

Universidade Federal do Rio de Janeiro

**SISTEMAS *FUZZY* HIERÁRQUICOS
COM ANÁLISE ESPECTRAL APLICADOS À
CLASSIFICAÇÃO SUPERVISIONADA**

Gustavo Eduardo Carnaval Barbosa

2012



SISTEMAS *FUZZY* HIERÁRQUICOS COM ANÁLISE ESPECTRAL APLICADOS À
CLASSIFICAÇÃO SUPERVISIONADA

Gustavo Eduardo Carnaval Barbosa

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Civil.

Orientador: Alexandre Gonçalves Evsukoff

Rio de Janeiro
Setembro de 2012

SISTEMAS *FUZZY* HIERÁRQUICOS COM ANÁLISE ESPECTRAL APLICADOS À
CLASSIFICAÇÃO SUPERVISIONADA

Gustavo Eduardo Carnaval Barbosa

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ
COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS
NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM
ENGENHARIA CIVIL.

Examinada por:

Prof. Alexandre Gonçalves Evsukoff, D.Sc.

Prof. Antônio Carlos Saraiva Branco, D.Sc.

Prof. Beatriz de Souza Leite Pires Lima, D.Sc.

Prof. Marley Maria Bernardes Rebuzzi Vellasco, Ph.D.

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2012

Barbosa, Gustavo Eduardo Carnaval

Sistemas Fuzzy Hierárquicos com Análise Espectral Aplicados à Classificação Supervisionada/Gustavo Eduardo Carnaval Barbosa. – Rio de Janeiro: UFRJ/COPPE, 2012.

XIII, 87 p.: il.; 29,7cm.

Orientador: Alexandre Gonçalves Evsukoff

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Civil, 2012.

Referências Bibliográficas: p. 49 – 51.

1. Sistemas *Fuzzy*. 2. Classificação Supervisionada. 3. Análise Espectral. 4. Modelos Hierarquizados. I. Evsukoff, Alexandre Gonçalves. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

"The closer one looks at a real-world problem, the fuzzier becomes its solution."

Zadeh

AGRADECIMENTOS

À minha família, pelo apoio incondicional em todos os momentos. À minha mãe (Maria Aparecida) pelo carinho e incentivo, minha irmã (Maíra) por todas as ajudas que já me fizeram perder a conta e à minha noiva (quase esposa) Liana, que é quem me dá forças para enfrentar o dia-a-dia, por toda a ajuda. Ao meu falecido pai (Eduardo) pelo exemplo de caráter, honestidade e humildade, com quem tive oportunidade de conviver por tão pouco tempo.

Ao meu orientador Alexandre, pela paciência e por toda a ajuda e apoio, mesmo com meus sumiços empresariais. Ao meu co-orientador, oficial ou não, Branco por toda a ajuda. E também à M^{lle} Galichet, que tive a oportunidade de conhecer, por ter feito juntamente com meu orientador e co-orientador o trabalho que é base desta dissertação.

Ao Programa de Engenharia Civil (PEC), à COPPE e à UFRJ, pela oportunidade de ter passado por um curso de mestrado com excelência em qualidade. Aos professores do PEC e da Escola Nacional de Ciências Estatísticas (ENCE), que me fizeram obter o conhecimento suficiente para a realização desta dissertação. À equipe do laboratório de informática do PEC, pela disponibilização do espaço e equipamentos.

À Brasilcap e Grupo Virtual, pela autorização de cursar o mestrado sem exigências ou muitas restrições e por me liberar muitas vezes para as atividades acadêmicas na UFRJ. Ao meu caro Watson, pela amizade e por segurar as pontas em ambas as empresas quando eu estava ausente.

A todos que contribuíram direta ou indiretamente para a conclusão deste trabalho.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários à obtenção do título de Mestre em Ciências (M. Sc.)

SISTEMAS *FUZZY* HIERÁRQUICOS COM ANÁLISE ESPECTRAL
APLICADOS À CLASSIFICAÇÃO SUPERVISIONADA

Gustavo Eduardo Carnaval Barbosa

Setembro/2012

Orientador: Alexandre Gonçalves Evsukoff

Programa: Engenharia Civil

Este trabalho apresenta uma metodologia para a geração de uma hierarquia de modelos de classificação supervisionada de forma que os modelos alocados em níveis hierarquicamente superiores sejam capazes de generalizar os modelos alocados em níveis inferiores. A metodologia fornece informações suficientes para a escolha do nível ótimo, podendo ponderar a relação entre acurácia e interpretabilidade do modelo. A quantidade de regras utilizadas no sistema *fuzzy* é gerada a partir da aplicação da análise espectral e os pesos de regra são calculados a partir da resolução de um problema de otimização quadrática. Os níveis hierárquicos são ajustados por um parâmetro que regula a similaridade entre os registros de treinamento, que influencia no número de regras obtido pela análise espectral. A metodologia foi aplicada a dez diferentes bases de *benchmark* e os resultados mostram que é favorável a análise da relação entre acurácia e interpretação a partir do conjunto de modelos hierárquicos apresentados.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

HIERARCHICAL FUZZY SYSTEMS WITH SPECTRAL ANALYSIS
FOR SUPERVISED CLASSIFICATION

Gustavo Eduardo Carnaval Barbosa

September/2012

Advisor: Alexandre Gonçalves Evsukoff

Department: Civil Engineering

This work presents a methodology that generates a hierarchy of models for supervised classification in the sense that the models allocated at higher hierarchical levels are able to generalize the models allocated at lower levels. The methodology provides enough information to choose the optimal level considering the balance between accuracy and interpretability of the model. The number of rules used in the fuzzy system is generated from the application of spectral analysis and the rule's weights are calculated by solving a quadratic optimization problem. The hierarchical levels are adjusted by a parameter that regulates the similarity between training data, which influences the number of rules obtained by spectral analysis. The methodology was applied to ten different benchmark datasets and the results show that it is favorable to analyze the balance between accuracy and interpretation from the set of hierarchical models presented.

SUMÁRIO

CAPÍTULO 1 – Introdução	1
1.1 Apresentação e Objetivo	1
1.2 Motivação	3
1.3 Estrutura da Dissertação	4
CAPÍTULO 2 – Metodologia	5
2.1 Classificação Supervisionada.....	5
2.2 Análise Espectral	5
2.3 Modelo <i>Fuzzy</i> Simbólico.....	8
2.3.1. Etapa de <i>Fuzzyficação</i>	9
2.3.2. Etapa de Inferência.....	10
2.3.3. Etapa de <i>Defuzzyficação</i>	11
2.4 Indução de Regras <i>Fuzzy</i>	11
2.5 Otimização de Pesos de Regras <i>Fuzzy</i>	13
2.6 Critérios de Avaliação de Classificação	14
CAPÍTULO 3 – Metodologia Proposta.....	15
3.1 Seleção de Modelos	15
3.2 Hierarquização	18
3.3 Visualização Completa dos Modelos.....	20
3.4 Algoritmo Proposto.....	20
CAPÍTULO 4 – Aplicação e Resultados	22
4.1 Descrição das Bases de <i>Benchmark</i>	22
4.2 Apresentação e Discussão dos Resultados Obtidos.....	23
4.2.1. Resultados da Base Iris	24
4.2.2. Resultados da Base Balance	26
4.2.3. Resultados da Base Diabetes.....	28

4.2.4.	Resultados da Base Cancer.....	30
4.2.5.	Resultados da Base Glass.....	32
4.2.6.	Resultados da Base Wine.....	34
4.2.7.	Resultados da Base Heart.....	36
4.2.8.	Resultados da Base Image.....	38
4.2.9.	Resultados da Base Ionosphere.....	40
4.2.10.	Resultados da Base Sonar.....	42
4.3	Dendrogramas.....	44
CAPÍTULO 5 – Conclusão e Próximos Passos.....		48
REFERÊNCIAS BIBLIOGRÁFICAS.....		49
APÊNDICE A – Descrição Detalhada das Bases de <i>Benchmark</i>		52
A.1.	Descrição da Base Iris.....	53
A.2.	Descrição da Base Balance.....	55
A.3.	Descrição da Base Diabetes.....	58
A.4.	Descrição da Base Cancer.....	61
A.5.	Descrição da Base Glass.....	64
A.6.	Descrição da Base Wine.....	67
A.7.	Descrição da Base Heart.....	69
A.8.	Descrição da Base Image.....	72
A.9.	Descrição da Base Ionosphere.....	74
A.10.	Descrição da Base Sonar.....	76
APÊNDICE B – Apresentação dos Gráficos de Melhor Modelo.....		80
APÊNDICE C – Algoritmo Para o Dendrograma no MATLAB.....		86

ÍNDICE DE FIGURAS

Figura 1 – Processo de aplicação da classificação supervisionada.	2
Figura 2 – Estrutura da metodologia proposta.	15
Figura 3 – Exemplo de relação entre a variação da dispersão e a quantidade de regras. ...	16
Figura 4 – Exemplo de dendrograma.	18
Figura 5 – Exemplo de Hierarquia de Modelos.....	19
Figura 6 – Forma de visualização completa.	20
Figura 7 – Relação do parâmetro de dispersão com resultados da base Iris.	24
Figura 8 – Visualização completa dos modelos da base Iris.	25
Figura 9 – Relação do parâmetro de dispersão com resultados da base Balance.	26
Figura 10 – Visualização completa dos modelos da base Balance.	27
Figura 11 – Relação do parâmetro de dispersão com resultados da base Diabetes.....	28
Figura 12 – Visualização completa dos modelos da base Diabetes.....	29
Figura 13 – Relação do parâmetro de dispersão com resultados da base Cancer.....	30
Figura 14 – Visualização completa dos modelos da base Cancer.	31
Figura 15 – Relação do parâmetro de dispersão com resultados da base Glass.....	32
Figura 16 – Visualização completa dos modelos da base Glass.....	33
Figura 17 – Relação do parâmetro de dispersão com resultados da base Wine.....	34
Figura 18 – Visualização completa dos modelos da base Wine.	35
Figura 19 – Relação do parâmetro de dispersão com resultados da base Heart.....	36
Figura 20 – Visualização completa dos modelos da base Heart.....	37
Figura 21 – Relação do parâmetro de dispersão com resultados da base Image.	38
Figura 22 – Visualização completa dos modelos da base Image.	39
Figura 23 – Relação do parâmetro de dispersão com resultados da base Ionosphere.	40
Figura 24 – Visualização completa dos modelos da base Ionosphere.	41
Figura 25 – Relação do parâmetro de dispersão com resultados da base Sonar.	42
Figura 26 – Visualização completa dos modelos da base Sonar.	43
Figura 27 – Dendrograma da Base Iris.....	45
Figura 28 – Dendrograma da Base Balance.....	46
Figura 29 – Dendrograma da Base Wine.....	47
Figura 30 – Gráfico de projeção da base Iris.....	53

