve ser diferente do usado nas discretizações para classificadores [3].

Alguns poucos métodos de discretização foram desenvolvidos seguindo essa filosofia, como a Discretização Multivariável (*Multivariable Discretization* - MD)[3] e a Discretização não Supervisionada Relativa (*Relative Unsupervided Discretization* - RUD) [40]. Ambas iniciam o processo de discretização usando EWD ou EFD e definem os intervalos de discretização de uma variável levando em conta seu relacionamento com todas as outras variáveis. Embora não levem em conta os relacionamentos entre os nós, o que seria ideal, já são bem mais adequados à aprendizagem Bayesiana que os métodos desenvolvidos para classificação. Esse tipo de discretização será abordado em mais detalhes no capítulo 3.

Existem ainda alguns outros métodos de discretização como Discretização Ordinal (*Ordinal Discretization*) [18], Discretização Proporcional (*Proportional Discretization* – PD) [65] e Discretização de Agrupamentos de média-K (*K-means Clustering Discretization*) [61], [40], que são variações, versões melhoradas ou casos específicos dos outros métodos.

Estudos comparativos feitos entre o uso de alguns desses métodos e a utilização de variáveis contínuas para aprendizagem em classificadores [16], [33], [39], [61] mostram que o erro de classificação não varia consideravelmente, que em vários casos uma rede discreta realmente apresenta um melhor desempenho que a rede contínua ou mista correspondente e que em média a discretização reduz o tempo de aprendizagem à metade [39].

Os erros de classificação obtidos a partir do uso de variáveis contínuas e de alguns métodos de discretização na aprendizagem Bayesiana são apresentados nas Tabelas 2.3 e 2.4. As bases de dados utilizadas são públicas e muito usadas na construção e comparação de classificadores de diversos tipos.

Em alguns casos, como para as bases de dados “Cleve”, “Diabetes” e “Heart”, Tabela 4.3, todos os métodos de discretização testados resultaram em um erro de classificação menor do que com a utilização das variáveis contínuas. Em outros, como para as bases “Anneal”, “Hypothyroid” e “Vehicle”, Tabela 4.3, o erro de classificação foi menor com a utilização das variáveis contínuas.

Tabela 2.3: Erro percentual de classificação para uso de variáveis contínuas e discretizadas para várias bases de dados [16].

Base de Dados	Variáveis Contínuas	EMD	1RD	EWD
Anneal	8,35	10,35	12,80	9,68
Australian	14,64	14,35	14,78	15,94
Breast	5,29	5,58	5,01	5,15
Cleve	26,38	20,76	20,77	23,43
Crx	13,91	15,22	14,49	15,22
Diabetes	29,16	23,96	27,60	26,56
German	27,70	26,00	29,90	28,90
Glass	34,11	30,38	40,69	40,18
Glass2	25,80	23,33	28,71	19,58
Heart	22,96	18,89	17,41	21,48
Hepatitis	21,94	24,52	20,65	20,00
Horse-colie	15,22	14,40	14,40	14,67
Hypothyroid	0,80	1,00	2,00	2,70
Iris	5,33	6,00	6,00	4,00
Sick-euthyroid	2,30	2,70	2,60	5,90
Vehicle	30,14	30,38	33,20	31,55
Média	17,75	16,74	18,19	17,81

Tabela 2.4: Erro percentual de classificação para uso de variáveis contínuas e discretizadas para várias bases de dados [39].

Base de Dados	Variáveis Contínuas	CMD	EFD	1RD	DBD	EMD
Australian	15,28	14,42	14,51	13,00	13,82	14,00
Breast	4,72	4,92	7,65	13,27	7,79	6,37
Bupa	33,13	33,99	43,96	36,32	31,90	34,29
Glass	1,86	1,90	22,43	18,79	2,77	2,31
Heart	22,16	20,21	22,86	20,00	16,07	20,35
Ionos	9,14	8,94	9,67	11,98	10,69	9,15
Iris	4,34	5,02	8,14	6,07	10,23	4,25
Pima	26,22	27,31	27,86	25,17	22,91	25,21
Thyroid	8,00	8,91	12,69	6,44	7,90	4,23
Vehicle	26,87	30,87	31,39	28,57	29,81	29,47
Wine	6,22	7,92	7,96	6,84	7,53	7,95
Média	14,36	14,95	18,99	16,95	14,69	14,33

Com raras exceções, como, por exemplo, para a base “Glass”, Tabela 2.4, as diferenças de desempenho entre as redes geradas a partir das variáveis contínuas e dos diferentes métodos de discretização foram muito pequenas.

Estudos comparativos feitos apenas entre métodos de discretização [6], [40], [65], sem a comparação com o desempenho de uma rede contínua ou mista, também mostram que o

desempenho da rede não varia muito com o método utilizado. Os resultados obtidos em uma dessas comparações são apresentados na Tabela 2.5.

Tabela 2.5: Erro percentual de classificação para uso de diferentes métodos de discretização para várias bases de dados [65].

Base de Dados	EWD	EFD	EMD	LD	PD
Adult	18,2	18,6	17,3	18,1	17,1
Australian-Sign-Language	38,3	37,7	36,5	36,4	35,8
Annealing	3,8	2,4	2,1	2,3	2,1
Breast-Cancer(Wisconsin)	2,5	2,6	2,7	2,6	2,7
Census-Income	24,5	24,5	23,6	24,6	23,3
Credid-Screening(Australia)	15,6	14,5	14,5	13,9	14,4
Echocardiogram	29,6	30,0	23,8	29,1	25,7
Forest-Covertype	32,4	33,0	32,1	32,3	31,7
German	25,1	25,2	25,0	25,1	24,7
Glass-Identification	39,3	33,7	34,9	32,0	32,6
Handwritten-Digits	12,5	13,2	13,5	12,8	12,0
Heart-Disease(Cleveland)	18,3	16,9	17,5	17,6	17,4
Hepatitis	14,3	14,2	13,9	13,7	14,1
Horse-Colie	20,5	20,8	20,7	20,8	20,3
Hypothyroid	3,6	2,8	1,7	2,4	1,8
Ionosphere	9,4	10,3	11,1	10,8	10,4
Ipums-la-99	21,0	21,1	21,3	20,4	20,6
Iris	5,7	7,7	6,8	6,7	6,4
Labor-Negotiation	12,3	8,9	9,5	9,6	7,4
Letter-Recognition	29,5	29,8	30,4	27,9	25,7
Liver-Disorders	37,1	36,4	37,4	37,0	38,9
Multiple-Features	31,0	31,8	32,9	31,0	31,2
Musk	13,7	18,4	9,4	15,4	8,2
Pima-Indians-Diabetes	24,9	25,6	26,0	25,4	26,0
Pionner-MobileRobot	13,5	15,0	19,3	15,3	4,6
Satimage	18,8	18,8	18,1	18,4	17,8
Sonar	25,6	25,1	25,5	25,8	25,7
Vehicle	38,7	38,8	38,9	38,1	38,1
Wine-Recognition	3,3	2,4	2,6	2,9	2,4
Média	20,1	20,0	19,6	19,6	18,6

Em algumas bases de dados como “Breast-Cancer(Wisconsin)”, “Heart-Disease(Cleveland)”, “Ionosphere”, “Iris” e “Liver-Disorders”, Tabela 2.5, os métodos mais simples, EWD e EFD se mostram mais efetivos que os outros, principalmente se for considerada a simplicidade de sua utilização.

Embora alguns métodos como o EMD se mostrem menos sensíveis às diferentes bases dados, com uma efetividade média melhor, não é possível elegê-lo como o melhor método de

discretização, uma vez que para algumas bases, como “Ionosphere”, “Letter-Recognition”, “Multiple-Features” e “Pionner-MobileRobot” , Tabela 2.5 o desempenho apresentado com o uso de EMD foi o pior do que com os outros métodos testados.

As variações não muito significativas pelo uso de um método e outro também dificultam a escolha de um método como sendo o melhor dentre os apresentados.

Comparando os resultados obtidos com o uso dos métodos de discretização a bases de dados comuns em estudos diferentes [16], [39], [65], é possível analisar a influência não só das diferentes bases, mas de diferentes procedimentos no desempenho dos métodos de discretização. A Tabela 2.6 reúne os resultados das Tabelas de 2.3 a 2.5.

A variação do desempenho de um mesmo método em testes diferentes foi algumas vezes maior que a variação do desempenho para métodos diferentes, como no caso da base “Vehic”, a variação do desempenho do EWD entre dois testes diferentes foi de mais de 7%, Tabela 2.6, enquanto que as variações entre EWD e outros métodos em um mesmo teste não foram superiores a 2%, Tabelas 2.3 e 2.5.

Tabela 2.6: Erro percentual de classificação para uso de diferentes métodos de discretização para várias bases de dados de várias referências.

Base	**Cont. [16]**	**Cont. [39]**	**EMD [16]**	**EMD [39]**	**EMD [65]**	**1RD [16]**	**1RD [39]**	**EWD [16]**	**EWD [65]**	**EFD [39]**	**EFD [65]**
Breast	5,29	4,72	5,58	6,37	2,7	5,01	13,27	5,15	2,5	7,65	2,6
Heart	22,96	22,16	18,89	20,35	17,5	17,41	20,00	21,48	18,3	22,86	16,9
Iris	5,33	4,34	6,00	4,25	6,8	6,00	6,07	4,00	5,7	8,14	7,7
Vehic	30,14	26,87	30,38	29,47	38,9	33,20	28,57	31,55	38,7	31,39	38,8

Nesse caso específico essa variação pode ser devida ao uso de uma quantidade diferente de intervalos de discretização nos dois testes, de qualquer modo, essas variações de desempenho entre diferentes utilizações do mesmo método, que não ocorrem somente com EWD, mostram que o desempenho dos métodos de discretização pode variar mais dependendo de decisões de implementação ou de execução de um mesmo método do que com o uso de métodos diferentes.

Desse modo com base nos resultados apresentados realmente não é possível eleger um dentre esses métodos como o melhor, mas é possível afirmar que até agora nenhum teste apresentou diferenças consideráveis entre a utilização dos métodos apresentados.

A maior parte das comparações entre os métodos de discretização é feita com base no erro de classificação, uma vez que grande parte das mesmas foi desenvolvida para esse contexto. Essa, porém, não é uma boa forma de avaliar o quanto esses métodos de discretização se

adequam para a descoberta de conhecimento, ou seja, para a descoberta de como as variáveis do domínio se relacionam.

Para avaliar isso é necessário que outros fatores diferentes do erro de classificação sejam usados como fatores de comparação de desempenho.

Um estudo comparativo realizado por Bay [3] entre EMD, que apresenta um dos melhores desempenhos em classificadores, e MD, que realiza a discretização baseando-se na distribuição conjunta das variáveis e não seguindo uma métrica relativa à uma variável de classe, mostra que a segunda cria intervalos com maior significado semântico, ou seja, que fazem mais sentido para as pessoas, como por exemplo separar a renda familiar em até U$1.400,00/mês, de U$1.400,00 a U$2.500,00, de U$2.500,00 a U$4.300,00, de U$4.300,00 a U$6.250,00 e acima de U$6.250,00, o que pode ser facilmente relacionado com o entendimento de riqueza e pobreza, pelo menos para os padrões americanos. Já o EMD cria intervalos em até U$3.000,00/mês, de U$3.000,00 a U$17.000,00, de U$17.000,00 a U$33.000,00 e mais 3 intervalos acima dos U$33.000,00, o que não tem sentido no mundo real.

Além disso, os intervalos encontrados pelo EMD podem variar bastante dependendo da variável escolhida como variável de classe.

Alguns relacionamentos bastante expressivos encontrados com o uso de MD como, por exemplo, entre *perdas de capital elevadas* e *salário elevado* não foram encontradas quando o EMD foi utilizado [3]. Esses resultados sugerem que EMD e outros métodos de discretização para classificação não são apropriados para descoberta de conhecimento.

Uma abordagem alternativa à discretização dos dados para serem usados em algoritmos de aprendizado é a discretização dos dados durante o aprendizado. Essa é a proposta da Discretização do Princípio de Tamanho Mínimo de Descrição (*Minimum Description Length Principle Discretization* – MDLPD) [19]. O princípio de Tamanho Mínimo de Descrição (*Minimum Description Lenght* – MDL) é usado em outros métodos de discretização como métrica ou critério de parada, mas, na aprendizagem com discretização esse princípio é usado para avaliar tanto a discretização quanto a BN gerada pela aprendizagem. O MDL avalia as BNs de acordo com o espaço necessário para armazenar o conhecimento adquirido. A rede que ocupe um menor espaço, em termos de armazenamento, e que possa representar o conjunto de treinamento é considerada a melhor.

Claramente esse princípio favorece discretizações com poucos intervalos e BNs com poucas ligações, uma vez que aumentando a quantidade de intervalos e a quantidade de

relacionamentos as tabelas de probabilidade condicional aumentam e conseqüentemente o espaço necessário para armazenar a rede também.

Na MDLPD a discretização de uma variável é feita levando-se em conta o relacionamento da variável com seus vizinhos na rede e não com a rede toda. A discretização é iniciada com uma EWD e então cada variável é tratada como se a discretização das outras já tivesse sido realizada.

Em comparações preliminares entre MDLPD e EMD para o aprendizado de uma rede naive-Bayes para classificação os resultados obtidos com EMD foram ligeiramente melhores [19]. Mesmo quando a MDLPD pode descobrir seu próprio DAG seus resultados não foram melhores do que os da EMD. A métrica usada pelo MDLPD para direcionar a busca pela melhor rede e pela melhor discretização, que procura minimizar o espaço de armazenamento e conseqüentemente privilegia redes mais simples está justamente abrindo mão da precisão preditiva em favor da minimização do tamanho da descrição. Como o desempenho apresentado na comparação com uma rede simples como a naive-Bayes já foi inferior ao EMD, os resultados em comparação com redes mais complexas devem ser ainda piores.

Podemos concluir então que apesar da existência de vários métodos de discretização poucos se adequariam para a aprendizagem Bayesiana para descoberta de conhecimento e não a construção de um classificador Bayesiano. Dos vários métodos propostos apenas a MD e a RUD se baseiam no relacionamento entre as variáveis como um todo e não com relação a uma única variável de classe.

No capítulo 3 um novo método de discretização é proposto e comparado ao cerne comum da MD e da RUD e no capítulo 6 são apresentados resultados comparativos entre o método proposto e EFD.

3. Proposta de Solução

No capítulo 2 foram apresentados vários métodos de discretização que não se adequam à aprendizagem Bayesiana para descoberta de conhecimento. Como a discretização pode justamente esconder os padrões que poderiam ser aprendidos, é necessário que o processo de discretização preserve ao máximo esses relacionamentos.

Nesse capítulo será apresentado um método de discretização; denominado Discretização via Tabela de Probabilidades – DTP, baseado nas relações das variáveis com seus sucessores e nas tabelas de probabilidade condicional.

O processo proposto é intercalado com a aprendizagem, mas não interage diretamente com ela, como no caso da DMDLP.

Inicialmente as variáveis são discretizadas por EWD ou EFD e os intervalos resultantes desse processo são então unidos pela DTP. Para a avaliação dos intervalos que devem ou não ser unidos foi definida uma métrica, denominada Diferença Média Quadrática – DMQ.

A DMQ entre dois intervalos, pode ser interpretada como sendo o Erro Médio Quadrático das probabilidades condicionadas pelo primeiro intervalo, considerando-se que as probabilidades condicionadas pelo segundo intervalo estão corretas.

Intuitivamente pode-se interpretar a DMQ como sendo a diferença entre as probabilidades condicionadas pelos dois intervalos sendo analisados. Um valor de DMQ próximo de zero indica que os intervalos condicionam probabilidades muito próximas e portanto podem ser unidos sem que a BN resultante perca informação sobre o relacionamento entre as variáveis. Valores de DMQ elevados indicam que os intervalos condicionam probabilidades consideravelmente distintas e portanto a união desses intervalos poderia acarretar a perda de relacionamentos entre as variáveis.

