

RESSALVA

Atendendo solicitação do(a)
autor(a), o texto completo desta tese
será disponibilizado somente a partir
de 05/08/2018.



PROGRAMA DE PÓS-GRADUAÇÃO EM ZOOLOGIA

ESTUDO DE COMUNIDADES DE DÍPTEROS NECRÓFAGOS SOB O FORMALISMO DE BICLUSTERS E ÁRVORES DE DECISÃO

CAROLINE RODRIGUES DE SOUZA

Tese apresentada ao Instituto de Biociências do Campus de Rio Claro, Universidade Estadual Paulista, como parte dos requisitos para obtenção do título de Doutora em Ciências Biológicas (Área de Concentração – Zoologia).

Agosto - 2016

CAROLINE RODRIGUES DE SOUZA

**ESTUDO DE COMUNIDADES DE DÍPTEROS NECRÓFAGOS SOB O
FORMALISMO DE GRAFOS E ÁRVORES DE DECISÃO**

Tese apresentada ao Instituto de
Biotecnologia do Campus de Rio Claro,
Universidade Estadual Paulista Júlio
de Mesquita Filho, como parte dos
requisitos para obtenção do título de
Doutora em Ciências Biológicas (Área
de Concentração – Zoologia)

Orientador: Prof. Dr. Cláudio José Von Zuben

Rio Claro – SP
2016

595.77 Souza, Caroline Rodrigues de
S729e Estudo de comunidades de dípteros necrófagos sob o
formalismo de biclusters e árvores de decisão / Caroline
Rodrigues de Souza. - Rio Claro, 2016
181 f. : il., figs., tabs.

Tese (doutorado) - Universidade Estadual Paulista,
Instituto de Biociências de Rio Claro
Orientador: Claudio José Von Zuben

1. Díptero. 2. Sinantropismo. 3. Moscas. 4. Ambientes
antrópicos. 5. Zoologia. I. Título.



CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: **ESTUDO DE COMUNIDADES DE DÍPTEROS NECRÓFAGOS SOB O FORMALISMO DE 8/CLUSTERS E ARVORES DE DECISÃO**

AUTORA: CAROLINE RODRIGUES DE SOUZA

ORIENTADOR: CLAUDIO JOSÉ VON ZUBEN

Aprovada como parte das exigências para obtenção do Título de Doutora em CIÊNCIAS BIOLÓGICAS (ZOOLOGIA), pela Comissão Examinadora:


k6

Prof. Dr. CLAUDIO JOSÉ VON ZUBEN

Departamento de Zoologia / Instituto de Biociências de Rio Claro - SP

Prof. Dr. JACQUELINE THYSSEN

Departamento de Biologia Animal - Instituto de Biologia / Universidade Estadual de Campinas - SP


Prof. Dr. RENATO CONDE GODOY

Departamento de Entomologia e Acarologia / USP - Escola Superior de Agricultura Luiz de Queiroz - Piracicaba/SP


Prof. Dra. MARIA JOSÉ DE OLIVEIRA CAMPOS

Departamento de Ecologia / Instituto de Biociências de Rio Claro - SP

Prof. Dr. RITO GIANNOTTI

Departamento de Zoologia / Instituto de Biociências de Rio Claro - SP

Rio Claro, 05 de agosto de 2016

Aos meus pais, Osmar e Floracy, pela dedicação, sustentação, coragem e amor à mim oferecidos durante esta caminhada.

À minha irmã, Leticia e ao meu irmão, Leonardo, pelo carinho, amor e exemplos de vida.

Ao meu marido, Ivanildo, pelo amor e paciência nos momentos mais difíceis.

Agradecimentos

Agradeço, primeiramente a Deus, que me criou e guiou os meus caminhos até aqui.

Agradeço, do fundo do meu coração, aos meus pais, Osmar e Floracy, que tanto se esforçaram para que eu chegasse até aqui. Também sempre me aconselharam a estudar e foram a base e a fortaleza nas horas mais difíceis. Também sou muito grata aos meus irmãos, Leonardo e Letícia, que sempre me apoiaram e estiveram torcendo por mim! Amo muito todos vocês!

Sou muito grata ao apoio, à paciência e ao amor que meu marido Ivanildo tanto dedicou à mim. Nas horas complicadas, era ele que sempre me acalmava e me dava a paz que eu tanto necessitava. Te amo! E não poderia me esquecer da nossa cachorrinha Penellope, que sempre me fez companhia nos vários dias de estudo em casa.

Agradeço também ao apoio de todos os meus amigos (Meire, Daliane, Stefania, Ana Cláudia, Carolinne, Marcela, Ana Paula...), estando perto ou longe, sempre estiveram presentes e preocupadas em todos esses anos de estudo. Também sou grata aos meus amigos e colegas de laboratório, que, de uma forma ou de outra, fizeram parte desses anos de aprendizado.

À UNESP – Universidade Estadual “Júlio Mesquita Filho”, pela oportunidade da realização deste trabalho.

Ao Prof. Dr. Cláudio José Von Zuben, pela orientação, pelos conhecimentos transmitidos, pelo apoio e pela amizade que se fortaleceu nesses anos de convivência.

Ao pesquisador Prof. Fernando José Von Zuben, agradeço, imensamente, pelo ensinamento das ferramentas matemáticas, pelo auxílio e pela paciência comigo. Também sou grata à Doutoranda Rosana Veronese que me ajudou muito na compreensão dos biclusters.

À FAPESP, pela bolsa de estudo concedida para realização deste estudo.

Portanto, agradeço a todos aqueles que de alguma forma puderam me auxiliar na realização deste trabalho.

RESUMO

O reconhecimento de padrões em comunidades ecológicas é um dos mais antigos e persistentes desafios da ciência ecológica. Nesse sentido, através da modelagem matemática a partir de grafos, utilizando *biclusters* e árvores de decisão, buscou-se encontrar padrões na ocorrência de dípteros necrófagos coletados em três diferentes áreas (rural, urbana e mata) e com três diferentes iscas (sardinha, fígado e carne moída) na cidade de Rio Claro-SP. A partir dos algoritmos (InClose e RInClose – *biclusters*, C4.5 – árvores de decisão através da análise dos *biclusters*, podemos destacar, por exemplo, aqueles que relacionaram *Atherigona orientalis* (Schiner) e *Musca domestica* Linnaeus em todos os ambientes de coleta, menos na área de floresta, provavelmente devido ao fato de serem espécies conhecidas por sua preferência por áreas habitadas pelo homem. Por outro lado, mesmo sendo espécies que possui alta dependência pela antropobiocenose, outros *biclusters* indicaram uma associação entre *Lucilia eximia* (Wiedemann) e *Chrysomya megacephala* (Fabricius) ocorrendo com maior frequência na floresta, atraídas pelas iscas de carne ou fígado, já conhecidas pela sua importância na atratividade de califorídeos. Isso pode ser explicado devido ao crescimento urbano ocorrido ao redor da floresta estudada, dando-lhe características próprias de local habitado pelo homem. No que se refere às árvores de decisão, foram obtidas no total 19 árvores de decisão (sendo 8 para a família Sarcophagidae, 4 para Muscidae e 7 para Calliphoridae) e a partir delas pode-se detectar vários padrões como: *A. orientalis*, *Synthesiomyia nudiseta* (Wulp, 1883) e *M. domestica* estando presente conjuntamente se a estação for Primavera, a isca for sardinha e o local for Urbana. *Oxysarcodexia thornax* (Walker) foi a única espécie, que para atingir a condição de presença, percorreu somente um nó, a estação Primavera, sendo, portanto, considerada generalista neste estudo, já que as variáveis local e isca foram indiferentes para determinar sua localização. Por tudo isso, considerou-se que os *biclusters* e as árvores de decisão são ferramentas úteis na estimativa de padrão na relação existentes entre a comunidade de dípteros necrófagos.

PALAVRAS-CHAVE: Sinantropismo, moscas e ambientes antrópicos.

ABSTRACT

The pattern recognition in ecological communities is one of the most antique and persistent challenges of ecological science. In this sense, through mathematical modelling from graphs, using biclusters and decision trees, the objective of the present work was to find patterns in the occurrence of necrophagous dipterans collected in three different areas (rural, urban and forest) and with three different baits (sardines, liver and beef) in the city of Rio Claro-SP. Algorithms (InClose and RInClose-biclusters, C 4.5-decision trees) were used to obtain the results. Through the analysis of biclusters, it can be highlighted those that related *Atherigona orientalis* (Schiner) and *Musca domestica* Linnaeus in all environments, less forest area, probably due to the fact of being species known for their preference for areas inhabited by humans. On the other hand, even being species possessing high dependence by antropobiosenose, other biclusters indicated an association between *Lucilia eximia* (Wiedemann) and *Chrysomya megacephala* (Fabricius) occurring with greater frequency in the forest, attracted by the bait of meat or liver, both known for their importance in the attractiveness of calliphorid fly. This may be explained due to the population growth occurred around the forest, giving the characteristics of place inhabited by man. With regard to decision trees, were obtained in total 19 decision trees (8 to the family Sarcophagidae, 4 for Muscidae and 7 to Calliphoridae) and from them various patterns can be detected such as: *A. orientalis*, *Synthesiomyia nudiseta* (Wulp, 1883) and *M. domestica* being present together if the season is Spring, the bait is sardines and the location is urban. *Oxysarcodexia thornax* (Walker) was the only specie in which to achieve the condition of presence, it is necessary to walk only one node (the spring), therefore, it was considered to be generalist in this study, since the local variables and bait were indifferent to determine this location. For all that, it was considered that the decision trees and biclusters are useful tools in the elucidation of existing relationship patterns between the community of dipterans collected.

KEY-WORDS : Synanthropism, Flies, Anthropic environments.