Figura 31 – <i>Data Image</i> da base Iris.....	54
Figura 32 – Matriz de Correlação da base Iris.....	54
Figura 33 – <i>Scatter Plot</i> da base Iris.....	55
Figura 34 – Gráfico de projeção da base Balance.....	56
Figura 35 – <i>Data Image</i> da base Balance.....	57
Figura 36 – Matriz de Correlação da base Balance.....	57
Figura 37 – <i>Scatter Plot</i> da base Balance.....	58
Figura 38 – Gráfico de projeção da base Diabetes.....	59
Figura 39 – <i>Data Image</i> da base Diabetes.....	60
Figura 40 – Matriz de Correlação da base Diabetes.....	60
Figura 41 – <i>Scatter Plot</i> da base Diabetes.....	61
Figura 42 – Gráfico de projeção da base Cancer.....	62
Figura 43 – <i>Data Image</i> da base Cancer.....	63
Figura 44 – Matriz de Correlação da base Cancer.....	63
Figura 45 – Gráfico de projeção da base Glass.....	65
Figura 46 – <i>Data Image</i> da base Glass.....	66
Figura 47 – Matriz de Correlação da base Glass.....	66
Figura 48 – Gráfico de projeção da base Wine.....	68
Figura 49 – <i>Data Image</i> da base Wine.....	68
Figura 50 – Matriz de Correlação da base Wine.....	69
Figura 51 – Gráfico de projeção da base Heart.....	70
Figura 52 – <i>Data Image</i> da base Heart.....	71
Figura 53 – Matriz de Correlação da base Heart.....	71
Figura 54 – Gráfico de projeção da base Image.....	72
Figura 55 – <i>Data Image</i> da base Image.....	73
Figura 56 – Matriz de Correlação da base Image.....	73
Figura 57 – Gráfico de projeção da base Ionosphere.....	75
Figura 58 – <i>Data Image</i> da base Ionosphere.....	75
Figura 59 – Matriz de Correlação da base Ionosphere.....	76
Figura 60 – Gráfico de projeção da base Sonar.....	77
Figura 61 – <i>Data Image</i> da base Sonar.....	78
Figura 62 – Matriz de Correlação da base Sonar.....	78

Figura 63 – Projeção da base Iris com os centros de regra do melhor modelo.	80
Figura 64 – Projeção da base Balance com os centros de regra do melhor modelo.	81
Figura 65 – Projeção da base Diabetes com os centros de regra do melhor modelo.	81
Figura 66 – Projeção da base Cancer com os centros de regra do melhor modelo.	82
Figura 67 – Projeção da base Glass com os centros de regra do melhor modelo.	82
Figura 68 – Projeção da base Wine com os centros de regra do melhor modelo.	83
Figura 69 – Projeção da base Heart com os centros de regra do melhor modelo.	83
Figura 70 – Projeção da base Image com os centros de regra do melhor modelo.	84
Figura 71 – Projeção da base Ionosphere com os centros de regra do melhor modelo.	84
Figura 72 – Projeção da base Sonar com os centros de regra do melhor modelo.	85

ÍNDICE DE TABELAS

Tabela 1 – Características gerais das bases de <i>Benchmark</i>	22
Tabela 2 – Melhores modelos encontrados no intervalo de dispersões	23
Tabela 3 – Parâmetros Dendrograma Base Iris.....	44
Tabela 4 – Parâmetros Dendrograma Base Balance.....	45
Tabela 5 – Parâmetros Dendrograma Base Wine.....	47
Tabela 6 – Características da base Iris.	53
Tabela 7 – Características da base Balance.	56
Tabela 8 – Características da base Diabetes.	59
Tabela 9 – Características da base Cancer.	62
Tabela 10 – Características da base Glass.....	64
Tabela 11 – Características da base Wine.....	67
Tabela 12 – Características da base Heart.....	70
Tabela 13 – Características da base Image.	72
Tabela 14 – Características da base Ionosphere.	74
Tabela 15 – Características da base Sonar.	77

CAPÍTULO 1 – Introdução

1.1 Apresentação e Objetivo

Durante o avanço da computação ocorrido nas últimas cinco décadas, é constante a necessidade de criação e aprimoramento de técnicas que permitam a manipulação e análise de dados com o objetivo de extração de informação relevante. Essa necessidade, com o volume e a diversidade de dados gerados atualmente, tornou-se multifatorial, uma vez que a discussão sobre a geração de qualquer novo algoritmo envolve diversas características tais como: precisão, escalabilidade, robustez, performance e interpretabilidade. A diversidade ocorre uma vez que várias áreas da ciência são envolvidas com diferentes origens de informação, seja através de obtenção por meio de sensores, por armazenamento sistêmico, pela extração da Web ou até mesmo pela simulação computacional.

É nesse contexto que a mineração de dados (ou Data Mining) se insere, e foi bem definida por HAN e KAMBER (2006) como sendo a extração automática de padrões que representam o conhecimento armazenado em grandes bases de dados. Em outras palavras, representa o processo de extração de informações implícitas, previamente desconhecidas e potencialmente úteis a partir de dados. A ideia seria a de se construir programas que consigam explorar bases de dados automaticamente, em busca de regularidades ou padrões. Padrões fortes encontrados provavelmente podem ser generalizados para que se façam previsões sobre dados futuros.

As atividades de mineração de dados são basicamente divididas em dois tipos: preditivas e descritivas. A classificação supervisionada e a regressão são exemplos de atividades preditivas assim como as regras de associação, o agrupamento (ou data clustering, ou ainda classificação não supervisionada) e a sumarização são exemplos de atividades descritivas. A classificação supervisionada, atividade explorada neste trabalho, tem por objetivo o desenvolvimento de um modelo (classificador) capaz de classificar um determinado objeto segundo classes previamente estabelecidas. Inicialmente, o conjunto de

dados é normalmente particionado em um “conjunto de treinamento ou aprendizagem” e um “conjunto de teste”. O ajuste do modelo é realizado no conjunto de treinamento, onde a classificação de cada registro é conhecida. O modelo é então avaliado em termos de acurácia, ou seja, por seu desempenho em classificar corretamente o conjunto de teste, composto por registros cuja classificação não é conhecida. Esse processo é ilustrado na *Figura 1*.

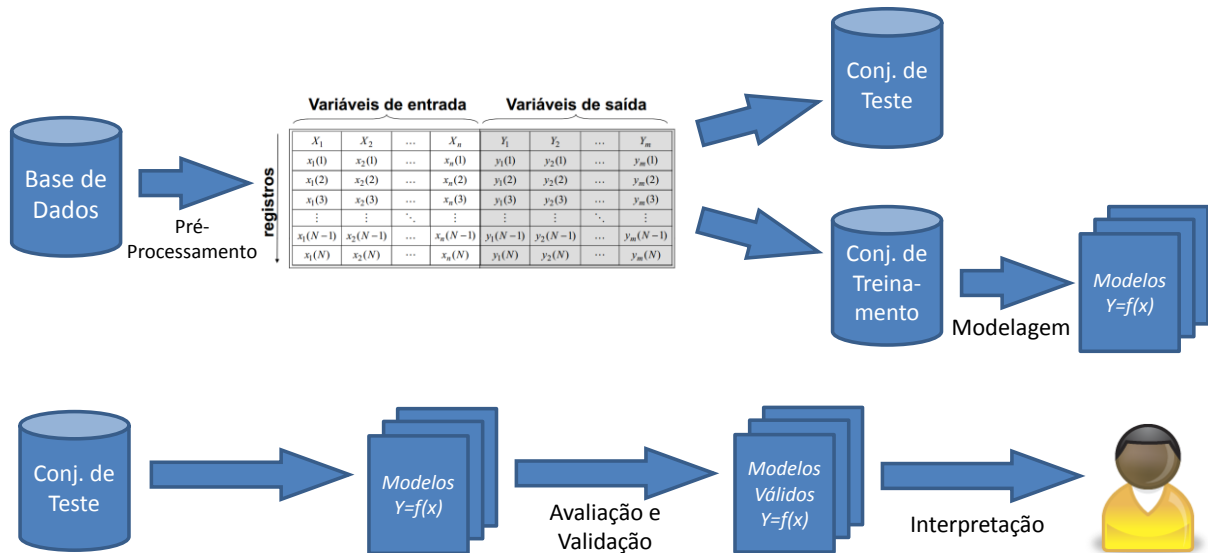


Figura 1 – Processo de aplicação da classificação supervisionada.

Apesar dos modelos geralmente serem desenvolvidos com foco na acurácia, a interpretabilidade do modelo também deve ser levada em consideração. A interpretabilidade refere-se ao nível de compreensão e discernimento que é fornecida sobre os dados pelo classificador. Por ser subjetiva, torna-se mais difícil de ser avaliada (HAN e KAMBER, 2006). Uma metodologia utilizada na classificação supervisionada que tem o seu fator de interpretabilidade discutido na literatura é a rede neural, pela complexidade envolvida na estrutura matemática do modelo de utilizar combinação linear de funções não lineares, apesar de haver consenso sobre o seu bom poder de acurácia em diversas aplicações. Essa relação entre acurácia e interpretabilidade foi definida por Zadeh já em 1973 como o princípio da incompatibilidade. À medida que a complexidade do problema aumenta, acurácia e interpretabilidade tornam-se fatores incompatíveis. Assim, a complexidade (no sentido de não-interpretabilidade) do modelo é função da complexidade do problema e/ou da qualidade dos dados disponíveis.

Portanto, o objetivo deste trabalho é criar uma metodologia que gere uma hierarquia de modelos no sentido de que os modelos alocados em níveis hierarquicamente superiores sejam capazes de generalizar os modelos alocados em níveis inferiores, em uma estrutura de modelos sequenciais que mantém a informação de relacionamento direto entre eles. Associadas a esse conjunto de modelos, serão apresentadas sempre três informações: a acurácia do classificador, a quantidade de grupos utilizados no modelo, que mantém relação com a interpretabilidade e um parâmetro de dispersão utilizado pelas funções de similaridade, de modo que se mantenha a capacidade de interpretação da variabilidade dos dados em função da aplicação da metodologia.

A estrutura da metodologia proposta é descrita a seguir: a quantidade de regras *fuzzy* é estimada através da análise espectral e, a seguir, os centros de regra são calculados a partir de um agrupamento utilizando o algoritmo *K-Means* (CASTRO, 2002; JAIN *et al.*, 1999; JAIN, 2010; XU, 2005). A partir dos centros, são extraídas as regras *fuzzy* de um sistema *fuzzy* simbólico e os pesos das regras são otimizados através da aplicação da otimização quadrática. A partir da aplicação dessas etapas em um intervalo de parâmetros de dispersão distintos, é possível selecionar o parâmetro de dispersão ótimo, que resulta no modelo de melhor acurácia. O conjunto de modelos com quantidade de regras inferior à quantidade do modelo ótimo em termos de acurácia é hierarquizado e é apresentado sob a forma de um gráfico que facilite a visualização. O usuário da metodologia terá então informações suficientes para escolher o nível com que quer trabalhar, podendo ponderar a relação entre acurácia e interpretabilidade do modelo e dos dados.

1.2 Motivação

A metodologia do trabalho é baseada na proposta de modelo feita por EVSUKOFF *et al.* (2011), que abrange da etapa de análise espectral à etapa de otimização quadrática apresentadas anteriormente, com diferenciação na avaliação e apresentação de resultados de um intervalo de dispersão e não na estimação de valor para a dispersão. A utilização da função Gaussiana para o cálculo de similaridade e também para o cálculo de pertinências foi preservada.