Cálculo da Diferença Média Quadrática:

Caso 1: Nó A com p intervalos, $\{a_1, a_2, \ldots, a_p\}$, e um sucessor B que tenha apenas A como antecessor e n intervalos, $\{b_1, b_2, \ldots, b_n\}$:

A diferença média quadrática, DMQ(a_k, a_l, B) , entre os intervalos a_k e a_l com relação à variável B será:

$$DMQ(a_k, a_l, B) = \frac{1}{n} \cdot \sum_{i=1}^{n} [\mathrm{P}(B = b_i \mid A = a_k) - \mathrm{P}(B = b_i \mid A = a_l)]^2 \qquad (3.1)$$

Caso 2: Nó A com p intervalos, $\{a_1, a_2, \ldots, a_p\}$, e um sucessor C que tem n intervalos, $\{c_1, c_2, \ldots, c_n\}$, e dois antecessores A e B, B com m intervalos, $\{b_1, b_2, \ldots, b_m\}$:

A diferença média quadrática, DMQ(a_k, a_l, C) , entre os intervalos a_k e a_l com relação à variável C será:

$$DMQ(a_k, a_l, C) = \frac{1}{n \cdot m} \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} [\mathrm{P}(C=c_i | B=b_j, A=a_k) - \mathrm{P}(C=c_i | B=b_j, A=a_l)]^2 \quad (3.2)$$

Caso geral: Nó A com um sucessor Z que tem n antecessores:

A fórmula da diferença média quadrática terá a mesma forma, com n somatórios e cada probabilidade condicional estará condicionada às n variáveis.

Podemos ainda no lugar da Diferença Média Quadrática usar a Diferença Média Quadrática Ponderada - DMQP entre as probabilidades condicionais. Essa ponderação irá fazer com que as diferenças das probabilidades condicionais que afetam uma maior quantidade de tuplas no conjunto de aprendizagem tenham um peso maior na composição da diferença. O cálculo é o mesmo da Diferença Média Quadrática com exceção de uma ponderação aplicada a cada termo da somatória.

Cálculo da Diferença Média Quadrática Ponderada:

Caso 1: Nó A com p intervalos, $\{a_1, a_2, \ldots, a_p\}$, e um sucessor B que tenha apenas A como antecessor e n intervalos, $\{b_1, b_2, \ldots, b_n\}$:

A diferença média quadrática ponderada, DMQP(a_k, a_l, B) , entre os intervalos a_k e a_l com relação à variável B será:

$$DMQP(a_k, a_l, B) = \frac{1}{n} \cdot \sum_{i=1}^{n} [\mathrm{P}(B = b_i \mid A = a_k) - \mathrm{P}(B = b_i \mid A = a_l)]^2 \cdot \mathrm{P}(B = b_i) \quad (3.3)$$

Caso 2: Nó A com p intervalos, $\{a_1, a_2, \ldots, a_p\}$, e um sucessor C que tem n intervalos, $\{c_1, c_2, \ldots, c_n\}$, e dois antecessores A e B, B com m intervalos, $\{b_1, b_2, \ldots, b_m\}$:

A diferença média quadrática, DMQ(a_k, a_l, C) , entre os intervalos a_k e a_l com relação à variável C será:

$$DMQP(a_k, a_l, C) = \frac{1}{n \cdot m} \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} [\mathrm{P}(C=c_i | B=b_j, A=a_k) - \mathrm{P}(C=c_i | B=b_j, A=a_l)]^2 \cdot \mathrm{P}(B=b_j, C=c_i) \quad (3.4)$$

Caso geral: Nó A com um sucessor Z que tem n antecessores:

A fórmula da diferença média quadrática ponderada terá a mesma forma, com n somatórios, cada probabilidade condicional estará condicionada às n variáveis e a probabilidade de ponderação também terá n variáveis.

Quando a variável sendo analisada tiver mais de um sucessor é necessário que as DMQs relativas a cada sucessor sejam combinadas em uma única DMQ. Para tanto é usada a norma euclidiana das DMQs com relação a cada sucessor. A fórmula da DMQ poderia ser

generalizada para contemplar essa combinação, mas seu cálculo fica mais claro e mais fácil de ser implementado se for realizado em duas etapas.

Tanto a DMQ quanto a DMQP foram definidas para quaisquer pares de intervalos a_k e a_l. Em um caso mais geral essa métrica pode ser usada para, por exemplo, unir categorias de uma variável qualitativa, mas no caso de discretização de variáveis contínuas, a discretização define intervalos de variação e, portanto, apenas intervalos adjacentes podem ser unidos.

Descrição dos passos da Discretização via Tabela de Probabilidades - DTP:

- Discretizar todas as variáveis contínuas usando EWD ou EFD.
- Utilizar um algoritmo de aprendizagem Bayesiana e gerar a BN que mais se adequa aos dados assim discretizados. Marcar os nós discretizados por EWD ou EFD como *não processados* e os nós que já eram discretos *como processados*.
- Para os nós folha, ou seja, para os nós que não tem sucessores, marcados como *não processados* unir intervalos adjacentes e refazer a aprendizagem. Repetir esse processo de união de intervalos enquanto os relacionamentos entre os nós folhas e seus antecessores se preservarem ou até que um número determinado de intervalos seja atingido.
- Repetir até que todos os nós discretizados tenham sido processados:
    - Selecionar apenas os nós marcados como *não processados* cujos sucessores já estejam marcados como *processados*.
    - Para cada nó selecionado, repetir até que nenhum par de intervalos possa ser unido:
        - Tomar as tabelas de probabilidade condicional de cada um de seus sucessores e calcular a Diferença Média Quadrática, ou a Diferença Média Quadrática ponderada das probabilidades condicionais entre cada intervalo e os intervalos adjacentes.
        - Unir os pares de intervalo cuja diferença média quadrática seja menor que um certo valor absoluto ou relativo, definido externamente. Esses valores podem ser uma função da quantidade de intervalos da variável.
        - Refazer a aprendizagem Bayesiana e tentar unir intervalos novamente.
- Caso a quantidade de intervalos das variáveis discretizadas ainda seja muito grande elas podem ser marcadas novamente como não processadas e o processo de união de intervalos pode ser repetido.

Exploraremos a seguir um exemplo numérico para ilustrar o método proposto e compará-lo com a filosofia dos métodos MD e RUD.

Os três métodos partem de uma discretização mais simples como EWD ou EFD e depois tentam unir os intervalos. As métricas para união dos intervalos também são parecidas, pois todas se baseiam na distribuição de probabilidade condicional ou conjunta. A grande diferença está no fato da discretização proposta, além de se basear nas probabilidades condicionais, discretizar as variáveis na seqüência inversa de dependência, enquanto que os outros dois não se preocupam com a seqüência de discretização.

Nesse exemplo iremos analisar o quanto a seqüência de discretização pode influenciar nos resultados, em outras palavras, estaremos comparando métodos como MD e RUD que não definem uma seqüência de discretização com a DTP proposta que discretiza os nós em uma ordem definida pelo DAG da BN gerada inicialmente. Para tanto iremos usar a mesma métrica, diferença média quadrática, e variar apenas a seqüência de discretização. No uso de MD ou RUD a métrica seria um pouco diferente, mas as diferenças causadas pela ordem de discretização dos nós continuariam existindo.

Exemplo Ilustrativo:

O domínio do problema é composto por 3 variáveis, *A*, *B* e *C*, onde *A* e *B* são variáveis contínuas e *C* é uma variável discreta que pode assumir apenas 2 valores.

Inicialmente as variáveis *A* e *B* são discretizadas em 10 intervalos por EWD ou EFD. O conjunto de dados é então utilizado na aprendizagem Bayesiana resultando no DAG da figura 3.1 e nas probabilidades condicionais das tabelas 3.1 e 3.2.

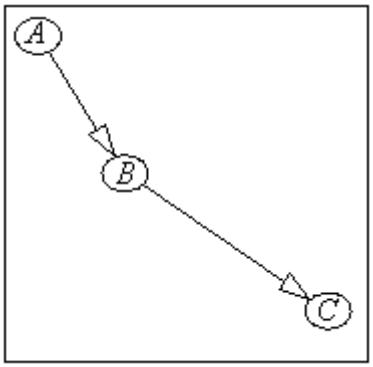

Figura 3.1: DAG do exemplo de discretização.

Tabela 3.1: Probabilidades condicionais de C em relação a B. Cada célula da tabela representa a Probabilidade da variável C estar no intervalo correspondente à linha dado que a variável B está no intervalo correspondente à coluna. Por exemplo, o valor 0.9 na última célula da tabela corresponde à P($C = c_2 \mid B = b_{10}$).

	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}
c_1	0.9	0.9	0.8	0.8	0.55	0.45	0.2	0.2	0.1	0.1
c_2	0.1	0.1	0.2	0.2	0.45	0.55	0.8	0.8	0.9	0.9

Tabela 3.2: Probabilidades condicionais de B em relação a A. Cada célula da tabela representa a Probabilidade da variável B estar no intervalo correspondente à linha dado que a variável A está no intervalo correspondente à coluna. Por exemplo, o valor 0.10 na última célula da tabela corresponde à P($B = b_{10} \mid A = a_{10}$).

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
b_1	0.10	0.04	0.08	0.06	0.13	0.06	0.06	0.08	0.12	0.18
b_2	0.02	0.04	0.08	0.06	0.06	0.13	0.06	0.08	0.12	0.10
b_3	0.02	0.04	0.08	0.06	0.08	0.06	0.06	0.08	0.12	0.10
b_4	0.02	0.04	0.08	0.14	0.05	0.07	0.14	0.08	0.12	0.10
b_5	0.15	0.20	0.20	0.25	0.15	0.15	0.25	0.20	0.18	0.13
b_6	0.25	0.20	0.20	0.15	0.25	0.25	0.15	0.20	0.18	0.23
b_7	0.13	0.11	0.07	0.05	0.06	0.05	0.13	0.07	0.04	0.02
b_8	0.13	0.11	0.07	0.05	0.12	0.05	0.05	0.07	0.04	0.02
b_9	0.13	0.11	0.07	0.05	0.05	0.12	0.05	0.07	0.04	0.02
b_{10}	0.05	0.11	0.07	0.13	0.05	0.06	0.05	0.07	0.04	0.10

Após a discretização inicial, os nós A e B que foram discretizadas por EWD ou EFD estão marcados como *não processados* e o nó C, que representa uma variável que já era discreta, está marcado como *processado*.

O método proposto, DTP, inicia a união dos intervalos pelo nó B, uma vez que é o único nó marcado como *não processado*, cujos sucessores estão marcados como *processados*. Para esse exemplo os intervalos serão unidos até que cada variável tenha apenas 3 intervalos. Estamos usando uma DMQ absoluta de 1.0 para mais de 3 intervalos e uma DMQ absoluta de 0.0 para 3 intervalos ou menos.

Calculando a DMQ de cada intervalo para o seguinte teremos os valores de DMQ da tabela 3.3.

Tabela 3.3: DMQ das probabilidades condicionadas pelos intervalos de B.

	b_1-b_2	b_2-b_3	b_3-b_4	b_4-b_5	b_5-b_6	b_6-b_7	b_7-b_8	b_8-b_9	b_9-b_{10}
DMQ	**0.0**	0.01	**0.0**	0.0625	**0.01**	0.0625	**0.0**	0.01	**0.0**

Por esses resultados os intervalos b_1 e b_2 serão unidos em um único intervalo, o mesmo acontecendo para os intervalos b_3 e b_4; b_5 e b_6; b_7 e b_8; e b_9 e b_{10}. Após essa união e uma nova aprendizagem temos as probabilidades condicionais da tabela 3.4.

Tabela 3.4: Probabilidades condicionais de *C* em relação a *B* após a primeira unificação de intervalos de *B*.

	b_{12}	b_{34}	b_{56}	b_{78}	b_{910}
c_1	0.9	0.8	0.5	0.2	0.1
c_2	0.1	0.2	0.5	0.8	0.9

Calculando mais uma vez os valores de DMQ teremos a tabela 3.5.

Tabela 3.5: DMQ das probabilidades condicionadas pelos intervalos de *B* após a primeira unificação de intervalos de *B*.

	b_{12}-b_{34}	b_{34}-b_{56}	b_{56}-b_{78}	b_{78}-b_{910}
DMQ	**0.01**	0.09	0.09	**0.01**

Os intervalos b_{12} e b_{34} serão unificados, o mesmo acontecendo para os intervalos b_{78} e b_{910}. Após uma nova aprendizagem teremos as probabilidades de tabela 3.6.

Tabela 3.6: Probabilidades condicionais de *C* em relação a *B* após a segunda unificação de intervalos de *B*.

	b_{1234}	b_{56}	b_{78910}
c_1	0.85	0.50	0.15
c_2	0.15	0.50	0.85

Calculando mais uma vez os valores de DMQ teremos a tabela 3.7.

Tabela 3.7: DMQ das probabilidades condicionadas pelos intervalos de *B* após a segunda unificação de intervalos de *B*.

	b_{1234}- b_{56}	b_{56}-b_{78910}
DMQ	0.1225	0.1225

Como nenhum par de intervalos pode ser unido, devido aos valores da DMQ, a unificação dos intervalos de *B* está terminada e será iniciada a unificação dos intervalos de *A*. Para tanto é necessário analisar as probabilidades condicionais de *B* dado *A* após a unificação dos intervalos de *B*, que são apresentados na tabela 3.8.

Tabela 3.8: Probabilidades condicionais de B em relação a A após a unificação dos intervalos de B.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
b_{1234}	0.16	0.16	0.32	0.32	0.32	0.32	0.32	0.32	0.48	0.48
b_{56}	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.36	0.36
b_{78910}	0.44	0.44	0.28	0.28	0.28	0.28	0.28	0.28	0.16	0.16

É fácil verificar visualmente que a DMQ entre a_1 e a_2 será zero, o mesmo ocorrendo para a DMQ entre a_3, a_4, a_5, a_6, a_7 e a_8 e quaisquer unificações de intervalos desse subconjunto; e para a DMQ entre a_9 e a_{10}.

Os intervalos da variável A serão então a_{12}, a_{345678} e a_{910}. E as probabilidades condicionais de B dado A são as da tabela 3.9.

Tabela 3.9: Probabilidades condicionais de B em relação a A após a unificação dos intervalos de B e de A.

	a_{12}	a_{345678}	a_{910}
b_{1234}	0.16	0.32	0.48
b_{56}	0.40	0.40	0.36
b_{78910}	0.44	0.28	0.16

Se os outros métodos iniciassem a discretização também pela variável B teríamos o mesmo resultado, vamos analisar então quais seriam os intervalos e as probabilidades condicionais caso a unificação dos intervalos fosse iniciada pela variável A ao invés da variável B.

Calculando a DMQ a partir das probabilidades da tabela 3.2 teríamos as DMQ apresentadas na tabela 3.10.

Tabela 3.10: DMQ das probabilidades condicionadas pelos intervalos de A.

	a_1-a_2	a_2-a_3	a_3-a_4	a_4-a_5	a_5- a_6	a_6-a_7	a_7-a_8	A_8- a_9	a_9-a_{10}
DMQ	0.0015	**0.0013**	0.0015	0.0045	0.0021	0.0041	0.0015	**0.0011**	0.0015

Os intervalos a_2 e a_3 seriam unidos em um único intervalo, o mesmo acontecendo para os intervalos a_8 e a_9. Após essa união e uma nova aprendizagem teríamos as probabilidades condicionais da tabela 3.11.

Tabela 3.11: Probabilidades condicionais de *B* em relação a *A* após a primeira unificação dos intervalos de *A*.

	a_1	a_{23}	a_4	a_5	a_6	a_7	a_{89}	a_{10}
b_1	0.10	0.06	0.06	0.13	0.06	0.06	0.100	0.18
b_2	0.02	0.06	0.06	0.06	0.13	0.06	0.100	0.10
b_3	0.02	0.06	0.06	0.08	0.06	0.06	0.100	0.10
b_4	0.02	0.06	0.14	0.05	0.07	0.14	0.100	0.10
b_5	0.15	0.20	0.25	0.15	0.15	0.25	0.190	0.13
b_6	0.25	0.20	0.15	0.25	0.25	0.15	0.190	0.23
b_7	0.13	0.09	0.05	0.06	0.05	0.13	0.055	0.02
b_8	0.13	0.09	0.05	0.12	0.05	0.05	0.055	0.02
b_9	0.13	0.09	0.05	0.05	0.12	0.05	0.055	0.02
b_{10}	0.05	0.09	0.13	0.05	0.06	0.05	0.055	0.10

As tabelas de 3.12 a 3.19 mostram as DMQs e as probabilidades condicionais de cada passo da unificação dos intervalos de *A*.