LISTA DE ILUSTRAÇÕES

Página

Figura 1 – Grafo com conjunto de vértices $V = \{a,b,c,d\}$ e arestas $A = \{ab,bc,cd\}$	17
Figura 2 - Diagrama das sete pontes de Königsberg.....	17
Figura 3 - Representação gráfica à esquerda de um grafo não direcionado (apenas uma aresta existe entre dois vértices) e à direita de um grafo direcionado (mais de uma aresta é permitida entre dois vértices). Os círculos representam os vértices (nós) e os segmentos de reta representam as ligações (arestas). Os números referem-se aos vértices (nós).....	18
Figura 4 - A aresta em vermelho é um laço e as arestas em verde são paralelas múltiplas.	18
Figura 5 - Dado um grafo G (a), os grafos $G1$ (b) e $G2$ (c) são subgrafos de G	19
Figura 6 - Grafos (a) bipartido e (b) grafo bipartido completo.	20
Figura 7 - Grafo F , para visualização de diferentes caminhos entre os nós 1 e 4.	21
Figura 8 - Comparação entre um grafo cíclico e uma árvore.....	22
Figura 9 - Estrutura de uma árvore.....	23
Figura 10 - Técnicas <i>bottom-up</i> e <i>top-down</i>	25
Figura 11 - Representação de uma árvore de decisão e sua respectiva representação no Espaço.....	26
Figura 12 - Uma árvore de decisão simples para o diagnóstico de um paciente.....	27
Figura 13 - Matriz X de Dados.....	46
Figura 14 – Multigrafo bipartido.....	47
Figura 15 - Estruturas dos <i>biclusters</i> . (a) <i>bicluster</i> único; (b) <i>biclusters</i> com linhas e colunas exclusivas; (c) estrutura de xadrez; (d) <i>biclusters</i> com linhas exclusivas; (e) <i>biclusters</i> com colunas exclusivas; (f) <i>biclusters</i> não-sobrepostos com estrutura de árvore; (g) <i>biclusters</i> não sobrepostos não-exclusivos; (h) <i>biclusters</i> sobrepostos com estrutura hierárquica; e (i) <i>biclusters</i> sobrepostos arbitrariamente posicionados.....	48
Figura 16 - Exemplo de um <i>bicluster</i> formado por um subconjunto de objetos e um subconjunto de atributos não-contíguos na matriz de dados original.....	50
Figura 17 - Exemplos de tipos diferentes de <i>biclusters</i> . (a) BVC. (b) BVCL. (c) BVCC. (d) BVCo, aditivo. (e) BVCo, multiplicativo. (f) BEC.....	51
Figure 18 – Final configuration of the decision tree designed to predict absence or presence of the species <i>Sarcodexia lambens</i> . Notice that the decision tree is automatically generated	

from the collected datasets, by a learning from data approach. The ellipsoidal nodes (internal nodes) perform a test over an attribute, thus being a decision point. The decision outcome will indicate which edge to follow. The square nodes (leaves of the tree) are the available classes. Starting from the root node, there is a single path to each leaf node, and the sequence of decisions from the root to each leaf may be expressed as a single IF-THEN classification rule, making the process highly interpretable..... 119

LISTA DE TABELAS

Tabela 1 - Espécies de três famílias de Diptera, coletadas em três ambientes diferentes utilizando três tipos de iscas, na cidade de Rio Claro-SP (continua).	40
Tabela 1 - Espécies de três famílias de Diptera, coletadas em três ambientes diferentes utilizando três tipos de iscas, na cidade de Rio Claro-SP	41
Tabela 2 – Ocorrência das espécies da família Muscidae, atraídas por três tipos de isca nos três ambientes de coleta Os números (1-9), presentes tanto nas linhas quanto nas colunas, são úteis para a geração dos <i>biclusters</i> e posteriormente sua interpretação.	54
Tabela 3 – Ocorrência das espécies da família Calliphoridae, atraídas por três tipos de isca nos três ambientes de coleta. Os números (1-9), presentes tanto nas linhas quanto nas colunas, são úteis para a geração dos <i>biclusters</i> e posteriormente sua interpretação.	54
Tabela 4 – Ocorrência das espécies da família Sarcophagidae, atraídas por três tipos de isca nos três ambientes de coleta Os números (1-9), presentes tanto nas linhas quanto nas colunas, são úteis para a geração dos <i>biclusters</i> e posteriormente sua interpretação.	55
Tabela 5 - Cinco espécies mais abundantes coletadas de Muscidae, distribuídas em seus locais de coleta e isca utilizadas.	56
Tabela 6 - Cinco espécies mais abundantes coletadas de Sarcophagidae, distribuídas em seus locais de coleta e isca utilizadas.	56
Tabela 7 - Cinco espécies mais abundantes coletadas de Calliphoridae, distribuídas em seus locais de coleta e isca utilizadas.	56
Tabela 8 - Cinco espécies mais abundantes coletadas de cada família, distribuídas em seus locais de coleta e isca utilizadas.	58
Tabela 9 – Demonstração em formato de tabela do bicluster OPSM de número 1, obtido através da Tabela 3.	65
Tabela 10 - Demonstração em formato de tabela do bicluster OPSM de número 2, obtido através da Tabela 3.	66
Tabela 11 - Demonstração em formato de tabela do bicluster OPSM de número 3, obtido através da Tabela 3.	66
Tabela 12 - Demonstração em formato de tabela do bicluster OPSM de número 1, obtido através da Tabela 4.	67
Tabela 13 - Demonstração em formato de tabela do bicluster OPSM de número 2, obtido através da Tabela 4.	67

Tabela 14 - Demonstração em formato de tabela do bicluster OPSM de número 3, obtido através da Tabela 4.	68
Tabela 15 - Demonstração em formato de tabela do bicluster OPSM de número 4, obtido através da Tabela 4.	68
Tabela 16 - Demonstração em formato de tabela do bicluster OPSM de número 5, obtido através da Tabela 4.	69
Tabela 17 - Demonstração em formato de tabela do bicluster OPSM de número 6, obtido através da Tabela 4.	69
Tabela 18 - Demonstração em formato de tabela do bicluster OPSM de número 7, obtido através da Tabela 4.	70
Tabela 19 - Demonstração em formato de tabela do bicluster OPSM de número 8, obtido através da Tabela 4.	70
Tabela 20 - Demonstração em formato de tabela do bicluster OPSM de número 1, obtido através da Tabela 5.	71
Tabela 21 - Demonstração em formato de tabela do bicluster OPSM de número 2, obtido através da Tabela 5.	71
Tabela 22 - Demonstração em formato de tabela do bicluster OPSM de número 3, obtido através da Tabela 5.	72
Tabela 23 - Demonstração em formato de tabela do bicluster OPSM de número 4 obtido através da Tabela 5.	72
Tabela 24 - Demonstração em formato de tabela do bicluster OPSM de número 5, obtido através da Tabela 5.	73
Tabela 25 - Demonstração em formato de tabela do bicluster OPSM de número 6, obtido através da Tabela 5.	73
Tabela 26 - Demonstração em formato de tabela do bicluster OPSM de número 7, obtido através da Tabela 5.	74
Tabela 27 - Demonstração em formato de tabela do bicluster OPSM de número 8, obtido através da Tabela 5.	74
Tabela 28 - Demonstração em formato de tabela do bicluster OPSM de número 9, obtido através da Tabela 5.	75
Tabela 29 - Demonstração em formato de tabela do bicluster OPSM de número 10, obtido através da Tabela 5.	75
Tabela 30 - Demonstração em formato de tabela do bicluster OPSM de número 11, obtido através da Tabela 5.	76

Tabela 31 - Demonstração em formato de tabela do bicluster OPSM de número 12, obtido através da Tabela 5.	76
Tabela 32 - Demonstração em formato de tabela do bicluster OPSM de número 13, obtido através da Tabela 5	77
Tabela 33 - Demonstração em formato de tabela do bicluster OPSM de número 14, obtido através da Tabela 5.	77
Tabela 34 - Demonstração em formato de tabela do bicluster OPSM de número 1, considerando o zero como dado faltante, obtido através da Tabela 6.	79
Tabela 35 - Demonstração em formato de tabela do bicluster OPSM de número 2, considerando o zero como dado faltante, obtido através da Tabela 6.	80
Tabela 36 - Demonstração em formato de tabela do bicluster OPSM de número 1, considerando o zero como dado faltante, obtido através da Tabela 7.	80
Tabela 37 - Demonstração em formato de tabela do bicluster OPSM de número 2, considerando o zero como dado faltante, obtido através da Tabela 7.	81
Tabela 38 - Demonstração em formato de tabela do bicluster OPSM de número 3, considerando o zero como dado faltante, obtido através da Tabela 7.	81
Tabela 39 - Demonstração em formato de tabela do bicluster OPSM de número 4, considerando o zero como dado faltante, obtido através da Tabela 7.	82
Tabela 40 - Demonstração em formato de tabela do bicluster OPSM de número 5, considerando o zero como dado faltante, obtido através da Tabela 7.	82
Tabela 41 - Demonstração em formato de tabela do bicluster OPSM de número 5, considerando o zero como dado faltante, obtido através da Tabela 7.	83
Tabela 42 - Demonstração em formato de tabela do bicluster OPSM de número 1, considerando o zero como dado faltante, obtido através da Tabela 9.	83
Tabela 43 - Demonstração em formato de tabela do bicluster OPSM de número 2, considerando o zero como dado faltante, obtido através da Tabela 9.	84
Tabela 44 - Demonstração em formato de tabela do bicluster OPSM de número 3, considerando o zero como dado faltante, obtido através da Tabela 9.	84
Tabela 45 - Demonstração em formato de tabela do bicluster OPSM de número 4, considerando o zero como dado faltante, obtido através da Tabela 9.	85
Tabela 46 - Demonstração em formato de tabela do bicluster OPSM de número 5, considerando o zero como dado faltante, obtido através da Tabela 9.	85