A utilização da análise espectral na metodologia que envolve a atividade de classificação é bem aceita na literatura a partir do que foi observado na revisão bibliográfica realizada. NG *et al.* (2002) mostraram que para se classificar pontos em \mathbb{R}^n as metodologias tradicionais são baseadas em algoritmos usados para aprendizagem de densidades mistas. Essa abordagem sofre de alguns pontos negativos: o primeiro é que usar estimadores de densidades paramétricas geralmente requer considerações fortes como, por exemplo, a respeito da função de densidade de cada grupo. O segundo, é que a função de log-verossimilhança tem como característica a multiplicidade de ótimos locais e, portanto, várias inicializações do algoritmo iterativo são necessárias.

Enquanto isso, as metodologias que utilizam a análise espectral utilizam os autovetores da matriz gerada pelo cálculo de distância entre cada par de registros da base dados, que seria equivalente à estruturação de um grafo a partir dos registros da base de dados. Ainda assim, essas metodologias caem no problema chamado de corte em grafo ou particionamento, que é NP-completo. Portanto, a abordagem que utiliza a análise espectral somente com o objetivo de estimação da quantidade de regras a ser considerada por um modelo *fuzzy* simbólico tornou-se a base deste trabalho. Ainda assim, foi identificada a necessidade de geração de uma metodologia que também explicitasse a relação entre a variação dos parâmetros de dispersão utilizados com os respectivos impactos em termos de acurácia.

1.3 Estrutura da Dissertação

No capítulo 2 deste trabalho, são detalhados os principais componentes da metodologia assim como a descrição matemática do problema de classificação supervisionada, a análise espectral e o sistema *fuzzy* utilizado. No capítulo 3, são discutidos conceitos relacionados a formas de visualização de conjuntos de modelos bem como a abordagem de hierarquização utilizada. No capítulo 4, são apresentadas as bases de *benchmark* utilizadas com a metodologia e também é feita uma análise dos resultados obtidos. Finalmente, no capítulo 5, são apresentadas as conclusões do trabalho.

CAPÍTULO 2 – Metodologia

2.1 Classificação Supervisionada

A atividade de classificação supervisionada, definida anteriormente de modo conceitual, é utilizada em problemas que envolvem a predição de variáveis categorizadas, como é o caso deste trabalho. Merece destaque o fato de que as variáveis categorizadas também podem ser numéricas com valores discretos, apesar da ordem dos valores não ter significado.

As soluções possíveis de um problema de classificação supervisionada são representadas pelos elementos do conjunto de classes ou grupos $\Omega = \{C_j, j = 1 \dots m\}$. A atividade consiste em identificar a classe correta de um conjunto de observações, representadas pelo vetor de atributos $\mathbf{x} = (x_1, \dots, x_n)$. O conjunto de treinamento é geralmente representado pela matriz \mathbf{X} , na qual cada linha $\mathbf{x}(t) = [x_1(t), \dots, x_n(t)]$ está associada a um registro t , $t = 1, \dots, N$, cuja classe é conhecida.

2.2 Análise Espectral

O teorema de decomposição espectral tem sido bastante utilizado na Análise de Componentes Principais – ACP (EVSUKOFF *et al.*, 2011). Esta pode ser empregada, entre outras coisas, para a extração da estrutura de dados de alta dimensionalidade, visando eliminar sua redundância, através da projeção linear das variáveis observadas em um espaço reduzido, gerado pelos autovetores da matriz de covariâncias entre variáveis.

Nesse estudo e conforme foi discutido, o uso da análise espectral visa prover uma estimativa do número de regras a ser utilizado, que pode não ser igual à quantidade de grupos real do problema. Esta metodologia é baseada na definição de um grafo $G(V, E)$ associado ao conjunto de treinamento dos dados. O conjunto de nós (ou V) representa os N registros do

conjunto de treinamento enquanto que o conjunto de arestas (ou E) é definido pela matriz de adjacências ou afinidade $A_{N \times N}$ (ABREU, 2005; LUXBURG, 2007; LUXBURG *et al.*, 2005; SANTOS *et al.*, 2009).

Por sua vez, a matriz de adjacências $A_{N \times N}$ é simétrica, de ordem $N \times N$ e seus elementos são valores reais que representam a similaridade entre cada par de registros do conjunto de treinamento, calculada pela função Gaussiana:

$$a_{ij} = \begin{cases} \exp\left(\frac{-\|x(i)-x(j)\|^2}{2\rho^2}\right) & \text{se } i \neq j \\ 0, \text{ c. c.}, & \end{cases} \quad (2.1)$$

onde i e j representam registros do conjunto de treinamento e ρ controla a variância (ou amplitude) da função. Dessa forma, registros próximos terão similaridade próxima a 1.

Na próxima seção, que descreverá o modelo *fuzzy*, será visto que a função gaussiana também será utilizada na etapa de *fuzzyficação*. Entretanto, as funções têm objetivos diferentes e, portanto, o parâmetro ρ de variância poderá ser diferente. A função Gaussiana utilizada para o cálculo de similaridade está relacionada à estrutura do modelo, ou seja, à quantidade de regras, enquanto que a função Gaussiana que será utilizada no cálculo das pertinências está relacionada à suavização do modelo.

A matriz diagonal D é calculada pela soma dos elementos da linha respectiva na matriz A (adjacências) e, portanto, seus valores são dados por:

$$d_{ii} = \sum_{j=1 \dots N} a_{ij}. \quad (2.2)$$

O problema do agrupamento espectral está relacionado ao problema do corte em grafo, no qual o objetivo é separar (ou cortar) o conjunto de nós do grafo em dois subconjuntos, minimizando o número de arestas entre os dois subconjuntos. A solução ótima para esse problema constitui um problema de otimização NP-completo (FILIPPONE *et al.*, 2008) e uma solução aproximada pode ser obtida pelo cálculo dos autovalores da matriz Laplaciana, definida por:

$$L = D - A. \quad (2.3)$$

Geralmente, a matriz Laplaciana é utilizada após normalização, que pode ser feita com diferentes critérios. Na nossa abordagem, a Laplaciana normalizada analisada por LI *et al.* (2007) foi adotada e é calculada diretamente a partir da normalização da matriz de adjacências:

$$L = D^{-1/2} A D^{-1/2}. \quad (2.4)$$

O mesmo critério de normalização é utilizado em NG *et al.* (2002), cujo algoritmo gerado compartilha a ideia de encontrar uma nova representação dos dados, baseada nos maiores autovalores da matriz Laplaciana normalizada e, posteriormente, aplicar o agrupamento nessa nova representação.

Com as ferramentas derivadas da teoria espectral dos grafos, é possível analisar a estrutura dos dados a partir da observação dos autovalores da Laplaciana normalizada, gerados a partir da decomposição:

$$L = Z\Lambda Z^T \quad (2.5)$$

onde Z é a matriz ortogonal dos autovetores e Λ é a matriz diagonal contendo os autovalores de L , que são todos reais já que a matriz Laplaciana normalizada é simétrica. As colunas de Z e Λ são ordenadas a partir dos valores dos autovalores, de modo decrescente.

Ao se calcular a matriz Laplaciana normalizada a partir do critério mostrado na Equação 2.4, são derivadas da teoria espectral do grafo as seguintes propriedades:

- 1) $\lambda_1 = 1$ e $\sum_{i=1 \dots N} \lambda_i = 0 \Rightarrow \text{traço}(L) = \text{traço}(\Lambda) = 0$;
- 2) $-1 \leq \lambda_i \leq 1$, $i = 1 \dots N$;
- 3) Se o grafo é conectado, então $\lambda_2 < 1$.

Baseado nas propriedades 1 e 2, é possível concluir que sempre haverá um inteiro K tal que:

$$\lambda_i \geq 0, \quad 1 < i \leq K \quad \text{e} \quad \lambda_j < 0, \quad K < j \leq N, \quad \text{geralmente tendo } K \ll N.$$

Além disso, como também foi mostrado em LI *et al.* (2007), num conjunto de dados contendo K grupos disjuntos onde a matriz de adjacências forneça um agrupamento perfeito como:

$$\hat{a}_{ij} = \begin{cases} 1, & \text{se } \mathbf{x}(i) \text{ e } \mathbf{x}(j) \text{ pertencem ao mesmo grupo com } i \neq j \\ 0, & \text{cc} \end{cases},$$

os autovalores da matriz Laplaciana normalizada calculado pela matriz de adjacências serão:

$$\begin{cases} \hat{\lambda}_i = 1, & 1 < i \leq K \\ \hat{\lambda}_i < 0, & K < j \leq N \end{cases}.$$

Caso outra função de similaridade seja usada para o cálculo da matriz de adjacências, tal como a Equação 2.1 que é baseada na função Gaussiana, então a solução da decomposição mostrada na Equação 2.5 é tal que $\lambda_i \rightarrow \hat{\lambda}_i$, $i = 1 \dots N$ (EVSUKOFF *et al.*, 2009).

Portanto, a quantidade de autovalores positivos da matriz Laplaciana pode ser associada à quantidade de regiões de dados regulares no espaço multidimensional das variáveis de entrada e, assim, pode ser usada para estimar a quantidade de grupos necessários para se representar os dados ou a quantidade de regras a serem usadas no modelo *fuzzy* a seguir.

A relação entre a quantidade de autovalores positivos, o número de variáveis do problema e o parâmetro de variância ou amplitude da Equação 2.1 é explicitado em EVSUKOFF *et al.* (2011). Para um valor fixo do parâmetro de variância, a quantidade de autovalores positivos aumenta com a quantidade de variáveis do problema. Com um valor baixo do parâmetro e um grande número de variáveis no problema, a quantidade de autovalores positivos será bem pequena. A variação do parâmetro de variância será explorada no próximo capítulo.

2.3 Modelo *Fuzzy* Simbólico

Considere um conjunto de dados com N registros e variáveis de entrada (ou independentes) e saída (ou dependentes) $T = \{(\mathbf{x}(t), \mathbf{y}(t)), t = 1 \dots N\}$, onde $\mathbf{x} \in \mathcal{R}^p$ representa o vetor de variáveis de entrada e $\mathbf{y} \in \mathcal{R}^q$ representa o vetor de variáveis de saída. Variáveis de entrada ou saída com valores nominais são modelados como variáveis discretas com valores pertencentes ao conjunto dos números naturais. Considere também um conjunto de símbolos de entrada $\mathcal{A} = \{A_i, i = 1 \dots n\}$ e um conjunto de símbolos de saída $\mathcal{B} = \{B_j, j = 1 \dots m\}$, onde cada elemento representa, de modo qualitativo, valores dos conjuntos de entrada e saída respectivamente. Os problemas abordados neste trabalho são compostos de apenas uma variável categorizada ($y \in \mathcal{N}$, sendo \mathcal{N} o conjunto dos números naturais) de saída ($q = 1$) e o valor $y(t) = j$ se refere à classe $B_j \in \mathcal{B}$.