Tabela 3.12: DMQ das probabilidades condicionadas pelos intervalos de *A* após a primeira unificação dos intervalos de *A*.

	a_1-a_{23}	a_{23}-a_4	a_4-a_5	a_5- a_6	a_6-a_7	a_7-a_{89}	a_{89}-a_{10}
DMQ	0.0018	**0.0018**	0.0045	0.0021	0.0041	0.0017	**0.0017**

Tabela 3.13: Probabilidades condicionais de *B* em relação a *A* após a segunda unificação dos intervalos de *A*.

	a_1	a_{234}	a_5	a_6	a_7	a_{8910}
b_1	0.10	0.060	0.13	0.06	0.06	0.127
b_2	0.02	0.060	0.06	0.13	0.06	0.100
b_3	0.02	0.060	0.08	0.06	0.06	0.100
b_4	0.02	0.087	0.05	0.07	0.14	0.100
b_5	0.15	0.216	0.15	0.15	0.25	0.170
b_6	0.25	0.183	0.25	0.25	0.15	0.204
b_7	0.13	0.077	0.06	0.05	0.13	0.043
b_8	0.13	0.077	0.12	0.05	0.05	0.043
b_9	0.13	0.077	0.05	0.12	0.05	0.043
b_{10}	0.05	0.103	0.05	0.06	0.05	0.070

Tabela 3.14: DMQ das probabilidades condicionadas pelos intervalos de *A* após a segunda unificação dos intervalos de *A*.

	a_1-a_{234}	a_{234}-a_5	a_5- a_6	a_6-a_7	a_7-a_{8910}
DMQ	0.00295	0.00212	**0.00206**	0.00412	0.00265

Tabela 3.15: Probabilidades condicionais de B em relação a A após a terceira unificação dos intervalos de A.

	a_1	a_{234}	a_{56}	a_7	a_{8910}
b_1	0.10	0.060	0.095	0.06	0.127
b_2	0.02	0.060	0.095	0.06	0.100
b_3	0.02	0.060	0.070	0.06	0.100
b_4	0.02	0.087	0.06	0.14	0.100
b_5	0.15	0.216	0.150	0.25	0.170
b_6	0.25	0.183	0.250	0.15	0.204
b_7	0.13	0.077	0.055	0.13	0.043
b_8	0.13	0.077	0.085	0.05	0.043
b_9	0.13	0.077	0.085	0.05	0.043
b_{10}	0.05	0.103	0.055	0.05	0.070

Tabela 3.16: DMQ das probabilidades condicionadas pelos intervalos de A após a terceira unificação dos intervalos de A.

	a_1-a_{234}	a_{234}-a_{56}	a_{56}-a_7	a_7-a_{8910}
DMQ	0.00295	**0.00151**	0.00371	0.00265

Tabela 3.17: Probabilidades condicionais de B em relação a A após a quarta unificação dos intervalos de A.

	a_1	a_{23456}	a_7	a_{8910}
b_1	0.10	0.074	0.06	0.127
b_2	0.02	0.074	0.06	0.100
b_3	0.02	0.064	0.06	0.100
b_4	0.02	0.076	0.14	0.100
b_5	0.15	0.190	0.25	0.170
b_6	0.25	0.210	0.15	0.204
B_7	0.13	0.068	0.13	0.043
b_8	0.13	0.080	0.05	0.043
b_9	0.13	0.080	0.05	0.043
b_{10}	0.05	0.084	0.05	0.070

Tabela 3.18: DMQ das probabilidades condicionadas pelos intervalos de A após a quarta unificação dos intervalos de A.

	a_1-a_{23456}	a_{23456}-a_7	a_7-a_{8910}
DMQ	0.00219	**0.00185**	0.00265

Tabela 3.19: Probabilidades condicionais de *B* em relação a *A* após a quinta unificação dos intervalos de *A*.

	a_1	a_{234567}	a_{8910}
b_1	0.10	0.072	0.127
b_2	0.02	0.072	0.100
b_3	0.02	0.063	0.100
b_4	0.02	0.087	0.100
b_5	0.15	0.200	0.170
b_6	0.25	0.200	0.204
b_7	0.13	0.078	0.043
b_8	0.13	0.075	0.043
b_9	0.13	0.075	0.043
b_{10}	0.05	0.078	0.070

Analisando os intervalos de *B*, as unificações seriam as mesmas encontradas anteriormente, e após a unificação dos intervalos de *B* e uma nova aprendizagem as probabilidades de *B* dado *A* seriam as da tabela 3.20.

Tabela 3.20: Probabilidades condicionais de *B* em relação a *A* após a unificação dos intervalos de *A* e de *B*.

	a_1	a_{234567}	a_{8910}
b_{1234}	0.16	0.293	0.427
b_{56}	0.40	0.400	0.373
b_{78910}	0.44	0.307	0.200

Comparando os intervalos obtidos para a variável *A* e as tabelas 3.20 e 3.9 notamos que apenas para valores pertencentes ao intervalo a_1 as probabilidades condicionais se mantiveram. A tabela 3.21 mostra a diferença entre as probabilidades condicionadas por cada um dos 10 intervalos originais de *A* para as duas abordagens. A tabela 3.22 mostra a variação percentual de cada probabilidade condicional.

Tabela 3.21: Diferença entre as probabilidades de *B* dado *A* para as 2 abordagens.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
b_{1234}	0	0.133	0.027	0.027	0.027	0.027	0.027	0.107	0.053	0.053
b_{56}	0	0	0	0	0	0	0	0.027	0.013	0.013
b_{78910}	0	0.133	0.027	0.027	0.027	0.027	0.027	0.080	0.040	0.040

Tabela 3.22: Variação percentual entre as probabilidades de *B* dado *A* para as 2 abordagens.

	a_1	a_2	a_3 a a_7	a_8	a_9 e a_{10}
b_{1234}	0%	83%	8.4%	33%	11%
b_{56}	0%	0%	0%	6.7%	3.6%
b_{78910}	0%	30%	9.6%	29%	25%

Em particular a discretização das variáveis fora da ordem inversa de dependência fez com que a probabilidade do valor de B estar no intervalo b_{1234} dado que o valor de A está no intervalo a_2 subisse de 16% para 29.3%. Ao mesmo tempo a probabilidade do valor de B estar no intervalo b_{78910} dado que o valor de A está no intervalo a_2 caiu de 44% para 30.7%.

Em casos mais complexos, com um número grande de variáveis e muitas dependências o efeito da ordem de discretização poderá ser ainda mais pronunciado. Desse modo o desempenho do método proposto, DPT, será melhor ou no mínimo igual ao desempenho de métodos que não se preocupam com a ordem de discretização das variáveis.

No Capítulo 4 será introduzido um sistema que será usado na comparação de resultados entre o método proposto e o método EFD, no Capítulo 5 são apresentados os detalhes de implementação e os procedimentos realizados e os resultados são apresentados no Capítulo 6.

4. Exemplo de Aplicação

4.1. Introdução

Conforme visto no capítulo 2, a grande maioria das comparações entre os métodos de discretização é feita através do uso das BNs como classificadores, mas essa comparação não se mostra adequada para os métodos de discretização que não tem como objetivo melhorar o desempenho em classificação.

Uma vez que simples interpretação dos intervalos gerados e dos relacionamentos encontrados é uma medida subjetiva, é necessária uma outra maneira de comparar o desempenho das redes construídas a partir de diferentes métodos de discretização.

Levando em conta a capacidade de uma BN de poder calcular as probabilidades de qualquer uma das variáveis apresentada uma evidência de todas as outras, podemos abstrair uma BN como sendo um conjunto de classificadores [38]. Usando a BN para classificar várias variáveis ao invés de apenas uma, estaremos explorando melhor o conhecimento adquirido pela aprendizagem Bayesiana e não apenas a efetividade de classificação. Para avaliar o desempenho do método de discretização proposto é necessário então que utilizemos a BN construída para classificar várias variáveis e não apenas uma.

Não há nenhum relato de uso de BNs nesse tipo de tarefa, desse modo fez-se necessária a construção de um novo tipo de aplicação para as BNs, a triagem para validação de dados.

4.2. Validação de Dados

A validação de dados pode ser definida como a inspeção de todos os dados coletados segundo sua integralidade e aceitabilidade, e a eliminação de valores errados [2].

A validação de dados pode ser manual ou automatizada, sendo a segunda preferível em alguns casos devido à maior velocidade e confiabilidade dos computadores em comparação a operadores humanos, entretanto uma revisão manual sempre é recomendável [2]. Desse modo um sistema de validação de dados age, em conjunto com um validador humano, para garantir a validade dos dados [15].

A validação automatizada pode ser dividida em duas etapas: triagem e verificação [2]. A triagem consiste na identificação de dados suspeitos, que podem ser apenas questionáveis ou estarem realmente errados. Essa etapa da validação pode ser totalmente automatizada e resulta na seleção ou ordenação dos dados suspeitos.

Como um dado suspeito pode ou não estar errado é necessário que seja feita uma verificação manual de cada um deles. A partir dessa verificação é tomada a decisão de validar o dado, invalidar o dado ou substituí-lo por um dado válido.

Quando se tem um bom conhecimento sobre o problema em questão, principalmente sobre os valores que os dados podem assumir e sobre seu comportamento alguns métodos simples podem ser usados para a triagem dos mesmos, como, por exemplo, o teste de faixa, o teste de relação e o teste de variação [2].

O teste de faixa consiste em identificar como suspeitos os dados cujos valores estejam fora de uma faixa considerada normal, mas para que isso seja feito é necessário que se tenha conhecimento sobre qual a faixa de valores esperada.

O teste de relação identifica como suspeitos os dados cujos valores deveriam se relacionar com outros dados de uma determinada forma que não ocorre. Quando temos conhecimento sobre leis físicas, por exemplo, que relacionam dois dados, é possível validá-los usando essas leis como base para o teste de relação [21].

E o teste de variação identifica como suspeitos os dados cujos valores apresentam uma variação fora do esperado de uma coleta para outra. Se os dados estão se referindo a uma grandeza contínua esse método pode ser usado, mas se as grandezas representadas pelos dados não apresentam uma continuidade qualquer variação pode ocorrer de uma coleta para outra.

Como nem sempre tais conhecimentos estão disponíveis faz-se necessário um método de triagem que independa desses fatores.

4.3. Caso de Estudo

O IPV, localizado no Centro Técnico Aeroespacial em São José dos Campos – SP é responsável por diversas estações coletoras de dados espalhadas pelo território brasileiro; nessas estações, diariamente, são coletados, calculados e armazenados centenas de dados relacionados à proteção ao vôo.

O grande problema com relação a esses dados, porém, não é sua utilização, mas sim sua validação, ou melhor, sua falta de validação. De hora em hora são registrados os valores de 32 variáveis de interesse, entre elas velocidade do vento, direção do vento, visibilidade, pressão atmosférica, temperatura, umidade relativa do ar, altitudes das camadas de nuvens, etc. A coleta, cálculo e registro das informações são feitas manualmente, através da leitura manual de sensores e da avaliação humana das condições atmosféricas. Os dados coletados são

inicialmente anotados em uma ficha e posteriormente inseridos em um banco de dados, com a identificação da estação coletora, do ano, mês, dia e hora da realização da coleta de dados.

Por ano em cada estação são realizadas, portanto, em média 8.766 coletas dos valores de 32 variáveis de interesse, totalizando em média 280.512 dados coletados por ano em cada estação.

O IPV já utiliza uma grande quantidade de pessoal na manutenção das estações coletoras, na coleta dos dados e em sua inserção em banco de dados, não podendo dispor de esforço humano suficiente para uma validação exaustiva dos dados, ainda mais se for considerada a utilização de métodos de validação redundante onde mais de uma pessoa realiza a verificação dos mesmos dados [42], [64].

Desse modo é necessário um método de validação de dados que demande um esforço bem menor que a verificação de todos os dados inseridos no banco.

Um exemplo do banco de dados do IPV é mostrado nas Figuras 4.1, 4.2 e 4.3 e os significados dos campos são apresentados na Tabela 4.1.

DIRVENTO	VELVENTO	RAJADA	TOTNUVEM	VISIB	QFE	QFF	QNH	BSECO	BUMIDO
10	4	0	5	1000	1011,5	1011,90002441406	1011,90002441406	23,7000007629395	22
0	0	0	5	1000	1011,70001220703	1012,09997558594	1012,09997558594	24	22,8999996185303
11	4	0	8	400	1006,59997558594	1007	1007	25	23,6000003814697
18	4	0	8	200	1006,20001220703	1006,59997558594	1006,59997558594	25,7999992370605	24,2000007629395
30	6	0	6	900	1006,40002441406	1006,79998779297	1006,79998779297	25	23,7999992370605
12	6	0	7	1500	1006,40002441406	1006,79998779297	1006,79998779297	25,6000003814697	24,5
30	5	0	7	1500	1007,79998779297	1008,20001220703	1008,20001220703	25,6000003814697	24,2999992370605
29	6	0	7	1500	1008,79998779297	1009,20001220703	1009,20001220703	24,6000003814697	23
29	10	0	7	2000	1010,09997558594	1010,5	1010,20001220703	25,6000003814697	23,2000007629395
6	6	0	7	1200	1005,5	1005,90002441406	1005,90002441406	31,2000007629395	24,1000003814697
27	10	0	7	1200	1006,70001220703	1007,09997558594	1007,09997558594	26,2000007629395	22,2000007629395
12	5	0	8	200	1004	1004,40002441406	1004,40002441406	27,3999996185303	24
8	3	0	8	1500	1013,5	1013,90002441406	1013,90002441406	25,3999996185303	22,7999992370605
7	6	0	8	1000	1011,09997558594	1011,5	1011,09997558594	23,3999996185303	22,5
27	7	0	7	400	1006,59997558594	1007	1007	25,8999996185303	23,7999992370605
30	4	0	8	800	1013	1013,40002441406	1013,40002441406	25,7999992370605	23,3999996185303
20	8	0	7	1500	1011,5	1011,90002441406	1011,90002441406	27,3999996185303	22,3999996185303
24	10	0	7	600	1009,09997558594	1009,5	1009,5	25,2000007629395	23
9	4	0	7	800	1006,79998779297	1007,20001220703	1007,20001220703	23,3999996185303	22,2000007629395
0	0	0	8	1200	1014,79998779297	1015,20001220703	1015,20001220703	20,7999992370605	20,3999996185303
14	6	0	7	1500	1018,40002441406	1018,79998779297	1018,79998779297	23,2000007629395	21,5
11	8	0	7	1000	1008,70001220703	1009,09997558594	1009,09997558594	23,6000003814697	21,8999996185303
27	3	0	7	2000	1007,79998779297	1008,20001220703	1008,20001220703	29,2000007629395	23,6000003814697
7	3	0	7	2000	1006,59997558594	1007	1007	29,2000007629395	24,2000007629395
33	9	0	7	2000	1004,40002441406	1004,79998779297	1004,79998779297	27	22,6000003814697
3	8	0	7	1200	1012,40002441406	1012,79998779297	1012,79998779297	25,2999992370605	21,5
0	0	0	8	1700	1014,70001220703	1015,09997558594	1015,09997558594	26	20,7999992370605

Figura 4.1: Base de Dados do IPV, variáveis de 1 a 10.

PO	UR	DIFPRESSAO	T	PRECIP	CGT	QTDNUVEM1	TIPONUVEM1	DIRNUVEM1	ALTNUVEM1	QTDNUVEM2	TIPONUVEM2
21	87	0,5	3	0	4	2	6	3	90	1	8
22	91	0,300000011920929	5	0	4	3	6	4	50	1	8
23	89	0	0	1,60000002384186	8	3	6	6	40	6	6
24	88	0,400000005960464	5	1	8	3	7	6	27	6	6
23	91	0,800000011920929	0	0	4	2	7	5	27	2	6
24	91	1	1	0	4	1	6	3	27	3	6
24	90	1,89999997615814	1	0,200000002980232	8	2	6	6	90	2	3
23	88	1,39999997615814	3	0	4	2	6	6	60	3	6
22	62	1,29999995231628	1	0	0	5	6	6	70	3	6
21	55	1,60000002384186	8	0	2	3	8	6	90	3	3
21	71	0,5	5	0,100000001490116	6	2	6	7	120	3	3
23	75	0,100000001490116	5	2,59999990463257	8	1	6	5	30	6	6
22	80	1,60000002384186	3	0	6	1	6	7	120	4	3
22	93	2,09999990463257	7	0,400000005960464	6	2	7	1	30	5	6
23	84	2,5	3	0,400000005960464	8	2	7	5	18	5	6
23	82	0,800000011920929	3	0	6	3	6	5	60	2	8
20	65	0,300000011920929	5	0	2	2	8	5	75	3	6
22	83	0,100000001490116	8	0,800000011920929	8	4	7	7	21	4	6
22	90	0,200000002980232	3	0	5	4	6	7	90	1	9
20	97	1,60000002384186	1	0	4	1	7	4	27	4	6
21	86	2,5	1	0	4	2	6	3	45	6	6
21	87	1	8	0	8	3	6	7	60	1	9
21	62	1,5	8	0	0	1	8	7	75	1	6
22	66	1,20000004768372	6	0	0	1	8	7	75	4	6
21	69	0,400000005960464	0	0	0	1	6	6	120	5	3
20	72	0,699999988079071	0	0	8	2	8	6	75	5	6
18	63	1,70000004768372	8	0	2	1	8	6	135	2	3

Figura 4.2: Base de Dados do IPV, variáveis de 11 a 22.