Tabela 47 - Demonstração em formato de tabela do bicluster OPSM de número 6, considerando o zero como dado faltante, obtido através da Tabela 9.	86
Tabela 48 - Demonstração em formato de tabela do bicluster OPSM de número 7, considerando o zero como dado faltante, obtido através da Tabela 9.	86
Tabela 49 - Demonstração em formato de tabela do bicluster OPSM de número 8 considerando o zero como dado faltante, obtido através da Tabela 9.	87
Tabela 50 - Demonstração em formato de tabela do bicluster OPSM de número 9 considerando o zero como dado faltante, obtido através da Tabela 9.	87
Tabela 51 - Demonstração em formato de tabela do bicluster OPSM de número 10 considerando o zero como dado faltante, obtido através da Tabela 9.	88
Tabela 52 - Demonstração em formato de tabela do bicluster OPSM de número 11 considerando o zero como dado faltante, obtido através da Tabela 9.	88
Tabela 53 - Demonstração em formato de tabela do bicluster OPSM de número 12 considerando o zero como dado faltante, obtido através da Tabela 9.	89
Tabela 54 - Demonstração em formato de tabela do bicluster OPSM de número 13 considerando o zero como dado faltante, obtido através da Tabela 9.	89
Tabela 55 - Demonstração em formato de tabela do bicluster OPSM de número 14 considerando o zero como dado faltante, obtido através da Tabela 9.	90
Tabela 56 - Demonstração em formato de tabela do bicluster OPSM de número 15 considerando o zero como dado faltante, obtido através da Tabela 9.	90
Tabela 57 - Demonstração em formato de tabela do bicluster OPSM de número 16 considerando o zero como dado faltante, obtido através da Tabela 9.	91
Tabela 58 - Demonstração em formato de tabela do bicluster OPSM de número 17 considerando o zero como dado faltante, obtido através da Tabela 9.	91
Tabela 59 - Demonstração em formato de tabela do bicluster OPSM de número 18 considerando o zero como dado faltante, obtido através da Tabela 9.	92
Tabela 60 - Demonstração em formato de tabela do bicluster OPSM de número 19 considerando o zero como dado faltante, obtido através da Tabela 9.	92
Tabela 61 - Demonstração em formato de tabela do bicluster OPSM de número 20 considerando o zero como dado faltante, obtido através da Tabela 9.	93
Table 62 – Abundance of species of blowflies collected along the four seasons, in three different locations, and using three types of baits. (L: liver, M: minced meat, S: Sardine, S1: <i>S. lambens</i> , S2: <i>O. thornax</i> , S3: <i>P (E) collusor</i> , S4: <i>O. avuncula</i> , S5: <i>P (S) ingens</i> , S6: <i>P (E) anguilla</i> , S7: <i>P (P) intermutans</i> , S8 <i>P: (S) florencioi</i> , S9 : <i>A. orientalis</i> , S10: <i>S. nudiseta</i> , S11:	

M. domestica, S12: *O. chalcogaster*, S13: *L. eximia*, S14: *H. segmentaria*, S15: *L. cuprina*, S16:
H. semidiaphana, S17: *C. megacephala*, S18: *C. albiceps*, S19:... *M.*
peregrina) 116

Table 63 - Tendencies in collections involving species of the Sarcophagidae family 117

Table 64- Tendencies in collections involving species of the Muscidae family 117

Table 65- Tendencies in collections involving species of the Calliphoridae family 118

SUMÁRIO

	Página
INTRODUÇÃO	14
DÍPTEROS NECRÓFAGOS	15
NOÇÕES SOBRE GRAFOS	16
ÁRVORES DE DECISÃO	22
REFERÊNCIAS	32
OBJETIVOS	38
DADOS UTILIZADOS	39
REFERÊNCIAS	42
CAPÍTULO I – Aplicação de biclusters para modelar a ocorrência de dípteros necrófagos, no sudeste do Brasil	43
RESUMO	43
INTRODUÇÃO	44
MATERIAL E MÉTODOS	52
RESULTADOS	57
DISCUSSÃO	93
CONCLUSÃO	95
REFERÊNCIAS	96
CAPÍTULO II - Application of decision trees to model the occurrence of dipterans in a monitored ecosystem	106
ABSTRACT	106
INTRODUCTION	107
MATERIAL AND METHODS	108
EXPERIMENTAL RESULTS	111
DISCUSSION	112
REFERENCES	120
CONSIDERAÇÕES FINAIS	127

APÊNDICE A – Biclusters com valores constantes iguais a 0, obtidos através da Tabela 3, ao se usar dados binários.....	128
APÊNDICE B – Biclusters do tipo BVCC e BVCL, obtidos a partir da Tabela 3.	130
APÊNDICE C – <i>Biclusters</i> do tipo OPSM, obtidos a partir da Tabela 3 (Muscidae).	133
APÊNDICE D – <i>Biclusters</i> do tipo OPSM, obtidos a partir da Tabela 4 (Calliphoridae).	136
APÊNDICE E – <i>Biclusters</i> do tipo OPSM, obtidos a partir da Tabela 5 (Sarcophagidae)....	140
APÊNDICE F – <i>Biclusters</i> do tipo OPSM, obtidos a partir da Tabela 6 (Muscidae).....	146
APÊNDICE G – <i>Biclusters</i> do tipo OPSM, obtidos a partir da Tabela 7 (Calliphoridae).	151
APÊNDICE H – <i>Biclusters</i> do tipo OPSM, obtidos a partir da Tabela 8 (Sarcophagidae). ..	158
APÊNDICE I – Árvores de decisão das espécies de dípteros necrófagos obtidas durante o presente trabalho.....	163

INTRODUÇÃO

O reconhecimento de padrões em comunidades ecológicas é um dos mais antigos e persistentes desafios da ciência ecológica. Por vários anos, ecólogos de comunidades concentraram-se em duas tarefas fundamentais: primeiro, desenvolver métodos adequados para a detecção de padrões em conjuntos de espécies; e segundo, identificar processos capazes de gerar determinados padrões (LEWINSOHN et al., 2006).

Uma grande variedade de sistemas, que se estendem desde redes de genes até redes sociais, tem sido modelada matematicamente a partir de grafos, os quais são estruturas de dados formadas por nós e arestas que interligam par-a-par esses nós (AGNARSSON; GREENLAW, 2006). A grande vantagem desses modelos está na possibilidade de evidenciar não apenas os elementos que compõem o sistema em estudo, mas também e principalmente as interações que eles estabelecem entre si (MASON; VERWOERD, 2007). De acordo com esta visão, os elementos do sistema são nós (genes, pessoas, espécies) conectados por ligações (corregulação de genes, interação de pessoas ou de espécies). Em ecologia, o uso de grafos tem sido aplicado há algum tempo, como nos estudos clássicos de teias tróficas de Connell (1961) e de Paine (1966). Como um processo biológico não é executado somente por um elemento, mas pela interação de múltiplos elementos (EISENBERG et al., 2000), a sua modelagem em grafos permite caracterizar adequadamente os principais aspectos de processos biológicos bastante complexos (BARABÁSI, 2002).

Além de compreender os processos ecológicos através da interação de seus elementos constituintes, muitas vezes busca-se desenvolver modelos de predição e de classificação a partir de dados observados. Dentre as abordagens mais empregadas, encontram-se os mecanismos indutivo-observacionais de obtenção de modelos matemáticos a partir de dados amostrados de fenômenos biológicos (EASTMAN, 2006). Sendo assim, em lugar de uma abordagem *top-down* de modelagem fenomenológica, adota-se aqui uma perspectiva *bottom-up*, centrada em dados e em ferramentas de aprendizado de máquina (DIEKS, 2009).

As árvores de decisão se destacam neste processo de modelagem a partir de dados, por produzirem modelos diretamente interpretáveis, por apresentarem baixo custo computacional, permitindo assim o tratamento de grandes bases de dados, e por obterem alto desempenho em tarefas de predição e classificação (BREIMAN et al., 1984).

Por outro lado, há também a tarefa de agrupamento de dados (*clustering*) que consiste em separar amostras de dados relativas a um determinado domínio de estudo em grupos (*clusters*) distintos, sendo que instâncias pertencentes a um mesmo grupo devem apresentar

características semelhantes (JAIN; MURTY; FLYNN, 1999). Já o agrupamento bidimensional de dados (*biclustering*), que foi aplicado neste trabalho, permite fazer esse agrupamento simultaneamente em linhas e colunas de uma Tabela Relacional e pode ser uma tarefa útil em diversos domínios, tais como em Bioinformática, Previsão de Séries Temporais, Reconhecimento de Padrões, Análise e Processamento de Imagens e Recuperação de Informação. Nomes como biclusterização, agrupamento duplo de dados, *simultaneous clustering*, *co-clustering*, *two-wayclustering*, *blockclustering*, *bidimensional clustering*, entre outros, também são frequentemente usados na literatura para referenciar a tarefa de *biclustering* (BARKOW et al., 2006; CHARRAD; BEM AHMED, 2011; CHENG; CHURCH, 2000; DE FRANÇA, 2010; MADEIRA; OLIVEIRA, 2004; NIE et al., 2011).

Sendo assim, o emprego de modelagem matemática em grafos (*biclusters* e árvores de decisão) tende a ampliar o conhecimento disponível sobre processos ecológicos e, particularmente, sobre a biodiversidade de insetos, além de permitir reconhecer possíveis padrões de relação em suas comunidades. Foram utilizados, no presente estudo, os dados de espécies de dípteros necrófagos coletados na cidade de Rio Claro – SP (SOUZA, 2011; SOUZA; ZUBEN, 2012; SOUZA; ZUBEN, 2016) e, a partir deles, estão sendo gerados modelos de grafos e árvores de decisão. Com os grafos gerados, o objetivo foi compreender melhor a interação dos seguintes fatores: local de coleta, tipo de isca e estação do ano, na determinação de qual (quais) espécie(s) de dípteros necrófagos podem ser mais facilmente encontradas conjuntamente.