Supondo a existência de uma relação entre variáveis de entrada e saída do tipo $\mathbf{y} = f(\mathbf{x})$, o objetivo geral do problema de aprendizagem é calcular uma aproximação

$\hat{y} = f(\mathbf{x})$ que pode ser representada como um conjunto de regras *fuzzy* $A_i \rightarrow B_j$ expressas na forma geral:

$$\text{se } \mathbf{x}(t) \text{ é } A_i \text{ então } y(t) \text{ é } B_j.$$

O modelo resultante é um sistema *fuzzy* chamado modelo *fuzzy* simbólico, calculado em três etapas que serão apresentadas a seguir: *fuzzyficação*, inferência e *defuzzyficação*.

2.3.1. Etapa de *Fuzzyficação*

Esta etapa pode ser entendida como o mapeamento $F: \mathcal{R}^p \rightarrow [0,1]^n$ do domínio do espaço p dimensional das variáveis de entrada para um espaço n dimensional. Cada função de pertinência *fuzzy* é calculada usando a função Gaussiana e é relacionada a um símbolo $A_i \in \mathcal{A}$ que representa uma região determinada do domínio multidimensional da variável de entrada. Conforme dito na seção 2.2, a Gaussiana utilizada para o cálculo da pertinência e que está relacionada à suavização do modelo é diferente da utilizada para o cálculo da similaridade que, por sua vez, está relacionada à estrutura do modelo.

O centro da função de pertinência $\omega_i \in \mathcal{R}^p$ pode ser tanto um registro do conjunto de treinamento dos dados quanto o centro de um grupo e representa um valor protótipo, ou seja, um valor representativo do símbolo A_i .

Um registro $\mathbf{x}(t) \in \mathcal{R}^p$ é mapeado no vetor *fuzzy* $\mathbf{u}(t) \in [0,1]^n$, de modo que:

$$\mathbf{u}(t) = [u_1(t), \dots, u_n(t)] = [\mu_{A_1}(\mathbf{x}(t)), \dots, \mu_{A_n}(\mathbf{x}(t))],$$

onde cada componente é calculado a partir de: $u_i(t) = g_i(\mathbf{x}, \omega_i, \sigma_i)$ e g é a função de pertinência Gaussiana:

$$g_i(\mathbf{x}, \omega_i, \sigma_i) = \exp\left(\frac{-\|\mathbf{x} - \omega_i\|^2}{2\sigma_i}\right)$$

que está relacionada à suavização do modelo, ω_i é o centro da função e σ_i é o seu parâmetro de dispersão, diferente do parâmetro de dispersão utilizado na função Gaussiana utilizada para o cálculo de similaridade na análise espectral e calculado por: $\sigma_i = \frac{1}{2} \|\hat{\omega}_i - \omega_i\|$, $\hat{\omega}_i =$

$$\arg \min_{\substack{j=1, \dots, n \\ i \neq j}} (\|\omega_j - \omega_i\|).$$

Geralmente, o resultado da *fuzzyficação* é normalizado antes de ser usado na próxima etapa, a partir da relação: $\mathbf{u}(t) = \frac{\mathbf{u}(t)}{\sum_{i=1,\dots,n} u_i(t)}$, que representa o quão próximo o registro $\mathbf{x}(t)$ está perto das regras.

2.3.2. Etapa de Inferência

É o mapeamento $I: [0,1]^n \rightarrow [0,1]^m$, onde n é o número regras e m é o número de símbolos de saída, que no caso da classificação supervisionada, representa o número de classes do problema.

O modelo é representado pela matriz de relação *fuzzy* $\Phi \in [0,1]^{n \times m}$, cujas componentes $\varphi_{ij} = \mu_{\Phi}(A_i, B_j)$ representam a confiança da regra $A_i \rightarrow B_j$ de tal modo que a regra ponderada pode ser expressa na forma geral:

$$\text{se } \mathbf{x}(t) \text{ é } A_i \text{ então } y(t) \text{ é } (B_1/\varphi_{i1}, \dots, B_m/\varphi_{im}).$$

A inferência *fuzzy* é então calculada usando o operador de composição, que produz um mapeamento linear do espaço de variáveis de entrada, escrito na forma vetor-matriz como:

$$\hat{\mathbf{v}}(t) = \mathbf{u}(t) \cdot \Phi.$$

Os símbolos *fuzzy* de saída representam as classes em problemas de classificação supervisionada e podem ser considerados independentemente, tal como $\Phi = (\boldsymbol{\varphi}_1 | \dots | \boldsymbol{\varphi}_m)$. Nesse caso, as componentes do vetor $\hat{\mathbf{v}} \in [0,1]^m$ são calculadas por $\hat{v}_j(t) = \mathbf{u}(t) \cdot \boldsymbol{\varphi}_j$, $j = 1, \dots, m$, onde $\boldsymbol{\varphi}_j$ é o vetor de confianças das regras relacionados à classe B_j e $\hat{v}_j(t)$ é a pertinência da classe calculada para a amostra $(\mathbf{x}(t), y(t))$, tal que $\hat{v}_j(t) = \mu_{B_j}(\mathbf{x}(t))$.

O uso de fatores de confiança para expressar a certeza de regras *fuzzy* tem sido investigado em problemas de classificação supervisionada (EVSUKOFF *et al.*, 2011). Os pesos de regra permitem flexibilidade no desenho do modelo e proveem informações adicionais a respeito da qualidade das regras.

2.3.3. Etapa de Defuzzyficação

Em problemas de classificação supervisionada, esta etapa pode ser entendida como o mapeamento $D: [0,1]^m \rightarrow \mathcal{N}$ que calcula o índice da classe de saída, baseado no símbolo *fuzzy* de saída. Geralmente, a regra de máximo é utilizada, de tal modo que o índice de classe é calculado como a componente com maior valor de pertinência:

$$\hat{y}(t) = j : \hat{v}_j(t) = \max(\hat{\mathbf{v}}(t)).$$

No modelo *fuzzy* Takagi-Sugeno (TSK), utilizado neste trabalho, as regras podem ser escritas como:

$$\text{se } \mathbf{x}(t) \text{ é } A_i \text{ então } y(t) = \theta_i, \text{ onde } \boldsymbol{\theta} = (\theta_1, \dots, \theta_n).$$

Nesse caso, a etapa de *defuzzyficação* é integrada à etapa de inferência e, portanto:

$$\hat{y}(t) = \mathbf{u}(t)\boldsymbol{\theta} \Rightarrow \hat{y}(t) = \sum_{i=1, \dots, n} u_i(t) \cdot \theta_i.$$

2.4 Indução de Regras Fuzzy

A partir da estimativa derivada da utilização da análise espectral da quantidade de grupos necessários para se representar o conjunto de dados que está sendo trabalhado, utiliza-se um algoritmo de agrupamento para gerar as regras *fuzzy* a partir da atribuição de uma regra para cada centro de grupo. A vantagem dessa abordagem é a flexibilidade devido ao grande número de algoritmos de agrupamento presentes na literatura. Conforme foi dito anteriormente, neste trabalho foi feita a aplicação do algoritmo *K-Means* para o cálculo dos centros de grupo devido a sua simplicidade e, conseqüentemente, ao seu fácil entendimento.

O *K-Means* é aplicado diretamente ao espaço das variáveis de entrada e é comum este algoritmo apresentar duas situações problema: a primeira é que os centros de grupo resultantes geralmente não representam registros presentes na base de dados. Essa questão é tratada a partir da aplicação de um algoritmo que captura os registros da base mais próximos aos centros gerados. Diferente do que foi realizado por EVSUKOFF *et al.* (2011), neste trabalho também foi utilizada uma restrição na escolha desses registros de modo que não haja repetição de registros no conjunto resultante, com o objetivo de facilitar a aplicação do método de visualização que será explicado posteriormente. A segunda situação problema é

que a inicialização aleatória do algoritmo pode gerar resultados bastante distintos. Portanto, foi fixada a inicialização através de uma matriz identidade multiplicada por uma constante que mantém os centros iniciais dentro da escala dos dados que estão sendo utilizados.

O algoritmo de indução utilizado está representado no *Algoritmo 1* a seguir.

Algoritmo 1: Indução de Regras

Entrada: $X = \{\mathbf{x}(t), t = 1 \dots N\}, \sigma_i, \rho, \delta$

Saída: Centros de regra, $W = [\omega_1, \dots, \omega_n]$.

1	INICIO
	-- <i>Análise Espectral</i>
2	Cálculo da Matriz A
3	Cálculo da Matriz D
4	Cálculo de $L = D^{-1/2}AD^{-1/2}$
5	Cálculo de $L = ZAZ^T$ -- $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$
6	$n \leftarrow K: \lambda_i \geq \delta, i = 1, \dots, K$ -- <i>Estimativa do n° de regras</i>
	-- <i>Agrupamento</i>
7	$W \leftarrow \text{agrupamento}(X, n)$
	-- <i>Atribuição das regras aos centros de grupos</i>
8	Cálculo dos protótipos pelos registros mais próximos aos centros
9	FIM

O número de regras é estimado na linha 6 e o parâmetro δ utilizado tem por objetivo evitar pequenas flutuações em torno do zero. Para as aplicações mostradas no capítulo 4, foi utilizado $\delta = 0,01$. A matriz W da linha 7 armazena as coordenadas dos centros de grupos formados pela aplicação do *K-Means*, conforme explicado anteriormente. Na linha 8, é feita a atribuição de regras aos centros dos grupos através dos registros mais próximos, com a premissa de que não haja repetição de registro.

2.5 Otimização de Pesos de Regras *Fuzzy*

A utilização de pesos de regras $\boldsymbol{\varphi}_j$ em problemas de classificação supervisionada pode representar o grau de confiança envolvido no relacionamento de um símbolo de entrada A_i com um símbolo de saída (classe) B_j . No caso da utilização do sistema TSK, os parâmetros de saída $\boldsymbol{\theta}$ serão derivados da solução do seguinte problema de otimização quadrática:

$$\begin{aligned} & \text{minimizar } \|\mathbf{U}\boldsymbol{\theta} - \mathbf{Y}\|^2 \\ & \text{sujeito a } \theta_L \leq \theta_i \leq \theta_U, \quad i = 1, \dots, n \end{aligned} \quad (2.6)$$

onde $\mathbf{U} = [\mathbf{u}(1) | \dots | \mathbf{u}(N)]^T$, $\mathbf{U} \in [0,1]^{N \times n}$ é a matriz cujas linhas são os vetores *fuzzy* calculados por $\mathbf{u}(t) = [u_1(t), \dots, u_n(t)] = [\mu_{A_1}(\mathbf{x}(t)), \dots, \mu_{A_n}(\mathbf{x}(t))]$ e $\mathbf{Y} = [y(1), \dots, y(N)]^T$ são os valores observados da variável de saída. Considerando que as variáveis são padronizadas, os limites para os valores de θ_i (ou θ_L e θ_U) são escolhidos tal que $\|\boldsymbol{\theta}\|^2 < n$, sendo n o número de regras.

Ao se expandir o termo quadrático e ignorar os termos constantes, a Equação 2.6 pode ser reescrita como:

$$\begin{aligned} & \text{minimizar } \frac{1}{2} \boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta} - \mathbf{C}_2^T \boldsymbol{\theta} \\ & \text{sujeito a } -\sqrt{n} \leq \theta_i \leq \sqrt{n}, \quad i = 1, \dots, n \end{aligned} \quad (2.7)$$

onde $\mathbf{K} = \mathbf{U}^T \mathbf{U}$ é uma matriz positiva definida e $\mathbf{C}_2^T = \mathbf{Y}^T \mathbf{U}$. Nos problemas de classificação supervisionada, um problema de otimização quadrática é resolvido para cada uma das classes, de modo independente e vários algoritmos de otimização podem ser utilizados.