DIRNUVEM2	ALTNUVEM2	QTDNUVEM3	TIPONUVEM3	DIRNUVEM3	ALTNUVEM3	QTDNUVEM4	TIPONUVEM4	DIRNUVEM4	ALTNUVEM4
3	100	4	3	6	270	1	1	6	750
4	60	2	3	6	240	1	1	6	750
6	60	1	9	6	120	7	4	6	240
6	60	1	9	6	120	7	4	6	270
6	90	5	3	6	270	3	1	6	750
6	60	7	3	6	240	3	1	6	750
6	210	5	4	6	270	6	1	6	750
6	120	3	3	6	240	6	4	6	300
6	120	6	3	6	270	3	1	6	750
6	270	5	4	6	360	3	1	6	750
7	240	6	4	6	400	1	1	6	600
5	120	1	9	5	150	8	4	5	300
7	300	6	4	7	400	8	1	7	750
1	60	4	3	8	240	8	4	8	300
5	60	1	9	6	120	6	3	7	270
5	75	4	3	7	210	8	4	7	270
5	120	6	3	7	240	2	1	7	600
7	45	1	9	7	120	7	3	7	240
7	150	6	3	7	240	7	2	7	750
5	90	7	3	6	240	8	2	7	750
4	90	3	3	4	210	1	1	6	750
7	90	6	3	7	270	7	1	7	750
7	90	5	3	7	240	6	1	8	750
7	90	6	3	7	240	5	1	7	750
7	300	5	4	7	360	7	1	7	750
6	100	6	3	6	270	5	1	6	750
6	300	6	4	6	450	8	2	6	750

Figura 4.3: Base de Dados do IPV, variáveis de 23 a 32.

Tabela 4.1: Significados dos campos da Base de Dados do IPV.

Campo	Significado
DIRVENTO	Direção do vento
VELVENTO	Velocidade do vento
RAJADA	Velocidade de rajadas de vento
TOTNUVEM	Cobertura de Nuvens no céu
VISIB	Visibilidade
QFE	Pressão atmosférica
QFF	Pressão ao nível do mar
QNH	Pressão para ajuste em vôo
BSECO	Temperatura medida em termômetro de bulbo seco
BUMIDO	Temperatura medida em termômetro de bulbo úmido
PO	Temperatura do ponto de orvalho
UR	Umidade relativa do ar
DIFPRESSAO	Variação de pressão atmosférica
T	Tendência do comportamento da pressão atmosférica
PRECIP	Quantidade de chuva
CGT	Condições Gerais de Tempo
QTDNUVEM1	Cobertura de nuvens na 1ª. camada de nuvens
TIPONUVEM1	Tipo de nuvens da 1ª. camada
DIRNUVEM1	Direção das nuvens da 1ª. camada
ALTNUVEM1	Altitude da 1ª. camada de nuvens
QTDNUVEM2	Cobertura de nuvens na 2ª. camada de nuvens
TIPONUVEM2	Tipo de nuvens da 2ª. camada
DIRNUVEM2	Direção das nuvens da 2ª. camada
ALTNUVEM2	Altitude da 2ª. camada de nuvens
QTDNUVEM3	Cobertura de nuvens na 2ª. camada de nuvens
TIPONUVEM3	Tipo de nuvens da 2ª. camada
DIRNUVEM3	Direção das nuvens da 2ª. camada
ALTNUVEM3	Altitude da 2ª. camada de nuvens
QTDNUVEM4	Cobertura de nuvens na 2ª. camada de nuvens
TIPONUVEM4	Tipo de nuvens da 2ª. camada
DIRNUVEM4	Direção das nuvens da 2ª. camada
ALTNUVEM4	Altitude da 2ª. camada de nuvens

4.4. Proposta de Solução

A proposta de solução para esse problema é a construção de uma rede Bayesiana a partir dos dados não validados e sua utilização na triagem dos mesmos.

Estaremos fazendo a triagem dos dados usando um teste de relação, mas ao invés de um especialista no domínio do problema criar as regras de relacionamento elas estarão sendo aprendidas pelo algoritmo de aprendizagem Bayesiana. Caso o método de discretização preserve os relacionamentos entre os dados teremos um desempenho melhor do que se o

método de discretização os destruísse. Desse modo, de acordo com a efetividade da triagem poderemos avaliar o desempenho do método de discretização.

Apresentando os valores de todas os variáveis menos uma como evidência, teremos as probabilidades da variável não apresentada assumir cada um de seus possíveis valores. Tomando a probabilidade dela assumir o seu valor apresentado teremos a probabilidade desse valor estar em concordância com a evidência, ou em outras palavras a probabilidade desse valor ser válido. Quanto menor o valor dessa probabilidade mais suspeito é o valor apresentado. Esse processo é feito para cada uma das variáveis do registro, podemos imaginar que estamos classificando cada um dos valores do registro como verdadeiro ou falso. Podemos tomar então a menor dessas probabilidades como a probabilidade do registro como um todo estar válido. Ordenando os registros em ordem crescente de probabilidade estaremos ordenando os registros do mais suspeito para o menos suspeito. Os registros suspeitos podem então passar por uma validação manual.

Se existirem dados errados, mas com uma grande probabilidade de ocorrência, que não serão classificados como suspeitos e, portanto, não serão verificados manualmente e nem corrigidos, os mesmos estarão seguindo as relações aprendidas pela rede Bayesiana e mesmo que fossem verificados manualmente dificilmente seriam invalidados por apresentarem valores coerentes com os valores que seriam esperados.

Podemos ainda assumir que como esses dados se relacionam com os demais obedecendo as relações aprendidas pela rede Bayesiana, ou seja, que como esses dados tem grande probabilidade de terem realmente ocorrido, os mesmos não influenciarão muito nos resultados de uma análise que os considere válidos.

Como a triagem dos dados é feita tomando cada variável como variável de classe, podemos usar a efetividade na triagem como medida de desempenho do aprendizado usando diferentes métodos de discretização.

No Capítulo 5 são descritos os procedimentos utilizados na implementação da aprendizagem e inferência Bayesiana para triagem dos dados de Proteção ao Vôo fornecidos pelo IPV e os procedimentos usados para a obtenção dos dados para comparação, que são apresentados e analisados no Capítulo 6.

5. Implementação e Procedimentos

O IPV forneceu os dados coletados nas estações de Porto Alegre e Manaus durante um ano para serem usados nesse trabalho. Os mesmos foram entregues em arquivos texto e antes de serem utilizados foi construído um programa em Borland Delphi para inserí-los em um banco de dados Interbase já preparado para essa finalidade. Tal operação foi necessária para facilitar a manipulação dos dados, principalmente no que se refere à discretização e à comparação dos resultados. Além das 32 variáveis medidas, foram consideradas mais duas variáveis para o domínio do problema, o dia do ano em que foi realizada a coleta e a hora do dia em que foi realizada a coleta, assim nosso domínio do problema contém 34 variáveis.

Para a aprendizagem Bayesiana foi escolhido o algoritmo K2 [13], que é o algoritmo mais representativo entre os baseados em busca heurística e se tornou bastante popular devido aos resultados obtidos quando aplicado ao conjunto de dados da rede ALARM, um "*benchmark*" amplamente aceito para os algoritmos de aprendizagem Bayesiana [31].

O algoritmo K2 avalia as possíveis topologias de uma rede bayesiana calculando a probabilidade dessa topologia gerar a base de dados em questão. O algoritmo começa assumindo que um nó não tem antecessores e incrementa o número de antecessores adicionando o antecessor que resulta no maior aumento de probabilidade da estrutura gerar a base de dados. Quando a adição de mais um antecessor ao nó não aumenta mais a probabilidade, o nó para de receber antecessores e o algoritmo faz o mesmo para o nó seguinte.

O fato de já existir uma implementação do mesmo em Matlab no Bayes Net Toolbox para Matlab [43] também contribuiu muito para sua escolha. Graças à versatilidade do Matlab foi bastante simples gravar a BN resultante da aprendizagem no Matlab num formato que pudesse ser lido por outras ferramentas. Isso foi necessário porque a inferência Bayesiana no Matlab era muito lenta.

Para o cálculo da DMQ foi criado um "*script*" em Matlab e os valores da DMQ entre cada intervalo foram analisados manualmente.

Para a propagação de crença foi construído um programa em Borland Delphi que utiliza a Interface de Programação de Aplicação (*Application Programming Interface* – API) do MSBNx – *"Microsoft Research's Bayesian network autoring and evaluation tool"* , que é um aplicativo Windows baseado em componentes para criação, edição e uso de BNs. Os componentes do MSBNx podem ser integrados a outros programas, dando a eles o poder da

inferência Bayesiana [30]. O programa construído abria a BN salva pelo Matlab e para cada registro do banco de dados calculava a probabilidade do valor de cada campo ocorrer, dados os valores de todos os outros campos, gravando essas probabilidades em um outro banco de dados.

Antes da aplicação do método de discretização foi realizado um estudo para verificar se os dados coletados em Manaus e Porto Alegre poderiam ser utilizados em conjunto ou se a união dos mesmos provocaria perda de informação, ou seja, se os dados coletados em Manaus e em Porto Alegre apresentavam relações entre as variáveis que não seriam encontrados caso os dados das duas estações fossem considerados em conjunto.

Para tanto os dados de Porto Alegre e Manaus foram discretizados em 10 intervalos pelo uso de EFD individualmente e considerando seus dados como um único conjunto de informações.

Foi realizada a aprendizagem Bayesiana com os dados de cada estação individualmente e com os dados das duas estações em conjunto. As BNs encontradas apresentaram diferenças nos DAGs, demonstrando que a união dos dados fez com que relacionamentos entre as variáveis que ocorrem localmente nas regiões de Porto Alegre e Manaus fossem perdidos. Os DAGs resultantes são apresentados nas figuras 5.1 a 5.3.

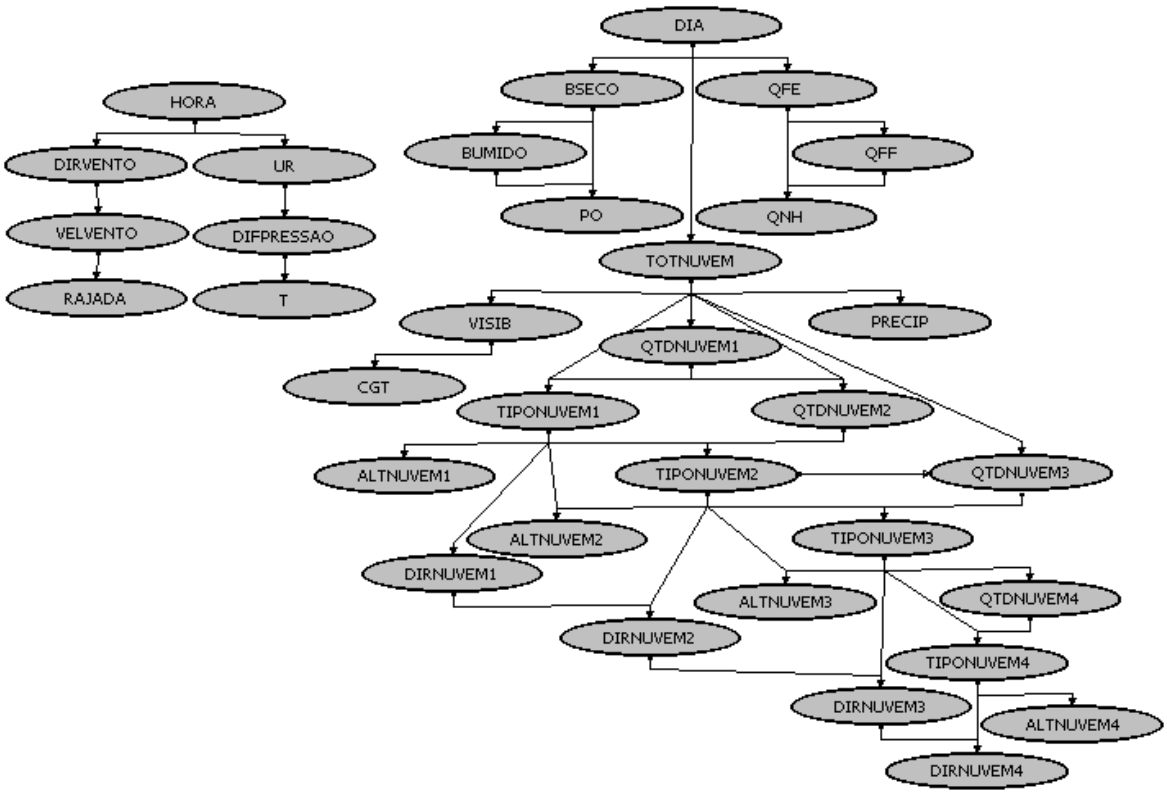

Figura 5.1: DAG da BN aprendida com os dados de Porto Alegre.

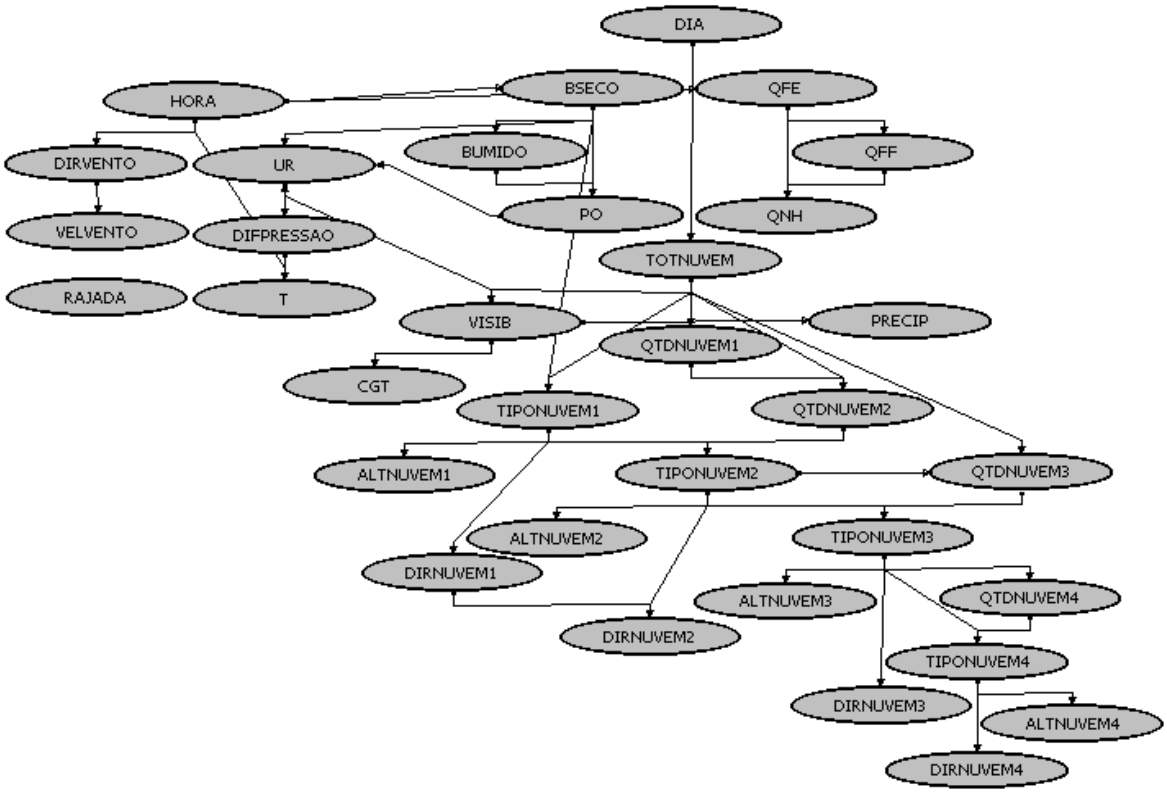

Figura 5.2: DAG da BN aprendida com os dados de Manaus.