DÍPTEROS NECRÓFAGOS

Os dípteros compõem uma das quatro ordens megadiversas de insetos, com mais de 160 mil espécies descritas (PAPE et al., 2011). Esta diversidade está associada a uma enorme variedade de hábitos alimentares, incluindo a necrofagia, em que os adultos e/ou imaturos de dípteros utilizam o recurso encontrado (matéria orgânica de origem animal) como fonte alimentar ou estimulante para cópula e oviposição (SMITH, 1986). Muitas espécies de dípteros necrófagos possuem importância médica e veterinária, uma vez que podem atuar na veiculação de patógenos, tais como bactérias, fungos, vírus, protozoários e ovos de helmintos ao homem e aos animais, além de serem causadoras de miíases (CHOW, 1940; GREENBERG, 1971; GUIMARÃES; PAPAVERO, 1999). Adicionalmente, sua associação ao ambiente humano ocorre devido às moscas explorarem substâncias e resíduos orgânicos

produzidos pela atividade humana e animal, especialmente fezes e resíduos vegetais como fonte energética (MONTEIRO, 1995).

Apesar dos hábitos necrófagos de alguns dípteros muscóides constituírem uma ameaça à saúde humana e animal, este comportamento torna-os consumidores importantes de carcaças de vertebrados e recicladores de nutrientes na natureza (BYRD; CASTNER, 2001). A maioria das fêmeas das espécies ovipositam em carcaças, tornando-as importantes agentes consumidoras desses recursos (VARGAS; WOOD, 2010). Os representantes de dípteros necrófagos mais comumente associados à colonização de carcaças são os muscóides, dentre eles podem-se citar as famílias: Calliphoridae, Sarcophagidae e Muscidae (PUJOL-LUZ et al., 2008).

Em virtude do hábito alimentar descrito acima, tais insetos podem ser de grande utilidade nos processos de investigação criminal, permitindo aos peritos estimarem o intervalo pós-morte, através da presença e frequência desses insetos e também da bionomia dos seus estágios imaturos (AMENDT et al., 2007). Essa ciência é conhecida como Entomologia Forense, relacionando o conhecimento sobre insetos e outros artrópodes às diversas questões criminais (OLIVEIRA-COSTA, 2003; PUJOL-LUZ et al., 2008)

A distribuição, abundância e composição da fauna necrófaga e seus padrões de dinâmica populacional, diferem de acordo com a área geográfica (ANDERSON; VANLAERHOVEN, 1996; ARNALDOS et al., 2001, CARVALHO et al. 2004). As causas dessas diferenças certamente estão ligadas a fatores ambientais, tais como a temperatura, além dos fatores biológicos intrínsecos do organismo (SERRA et al., 2007). É relevante observar a influência das interações entre esses fatores sobre a composição da fauna de dípteros necrófagos, ainda mais no que se refere às moscas-varejeiras que são os primeiros artrópodes a colonizarem corpos em decomposição (SMITH, 1986).

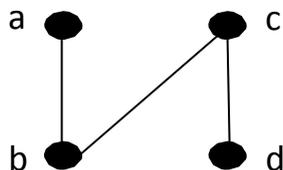
NOÇÕES SOBRE GRAFOS

Um grafo é uma estrutura de abstração bastante útil na representação e na solução de diversos tipos de problemas (GOLDBARG; GOLDBARG, 2012), sendo essas suas maiores vantagens. O termo grafo provém de uma expressão de notação gráfica e foi introduzido pela primeira vez pelo químico Edward Frankland e posteriormente usado por Alexander Crum Brown (1884) (FOULDS, 1992).

Usando uma definição simples, tem-se que um grafo é um modelo matemático que representa uma coleção de objetos (chamados vértices), que são ligados aos pares por outra

coleção de objetos (chamados arestas ou arcos). Na forma de ilustração, os vértices podem ser representados por pontos, caixas ou círculos, e as arestas por linhas conectando os vértices (Figura 1).

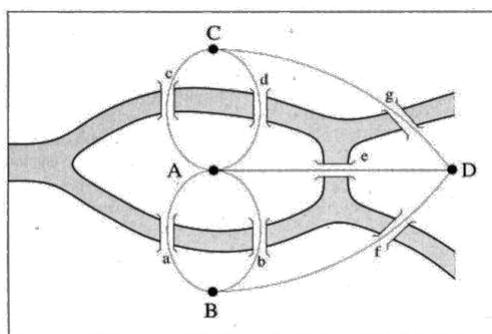
Figura 1 – Grafo com conjunto de vértices $V = \{a,b,c,d\}$ e arestas $A = \{ab, bc, cd\}$.



Grafo é um par ordenado de conjuntos disjuntos $(V; A)$ tal que A é um subconjunto do conjunto de pares não ordenados de V . O conjunto V é chamado de conjunto dos vértices (ou nós) e o conjunto A de conjunto das ligações (ou arestas). Diz-se que um elemento $(v1; v2)$ de A liga os vértices $v1$ e $v2$ e que $v1$ e $v2$ são vértices adjacentes. Duas ligações são adjacentes se elas compartilham um mesmo vértice (BALAKRISHNAN; RANGANATHAN, 2012).

O primeiro e mais famoso problema em Teoria dos Grafos foi o das Pontes de Königsberg (na Antiga Prússia) resolvido por Euler (1736). O quebra-cabeça famoso na época era encontrar um passeio que visitasse todas as pontes da cidade de Königsberg (Figura 2), passando uma única vez em cada ponte. Analisando esse diagrama abstrato, Euler provou que tal passeio era impossível.

Figura 2 - Diagrama das sete pontes de Königsberg



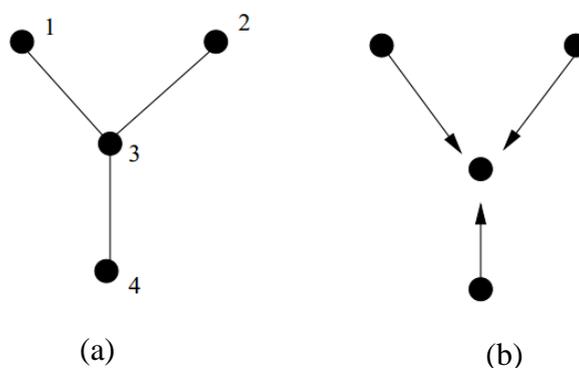
Fonte: Arquivo Escolar (2010).

A seguir, encontram-se algumas definições sobre grafos.

a) Grafos orientados e não orientados

Em alguns casos, os grafos podem possuir uma orientação específica em cada aresta. Os que incluem essa informação são ditos **grafos orientados** (Figura 3b) e aqueles que não a apresentam, são ditos **não orientados** (Figura 3a). Em um grafo orientado, as arestas são pares ordenados, ou seja, um vértice é considerado a “origem” da aresta e o outro seu “destino”. Nesse caso, as arestas possuem um sentido marcado por uma seta e recebem o nome de **arcos** (GOLDBARG; GOLDBARG, 2012).

Figura 3 - Representação gráfica à esquerda de um grafo não orientado (apenas uma aresta existe entre dois vértices) e à direita de um grafo orientado (mais de uma aresta é permitida entre dois vértices). Os círculos representam os vértices (nós) e os segmentos de reta representam as ligações (arestas). Os números referem-se aos vértices (nós).



b) Laço

Os nós constituintes de uma aresta podem ou não ser diferentes. No caso de não serem, a aresta forma um **laço** (Figura 4 - aresta no número 3). Algumas definições permitem laços no grafo; outras proíbem laços, exigindo que os dois extremos de cada aresta sejam vértices distintos. E quando as arestas ligam os mesmos pares de nós, elas são conhecidas como **arestas paralelas** (Figura 4 – arestas ligando os nós 4 e 5).

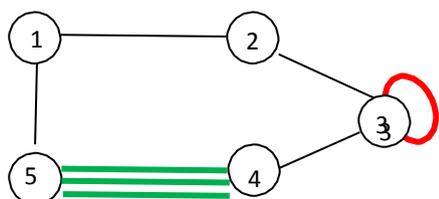


Figura 4 – A aresta no número 3 é um laço e as arestas entre os números 4 e 5 são paralelas múltiplas.

c) Grafo simples, pseudografo, multigrafo

Quando um grafo não possui laços nem arestas paralelas, ele é conhecido como **grafo simples**. Se houver no mínimo um laço, o grafo pode ser chamado de **pseudografo**. Se esse possuir todos os vértices com um laço associado, ele é denominado **pseudografo reflexivo**.

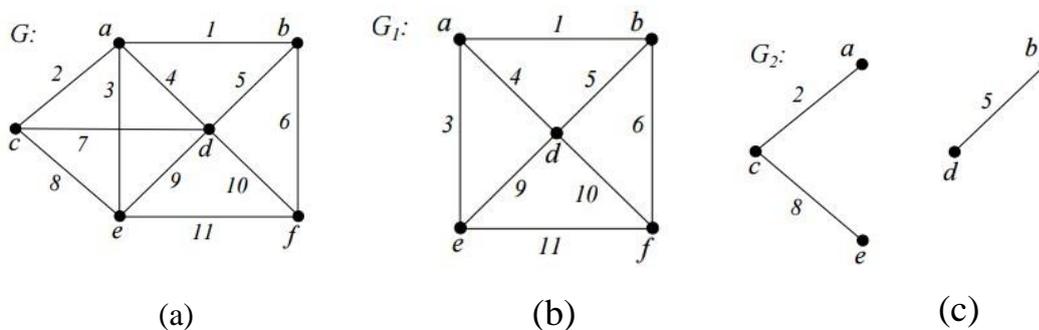
Um grafo não direcionado que possui no mínimo duas arestas paralelas é denominado **multigrafo**. Quando o multigrafo possuir dois ou mais arcos de mesma direção ligando um mesmo par de vértices, ele é conhecido como **multigrafo direcional** (GOLDBARG;GOLDBARG, 2012).

d) Subgrafo

Um grafo $H(V', A')$ é um subgrafo de um grafo $G(V, A)$ se todos os vértices e todas as arestas de H pertencem a G ($V' \subseteq V, A' \subseteq A$), e cada aresta de H possui as mesmas extremidades que em G (Figura 5). Denotamos um subgrafo através da mesma notação usada

para conjuntos, isto é $H \subset G$.

Figura 5 – Dado um grafo $G(a)$, os grafos $G_1(b)$ e $G_2(c)$ são subgrafos de G .