Com a restrição de valores feita para os pesos, eles poderiam então ser utilizados como auxílio na interpretação do modelo. Peso igual a um significa que a regra associada é certa e que a região definida por A_i não pode ser relacionada à classe B_j . Peso entre zero e um significa que a regra associada é incerta e que A_i representa uma região parcialmente relacionada a B_j . Por fim, peso igual a zero significa que a regra associada é certa e que a região definida por A_i está totalmente relacionada à classe B_j .

No próximo capítulo, serão discutidos conceitos relacionados a formas de visualização de conjuntos de modelos bem como a abordagem de hierarquização utilizada.

2.6 Critérios de Avaliação de Classificação

Pode-se contar com medidas de qualidade de ajuste para a atividade de regressão, como é o caso do coeficiente de determinação ou o critério de informação de Akaike. Associadas a tais medidas, geralmente são utilizadas formas de mensuração da qualidade de classificação, como por exemplo:

- **Matriz de Confusão (HAN e KAMBER, 2006):** Trata-se de uma tabela que descreve, de modo quantitativo, o desempenho do modelo de classificação no conjunto de testes realizados, associados à prévia identificação das classes. Em cada célula é informada a quantidade de registros que pertenciam originalmente à classe da linha respectiva e que foram classificados pelo modelo como pertencentes à classe da coluna respectiva. Essa matriz pode ser calculada para problemas com qualquer quantidade de classes.
- **Métricas de Erro (HAN e KAMBER, 2006):** Estas medidas utilizam as quantidades da Matriz de Confusão e têm por objetivo destacar alguma determinada característica do modelo utilizado. Pode-se considerar como exemplo desse tipo de métrica a sensibilidade, a especificidade, a acurácia ou a medida F.

Neste trabalho, devido à grande quantidade de modelos gerados pela variação do parâmetro de dispersão, optou-se pelo uso de uma métrica de erro ao invés da apresentação de Matrizes de Confusão para cada modelo avaliado. Portanto, o critério de avaliação utilizado neste trabalho foi a acurácia.

CAPÍTULO 3 – Metodologia Proposta

3.1 Seleção de Modelos

A metodologia apresentada no capítulo anterior foi a utilizada em EVSUKOFF *et al.*, 2011 e consiste na primeira fase da metodologia proposta neste trabalho, conforme ilustrado na *Figura 2*.

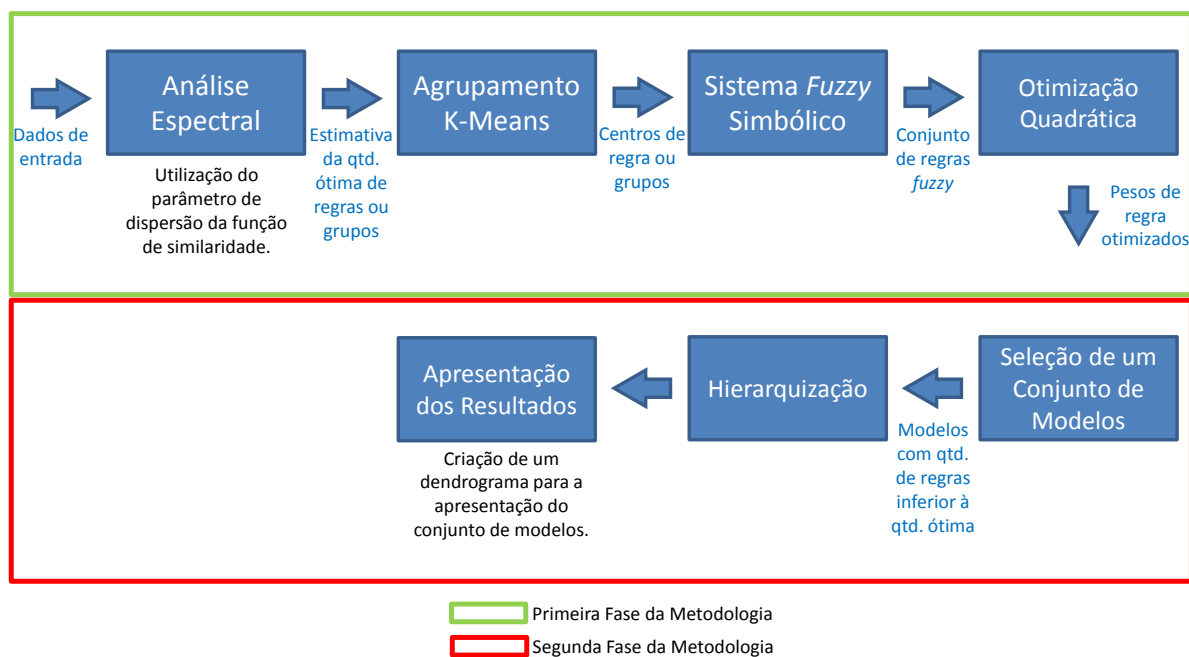


Figura 2 – Estrutura da metodologia proposta.

Já a segunda fase, constitui o aspecto diferencial e complementar deste estudo. Após a otimização quadrática, a primeira fase é aplicada usando-se um conjunto de valores do parâmetro de dispersão pertencentes a um intervalo calculado em função do desvio padrão geral das variáveis de entrada padronizadas. A cada variação do parâmetro de dispersão, é gerado um novo modelo que pode chegar a uma quantidade de regras diferente. A *Figura 3*

ilustra um exemplo da relação entre a variação da dispersão e a quantidade de regras geradas de uma das bases de teste utilizadas no trabalho, que será detalhada posteriormente.

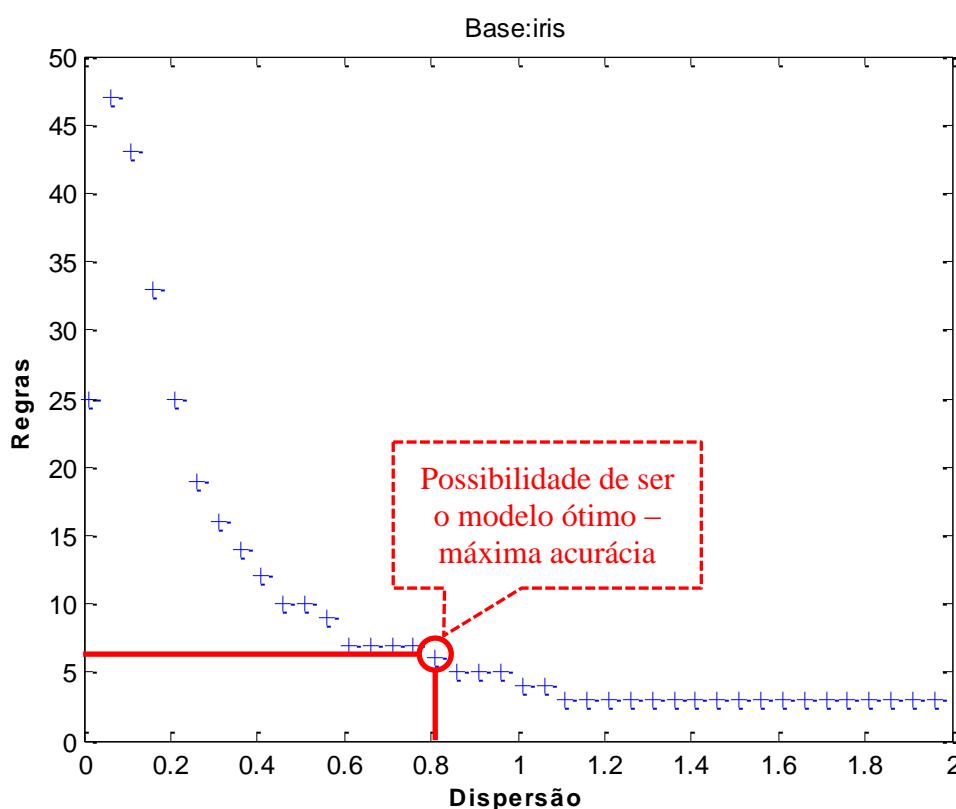


Figura 3 – Exemplo de relação entre a variação da dispersão e a quantidade de regras.

Para a escolha do melhor modelo do intervalo, é utilizado o critério de maior acurácia calculada em validação cruzada (KOHAVI, 1995). A validação cruzada consiste na divisão aleatória dos dados em k partições mutuamente excludentes ou "*folds*", D_1, D_2, \dots, D_k , de modo que cada uma fique com tamanho aproximadamente igual. A partir disso, o treinamento e o teste são realizados k vezes: a cada iteração, uma partição é reservada como o conjunto de teste e as partições restantes são coletivamente usadas como conjunto de treinamento, de modo que são obtidos k resultados de modelo. Ao contrário da partição tradicional de treinamento e teste, com essa metodologia cada partição é utilizada o mesmo número de vezes para treinamento e todas são utilizadas uma vez para teste. Nos problemas de classificação supervisionada, a acurácia pode ser calculada pela razão entre o número total de classificações corretas nas k iterações e a quantidade total de registros da base de dados (HAN e KAMBER, 2006).

Os modelos selecionados para a hierarquização são os que foram calculados com valor de dispersão maior que o valor do melhor modelo em acurácia e que, por consequência e como ilustrado na *Figura 3*, apresentam quantidade inferior de número de regras. A primeira fase da metodologia é então rodada novamente para esse conjunto de dispersões, dessa vez sem o uso da validação cruzada, porém mantendo o resultado em termos de acurácia que havia sido calculado. Essa nova rodada se torna necessária para que se obtenha uma quantidade inteira de número de regras, visto que na validação cruzada esse resultado é dado pela média obtida nas k iterações.

A partir desse ponto, ou seja, com um conjunto de modelos selecionados em mãos, a abordagem mais encontrada na literatura é a de se definir métodos de combinar os modelos gerados em um único, com o objetivo de incrementar o poder de acurácia. Os trabalhos resumidos a seguir são exemplos deste tipo de aplicação.

FRED e JAIN (2005) propuseram um método de combinação de múltiplos agrupamentos através do que denominaram acumulação de evidências. A ideia apresentada foi a de combinar agrupamentos derivados da aplicação de diferentes algoritmos e também a de utilizar o mesmo algoritmo com variações de parâmetros.

MIRZAEI e RAHMATI (2010) mostraram que os métodos de combinação de agrupamento têm recebido considerável atenção nos últimos anos. Propuseram então uma metodologia para a combinação de agrupamentos hierárquicos. Essa escolha foi motivada pela vantagem da metodologia hierárquica produzir não somente um agrupamento, mas sim um conjunto hierarquizado, característica também explorada neste trabalho. Com o uso de bases de *benchmark*, confirmaram a superioridade da metodologia de combinação em relação aos métodos hierárquicos tradicionais.