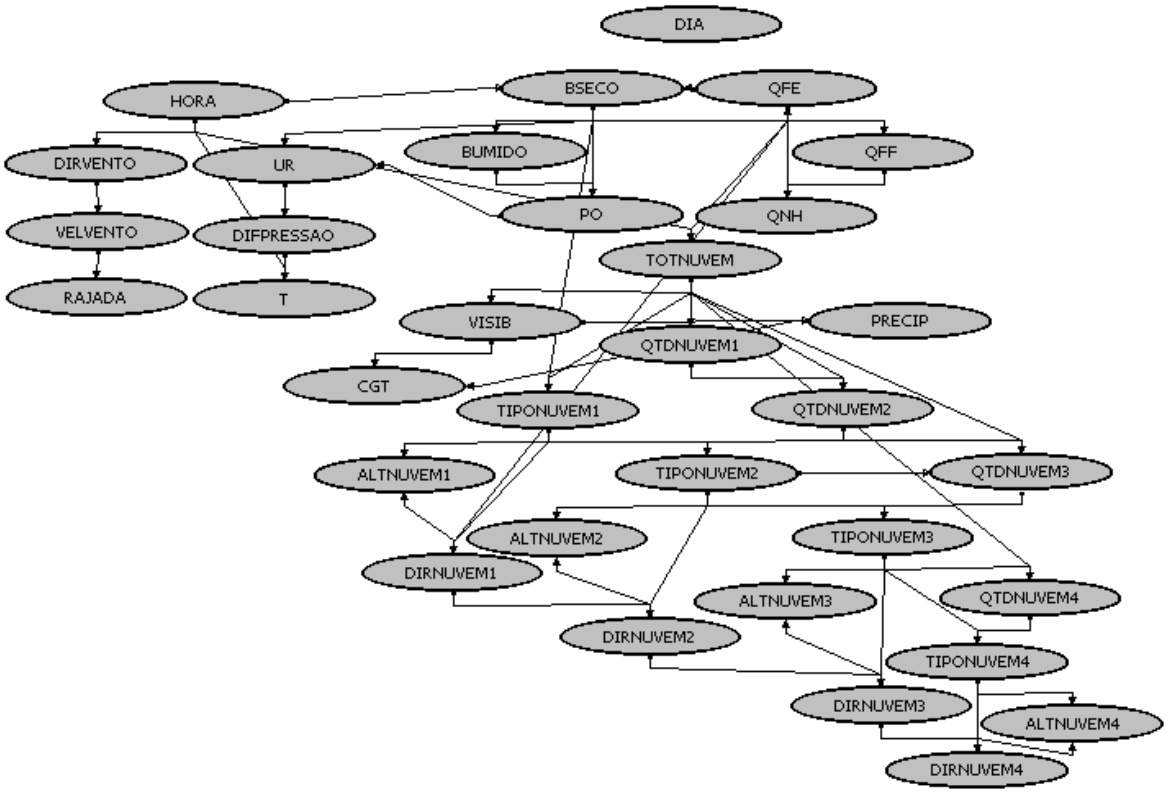

Figura 5.3: DAG da BN aprendida com os dados de Porto Alegre e Manaus em conjunto.

A BN construída a partir dos dados de Porto Alegre, Figura 5.1, apresenta relacionamentos entre a variável *DIA* e as variáveis *BSECO*, *QFE* e *TOTNUVEM*. Já a BN

construída a partir dos dados de Manaus, Figura 5.2, apresenta relacionamentos entre a variável *DIA* e apenas a variável *TOTNUVEM*. Enquanto a variável *HORA* se relaciona com as variáveis *BSECO* e *QFE*. Isso sugere que em Porto Alegre as variáveis *BSECO* (temperatura) e *QFE* (pressão) variam com a variação do *DIA* do ano, enquanto que em Manaus essas variáveis estão relacionadas com a *HORA* do dia. Essa diferença de relacionamentos é compreensível, uma vez que em Porto Alegre as variações climáticas com as estações do ano são mais pronunciadas do que em Manaus, onde as variações com a hora do dia são mais perceptíveis.

Comparando ainda os dois DAGs anteriores com o DAG da BN construída a partir dos dados conjuntos das duas estações, Figura 5.3, notamos que a variável *DIA* não se relaciona com nenhuma outra variável nesse último, nem mesmo com *TOTNUVEM*, com a qual a variável *DIA* estava relacionada nas duas BNs anteriores. Isso indica que o relacionamento existente entre essas duas variáveis em Porto Alegre é diferente do relacionamento entre elas em Manaus e que se os dados forem considerados em conjunto a informação sobre esse relacionamento provavelmente será perdida.

Essas e outras diferenças entre os DAGs das BNs indicam que os dados não devem ser considerados em conjunto e que para cada estação deve ser gerada e utilizada uma BN diferente que irá aprender os relacionamentos entre as variáveis no local de coleta das mesmas. Por esse motivo, os dados de apenas uma das estações, Porto Alegre, foram utilizados no restante deste trabalho.

As variáveis foram então inicialmente discretizadas em 10 intervalos pelo uso de EFD. Uma quantidade muito grande de intervalos para a discretização inicial tornaria o processo muito lento e trabalhoso, uma vez que tanto a discretização inicial quanto a unificação dos intervalos não seriam totalmente automatizada. Por outro lado um número muito pequeno de intervalos poderia não ser suficiente para avaliar o método proposto, uma vez que com poucos intervalos iniciais ocorreriam poucas unificações.

O objetivo era transformar os 10 intervalos em no máximo 4 intervalos, preservando os padrões de relacionamento existentes entre as variáveis. Um número reduzido de intervalos favorece tanto a aprendizagem quanto a inferência Bayesiana e está mais próximo da compreensão humana.

Os intervalos das variáveis folha: *RAJADA*, *T*, *CGT*, *PO*, *PRECIP*, *DIRNUVEM4*, *QNH*, *ALTNUVEM1*, *ALTNUVEM2*, *ALTNUVEM3* e *ALTNUVEM4* foram unidos enquanto os relacionamentos eram preservados. Vale ressaltar que os significados das variáveis não estão sendo levados em conta uma vez que a aprendizagem Bayesiana está sendo usada justamente

para aprender os relacionamentos entre elas. Numa aplicação comercial, no entanto, esse significado poderia ser utilizado para direcionar a aprendizagem Bayesiana.

A unificação de intervalos pode aumentar a relação da variável discretizada com outras variáveis antes não relacionadas à variável discretizada, sem diminuir sua relação com os antigos antecessores, desse modo ao longo do processo de unificação novos relacionamentos podem surgir e alguns antecessores podem mudar, mas uma variável não pode perder todos seus antecessores, o que caracterizaria a perda de todas as relações das outras variáveis com a variável discretizada e não o aumento de uma ou mais relações.

Como o algoritmo de busca parte de um DAG sem arcos e vai adicionado os que mais contribuem com a melhoria da rede, um evidenciamento de um outro relacionamento devido aos novos intervalos pode fazer com esse novo relacionamento contribua mais para a melhoria da rede do que o antigo, e uma vez adicionado esse relacionamento a adição do relacionamento antigo pode não melhorar a rede consideravelmente. O mesmo pode acontecer com uma variável que inicialmente tinha vários sucessores e depois os perdeu. Os novos intervalos podem ter evidenciado muito o relacionamento dos sucessores com apenas um ou um subconjunto dos antecessores, desse modo os relacionamentos com os demais antecessores não contribuem mais significativamente para a melhoria da rede.

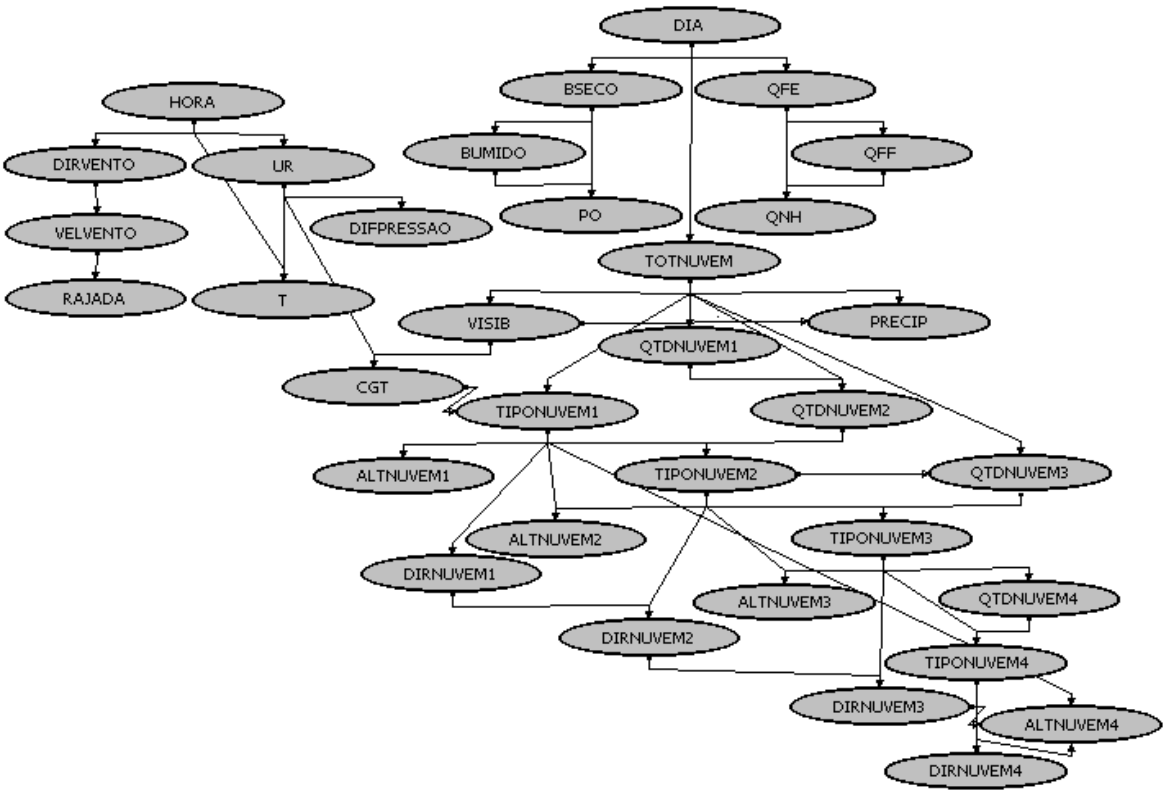

Figura 5.4: DAG da BN aprendida após a discretização dos nós folha.

No DAG da Figura 5.4 verificamos que a variável *DIFPRESSAO* passou a ser folha, uma vez que o relacionamento que existia entre a variável *DIFPRESSAO* e a variável *T* foi substituído por relações entre as variáveis *HORA* e *UR* e a variável *T*. Mais uma vez, como já exposto acima, isso ocorre porque com os novos intervalos esse padrão de relacionamento foi evidenciado e não porque o padrão de relacionamento entre *DIFPRESSAO* e *T* foi perdido.

A nova variável folha *DIFPRESSAO* e as variáveis cujos sucessores já foram processados: *VELVENTO*, *DIFPRESSAO*, *QFF*, *BUMIDO*, *TIPONUVEM4*, *VISIB* e *DIRNUVEM3* são então processadas e a discretização continua até que todos as variáveis sejam processadas. Como critério de unificação de intervalos foi utilizada uma DMQ absoluta de 0.04 para mais de 4 intervalos e uma DMQ absoluta de 0.01 para 4 intervalos ou menos. O DAG da BN aprendida ao final do processo é apresentado na Figura 5.5.

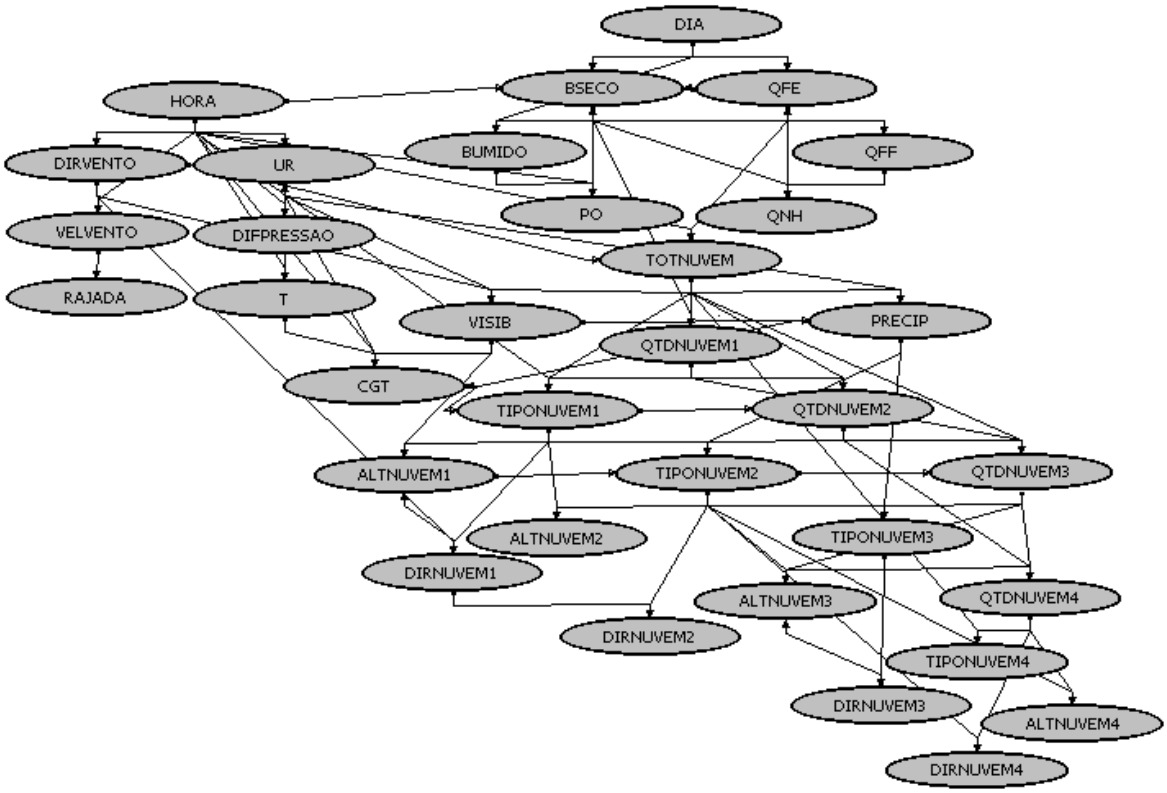

Figura 5.5: DAG da BN aprendida após a discretização de todas as variáveis.

A cada passo do método de discretização os relacionamentos já existentes foram mantidos ou substituídos por relacionamentos mais fortes e novos relacionamentos surgiram. Devido ao surgimento desses novos relacionamentos o DAG da aprendizagem final apresenta muito mais arcos do que o DAG inicialmente obtido.

São apresentadas na tabela 5.1 as DMQs calculadas para a unificação dos intervalos da variável *QTDNUVEM4*, que tinha como sucessor apenas a variável *TIPONUVEM4*.

Tabela 5.1: DMQs das probabilidades da variável *TIPONUVEM4,* condicionadas pelos intervalos da variável *QTDNUVEM4*.

	i_1-i_2	i_2-i_3	i_3- i_4	i_4-i_5	i_5-i_6	i_6-i_7	i_7-i_8	i_8-i_9
DMQ	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.0

Como base nos valores de DMQ os intervalos de i_1 a i_8 foram unidos em um único intervalo e o intervalo i_9 ficou isolado.

Outro exemplo é o da variável *HORA*, onde o último intervalo poderia ser unido ao primeiro, uma vez que as horas do dia formam um ciclo. A variável *DIA* tinha dois sucessores, *DIRVENTO* e *UR*. Os valores das DMQs das probabilidades dessas duas variáveis condicionadas pela variável *HORA* são apresentadas na Tabela 5.2.