Fonte: OLIVEIRA; RANGEL, 2013.

Nesse contexto, podemos afirmar que:

- Todo grafo é um subgrafo de si próprio.
- Um subgrafo de um subgrafo de um grafo G também é um subgrafo de G .
- Um vértice de um grafo G é um subgrafo de G .

- Uma aresta de um grafo G é um subgrafo de G .

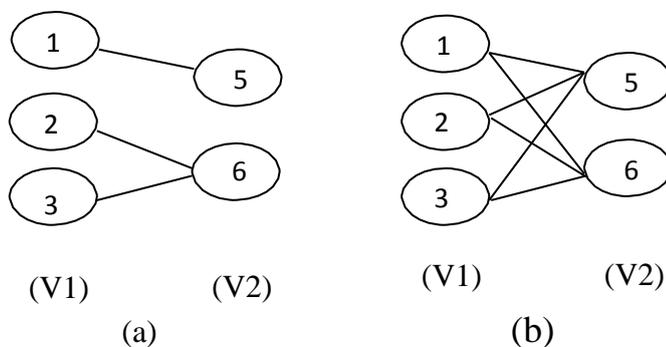
e) Grau do vértice

Dado um grafo G e um vértice $v \in V(G)$, designa-se por *grau* ou *valência* de v e denota-se por $d_G(v)$ como sendo o número de arestas de G incidentes em relação a v . Nesta definição, cada laço deve ser contado duas vezes.

f) Grafos bipartidos

Um grafo bipartido é aquele em que o conjunto V de vértices pode ser dividido em dois subconjuntos disjuntos V_1 e V_2 , tais que todas as arestas têm uma terminação em V_1 e uma em V_2 . Não há arestas entre elementos do mesmo conjunto. E um grafo bipartido completo ocorre quando todos os vértices de V_1 são ligados a todos os vértices de V_2 (Figura 6).

Figura 6 - Grafos (a) bipartido e (b) grafo bipartido completo.



g) Grafo Conexo

Um grafo é dito **conexo** se para todo par de vértices i e j existe pelo menos um caminho entre i e j (GOLDBARG ;GOLDBARG, 2012). Caso contrário o grafo é **desconexo**. Um grafo é **totalmente desconexo** quando não existe nenhuma aresta.

Em virtude da simetria dos grafos, a existência de um caminho de i a j equivale à existência de um caminho de j a i . Portanto, um grafo é conexo se e somente se quaisquer dois de seus vértices são ligados por um caminho.

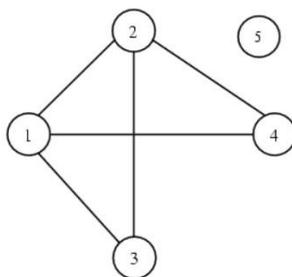
h) Passeio, Caminho e Ciclo

Um **passeio** em um grafo $F = (V, E)$ é uma seqüência alternada de vértices e arestas que começa e termina com vértices.

Um **caminho** é uma cadeia sem repetição de vértices (GOLDBARG; GOLDBARG, 2012). Ou seja, um caminho é um passeio que não possui nós repetidos. Na Figura 7, podemos notar que no grafo F , entre os nós 1 e 4, temos os seguintes caminhos $(1,4)$, $(1,2,4)$, $(1,3,2,4)$. E o **comprimento de um caminho** é o número de arcos que ele contém. Ciclos de comprimento 1 são laços (*loops*).

Um **caminho** considerado **fechado** é conhecido por **ciclo**, isto é, um passeio que contém exatamente dois nós iguais: o primeiro e o último. Um grafo sem ciclos é dito **acíclico**.

Figura 7 - Grafo F , para visualização de diferentes caminhos entre os nós 1 e 4.



Fonte: Prestes (2012)

i) Árvore

A **árvore** é um tipo especial de **grafo**, útil na representação de dados. Cada elemento (nó) tem zero ou mais sucessores, mas tem apenas um predecessor, exceto o primeiro nó, a

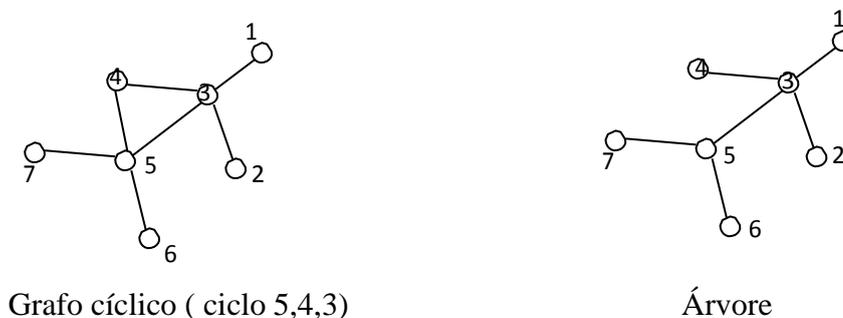
raiz da árvore. Ela é constituída por grafos que não admitem ciclos (**grafos acíclicos**) e que são **conexos** (GOLDBARG; GOLDBARG, 2012). Na Figura 8, têm-se uma comparação entre árvore e um grafo cíclico.

Para um grafo $G=(V, A)$ de n - vértices ($n > 0$), as seguintes afirmações são verdadeiras (e caracterizam uma árvore com n vértices):

- G é conexo e não possui ciclos;
- G é conexo e tem $n-1$ arestas;
- G tem $n-1$ arestas e nenhum ciclo;
- Para dois vértices u, v , há exatamente um caminho entre u e v .

Em outras palavras, uma árvore é um grafo constituído por um conjunto de nós e um conjunto de arcos que ligam pares de nós, em que: cada arco liga um **nó-pai** a um ou mais **nós-filhos** e todos os nós, com exceção da raiz, têm um nó-pai. Ao predecessor (único) de um nó, chama-se **nó-pai** e os seus sucessores são os **nós-filhos**. O **grau de um nó** é o número de sub-árvores (ou nós-filhos) que descendem desse nó. Um **nó-folha** não tem filhos, tem grau 0. Um **nó-raiz** não tem pai. Os arcos que ligam os nós chamam-se **ramos** (Figura 9).

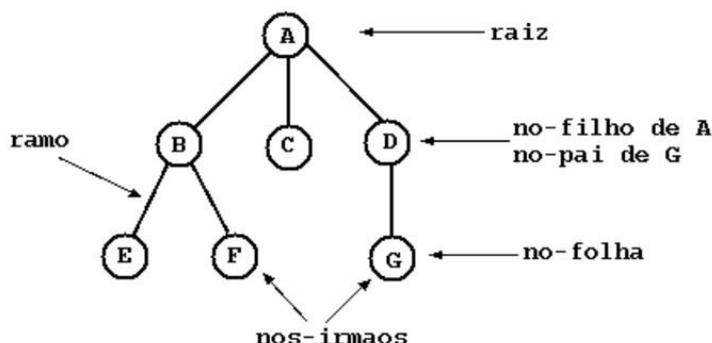
Figura 8- Comparação entre um grafo cíclico e uma árvore



ÁRVORES DE DECISÃO

Árvores de decisão são modelos estatísticos que utilizam um treinamento supervisionado para a classificação e previsão de dados. Em outras palavras, em sua construção é utilizado um conjunto de treinamento formado por entradas e saídas (são conhecidas como classes).

Figura 9 – Estrutura de uma árvore.



Embora diversos modelos matemáticos tenham surgido nas diferentes áreas do conhecimento relacionadas à classificação, as árvores de decisão têm sido consideradas como um dos modelos mais adequados para a extração do conhecimento para a Aprendizagem Automática, pois são simples e de fácil compreensão; e podem ser expressas numa sub-rotina em qualquer linguagem de programação (MICHIE et al., 1994).

Amplamente utilizadas em algoritmos de classificação, as árvores de decisão são representações simples do conhecimento, utilizam a estratégia de dividir para conquistar: um problema complexo é decomposto em sub-problemas mais simples e recursivamente esta técnica é aplicada a cada sub-problema. A classificação, através de uma árvore de decisão, ocorre à medida que são percorridos os caminhos descritos pelos nós (nodos), até ser encontrado um nó que contém a característica determinante do caminho seguido, recebendo então o nome de folha (GAMA, 2004).

O ponto principal das Árvores de Decisão é a tomada de decisões considerando para isso os atributos mais relevantes e compreensíveis para a maioria das pessoas. Ao escolher e apresentar os atributos em uma ordem de importância, as Árvores de Decisão permitem conhecer quais fatores que mais influenciam na sua construção.

Elas são muito úteis em atividades de mineração de dados, isto é, o processo de extração de informações previamente desconhecida, a partir de grandes bases de dados. Aplicações desta técnica podem ser vista em diversas áreas, desde cenários de negócios até sistemas de piloto automático de aeronaves e diagnósticos médicos. São também uma forma

de construir classificadores que predizem ou revelam classes ou informações úteis baseadas nos valores de atributos de um conjunto de dados.

Construção de uma árvore de decisão

A construção dos modelos computacionais de classificação emprega comumente um dos paradigmas relacionados abaixo (HEIJST et al., 1997) e eles podem ser melhor compreendidos através da Figura 10.