JIA *et al.* (2011) mostraram que frequentemente é possível se obter melhores agrupamentos quando partes de agrupamentos disponíveis são combinadas e, por isso, propuseram um novo método de combinação. Os agrupamentos utilizados foram gerados pela metodologia de agrupamento espectral com variação dos parâmetros de dispersão e da inicialização do *K-Means*. Mostrou também que os dois grandes desafios das metodologias de combinação de agrupamento são a geração de agrupamentos componentes e a definição da função de consenso.

Ao contrário dos métodos de combinação descritos anteriormente, a abordagem de hierarquização utilizada tem o objetivo de facilitar a interpretabilidade dos dados e do processo de modelagem e, portanto, busca encontrar uma forma de apresentação dos modelos gerados que ajude a visualização, de modo que o usuário da metodologia possa escolher o nível que ele quer trabalhar, podendo ponderar a relação entre acurácia e interpretabilidade.

3.2 Hierarquização

Os métodos de agrupamento hierárquico tradicionais funcionam agrupando objetos de dados dois a dois no que é chamado de árvore de grupos. Podem ser classificados como aglomerativos ou divisivos, dependendo da decomposição hierárquica ser formada no sentido baixo-cima (ou de fusão de objetos) ou no sentido cima-baixo (divisão de objetos). A vantagem da metodologia é a geração ou apresentação de um conjunto de agrupamentos e não somente uma possibilidade. A deficiência geralmente reside no fato de que, uma vez realizada uma operação de fusão ou divisão, não há uma maneira simples de correção a menos que toda a estrutura hierárquica inferior seja revista.

No contexto do agrupamento hierárquico, é comum a utilização de uma ferramenta de visualização chamada dendrograma. Um dendrograma é uma estrutura de árvore de grupos que facilita a representação do processo de agrupamento, uma vez que mostra passo-a-passo como os objetos de dados foram agrupados. Um exemplo de dendrograma é apresentado na *Figura 4*.

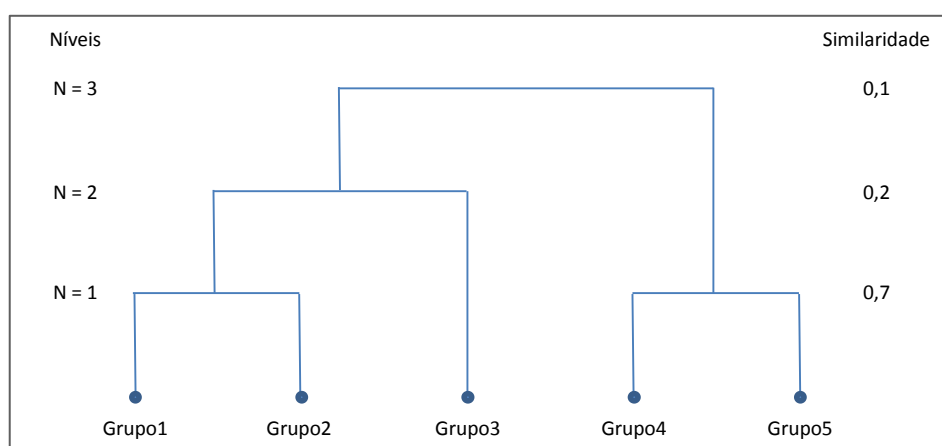


Figura 4 – Exemplo de dendrograma.

Após a seleção dos modelos a partir do modelo com maior acurácia, foi utilizada uma hierarquização aglomerativa dos mesmos em função da distância dos centros de mesma classe. Em outras palavras, considerando que são conhecidos: 1) as quantidades de regras de cada modelo por classe e 2) suas respectivas posições, é possível passar de um modelo para o seguinte tendo como critério de junção as regras mais próximas de mesma classe. Dessa forma e assim como é ilustrado na *Figura 5*, é criada a hierarquia de modelos no sentido de que os modelos alocados em níveis hierarquicamente superiores sejam capazes de generalizar os modelos alocados em níveis inferiores, em uma estrutura de modelos sequenciais que mantém a informação de relacionamento direto entre eles.

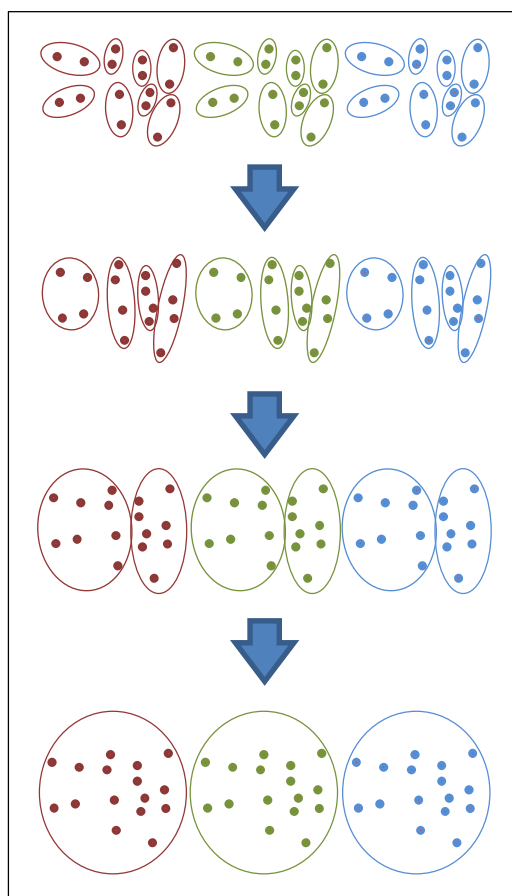


Figura 5 – Exemplo de Hierarquia de Modelos

A hierarquização produzida será apresentada sob a forma de um dendrograma, descrito anteriormente, de modo a facilitar ainda mais a interpretação da relação entre os modelos no sentido da complexidade e da acurácia associada.

3.3 Visualização Completa dos Modelos

O conjunto completo de modelos selecionados será também apresentado como ilustrado na *Figura 6*, de forma ordenada em função do valor de dispersão.

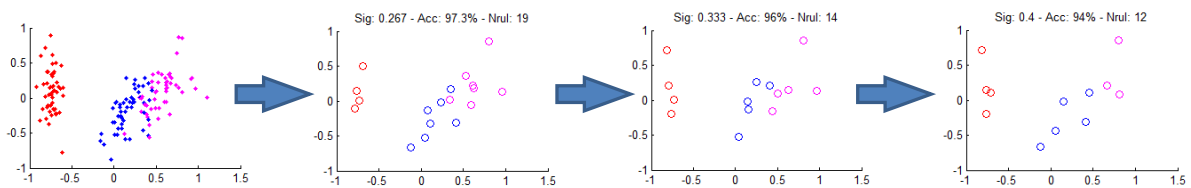


Figura 6 – Forma de visualização completa.

Essa forma de apresentação será dada partindo de uma visualização do conjunto de dados com o uso da análise de componentes principais (ACP) para gerar a projeção em duas dimensões e, posteriormente, os resultados dos modelos selecionados. São apresentados: o parâmetro de dispersão utilizado, a acurácia do modelo, a quantidade de regras geradas e uma visualização dos centros de regras geradas também por meio de projeção em duas dimensões da ACP.

3.4 Algoritmo Proposto

Dado o Algoritmo 1 de indução de regras apresentado anteriormente, o algoritmo completo proposto pode ser descrito como:

Algoritmo 2: Hierarquização

Entrada: Conjunto de dados, com variáveis de entrada e variável de saída.

Saída: Hierarquia de modelos.

1	INICIO DA FASE 1
2	Definição de um intervalo de valores para o parâmetro de dispersão;
3	PARA cada parâmetro de dispersão:
4	Algoritmo 1;
5	Execução da Otimização Quadrática;
6	Armazenamento de acurácia obtido em validação cruzada;
7	FIM PARA
8	FIM DA FASE 1
1	INICIO DA FASE 2
2	Seleção do conjunto de modelos → dispersão maior do que o modelo de melhor acurácia;
3	PARA cada modelo selecionado:
4	Reprocessar do Algoritmo 1 fora de validação cruzada; --Obter: quantidade de regras de cada classe e respectivos posicionamentos
5	Avaliar cada modelo do conjunto de modelos selecionados no sentido da quantidade de regras de cada classe do problema → Nova Seleção;
6	FIM PARA
7	PARA cada modelo da Nova Seleção:
8	Agrupa centros mais próximos (2 a 2) que pertençam a mesma classe até que a quantidade de centros do próximo modelo seja alcançada;
9	FIM PARA
10	Apresenta resultado hierarquizado associado cada nível a sua respectiva dispersão e acurácia;
11	FIM DA FASE 2

CAPÍTULO 4 – Aplicação e Resultados

4.1 Descrição das Bases de *Benchmark*

Foram utilizadas dez bases de *benchmark* para aplicação da metodologia proposta neste estudo. Essas bases são provenientes do *UCI Machine Learning Repository*¹ e são bastante exploradas na literatura para teste de algoritmos de classificação supervisionada.

Características gerais das bases utilizadas			
Base	Quantidade de Variáveis	Quantidade de Registros	Quantidade de Grupos
Iris	4	150	3
Balance	4	625	3
Diabetes	8	768	2
Cancer	9	286	2
Glass	9	214	6
Wine	13	178	3
Heart	13	270	2
Image	18	210	7
Ionosphere	34	351	2
Sonar	60	208	2

Tabela 1 – Características gerais das bases de *Benchmark*

A análise detalhada das informações de cada base de *benchmark* encontra-se no Apêndice A.

¹ Bases disponíveis para *download* em <http://archive.ics.uci.edu/ml/>.

4.2 Apresentação e Discussão dos Resultados Obtidos

Para todas as bases utilizadas, foi utilizado o mesmo critério de intervalo de dispersões que é calculado, para cada base, em função do desvio padrão de suas variáveis. Conforme foi observado ao longo do desenvolvimento do trabalho, para valores de dispersão próximos a zero, a quantidade de regras geradas também tende a zero. Portanto, o intervalo foi iniciado a partir do valor do desvio padrão da base multiplicado por um fator = 0,4. As variáveis de todas as bases foram padronizadas previamente.

Os melhores modelos encontrados são apresentados na *Tabela 2*. A projeção em duas dimensões por ACP do posicionamento dos centros de regra dos melhores modelos é apresentado para cada uma das bases no Apêndice B.

Resultados dos Melhores Modelos Gerados				
Base	Acurácia	Qtd. Regras	Dispersão	Tempo de Processamento
Iris	97,33	19	0,2667	0,0983
Balance	90,72	32	0,4667	0,7466
Diabetes	77,73	126	0,4000	1,6952
Cancer	97,36	3	2,4667	0,6693
Glass	69,16	52	0,2000	0,3604
Wine	98,31	60	0,1333	0,1829
Heart	84,81	27	0,9333	0,1678
Image	85,71	65	0,1333	0,4705
Ionosphere	96,01	51	0,6667	0,3549
Sonar	84,62	46	1,2000	0,1994

Tabela 2 – Melhores modelos encontrados no intervalo de dispersões

Para cada base, serão apresentados três gráficos de resultados, que representam respectivamente:

- A relação do parâmetro de dispersão (denominado sig nos gráficos de resultados) com resultados em termos de número de regras (nrul nos gráficos), tempo e acurácia (acc nos gráficos);

- A forma de visualização completa dos modelos de cada base.