Tabela 5.2: DMQs das probabilidades condicionadas pelos intervalos da variável *HORA*.

	i_1-i_2	i_2-i_3	i_3- i_4	i_4-i_5	i_5-i_6	i_6-i_7	i_7-i_8	i_8-i_9	i_9-i_{10}	i_{10}-i_1
DIRVE.	0.0046	0.0028	0.0015	0.0080	0.0046	0.0075	0.0082	0.0025	0.0041	0.0098
UR	0.0075	0.0078	0.0740	0.0256	0.0170	0.0087	0.0087	0.0633	0.0292	0.0121

Combinando os valores de DMQ da Tabela 5.2, utilizando a norma euclidiana, temos as DMQs combinadas dos intervalos da variável *HORA*, que são apresentadas na tabela 5.3.

Tabela 5.3: DMQs combinadas das probabilidades condicionadas pelos intervalos da variável *HORA*.

	i_1-i_2	i_2-i_3	i_3- i_4	i_4-i_5	I_5-i_6	i_6-i_7	i_7-i_8	i_8-i_9	I_9-i_{10}	i_{10}-i_1
DMQ	0.0088	**0.0083**	0.0740	0.0268	0.0176	**0.0115**	0.0120	0.0633	0.0295	0.0156

Pelos valores de DMQ da Tabela 5.3, os intervalos i_2 e i_3 foram unidos em um único intervalo, o mesmo acontecendo para os intervalos i_6 e i_7. Após um novo aprendizado as DMQs foram calculadas novamente e as DMQs combinadas são apresentadas na Tabela 5.4.

Tabela 5.4: DMQs combinadas das probabilidades condicionadas pelos intervalos da variável *HORA* após a primeira unificação.

	i_1-i_{23}	i_{23}- i_4	i_4-i_5	i_5-i_{67}	i_{67}-i_8	i_8-i_9	i_9-i_{10}	i_{10}-i_1
DMQ	**0.0090**	0.0488	0.0268	**0.0136**	0.0144	0.0633	0.0295	0.0156

Os intervalo i_1 foi então unido ao intervalo i_{23} e o intervalo i_5 foi unido ao intervalo i_{67}. Após um novo aprendizado as DMQs foram calculadas e a DMQ combinada para cada intervalo é apresentada na Tabela 5.5.

Tabela 5.5: DMQs combinadas das probabilidades condicionadas pelos intervalos da variável *HORA* após a segunda unificação.

	i_{123}- i_4	i_4-i_{567}	i_{567}-i_8	i_8-i_9	i_9-i_{10}	i_{10}-i_{123}
DMQ	0.0534	0.0386	**0.0198**	0.0633	0.0295	**0.0284**

Os intervalo i_{10} foi então unido ao intervalo i_{123} e o intervalo i_8 foi unido ao intervalo i_{567}. Após um novo aprendizado as DMQs foram calculadas e a DMQ combinada para cada intervalo é apresentada na Tabela 5.6.

Tabela 5.6: DMQs combinadas das probabilidades condicionadas pelos intervalos da variável *HORA* após a terceira unificação.

	i_{10123}- i_4	i_4-i_{5678}	i_{5678}-i_9	i_9-i_{101234}
DMQ	0.0446	0.0444	0.0458	0.0534

Como nenhuma das DMQs da Tabela 5.6 atende os critérios de DMQ absoluta e DMQ relativa a unificação para a variável *HORA* está completa.

Um caso mais interessante é o da variável *DIA*, onde o último intervalo também poderia ser unido ao primeiro, uma vez que os dias do ano formam um ciclo. Quando foi processada a variável dia tinha 3 sucessores já processados, *QFE*, *BSECO* e *UR*, sendo que *BSECO* e *UR* além de *DIA* tinham a variável *HORA* como antecessor. A DMQ foi calculada com relação a cada um desses sucessores, sendo que no cálculo da DMQ com relação a *BSECO* e *UR* foi considerada também a variável *HORA*, conforme a equação (3.2). Os valores das DMQs são apresentados na Tabela 5.7.

Tabela 5.7: DMQs das probabilidades condicionadas pelos intervalos da variável *DIA*.

	i_1-i_2	i_2-i_3	i_3- i_4	i_4-i_5	i_5-i_6	i_6-i_7	i_7-i_8	i_8-i_9	i_9-i_{10}	i_{10}-i_1
QFE	0.0032	0.0164	0.0038	0.0348	0.0036	0.0007	0.0036	0.0199	0.0308	0.0019
BSECO	0.0024	0.0312	0.0362	0.0188	0.0033	0.0045	0.0309	0.0389	0.0348	0.0119
UR	0.0011	0.0008	0.0001	0.0005	0.0001	0.0009	0.0004	0.0008	0.0005	0.0015

Combinando os valores de DMQ da Tabela 5.7, utilizando a norma euclidiana, temos as DMQs combinadas dos intervalos da variável *DIA*, que são apresentadas na tabela 5.8.

Tabela 5.8: DMQs combinadas das probabilidades condicionadas pelos intervalos da variável *DIA*.

	i_1-i_2	i_2-i_3	i_3- i_4	i_4-i_5	i_5-i_6	i_6-i_7	i_7-i_8	i_8-i_9	i_9-i_{10}	i_{10}-i_1
DMQ	**0.0042**	0.0353	0.0364	0.0395	0.0048	**0.0047**	0.0311	0.0437	0.0465	0.0122

Pelos valores de DMQ da Tabela 5.8 os intervalos i_1 e i_2 foram unidos em um único intervalo, o mesmo acontecendo para os intervalos i_6 e i_7. As DMQs combinadas, calculadas a cada unificação, são apresentadas nas Tabelas 5.9 a 5.12.

Tabela 5.9: DMQs combinadas das probabilidades condicionadas pelos intervalos da variável *DIA* após a primeira unificação.

	i_{12}-i_3	i_3- i_4	i_4-i_5	i_5-i_{67}	i_{67}-i_8	i_8-i_9	i_9-i_{10}	i_{10}-i_{12}
DMQ	0.0436	0.0364	0.0395	**0.0039**	0.0267	0.0437	0.0465	**0.0101**

Tabela 5.10: DMQs combinadas das probabilidades condicionadas pelos intervalos da variável *DIA* após a segunda unificação.

	i_{1012}-i_3	i_3- i_4	i_4-i_{567}	i_{567}-i_8	i_8-i_9	i_9-i_{1012}
DMQ	0.0401	0.0364	0.0451	**0.0217**	0.0437	0.0580

Tabela 5.11: DMQs combinadas das probabilidades condicionadas pelos intervalos da variável *DIA* após a terceira unificação.

	i_{1012}-i_3	i_3- i_4	i_4-i_{5678}	i_{5678}-i_9	i_9-i_{1012}
DMQ	0.0401	**0.0364**	0.0372	0.0610	0.0580

Tabela 5.12: DMQs combinadas das probabilidades condicionadas pelos intervalos da variável *DIA* após a terceira unificação.

	i_{1012}-i_{34}	i_{34}-i_{5678}	i_{5678}-i_9	i_9-i_{1012}
DMQ	0.0580	0.0515	0.0568	0.0587

Como nenhuma das DMQs da Tabela 5.12 atende os critérios de DMQ absoluta e DMQ relativa a unificação para a variável *DIA* está completa.

Com a BN definida foi criado um outro "*script*" em Matlab para gravar a BN no formato aceito pelo MSBNx. Foi construído também um aplicativo em Borland Delphi que utilizando a API do MSBNx processava cada registro do banco de dados. O aplicativo tomava cada registro de 34 campos do banco e para cada campo do mesmo apresentava os valores dos outros 33 campos à BN via API do MSBNx e recebia a probabilidade do valor do campo estar correto dada a evidência. As probabilidades de cada campo de cada registro foram também armazenadas em um banco de dados.

No banco de dados resultante foi criado um campo extra que recebeu o mínimo entre os valores apresentados por cada campo e os registros foram então ordenados crescentemente por esse valor.

Para verificar a efetividade do método proposto de discretização na conservação dos padrões de relacionamento entre as variáveis, uma cópia dos dados de treinamento foi discretizada pelo uso de EFD e uma outra BN foi construída a partir da aprendizagem desses dados. Para cada variável, a quantidade de intervalos para a EFD foi igual à da discretização pelo método proposto, ou seja, as variáveis das duas BNs tinham a mesma quantidade de intervalos.

A EFD foi escolhida devido à sua simplicidade de implementação e por ter sido usada como discretização inicial no método proposto. A EFD não apresenta os melhores resultados quando comparada aos métodos mais elaborados, mas a diferença dos resultados em sistemas de classificação não é muito significativa, por esse motivo o método EFD pode ser considerado como representativo do desempenho geral dos métodos atuais de discretização. O DAG da BN obtida a partir da EFD é apresentado na Figura 5.6.

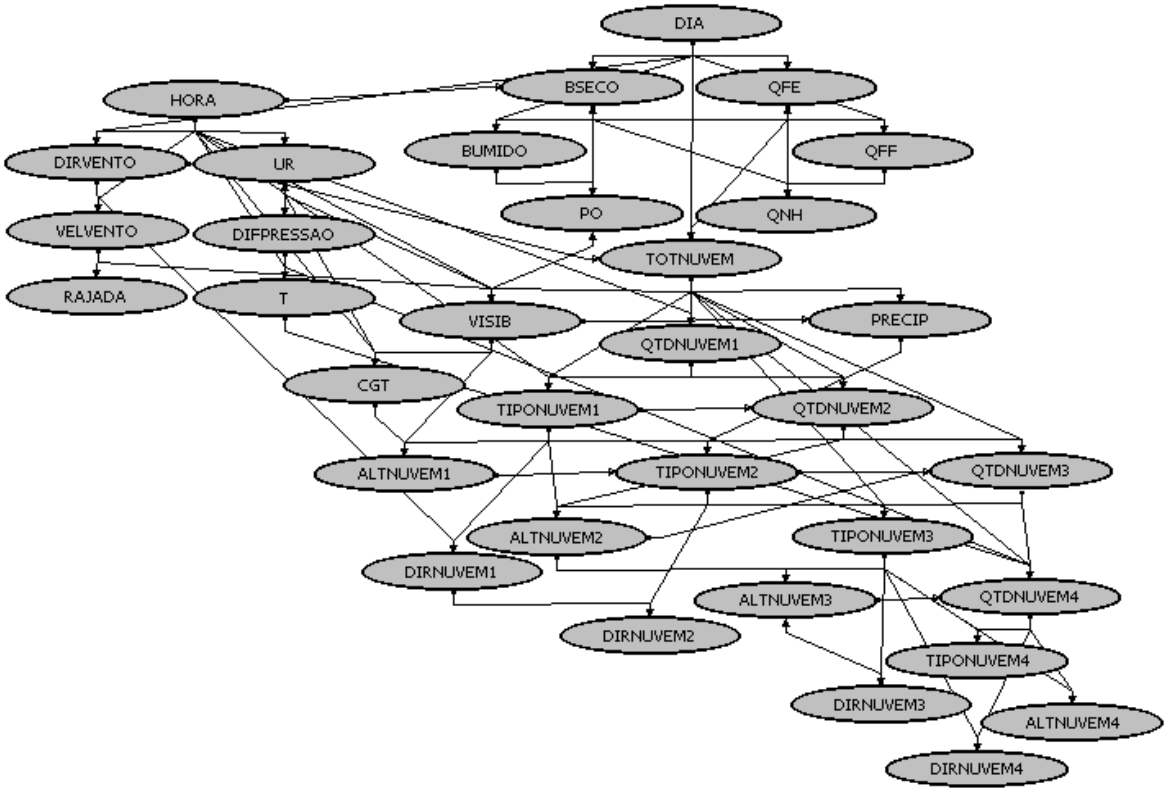

Figura 5.6: DAG da BN aprendida com os dados discretizados por EFD.

Facilmente podemos verificar que o DAG da BN aprendida como o método proposto, Figura 5.5, e o DAG da BN aprendida com EFD apresentam relacionamentos diferentes, mas como a simples análise desses relacionamentos seria uma comparação subjetiva, as redes resultantes dos dois processos de discretização foram comparadas pela sua efetividade em realizar uma correta triagem de dados incorretos, como descrito no próximo capítulo.

6. Resultados Obtidos.

6.1. Método de Discretização

Para comparar os resultados obtidos com a DTP e com a EFD, as BNs geradas pela aprendizagem Bayesiana a partir dessas discretizações foram utilizadas para a triagem de dados incorretos. Conforme citado no Capítulo 4, estaremos usando uma mesma BN para classificar várias variáveis e o desempenho nessa classificação será utilizado como medida da aprendizagem.

Foram escolhidas seis variáveis entre as 34 existentes, *DIRVENTO*, *TOTNUVEM*, *VISIB*, *BSECO*, *BUMIDO* e *PRECIP* para serem utilizadas nessa comparação.

Em cada base discretizada, para cada variável escolhida, foram selecionados aleatoriamente 50 registros. Na seleção dos registros foi considerada a distribuição dos mesmos segundo os possíveis intervalos da variável escolhida.

Por exemplo para a variável *DIRVENTO*, na base discretizada com DTP, foram selecionados 18 registros com o valor de *DIRVENTO* no 1º. intervalo de discretização, 18 registros com o valor de *DIRVENTO* no 2º. intervalo de discretização, 9 registros com o valor de *DIRVENTO* no 3º. intervalo de discretização e 5 registros com o valor de *DIRVENTO* no 4º. intervalo de discretização. A distribuição dos registros segundo os intervalos, que justifica a quantidade de registros escolhidos para cada intervalo é apresentada na tabela 6.1.

Tabela 6.1: Distribuição dos registros da Base discretizada por DTP segundo os intervalos da variável *DIRVENTO*.

Intervalo	1º.	2º.	3º.	4º.
Porcentagem dos registros	36%	37%	17%	10%

Em cada um dos 50 registros selecionados o valor da variável escolhida foi alterado para um dos intervalos adjacentes. Na escolha do intervalo adjacente, para qual o valor da variável escolhida foi alterado, foi considerada também a distribuição dos registros da base segundo os intervalos da variável escolhida.

Usando a mesma variável, *DIRVENTO*, como exemplo, os 18 registros selecionados com o valor da variável *DIRVENTO* no 1º. intervalo tiveram o valor da variável *DIRVENTO* alterado para o 2º. intervalo (único intervalo adjacente). Dos 18 registros selecionados com o valor da variável *DIRVENTO* no 2º. intervalo, 12 tiveram o valor da variável *DIRVENTO* alterado para o 1º. intervalo e os outros 6 registros tiveram o valor da variável *DIRVENTO*

alterado para o 3º. intervalo. Dos 9 registros selecionados com o valor da variável *DIRVENTO* no 3º. intervalo, 7 tiveram o valor da variável *DIRVENTO* alterado para o 2º. intervalo e os outros 2 registros tiveram o valor da variável *DIRVENTO* alterado para o 4º. intervalo. E os 5 registros selecionados com o valor da variável *DIRVENTO* no 4º. intervalo tiveram o valor da variável *DIRVENTO* alterado para o 3º. intervalo.

Desse modo a seleção resultou em mais registros selecionados com o valor da variável escolhida sendo o do intervalo com um maior número de registros do total. E as alterações resultaram em mais registros com o valor da variável escolhida sendo alterado para o valor do intervalo adjacente com um maior número de registros do total.

As BNs foram então utilizadas na triagem desses registros com relação à base original. Para cada uma das variáveis escolhidas para o teste foi verificada qual a porcentagem de registros alterados estava contida entre os 2.5%, 5%, 10%, 15% e 20% registros mais suspeitos. As triagens foram feitas utilizando-se as BNs geradas usando os dois métodos de discretização, o método proposto, Discretização pela Tabela de Probabilidades – DTP e o método de Discretização em Igual Freqüência – EFD. Os resultados são apresentados nas Tabelas 6.2 a 6.7. As diferenças superiores a 10% estão destacadas.

Tabela 6.2: Porcentagem dos registros com a variável *DIRVENTO* alterada selecionados pela triagem.

Dados Suspeitos	2.5%	5%	10%	15%	20%
Triagem DTP	16%	20%	26%	**42%**	42%
Triagem EFD	16%	20%	26%	**28%**	44%

Tabela 6.3: Porcentagem dos registros com a variável *TOTNUVEM* alterada selecionados pela triagem.

Dados Suspeitos	2.5%	5%	10%	15%	20%
Triagem DTP	**92%**	**94%**	**99%**	**100%**	**100%**
Triagem EFD	**64%**	**64%**	**70%**	**72%**	**76%**

Tabela 6.4: Porcentagem dos registros com a variável *VISIB* alterada selecionados pela triagem.