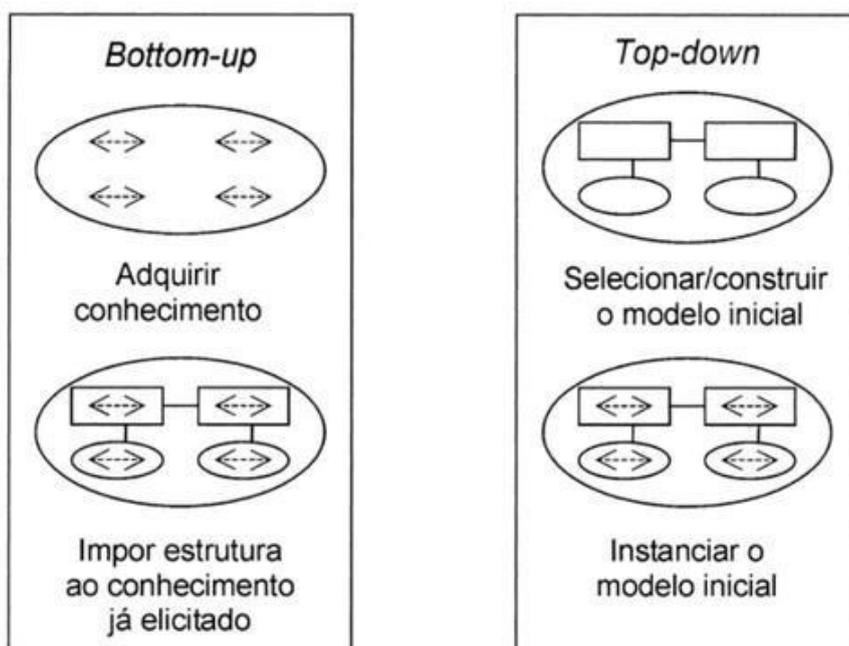
- *Top-down*: refere-se ao processo de construção do modelo de classificação a partir de informações fornecidas por especialistas;
- *Bottom-up*: refere-se ao processo de refinamento em que um modelo abstrato é selecionado ou construído, e depois instanciado com conhecimento específico da aplicação.

Como afirmado por Rezende (2003), os algoritmos que induzem árvores de decisão pertencem à família de algoritmos *Top Down Induction of Decision Trees* – TDIDT, ou seja, do nó raiz em direção às folhas. Embora haja diferenças significativas na forma de efetuar os passos, qualquer algoritmo desta categoria utiliza a técnica de dividir para conquistar. Como já explicado anteriormente, esta filosofia baseia-se na sucessiva divisão do problema em vários sub-problemas de menores dimensões, até que uma solução para cada um dos problemas mais simples seja encontrada (CASTANHEIRA, 2008). O TDIDT produz regras de decisão de forma implícita numa árvore de decisão, a qual é construída por sucessivas divisões dos exemplos de acordo com os valores de seus atributos preditivos. De acordo com BRAMER (2007), esse processo é conhecido como particionamento recursivo (uma árvore de decisão particiona recursivamente um conjunto de dados até que cada subconjunto obtido deste particionamento possua casos de uma única classe).

Para atingir este objetivo, esse tipo de algoritmo examina e compara a distribuição de classes durante o processo de construção da árvore de decisão. Através disso, os dados ficam organizados de maneira compacta, sendo utilizados para classificar novos casos (GOLDSCHMIDT, 2010). Uma vez que o número de árvores de decisão possíveis aumenta à medida que o número de atributos também aumenta, torna-se impraticável buscar a estrutura

da árvore de decisão ótima para um determinado problema, devido ao elevado custo computacional envolvido (problema NP-Completo). Nesse sentido, há algoritmos baseados em heurísticas que, mesmo não garantindo uma solução ótima, apresentam resultados aceitáveis (VON ZUBEN; ATTUX, 2010)

Figura 10 – Técnicas *bottom-up* e *top-down*.

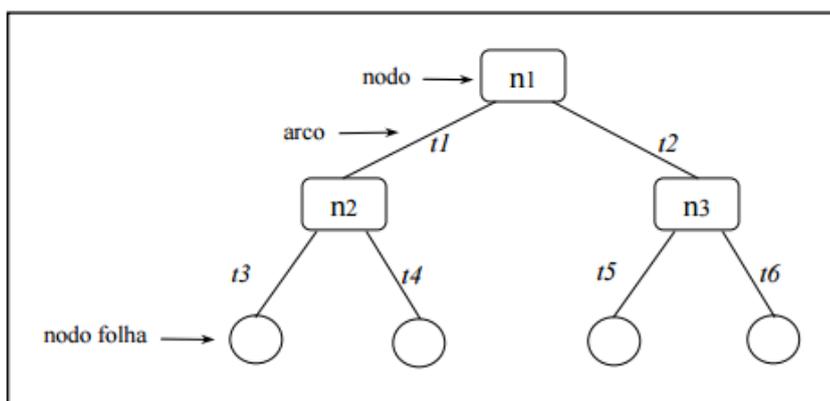


Fonte : Heijst; Shreiber; Wielinga (1997)

Representação e exemplo de uma árvore de decisão

A Figura 11, a seguir, representa uma árvore de decisão. Nota-se que cada nó (nodo) de decisão possui um teste para algum atributo e cada ramo descendente corresponde a um possível valor deste atributo. Além disso, o conjunto de ramos é distinto e cada nó folha está associada a uma classe. E a regra de classificação é cada percurso que árvore faz da raiz à folha (GAMA, 2004; GARCIA, 2003).

Figura 11 - Representação de uma árvore de decisão e sua respectiva representação no espaço.



Fonte: Gama (2004)

Na Figura 12, têm-se um exemplo de uma árvore de decisão para o diagnóstico de um paciente. Cada elipse é uma pergunta (teste) em um atributo para um dado conjunto de pacientes. Cada retângulo representa uma classe, nesse caso, o diagnóstico. Para classificar (diagnosticar) um paciente, é necessário começar pela raiz e seguir os testes até alcançar a folha.

Como as regras que representam uma árvore de decisão são disjuntas, isto é, apenas uma única regra dispara quando um novo exemplo é classificado, uma outra forma de representar tais regras é escrevê-las de forma separada para cada nó-folha, iniciando pela raiz (REZENDE, 2003). As regras de decisão, ou regras de classificação, são estruturas do tipo “Se <condição> então <conclusão>”, em que <condição> é um conjunto de atributos e seus valores e <conclusão> é uma classe do conjunto de dados.

Abaixo, têm-se as regras da árvore exemplificada anteriormente:

Se Paciente se sente bem = Sim **Então**

Classe = Saudável

Fim Se

Se Paciente se sente bem = não **and** Paciente tem dor= não

And Temperatura do Paciente ≤ 37 **Então**

Classe=Saudável

Fim Se

Se Paciente se sente bem = não **and** Paciente tem dor= não

And Temperatura do Paciente > 37 **Então**

Classe = Doente

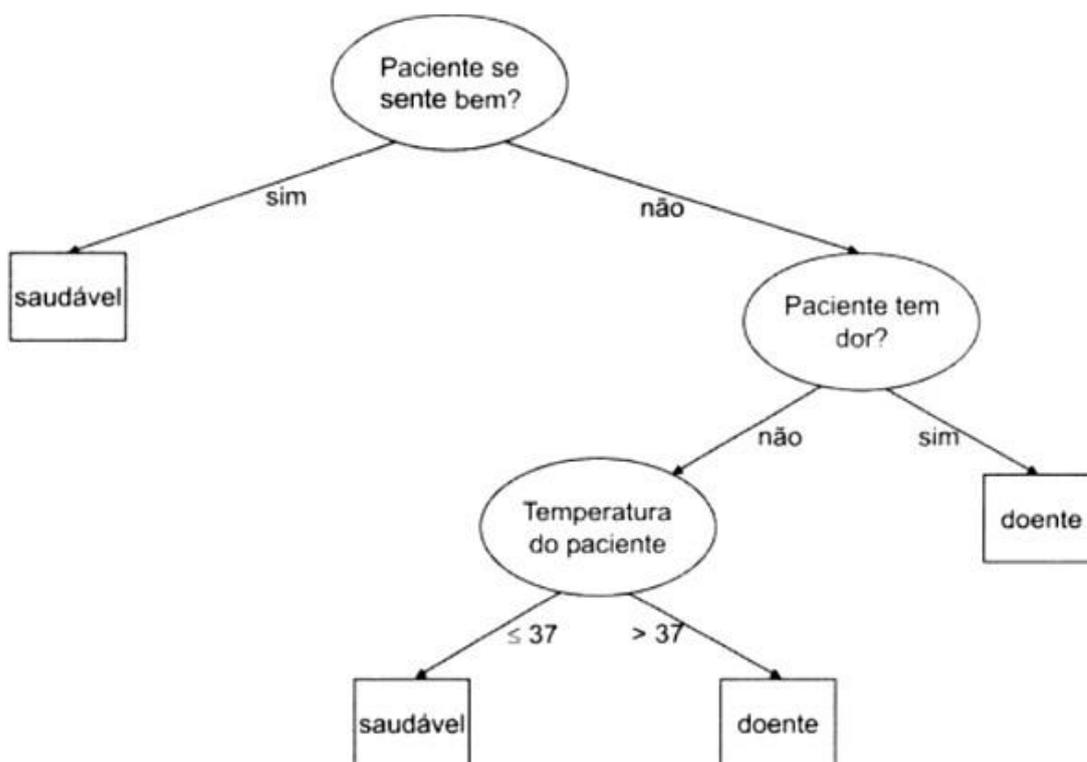
Fim Se

Se Paciente se sente bem = não **and** Paciente tem dor= sim

Classe = Doente

Fim Se

Figura 12- Uma árvore de decisão simples para o diagnóstico de um paciente



Fonte: Rezende (2003)

Indução de árvores de decisão

De uma forma geral, Quinlan (1986) afirma que se os atributos de uma árvore de decisão estão adequados, é sempre possível construir uma árvore de decisão que classifica de forma correta cada objeto no conjunto de treinamento. De acordo com Tan et al. (2005), durante o processo de construção de modelo através de um algoritmo de aprendizado, é necessário construí-lo com uma boa capacidade de generalização, ou seja, que consiga prever, com alta

taxa de acerto, rótulos de classes para exemplos que não foram utilizados na construção do modelo.

No intuito de fazer isso, a árvore de decisão deve capturar alguma relação significativa entre a classe de um objeto e os valores dos seus atributos. Dada a escolha entre duas árvores de decisão, sendo ambas corretas em relação ao conjunto de treinamento, parece sensato preferir a mais simples, que é a provável em melhor capturar a estrutura inerente ao problema. É esperado que árvore mais simples seria, portanto, a melhor em classificar corretamente mais objetos fora do conjunto de treinamento.