É possível notar que em todas as bases, com exceção da base Cancer, o melhor modelo em termos de acurácia possui uma quantidade de regras muito maior do que a quantidade de classes do problema. Isso ocorre geralmente pela dificuldade de atuação do classificador quando há interseção entre classes. Entretanto, sob a ótica da interpretabilidade dos dados e do modelo, é muito melhor a leitura do modelo com o uso de poucas regras.

4.2.1. Resultados da Base Iris

Com o gráfico de visualização da base Iris (*Figura 8*), por exemplo, fica fácil perceber que ao nível de acurácia de 86%, se tem um modelo muito simples, com apenas 3 regras. Pela simplicidade da base, o tempo de processamento em todas as iterações é baixo e não passa o valor de 0,15 segundo, como pode ser visto na *Figura 7*.

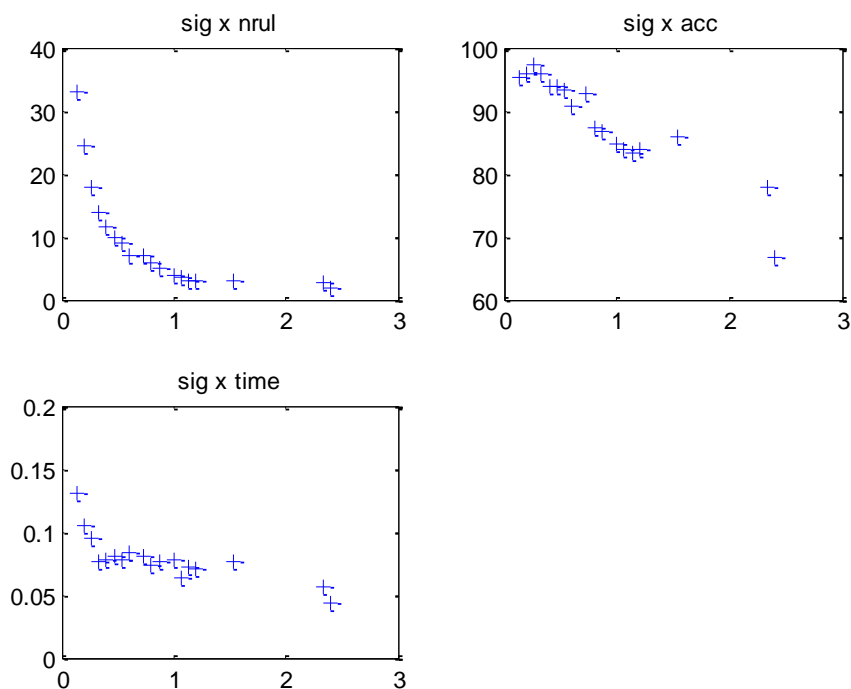


Figura 7 – Relação do parâmetro de dispersão com resultados da base Iris.

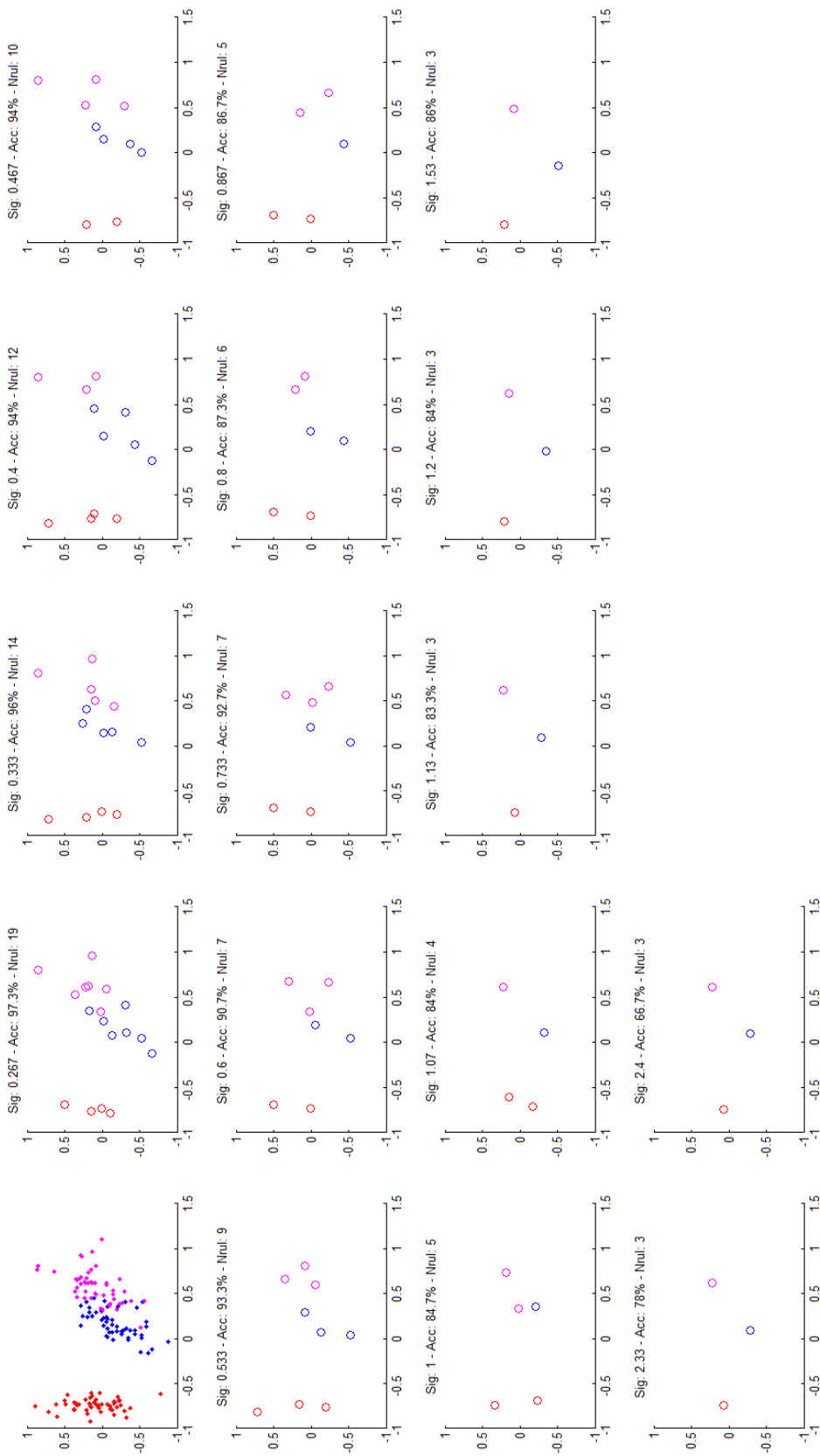


Figura 8 – Visualização completa dos modelos da base Iris.

4.2.2. Resultados da Base Balance

Já no caso da base Balance, por ser uma base de 625 registros, é notável o alto tempo de processamento para dispersões pequenas, que são relacionadas a grandes quantidades de regras, conforme mostrado na *Figura 9*. Pelo comportamento completo dos dados, não há coesão evidente no gráfico de acurácia contra dispersão. Na visualização completa (*Figura 10*), é possível notar um modelo simples ao nível de acurácia de 86,6%. A distribuição dos centros nesse nível enfatiza a separação a esquerda e a direita das duas principais classes do problema.

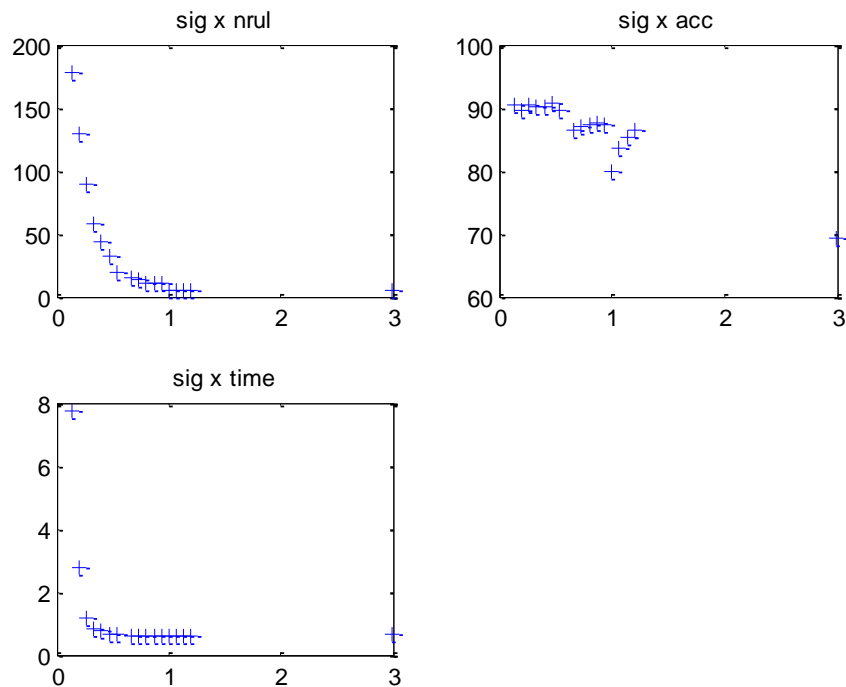


Figura 9 – Relação do parâmetro de dispersão com resultados da base Balance.