Dados Suspeitos	2.5%	5%	10%	15%	20%
Triagem DTP	36%	40%	44%	44%	54%
Triagem EFD	40%	44%	50%	54%	60%

Tabela 6.5: Porcentagem dos registros com a variável *BSECO* alterada selecionados pela triagem.

Dados Suspeitos	2.5%	5%	10%	15%	20%
Triagem DTP	**46%**	**50%**	**68%**	**70%**	**74%**
Triagem EFD	**20%**	**24%**	**40%**	**40%**	**50%**

Tabela 6.6: Porcentagem dos registros com a variável *BUMIDO* alterada selecionados pela triagem.

Dados Suspeitos	2.5%	5%	10%	15%	20%
Triagem DTP	60%	68%	68%	78%	**88%**
Triagem EFD	54%	60%	70%	74%	**76%**

Tabela 6.7: Porcentagem dos registros com a variável *PRECIP* alterada selecionados pela triagem.

Dados Suspeitos	2.5%	5%	10%	15%	20%
Triagem DTP	**62%**	**64%**	**70%**	**76%**	**84%**
Triagem EFD	**20%**	**26%**	**38%**	**50%**	**58%**

As Tabelas 6.2 a 6.7 mostram que, para 3 das 6 variáveis escolhidas para comparação (*DIRVENTO*, *TOTNUVEM*, *VISIB*, *BSECO*, *BUMIDO* e *PRECIP*), o método proposto resultou em uma BN capaz de selecionar pelo menos 24% a mais de registros errados quando comparada à BN resultante da EFD, sendo que essa diferença chegou a ser de 42%, Tabela 6.6. Já para as outras 3 variáveis os dois métodos de discretização apresentam resultados bastante próximos, a maioria ficando dentro de uma faixa de variação de 10%, mas mesmo assim com uma leve vantagem para o método proposto.

Esses resultados indicam que para as variáveis *DIRVENTO*, *VISIB* e *BUMIDO* os dois métodos de discretização mantiveram, perderam ou evidenciaram na mesma proporção seus relacionamentos com as outras variáveis. Já no caso das variáveis *TOTNUVEM*, *BSECO* e *PRECIP* o método proposto foi capaz de manter ou evidenciar os relacionamentos com as outras variáveis enquanto que a EFD fez com que esses relacionamentos fossem enfraquecidos ou mesmo perdidos.

Em particular para *TOTNUVEM*, um de seus antecessores foi alterado de uma discretização para outra. Na rede gerada com o uso de DTP *TOTNUVEM* tem como antecessores *HORA* e *DIRVENTO*, enquanto que na rede gerada com o uso de EFD *TOTNUVEM* tem como antecessores *DIA* e *DIRVENTO*, ou seja, a EFD fez com que o relacionamento entre *HORA* e *TOTNUVEM* fosse perdido, o que fez com que o relacionamento entre *DIA* e *TOTNUVEM* fosse escolhido pelo algoritmo de aprendizagem como mais representativo. Pelos resultados apresentados na Tabela 6.2 esse relacionamento

entre *HORA* e *TOTNUVEM* que foi perdido era bem mais representativo que o entre *DIA* e *TOTNUVEM*.

Para a variável *BSECO* a análise é mais simples, a *EFD* fez com que os relacionamentos entre *QFE* e *BSECO*, e entre *BSECO* e *QTDNUVEM1* não fossem encontrados. Sem esses relacionamentos a inferência da BN resultante perante a evidência foi menos efetiva, conforme mostra a Tabela 6.4.

Para a variável *PRECIP* a EFD fez com que três relacionamentos não fossem encontrados, entre *UR* e *PRECIP*, entre *PRECIP* e *TIPONUVEM3* e entre *PRECIP* e *CGT*. A falta desses relacionamentos também fez com que a BN fosse menos efetiva na triagem dos registros alterados.

Já para a variável *DIRVENTO* um relacionamento deixou de ser encontrado e um outro surgiu com o uso de EFD. O relacionamento que foi perdido foi entre *DIRVENTO* e *VISIB* e o relacionamento encontrado foi entre *HORA* e *DIRVENTO*. Pelos resultados apresentados na Tabela 6.1 esse novo relacionamento foi capaz de manter a representatividade da BN encontrada.

A variável *VISIB* merece uma análise um pouco mais detalhada, visto que os resultados obtidos a partir da EFD são um pouco melhores que os resultados obtidos a partir da DTP. O fato da variável *VISIB* ter um antecessor a mais no DAG gerado a partir da EFD do que no DAG gerado pela DTP o fato do algoritmo de aprendizagem Bayesiana parar de adicionar antecessores a um nó caso essa adição não melhore significativamente a adequação da BN gerada podem explicar esse melhor desempenho da rede gerada a partir da EFD.

Comparando os DAGs verificamos que da BN gerada a partir da DTP para a BN gerada a partir da EFD a variável *VISIB* deixou de ter *DIRVENTO* como antecessor e passou a ter *HORA* e *VELVENTO* como antecessores. Além disso, a variável *VISIB* deixou de ter *TOTNUVEM1* como sucessor e passou a ter *ALTNUVEM1* e *PO* como sucessores.

Com a DTP o relacionamento entre *DIRVENTO* e *VISIB* contribuía mais para a adequação da BN aos dados que o relacionamento entre *HORA* e *VISIB* e que o relacionamento entre *VELVENTO* e *VISIB*. Desse modo, mesmo que o relacionamento entre o par *HORA-VELVENTO* e a variável *VISIB* fosse mais evidente que entre *DIRVENTO* e *VISIB*, este não seria encontrado, uma vez que o algoritmo primeiro adicionaria o relacionamento entre *DIRVENTO* e *VISIB* à rede. E uma vez este relacionamento adicionado, a adição de um relacionamento entre *HORA* e *VISIB* ou entre *VELVENTO* e *VISIB* não melhoraria muito a adequação da rede.

Com a EFD o relacionamento entre *DIRVENTO* e *VISIB* pode ter sido mais afetado que os outros, de forma que o relacionamento entre *HORA* e *VISIB* ou entre *VELVENTO* e *VISIB* passou a contribuir mais para a adequação da rede do que o relacionamento entre *DIRVENTO* e *VISIB*. Desse modo um desses relacionamentos foi adicionado à rede no lugar do relacionamento entre *DIRVENTO* e *VISIB* e posteriormente o outro foi adicionado por aumentar consideravelmente a adequação da BN.

Caso a EFD tenha afetado pouco o relacionamento entre o par *HORA*-*VELVENTO* e a variável *VISIB*, de modo que esse ainda fosse mais representativo que o relacionamento original entre *DIRVENTO* e *VISIB* a BN gerada a partir da EFD apresentaria um desempenho melhor que a gerada a partir da DTP. Note-se que essa diferença de desempenho em favor da EFD é devido a uma restrição do algoritmo de aprendizagem Bayesiana, que não é capaz de verificar se um condicionante duplo é melhor do que um condicionante simples onde o condicionante simples não faz parte do condicionamento duplo. Caso o algoritmo de aprendizagem fosse capaz de reconhecer essa situação os relacionamentos entre *HORA* e *VISIB* e entre *VELVENTO* e *VISIB* teriam sido encontrados no caso da DTP no lugar do relacionamento entre *DIRVENTO* e *VISIB*.

Já para a variável *BUMIDO* os relacionamentos encontrados foram os mesmos, mas a EFD provavelmente diminuiu um pouco sua representatividade, desse modo os resultados com DTP foram um pouco melhores conforme a Tabela 6.5.

Pelos resultados obtidos verificamos que com exceção da variável *VISIB*, a BN construída a partir do método de discretização proposto apresentou sempre um desempenho superior ao da BN construída a partir da EFD, sendo alguma vezes consideravelmente superior. O melhor desempenho da segunda no caso específico da variável *VISIB* pode ainda ser atribuído a uma limitação do algoritmo de aprendizagem e não a uma perda do padrão de relacionamento por parte do método de discretização proposto.

Como nas comparações realizadas entre a EFD e outros métodos de discretização mais elaborados esses não obtiveram resultados significativamente melhores do que a primeira, podemos inferir, devido aos resultados expressivamente melhores da DTP com relação à EFD na metade dos casos analisados e pouco melhores nos restantes, que o método proposto obterá resultados significativamente melhores que os métodos mais elaborados de discretização. Essa comprovação fica, no entanto, como sugestão para trabalhos posteriores.

Vale ressaltar que os resultados foram obtidos a partir de uma base de dados real fornecida que, apesar de por hipótese ser considerada como válida, poderia conter erros e ruído nas medidas. Desse modo o conjunto de aprendizagem poderia conter ruído e erros e mesmo

assim foi possível verificar claramente a diferença de desempenho resultante da aplicação do método proposto de discretização para aprendizagem Bayesiana.

6.2. Triagem de Dados

Para verificar a aplicabilidade das BNs na validação de dados foi feito mais um teste. Em ambas as bases foram escolhidos 100 registros aleatoriamente. Em 67 desses registros o valor de uma das variáveis foi alterado para um intervalo adjacente e nos outros 33 registros foram alterados os valores de duas variáveis. A aprendizagem para os dois métodos de discretização foi refeita com a inclusão dos registros alterados na base e a triagem foi realizada em ambas as bases.

Do mesmo modo que para a comparação entre os métodos de discretização, foi verificada qual a porcentagem de registros alterados estava contida entre os 2.5%, 5%, 10%, 15% e 20% registros mais suspeitos. Os resultados são apresentados na Tabela 6.8.

Tabela 6.8: Porcentagem dos registros alterados selecionados pela triagem.

Dados Suspeitos	2.5%	5%	10%	15%	20%
Triagem DTP	76%	81%	83%	88%	88%
Triagem EFD	77%	79%	80%	81%	82%

Por esses resultados, pela análise de apenas 5% dos registros da base de dados uma pessoa seria capaz de identificar 81% dos registros errados. Caso esses 81% estivessem uniformemente espalhados pela base seria necessário que 81% da base fosse analisada, desse modo, para identificar a mesma quantidade de registros errados uma pessoa analisaria 16 vezes menos registros se usasse a abordagem Bayesiana para a triagem dos dados.

Comparando os resultados obtidos com a DTP e com a EFD notamos que a DTP apresenta resultados um pouco melhores. Como para cada variável apenas 4 ou no máximo 5 registros foram alterados essas variações são menos significativas que as apresentadas anteriormente.

Foi verificada também qual a porcentagem de registros alterados com a correta identificação do campo alterado estava contida entre os 2.5%, 5%, 10%, 15% e 20% registros mais suspeitos. Os resultados são apresentados na Tabela 6.9.

Tabela 6.9: Porcentagem dos registros alterados selecionados pela triagem com a correta identificação do campo alterado.

Dados Suspeitos	2.5%	5%	10%	15%	20%
Triagem DTP	75%	79%	81%	84%	84%
Triagem EFD	73%	75%	76%	76%	76%

Pelos resultados notamos que uma pessoa seria capaz de identificar 79% dos registros alterados analisando apenas um campo de apenas 5% do total de registros. Considerando novamente que os registros com erro estejam uniformemente espalhados pela base, sem a triagem uma pessoa teria que analisar todos os campos de 79% dos registros para identificar a mesma quantidade de registros errados, o que daria cerca de 500 vezes a quantidade de campos analisados com o auxílio da triagem Bayesiana.

Novamente a triagem de dados com a BN construída a partir da EFD mostra-se um pouco menos eficiente que a BN construída a partir da DTP.

É interessante notar que os registros alterados que não foram selecionados pela triagem Bayesiana apresentam valores com probabilidade de ocorrência elevada, uma vez que não foram selecionados entre os 5%, 10% ou 20% mais suspeitos. Desse modo esses registros, apesar de terem sido alterados seguem os padrões de relacionamento da base de dados e mesmo que fossem selecionados por uma triagem, dificilmente seriam identificados por uma pessoa, justamente por seguirem os padrões da base.

O uso desses registros alterados, que em uma situação real seriam registros que contém erros, que não foram selecionados como suspeitos, em levantamentos estatísticos, desenvolvimento de modelos, extração de conhecimento ou qualquer outra forma de análise de dados não deve ser ainda muito prejudicial, uma vez que seus valores estão de acordo com os padrões da base de dados.

Apesar de não ser o objetivo principal deste trabalho validar a aplicação de BNs à triagem e validação de dados, os resultados obtidos apontam para um grande potencial de aplicação. Cabe ressaltar mais uma vez que a base de dados utilizada é uma base de dados real, que provavelmente já contém registros com erro e que mesmo assim a BN foi capaz de selecionar corretamente a maior parte dos registros alterados como suspeitos.

Isso mostra que as BNs tem um grande potencial de aplicação no auxílio à validação de dados, através da aprendizagem Bayesiana, preferencialmente com a utilização de um método de discretização apropriado à extração de conhecimento, e da inferência Bayesiana para a triagem automática de dados.

Os resultados apontam também para um uso promissor das BNs para extração de conhecimento, análise e predição de dados em Proteção ao Vôo, uma vez que os resultados obtidos, no geral, demonstraram que as BNs geradas foram capazes de aprender corretamente os padrões de relacionamento entre as variáveis do domínio.

7. Conclusões

7.1. Conclusões Gerais

Os resultados obtidos comprovam que o método proposto de discretização se adequa melhor à aprendizagem Bayesiana que a EFD. Como nenhum dos outros métodos de discretização conhecidos se mostrou substancialmente melhor que a EFD, inferimos que o método proposto se adequa melhor à aprendizagem Bayesiana que os métodos conhecidos.

Pela utilização desse novo método a aprendizagem Bayesiana será capaz de extrair mais informação sobre os padrões de relacionamento entre as variáveis de um conjunto de treinamento, melhorando consideravelmente o desempenho de aplicações construídas a partir dessas redes. Essa melhoria no desempenho das redes fará com que os resultados obtidos onde elas já são aplicadas sejam ainda melhores e possivelmente habilitará novos usos para as mesmas.

Pelos resultados obtidos é possível também vislumbrar o potencial da aplicação das redes Bayesianas no auxílio à validação de dados. O trabalho realizado no desenvolvimento do sistema utilizado na comparação dos resultados é um excelente passo inicial na construção de um sistema que possa ser utilizado, na prática, pelo IPV ou por outras entidades que necessitem de um processo de triagem para diminuir a necessidade de validações manuais.

Com os dados de Proteção ao Vôo validados, aplicativos construídos ou análises feitas a partir dos mesmos serão mais confiáveis e apresentarão melhores resultados.

7.2. Contribuições do trabalho

São contribuições originais deste trabalho:

- A proposição de um método de discretização de dados para aprendizagem Bayesiana (Capítulo 3);
- Uma análise imparcial dos resultados obtidos em comparações entre outros métodos de discretização (Capítulo 2);
- Proposição da utilização de aprendizagem e inferência Bayesiana no auxílio à validação de dados, especificamente na triagem automática de dados (Capítulo 4);
- Proposição e utilização de um método de comparação quantitativo entre redes Bayesianas geradas a partir de diferentes métodos de discretização(Capítulo 4). O

método proposto pode ser usado também na comparação entre redes Bayesianas geradas por diferentes métodos de aprendizagem;

- Comparação entre os resultados obtidos pelo uso do método de discretização proposto e da Discretização em Igual Freqüência (Capítulo 6);

7.3. Trabalhos Futuros

Ficam como sugestões de trabalhos futuros as seguintes possibilidades:

- Demonstração formal da efetividade do método proposto baseado na teoria de probabilidade.
- Construção de um sistema especialista probabilístico Bayesiano na área de Proteção ao Vôo.
- Comparação entre o método de discretização proposto e métodos mais elaborados que a EFD.
- Aprimoramento do método proposto pela utilização de heurísticas ou de outros métodos de discretização na escolha dos intervalos iniciais.
- Aprimoramento do método proposto especificamente na discretização dos nós folha, que poderia ser feita pela utilização de um ou mais métodos de discretização já existentes.
- Aprimoramento do método proposto pela proposição de possíveis medidas mais adequadas do que a DMQ e a DMQP.
- Comparação do método proposto com outros métodos de discretização no contexto de classificadores Bayesianos, inclusive naive-Bayes.
- Construção de um software que faça a discretização e a aprendizagem Bayesiana automaticamente.
- Proposição de um novo algoritmo de aprendizagem Bayesiana, possivelmente interagindo com o método de discretização proposto.
- Validação da utilização da aprendizagem e da inferência Bayesiana na triagem de dados com estudos de outros domínios de problemas e comparação com outros métodos de triagem de dados.
- Utilização de redes Bayesianas na sugestão de valores para correção/substituição de dados errados.
- Utilização do método de discretização proposto, aprendizagem e inferência Bayesiana na supervisão de sistemas de controle automático.