Poda

No processo de construção de árvores de decisão, algumas arestas ou sub-árvores podem refletir ruídos ou erros, gerando um problema conhecido como sobreajuste, que significa um aprendizado muito específico do conjunto de treinamento, não permitindo ao modelo generalizar. No intuito de se evitar isso, recorre-se à limitação e à redução do tamanho da árvore, respectivamente, através das abordagens de pré-poda e pós-poda (ESPOSITO et al., 1997), cujo objetivo é melhorar a taxa de acerto do modelo para novos exemplos, os quais não foram utilizados no conjunto de treinamento (HAN, 2001).

De acordo com Von Zuben; Attux (2010), o processo de podagem pode ser feito a partir dos seguintes passos:

1. Percorre a árvore em profundidade.
2. Para cada nó de decisão calcula: erro no nó • soma dos erros nos nós descendentes
3. Se o erro do nó é menor ou igual à soma dos erros dos nós descendentes então o nó é transformado em folha.

A medida da diferença dada por uma função baseada nas proporções das classes entre o nó corrente e seus descendentes valoriza a pureza das partições. Na pós-podagem a árvore é gerada no tamanho máximo e então a árvore é podada aplicando métodos de evolução confiáveis. A sub-árvore com o melhor desempenho será a escolhida. Este processo pode ser computacionalmente ineficiente pelo fato de gerar uma árvore muito grande e depois esta mesma árvore é reduzida a uma árvore mínima. Para interromper o crescimento da árvore, verifica-se se a divisão é confiável ou não. Caso seja confiável, interrompe-se o crescimento

da árvore. Este processo é conhecido como pré-podagem da árvore. A pré-podagem é mais rápida porém menos eficiente que a pós-podagem pelo fato do risco de interromper o crescimento da árvore ao selecionar uma árvore sub-ótima (BREIMAN, 1984).

A pós-poda é a abordagem mais utilizada e mais confiável, mas requer um processo mais lento, enquanto que a pré-poda tem a vantagem de não gastar tempo na construção de uma estrutura que não será utilizada no final da árvore. Nem sempre a árvore podada é mais precisa que a correspondente gerada, mas a poda ajuda a simplificar a árvore, o que é essencial em árvores muito complexas.

Algoritmos

Há muitos algoritmos de classificação que elaboram as árvores de decisão. Não há uma forma de determinar qual é o melhor, tendo em vista que o seu desempenho pode variar de acordo com a situação (GOLDSCHMIDT, 2010).

Os estudos que permitiram o aparecimento das árvores de decisão tiveram início com o professor Ross Quinlan da Universidade de Sidney. A sua contribuição foi a elaboração de um algoritmo chamado ID3 – Iterative Dichotomiser 3 - desenvolvido em 1983 (QUINLAM, 1993), baseado em sistemas de inferência e em conceitos de sistemas de aprendizagem. A partir de um conjunto de exemplos, ele constrói árvores de decisão que serão usadas para classificar amostras futuras. Além disso, o ID3 separa um conjunto de treinamento em subconjuntos, de forma que estes contenham exemplos de uma única classe. Essa divisão é feita através de um único atributo selecionado a partir de uma propriedade estatística, conhecida como **ganho de informação**, que mede quanto informativo é um atributo. Um ponto importante desse algoritmo é que ele é limitado, já que ele não trabalha com atributos do tipo contínuo (CASTANHEIRA, 2008).

O ID3 usa o **ganho de informação** para selecionar, entre os candidatos, os atributos que serão utilizados a cada passo, enquanto constrói a árvore. E para se obter o ganho de informação, é necessário, primeiramente, calcular a **entropia** que caracteriza a pureza/impureza dos dados: em um conjunto de dados e é uma medida da falta de homogeneidade dos dados de entrada em relação a sua classificação (MITCHELL, 1997). Nos casos em que a árvore é usada para classificação, os critérios de partição mais conhecidos são baseados na entropia.

Para determinar se uma condição de teste realizada é realmente boa, é preciso comparar o grau de entropia do nó-pai (antes da divisão) com o grau de entropia dos nós-

filhos (após a divisão). O atributo que gerar uma maior diferença é escolhido como condição de teste. O ganho é dado pela Equação (1), na forma:

$$\text{Ganho} = \text{entropia (pai)} - \sum_{j=1}^n \left[\frac{N(v_j)}{N} \text{entropia}(v_j) \right] \quad (1)$$

onde n é o número de valores do atributo, ou seja, o número de nós-filhos, N é o número total de objetos do nó-pai e $N(v_j)$ é o número de exemplos associados ao nó filho v_j .

O grau de **entropia** é definido pela Equação (2) a seguir:

$$\text{Entropia (nó)} = - \sum_{i=1}^c p(i/\text{nó}) \cdot \log_2[p(i/\text{nó})] \quad (2)$$

onde $p(i/\text{nó})$ é a fração dos registros pertencentes à classe i no nó, e c é o número de classes.

A entropia terá valor máximo (igual a 1) quando o conjunto de dados for heterogêneo, ou seja, quando x predizer totalmente y , e será 0 quando x e y não apresentarem nenhuma associação (MITCHEL, 1997). O critério de ganho seleciona como atributo-teste aquele que maximiza o ganho de informação. Ao se utilizar o ganho de informação, pode-se ter um grande problema: dar preferência a atributos com muitos valores possíveis (número de arestas). Buscando evitar isso, Quinlan (1993) criou a **Razão de Ganho** (do inglês Gain Ratio), que nada mais é do que o ganho de informação relativo (ponderado) como critério de avaliação. A **razão de ganho** é definida pela Equação (3), na forma:

$$\text{Razão de ganho (nó)} = \frac{\text{ganho}}{\text{entropia (nó)}} \quad (3)$$

Quinlan (1988) sugere ainda que isso seja feito em duas etapas: (1) calcular o ganho de informação para todos os atributos, considerando apenas aqueles atributos que obtiveram um ganho de informação acima da média, (2) a partir disso, escolher aquele que apresentar a melhor razão de ganho. Sendo assim, Quinlan mostrou que a razão de ganho supera o ganho de informação tanto em termos de acurácia quanto em termos de complexidade das árvores de decisão geradas.

O algoritmo C4.5 (QUINLAN, 1993), utilizado no presente trabalho, representa uma significativa evolução do ID3 (QUINLAN, 1986). Por ter mostrado excelentes resultados em problemas de classificação, ele é um dos algoritmos mais utilizados na literatura. As principais contribuições que ele trouxe em relação ao ID3 são (DE ANDRADE,2013):

- Trabalha tanto com atributos categóricos (ordinais ou não-ordinais) como também com atributos contínuos. Para lidar com os últimos, esse algoritmo C4.5 define um limiar e divide os exemplos de forma binária: aqueles cujo valor do atributo é maior que o limiar e aqueles cujo valor do atributo é menor ou igual ao limiar;
- O algoritmo C4.5 permite a utilização de valores desconhecidos que são representados como '?' e ele trata esses valores de forma especial. Esses valores não são utilizados nos cálculos de ganho e entropia;
- Utiliza a medida de razão de ganho para selecionar o atributo que melhor divide os exemplos. Essa medida se mostrou superior ao ganho de informação, gerando árvores mais precisas e menos complexas;
- Lida com problemas em que os atributos possuem custos diferenciados;
- Apresenta um método de pós-poda das árvores geradas. O algoritmo C4.5 faz uma busca na árvore, de baixo para cima, e transforma em nós folha aqueles ramos que não apresentam nenhum ganho significativo.

A ferramenta de mineração de dados WEKA (WITTEN & FRANK, 1999) (<http://www.cs.waikato.ac.nz/~ml/weka/index.html>) (*Waikato Environment for Knowledge Analysis*) disponibiliza a implementação do algoritmo C4.5, porém o mesmo é chamado de J48 nessa ferramenta. De acordo com De Andrade (2013), o C4.5 é do tipo:

- Guloso: executa sempre o melhor passo avaliado localmente, sem se preocupar se este passo, junto à sequência completa de passos, vai produzir a melhor solução ao final;
- “Dividir para conquistar”: partindo da raiz, criam-se sub-árvores até chegar nas folhas, o que implica em uma divisão hierárquica em múltiplos subproblemas de decisão, os quais tendem a ser mais simples que o problema original.

REFERÊNCIAS

AGNARSSON, G.; GREENLAW, R. **Graph Theory: Modeling, Applications, and Algorithms** Prentice Hall; 1 edition ,2006.

AMENDT, J. et al. Best practice in forensic entomology – standards and guidelines. **Int. J. Legal Med.**, v.121, p.90-104, 2007.

ANDERSON, G.; VANLAERHOVEN, S.H.. Initial studies on succession on carrion in the carrion in Southwestern British Columbia. **J. Foren. Sci.** 41: 617-625,1996.

ARAÚJO, J. M. F. R. **Inteligência Artificial I - Aprendizagem (Parte II)**, Disponível em http://www.dsc.ufcg.edu.br/~joseana/IAPos_NA16_2.pdf Acesso em 6 de janeiro de 2015.

ARNALDOS I, et al. An initial study on the succession of sarcosaprophagous Diptera (Insecta) on carrion in the southeastern Iberian Peninsula. **Int. J. of Legal Medi.** 114, 156-162, 2001.

ARQUIVO ESCOLAR. **Capítulo 3: Introdução a Teoria de Grafos.** 2010. Disponível em : http://arquivoescolar.org/bitstream/arquivo-e/45/2/metodos_finitos_II.pdf. Acesso em 10 jan 2014.

BALAKRISHNAN R; RANGANATHAN K (2012). **A textbook of graph theory.** Springer, Berlin.

BARABÁSI, A-L; et al. **Evolution of the Social Network of Scientific Collaborations.** Physica A, 311. 2002. pág. 590-614.

BARKOW, S. et al. **BicAT: a biclustering analysis toolbox.** Bioinformatics 22, 1282-1283, 2006.

BRAMER, M. **Principles of data mining.** Springer, London, 2007.

BREIMAN L, FRIEDMAN JH, OLSHEN RA, STONE CJ. **Classification and regression trees**. Pacific Grove: Cole Advanced Books and Software; 1984.