- Proposição de métodos não semânticos de comparação entre redes Bayesianas geradas por algoritmos de aprendizagem diferentes ou por métodos de discretização diferentes fora do contexto de classificadores.

Tanto a área de aplicações de redes Bayesianas quanto as áreas de discretização de dados e de validação de dados estão repletas de possíveis trabalhos nos mais diversos níveis. As sugestões apresentadas, apenas uma fração dessas possibilidades; são apenas algumas das direções que podem ser seguidas na descoberta do conhecimento.

Referências Bibliográficas

[1] BACH, F. R.; JORDAN, M. I. Learning graphical models with Mercer kernels. In: *ADVANCES in Neural Information Processing Systems.* Cambridge, MA: MIT Press, 2003. n. 15

[2] BAILEY, B. H.; MCDONALD, S. L. *Wind Resource Assessment Handbook: Fundamentals for Conducting a Successful Monitoring Program*. Albany, NY: WindBooks Incorporated, 1997.

[3] BAY, S. D. Multivariate discretization of continuous variables for set mining. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 6., Boston, 2000. *Proceedings*... Boston: ACM, 2000. p.315-319.

[4] CARNEIRO, A. L.; SILVA, W. T. *Introdução a Redes Bayesianas.* Brasília: Universidade de Brasília, Departamento de Ciência da Computação, 1999. (Relatório de Pesquisa CIC/UnB – 09/99).

[5] CASTILLO, E.; GUTIERREZ, J. M.; HADI, A. S. *Expert Systems and Probabilistic Network Models*. New York: Springer Verlag, 1997.

[6] CERQUIDES, J.; MANTARAS, R. L.. Proposal and empirical comparison of a parallelizable distance-based discretization method. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 3., 1997. *Proceedings*... [S.l.n.], 1997. p. 139-142. (KDD97).

[7] CHAN, P. P. F. *An Expert System for Diagnosis of Problems in Reinforced Concrete Structures*. 1996. Dissertação (Mestrado em Ciências Aplicadas em Tecnologia da Informação) - Department of Computer Science, Royal Melbourne Institute of Technology, Melbourne.

[8] CHENG, J.; BELL, D. A. ; LIU, W. Learning belief networks from data: an information theory based approach. In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 6. , 1997. *Proceedings*...[S.l.]: ACM, 1997.

[9] CHICKERING, D. M.; HECKERMAN, D. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, n. 29, p.181-212, 1997.

[10] CHICKERING, D. M.; HECKERMAN, D.; MEEK, A. Bayesian approach to learning bayesian networks with local structure. In: CONFERENCE ON UNCERTAINTY IN

ARTIFICIAL INTELLIGENCE, 30. , 1997, Providence, RI. *Proceedings*... Providence, RI: Morgan Kaufmann, 1997. p.80-89. (UAI-97).

[11] CHOUDREY, R. A. *Variational Methods for Bayesian Independent Component Analysis*. 2002. Tese (Doutorado of Department of Engineering Science) - University of Oxford, Oxford.

[12] COELHO, A. S. G. *Abordagem Bayesiana na análise genética de populações utilizando dados de marcadores moleculares*. 2002. Tese (Doutorado em Agronomia) – USP. Escola Superior de Agricultura, Piracicaba.

[13] COOPER, G. F.; HERSKOVITZ, E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, n.9, p.309-347, 1992.

[14] COZMAN, F. G.. Generalizing Variable Elimination in Bayesian Networks. In: IBERAMIA/SBIA WORKSHOPS, 2000, São Paulo. *Proceedings*... São Paulo: Tec Art, 2000. p.27-32

[15] DAVIS, H. R. *A model based data validation system*. 1995. Dissertação (Mestrado em Engenharia Elétrica) - Vanderbilt University, Nashville.

[16] DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and unsupervised discretization of continuous features. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 12., 1995. *Proceedings*...[S.l.]: Morgan-Kaufmann, 1995. p.194-202.

[17] DOSHI, P. J. *Effective methods for building probabilistic models from large noisy data sets*. 2001. Dissertação (Mestrado em Ciências da Computação) - Drexel University, Philadelphia.

[18] FRANK, E.; WITTEN, I. H. Making better use of global discretization. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 17., Bled, Slovenia, 1999. *Proceedings*...Bled: [s.n.], 1999.

[19] FRIEDMAN, N.; GOLDSZMIDT, M. Discretizing continuous attributes while learning Bayesian networks. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 13. , 1996. *Proceedings*... [S.l.n.], 1996. p.157-165. (ICML 1996).

[20] FRIEDMAN, N.; GOLDSZMIDT, M.; LEE, T. J. Bayesian network classification with continuous attributes: getting the best of both discretization and parametric fitting. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 15., 1998. *Proceedings* … [S.l.]: Morgan Kaufmann, 1998. p.179-187.

[21] GÖTZ W.; REISACHER, S. Data validation improves reliability and accuracy of performance monitoring and field acceptance tests. In: EUROPEAN CONFERENCE ON TURBO MACHINERY, 3., 1998, Orlando. *Proceedings* … [S.l.n.], 1998

[22] HARTIGAN, J. A. *Bayes theory.* New York: Springer-Verlag, 1983.

[23] HECKERMAN, D. A Bayesian approach to learning causal networks. In: CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 11., San Francisco, CA, 1995. *Proceedings* … San Francisco, CA: Morgan Kaufmann, 1995.

[24] HECKERMAN, D. *A tutorial on learning with Bayesian networks*: *learning in graphical models*. Cambridge, MA: MIT Press, 1999. p.301-354.

[25] HECKERMAN, D.; GEIGER, D.; CHICKERING, D. M. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, n. 20, p.197-243, 1995.

[26] HECKERMAN, D.; MEEK, C.; COOPER, G. A Bayesian approach to causal discovery. In: GLYMOUR, G.; COOPER, G. (Ed.). *Computation, causation and discovery*. Cambridge, MA: MIT Press, 1999.

[27] HOLST, A. *The use of a Bayesian neural network model for classification tasks*. 1997. Dissertação (Mestrado of Department of Numerical Analysis and Computing Science) - Royal Institute of Technology, Stockholm.

[28] HRUSCHKA JR., E. R.; SILVA, W. T. *Propagação de crença e aprendizado em redes Bayesianas*. Brasília: Universidade de Brasília, Departamento de Ciência da Computação 1995. (Relatório de Pesquisa CIC/UnB – 03/96).

[29] HRUSCHKA JR., E. R.; SILVA, W. T. *Propagação de crenças em redes Bayesianas*. Brasília: Universidade de Brasília, Departamento de Ciência da Computação, 1997. (Relatório de Pesquisa CIC/UnB – 02/97).

[30] KADIE, C. M.; HOVEL, D.; HORVITZ, E. *MSBNx:* a component-centric toolkit for modeling and inference with Bayesian networks. Redmond : Microsoft, 2001. (Microsoft Research Technical Report 2001-67).

[31] KOEHLER, C.; NASSAR, S. M. Modelagem de redes Bayesianas a partir de bases de dados médicas. In: JORNADAS ARGENTINAS DE INFORMÁTICA E INVESTIGACIÓN OPERATIVA, 31., 2002. *Anais...* [S.l.n.], 2002. p.164-176 (EVENTO SIMULTANEO SIMPOSIO ARGENTINO DE INFORMÁTICA EN SALUD, 5.)

[32] KWEE, I. W. *Towards a Bayesian framework for optical tomography*. 1999. Tese (Doutorado of Department of Medical Physics and Bioengineering) - University College London, Londres.

[33] KOHAVI, R; SAHAMI, M. Error-based and entropy-based discretization of continuous features. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2., 1996, Portland. *Proceedings*... Portland: [s.n.], 1996. p.114-119.

[34] LADEIRA, M.; VICARRI, R. M.; COELHO, H. *Raciocínio probabilístico em sistemas inteligentes*. Brasília: Universidade de Brasília, Departamento de Ciência da Computação, 1999. (Relatório de Pesquisa CIC/UnB – 03/99).

[35] LADEIRA, M.; VICARRI, R. M.; COELHO, H. *Redes Bayesianas multiagentes*. Brasília: Universidade de Brasília, Departamento de Ciência da Computação, 1999. (Relatório de Pesquisa CIC/UnB – 08/99).

[36] LAM, W. ; BACCHUS, F. Learning Bayesian belief networks: an approach based on the MDL principle. *Computational Intelligence*, n.10, p.269-293, 1994.

[37] LARSON, H. J. *Introduction to probability theory and statistical inference*. 3rd ed. New York: John Wiley, 1982.

[38] LING, C. X.; ZHANG, H. The representational power of discrete Bayesian networks. *The Journal of Machine Learning Research*, v.3, p.709-721, 2003.

[39] LIU, H. et al. Discretization: an enabling technique. *Journal of Data Mining and Knowledge Discovery*, v.6, n.4, p 393-423, 2002.

[40] LUDL, M.; WIDMER, G. Relative unsupervised discretization *for regression problems*. In: EUROPEAN CONFERENCE ON MACHINE LEARNING, 11., 2000, Barcelona. *Proceedings*... Barcelona: [s.n.], 2000. (ECML 2000).

[41] MACKAY, D. J. C. *Bayesian methods for adaptive models*. 1992. Tese (Doutorado). California Institute of Tecnology, Pasadena.

[42] MADIGAN, D.; YORK, J. Bayesian graphical models for discrete data. *International Statistical Review*, v.63, n.2, p.215-232, 1995.

[43] MURPHY, K. The Bayes net toolbox for Matlab. *Computing Science and Statistics*, v.33, 2001.

[44] NASCIMENTO JR., C. L.; YONEYAMA, T. *Inteligência artificial em controle e automação*. São Paulo: Edgard Blücher, 2000.

[45] NEAL, R. M. *Monte Carlo implementation of Gaussian process models for Bayesian classification and regression*. Toronto: University of Toronto, Department of Statistics, 1997. (Technical Report 9702).

[46] NEAL, R. M. *Probabilistic inference using Markov chain Monte Carlo methods*. Toronto: University of Toronto, Department of Computer Science, 1993. (Technical Report CRG-TR-93-1).

[47] PAPOULIS, A. *Probability, random variables, and stochastic processes*. 3rd Ed. New York: McGraw-Hill, 1991.

[48] PEARL, J. *Probabilistic reasoning in intelligent systems*: *networks of plausible inference.* San Mateo, CA: Morgan Kaufmann, 1988.

[49] RAMONI, M.; SEBASTIANI, P. Learning Bayesian networks from incomplete databases. In: CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 11., 1995. *Proceedings*... [S.l.]: Morgan Kaufman, 1995. p.401-408.

[50] REIS, L. A. *SANEP* – Sistema Especialista Probabilístico de Apoio a Nutrição Enteral Pediátrica. 2001. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Santa Catarina, Florianópolis.

[51] RICH, E.R.; KNIGHT, K. *Inteligência artificial*. 2 ed. Rio de Janeiro: Makron Books, 1994.

[52] ROUSU, J. *Efficient range Partitioning in Classification Learning*. 2001. Tese (Ph.D. Department of Computer Science) - University of Helsinki, Helsinki.

[53] SAHAMI, M. et al. A Bayesian approach to filtering junk e-mail. In: WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION, 1998, Madison. *Proceedings*… Madison: AIAA, 1998.

[54] SANTOS, L. P. P. *Application level runtime load management*: a Bayesian approach. 2001. Tese (Doutorado em Informática) - Escola de Engenharia da Universidade do Ninho, Braga.

[55] SCHMIDT, A. M. *Bayesian spatial interpolation of environmental monitoring stations*. 2001. Tese (Doutorado of Department of Probability and Statistics) - School of Mathematics, University of Sheffield, Sheffield.

[56] SEEGER, M. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, n.3, p.233-269, 2002.

[57] SILVESTRE, A. M. *Raciocínio probabilístico aplicado ao diagnóstico de insuficiência cardíaca congestiva (ICC)*. 2003. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre.

[58] TIPPING, M. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning*, v. 1, p. 211-244, 2001.

[59] VALLDEPERAS, E. M. *Sistema Bayesiano para modelado del alumno*. 2000. Tese (Doutorado do Departamento de Lenguages y Ciencias de la Computación) - Universidad de Málaga, Málaga.

[60] VEGAS, F. J. D. *Sistema experto Bayesiano para ecocardiografía*. 1994. Tese (Doutorado em Ciências Físicas) - Faculdad de Ciencias de la Universidad Nacional de Educacion a Distancia, Madrid.

[61] VENTURA, D.; MARTINEZ, T. R. An empirical comparison of discretization methods. In: INTERNATIONAL SYMPOSIUM ON COMPUTER AND INFORMATION SCIENCES, 10., 1995. *Proceedings*... [S.l.n.], 1995. p. 443-450.

[62] WANG, H.; ZANIOLO, C. CMP: a fast decision tree classifier using multivariate predictions. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 16., 2000, San Diego. *Proceedings*... San Diego: ICDE, 2000. p.449-460. (ICDE'2000).

[63] WANG, K.; LIU, B. Concurrent discretization of multiple attributes. In: THE PACIFIC RIM INTERNATIONAL CONF. ON ARTIFICIAL INTELLIGENCE, 1998. *Proceedings*... [S.l.n.], 1998. p. 250-259.

[64] WEST, M.; WINKLER, R.L. Database error trapping and prediction. *Journal of the American Statistical Association*, n.86, p.987-996, 1991.

[65] YANG, Y. *Discretization for naive-Bayes learning*. 2003. Tese (Doutorado em Ciência da Computação) - School of Computer Science and Software Engineering of Monash University, Melbourne.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO N°	4. N° DE PÁGINAS
TM	19 de dezembro de 2003	CTA/ITA-IEE/TM-018/2003	81

5. TÍTULO E SUBTÍTULO:

Discretização para Aprendizagem Bayesiana: Aplicação no Auxílio à Validação de Dados em Proteção ao Vôo

6. AUTOR(ES):

Jackson Paul Matsuura

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica. Divisão de Engenharia Eletrônica – ITA/IEE

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Discretização, redes Bayesianas, validação de dados, aprendizagem Bayesiana.

9.PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Teorema de Bayes; Funções de distribuição de probabilidades; Análise estatística multivariada; Validação; Matemática

10. APRESENTAÇÃO: **X Nacional** **Internacional**

ITA, São José dos Campos, 2003, 81 páginas.

11. RESUMO:

A utilização de redes Bayesianas vem crescendo em diversas áreas e aplicações. As redes Bayesianas podem ser construídas a partir do conhecimento de especialistas ou por algoritmos de aprendizagem que inferem as relações entre as variáveis do domínio a partir de um conjunto de dados de treinamento.

A construção manual de redes Bayesianas vem cada vez mais sendo preterida pelo uso de algoritmos de aprendizagem que,em geral, pressupõem que as variáveis utilizadas na aprendizagem sejam discretas ou, caso sejam contínuas, apresentem uma distribuição Gaussiana, o que normalmente não ocorre na prática.

Portanto para o uso de algoritmos de aprendizagem é necessário que as variáveis sejam discretizadas segundo algum critério, que no caso mais simples pode ser uma discretização uniforme.

A grande maioria dos métodos de discretização existentes, porém, não são adequados à aprendizagem de redes Bayesianas, pois foram desenvolvidos no contexto de classificação e não de descoberta de conhecimento.

Nesse trabalho é proposto e utilizado um método de discretização de variáveis que leva em conta as distribuições condicionais das mesmas no processo de discretização, objetivando um melhor resultado do processo de aprendizagem Bayesiana.

O método proposto foi utilizado em uma base de dados real de informações de Proteção ao Vôo e a rede Bayesiana construída foi utilizada no auxílio à validação de dados, realizando uma triagem automatizada dos dados.

Foi realizada uma comparação entre o método proposto de discretização e um dos métodos mais comuns.

Os resultados obtidos mostram a efetividade do método de discretização proposto e apontam para um grande potencial dessa nova aplicação da aprendizagem e inferência Bayesiana.

12. GRAU DE SIGILO:

(X) OSTENSIVO **( ) RESERVADO** **( ) CONFIDENCIAL** **( ) SECRETO**