BYRD, J. H.; CASTNER, J. L. **Insects of forensic importance**, p. 43–80. In: J. H. Byrd & J. L. Castner (eds.). *Forensic Entomology: the utility of arthropods in legal investigations*. Boca Raton, CRC Press, xvi + 418 p, 2001

CARVALHO L.M.L., et al..Observations on the succession patterns of necrophagous insects on a pig carcass in an urban area of Southeastern Brazil. **Aggrawal's Int. J. Foren. Med. Toxic.** 5, 33-39, 2004.

CASTANHEIRA, L. C. **Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões**. UFMG, Belo Horizonte, 2008 (Dissertação). Disponível em: <https://online.unisc.br/seer/index.php/tecnologia/article/viewFile/3283/2692>, Acesso em 10 de abril de 2014.

CHARRAD M.; BEN AHMED, M.**Simultaneous clustering: A survey**. In: Kuznetsov SO, Mandal DP, Kundu MK, Pal SK (eds) *Pattern Recognition and Machine Intelligence, Lecture Notes in Computer Science*, vol 6744, Springer, Berlin Heidelberg, pp 370–375, DOI 10.1007/978-3-642-21786-9 60, 2011.

CHENG, Y.; CHURCH, G. M. **Biclustering of expression data**. In Proc. 8th Intl. Conf. on Intelligent Systems for Molecular Biology, pages 93–103. AAAI Press, 2000.

CHOW, C.Y. The commom blue bottle fly *Chrysomya megacephala* as a carrier of pathogenic bacteria in Peiping. **China Chin. Med.**, v.57, p.145-153, 1940.

DE ANDRADE, CESAR A. B. **Análise Automática de Malwares Utilizando as Técnicas de Sandbox e Aprendizado de Máquina**. Dissertação (Mestrado em Sistemas e Computação) - Instituto Militar de Engenharia, 2013.

DE FRANÇA, F. O. **Biclusterização na análise de dados incertos**. Tese de Doutorado, Faculdade de Engenharia Elétrica e de Computação, UNICAMP, Campinas, 2010.

DIEKS, D. (2009), “Bottom-Up and Top-Down: The Plurality of Explanation and Understanding in Science,” in H. de Regt, K. Eigner & S. Leonelli (eds.). **Scientific Understanding: Philosophical Perspectives**. Pittsburgh: University of Pittsburgh Press.

EASTMAN, J.R. Idrisi 15: The Andes Edition. Worcester, MA: Clark University, 2006.

EULER, L. Solutio problematis ad geometriam situs pertinentis, Comment. **Acad. Sci. Imp. Petropol.** 8 (1736), 128–140.

EISENBERG, D.; MARCOTTE, E. M.; XENARIOS, J.; YEATES, T. O. Protein function in the post-genomic era. **Nature**, 405:823826, 2000.

ESPOSITO, F.; MALERBA, D.; SEMERARO, G. A comparative analysis of methods for pruning decision trees. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, New York, v. 19, n. 5, p. 476-491, May 1997.

FOULDS, L. R. **Graph Theory Applications**. Springer-Verlag, 1992.

GAMA, J. (2004) **Árvores de Decisão, Universidade do Porto**. Disponível em http://www.liaad.up.pt/~jgama/Aulas_ECD/arv.pdf, acesso em 14/06/14.

GARCIA, S.C. **O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde**. UFRGS. Dissertação, 2003.

GOLDBARG, M.; GOLDBARG, E. **Grafos : conceitos, algoritmos e aplicações**. Tradução . Rio de Janeiro: Elsevier, 2012.

GOLDSCHMIDT, R. R.. **Uma Introdução à Inteligência Computacional: Fundamentos, Ferramentas e Aplicações**. 1 ed. Rio de Janeiro: IST-Rio, 2010. v. 1, p. 142.

GOMES, L. Entomologia Forense: Novas tendências e tecnologias nas ciências criminais . Rio de Janeiro. Ed. Technical Books, 2010.

GREENBERG, B. 1971. **Flies and disease**, Vol. 1. Princeton University Press, 856 pp.

GUIMARÃES J. H.; PAPAVERO, N. 1999. **Myiasis in man and animals in the Neotropical region**. Bibliographic database. São Paulo, FAPESP/ Editora Plêiade, 308p.

HAN, J. **Data Mining: Concepts and Techniques**. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

HEIJST, V. G.; SCHREIBER, A. T.; WIELINGA, B. J. Using explicit ontologies in KBS development. **International Journal of Human-Computer Studies** , v. 46, n. 2-3, p. 183-192, feb./mar 1997

JAIN, A. K., MURTY, M. N., FLYNN, P. J. **Data Clustering: A Review**. ACM Computing Surveys, vol. 31, no. 3, pp. 254-323, Sep., 1999.

LEWINSOHN, T. M.; PRADO, P. I.; JORDANO, P.; BASCOMPTE, J.; OLESEN, J. Structure in plant animal interactions assemblages. **Oikos**, 113: 174-184, 2006.

MADEIRA, S. C.; OLIVEIRA, A. L. (2004). **Biclustering algorithms for biological data analysis: A survey**. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 1(1):24–45.

MASON, O.; VERWOERD, M.. Graph theory and networks in Biology. **IET Systems Biology** . 1, 89-119, 2007.

MICHIE D., SPIEGELHALTER D. J., TAYLOR C. C. **Machine learning, Neural and Statistical Classification** . Ellis Horwood, 1994

MITCHELL, T.M. **Machine Learning**. WCB/McGraw-Hill, 1997.

MONTEIRO, R.M. **Microhimenópteros (Insecta: Hymenoptera) parasitóides e insetos predadores de moscas sinantrópicas (Insecta: Diptera) na Granja Capuavinha, Monte-Mor, SP**. Campinas: UNICAMP, 1995. 99p. (Dissertação, Mestrado).

NIE D.; YAN FU; JUNLIN ZHOU; YUKE FANG. Predicting Time Series with Multiple Mixed Models. **J. Comput. Informat. Systems** 7:4 (2011) 1092-1099.

OLIVEIRA, V A; RANGEL, S. 2013. **Teoria dos Grafos: Subgrafos, Operações com Grafos.** Disponível em:

http://www.ibilce.unesp.br/Home/Departamentos/MatematicaAplicada/socorro4029/aula2_operacoesrev2014.pdf. Acesso em ago 2015.

OLIVEIRA-COSTA, J. 2003. **Entomologia Forense: Quando os insetos são vestígios.** Editora Millenium, Campinas, 258p.

PAPE, T.; BLAGODEROV, V.; MOSTOVSKI, M. B. 2011. Order DIPTERA Linnaeus, 1758. In: Zhang, Z. Q. (Ed.) Animal biodiversity: An outline of higherlevel classification and survey of taxonomic richness. **Zootaxa**, 3148: 1-237.

PAINE, R .T. 1966. **Food web complexity and species diversity.** The American Naturalist 100: 65 - 75.

PRESTES, E. **Teoria dos Grafos.** (2012). Disponível em: <http://www.inf.ufrgs.br/~prestes/Courses/Graph%20Theory/GrafosA2.pdf>. Acesso em 10 de jan de 2014.

PUJOL-LUZ, J. R.; ARANTES, L. C.; CONSTANTINO, R. Cem anos da entomologia forense no Brasil (1908-2008). **Rev. Bras. Entomol.**, v.52, p.485-492, 2008.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações.** Ed. Manole, São Paulo, 2003.

QUINLAN, J. R. **Induction of Decision Trees**, in Machine Learning, Volume 1 , pages 81-106, 1986.

QUINLAN, J. **Decision trees and multivalued attributes.** Machine Intelligence, 11:305-318, 1988.

QUINLAN, J. R. **C4.5: Programs for machine learning.** San Mateo, CA: Morgan Kaufmann Publishers, 1993.

SERRA H, GODOY WAC, VON ZUBEN FJ, VON ZUBEN CJ, REIS SF (2007.) Sex ratio and dynamic behavior in populations of the exotic blowfly *Chrysomya albiceps* (Diptera, Calliphoridae). **Braz. J. Biol.** 67, 347-353.

SMITH, KGV, 1986. **A manual of forensic entomology**. Cornell Univ. Press, Ithaca, NT.

SOUZA, C.R. **Sazonalidade, sinantropia e preferência por iscas de dípteros necrófagos da região de Rio Claro, SP**. 2011. 60 f. Dissertação - (mestrado) - Universidade Estadual Paulista, Instituto de Biociências de Rio Claro, 2011. Disponível em: <<http://hdl.handle.net/11449/99577>>. Acesso em 10 fev 2013.

SOUZA , C. R.; ZUBEN , C. J. V. 2012. Diversity and synanthropy of Calliphoridae (Diptera) in the region of Rio Claro, SP, Brazil. **Neot. Entomol.** 41 :243-248.

SOUZA , C. R.; ZUBEN , C. J. V. 2016. Synanthropy of Sarcophagidae (Diptera) in southeastern Brazil. **Neot. Entomol.** Doi: 10.1007/s13744-016-0411-0.

TAN, P. N.; STEINBACH, M.; KUMAR,V. 2005. **Introduction to Data Mining** (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

VARGAS, J.; WOOD, D. M. 2010. Calliphoridae, p. 1297–1304. In: B. V. Brown, A. Borkent, J. M. Cumming, D. M. Wood, N. E. Woodley & M. A. Zumbado (eds.). **Manual of Central American Diptera**. Vol. 2. Canada, Ontario, NCR Research Press, 728 p.

VON ZUBEN, F. J.; ATTUX, R. R. F. **Árvores de Decisão**, Tópico 7 das Notas de Aula do Curso de Pós-Graduação IA004 - Redes Neurais II, FEEC/Unicamp, 2010. Disponível em ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia004_1s10/notas_de_aula/topico7_IA004_1s10.pdf, Acesso em 10 de setembro de 2012.

WITTEN, I.; FRANK, E. **Practical machine learning tools and techniques with Java implementations**. San Francisco, CA: Morgan Kaufmann, 1999.