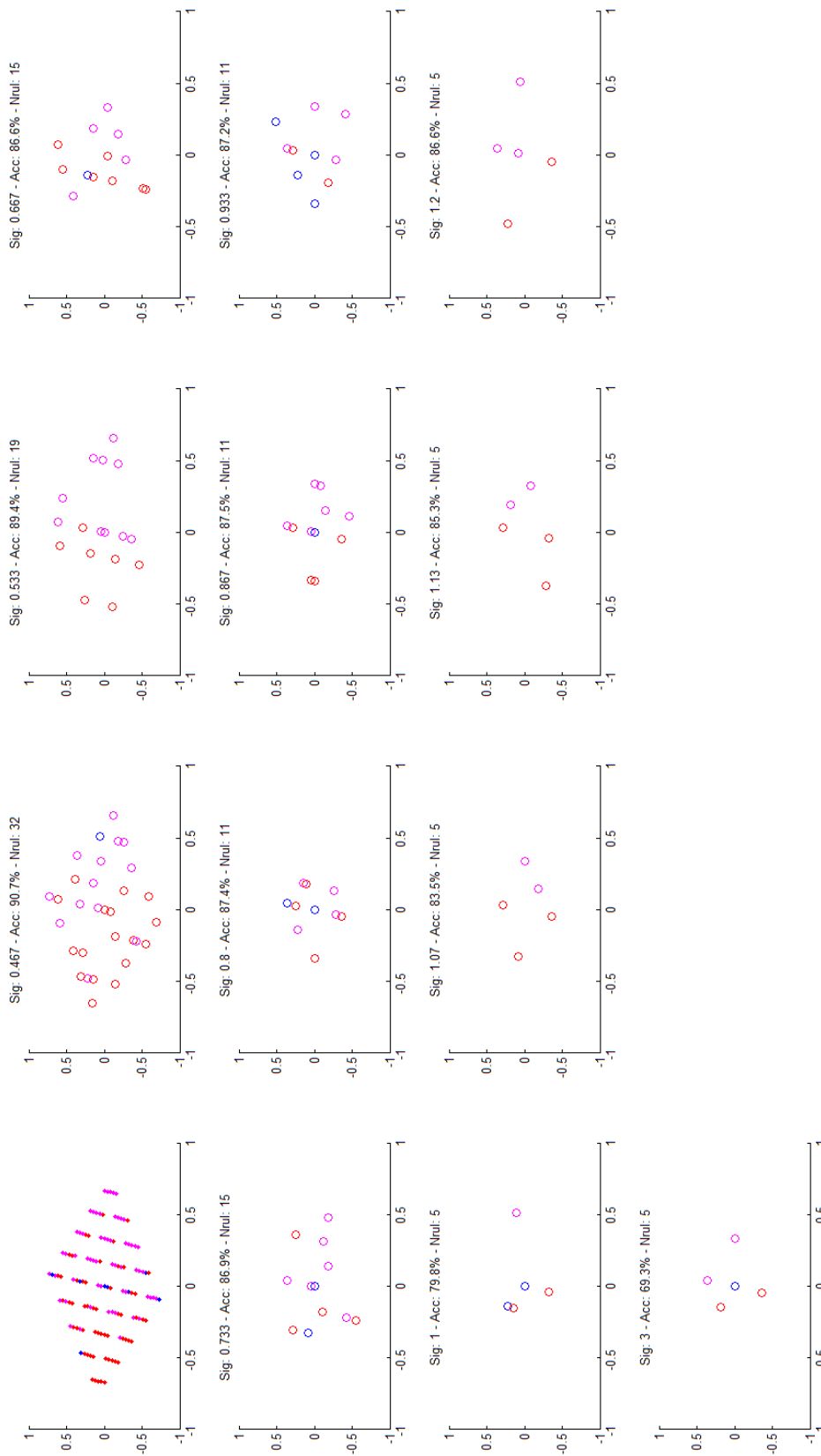


Figura 10 – Visualização completa dos modelos da base Balance.

4.2.3. Resultados da Base Diabetes

A base Diabetes também tem como característica o número maior de registros comparado às outras bases utilizadas. Na *Figura 11*, é possível notar que algumas iterações passaram do tempo de execução de 15s. Pelo alto grau de mistura entre as 2 classes do problema, o melhor modelo alcançado com o uso de 126 regras já está ao nível de acurácia de 77,7%, conforme pode ser observado na visualização completa (*Figura 12*). Entretanto, ao nível de 72% de acurácia se tem um modelo muito mais simples, que conta com o uso de 4 regras somente e baixo tempo de execução.

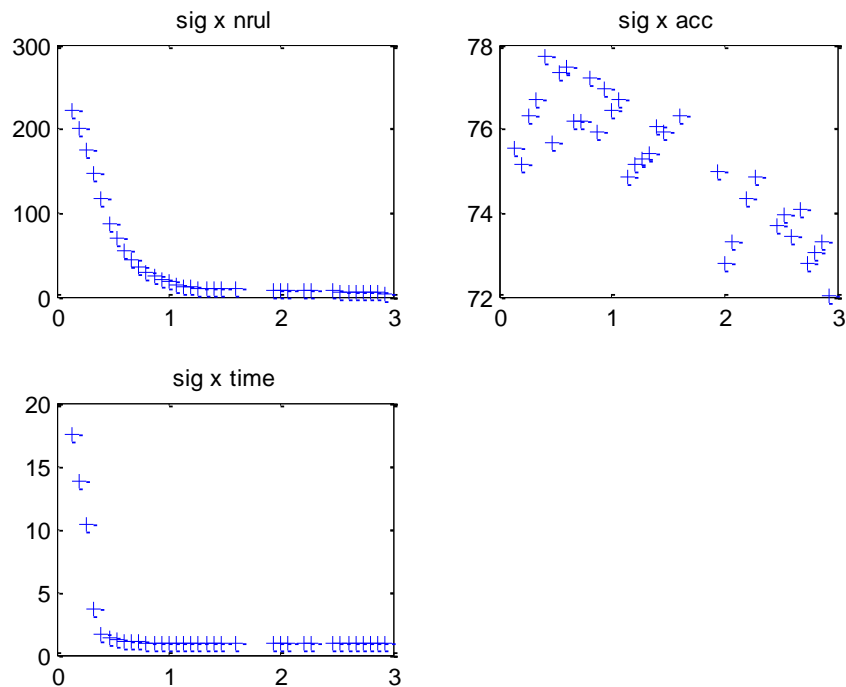


Figura 11 – Relação do parâmetro de dispersão com resultados da base Diabetes.

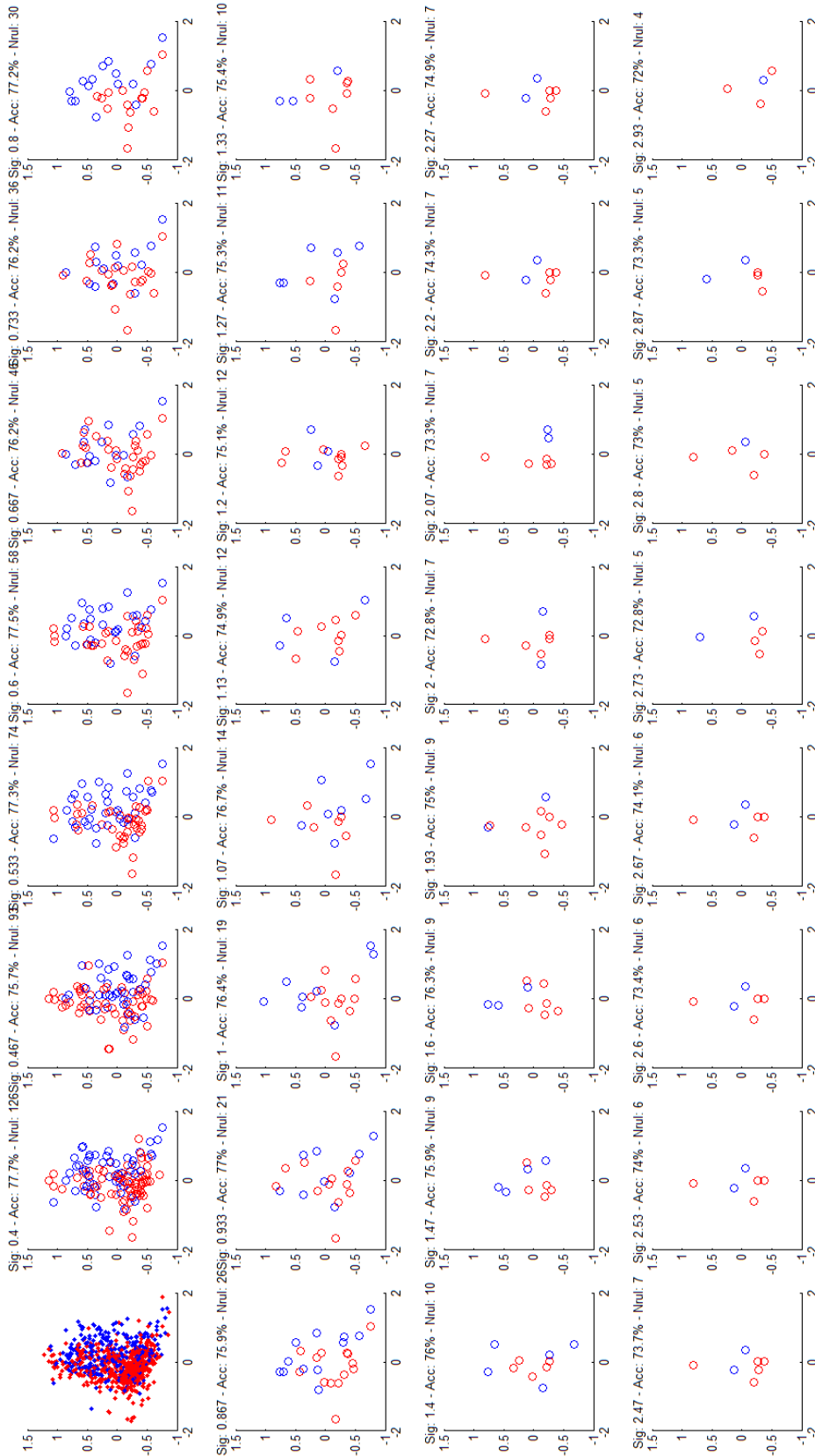


Figura 12 – Visualização completa dos modelos da base Diabetes.

4.2.4. Resultados da Base Cancer

A base Cancer também conta com elevada quantidade de registros em comparação às demais bases e, por isso, também cai na mesma observação com relação ao tempo de processamento que ocorreu nas bases Diabetes e Balance, como pode ser visto na *Figura 13*. O comportamento do gráfico de acurácia contra dispersão apresenta um comportamento sem padrão, diferente das outras bases analisadas. Isso se dá pelo comportamento de fácil separação entre classes da base, pois a acurácia está variando num intervalo muito pequeno, de apenas 2% como pode ser visto no eixo. Entretanto, trata-se de uma base de mais fácil classificação e com somente 2 regras já se tem um nível de acurácia de 95,8%, como pode ser visto na visualização completa da base (*Figura 14*).

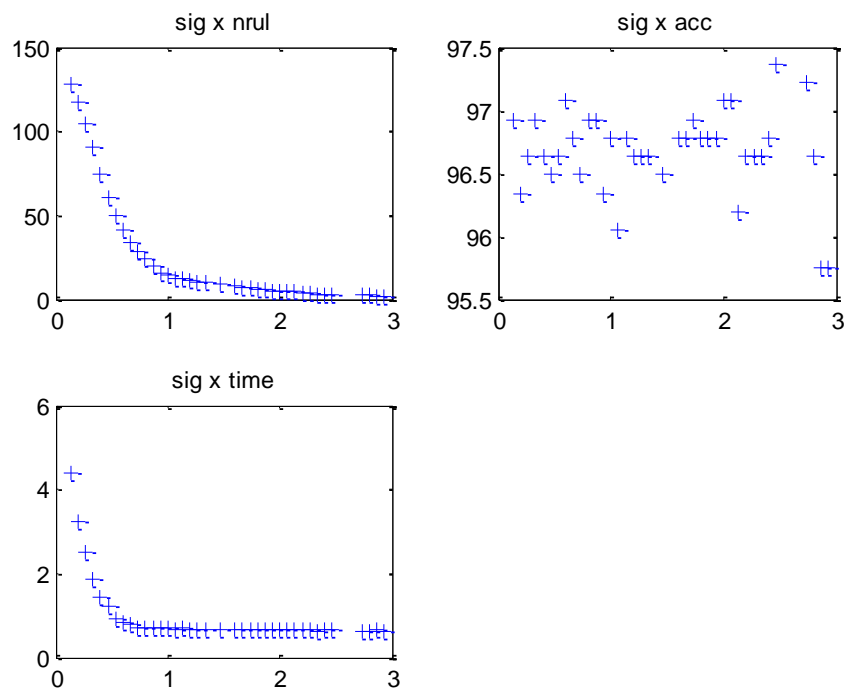


Figura 13 – Relação do parâmetro de dispersão com resultados da base Cancer.

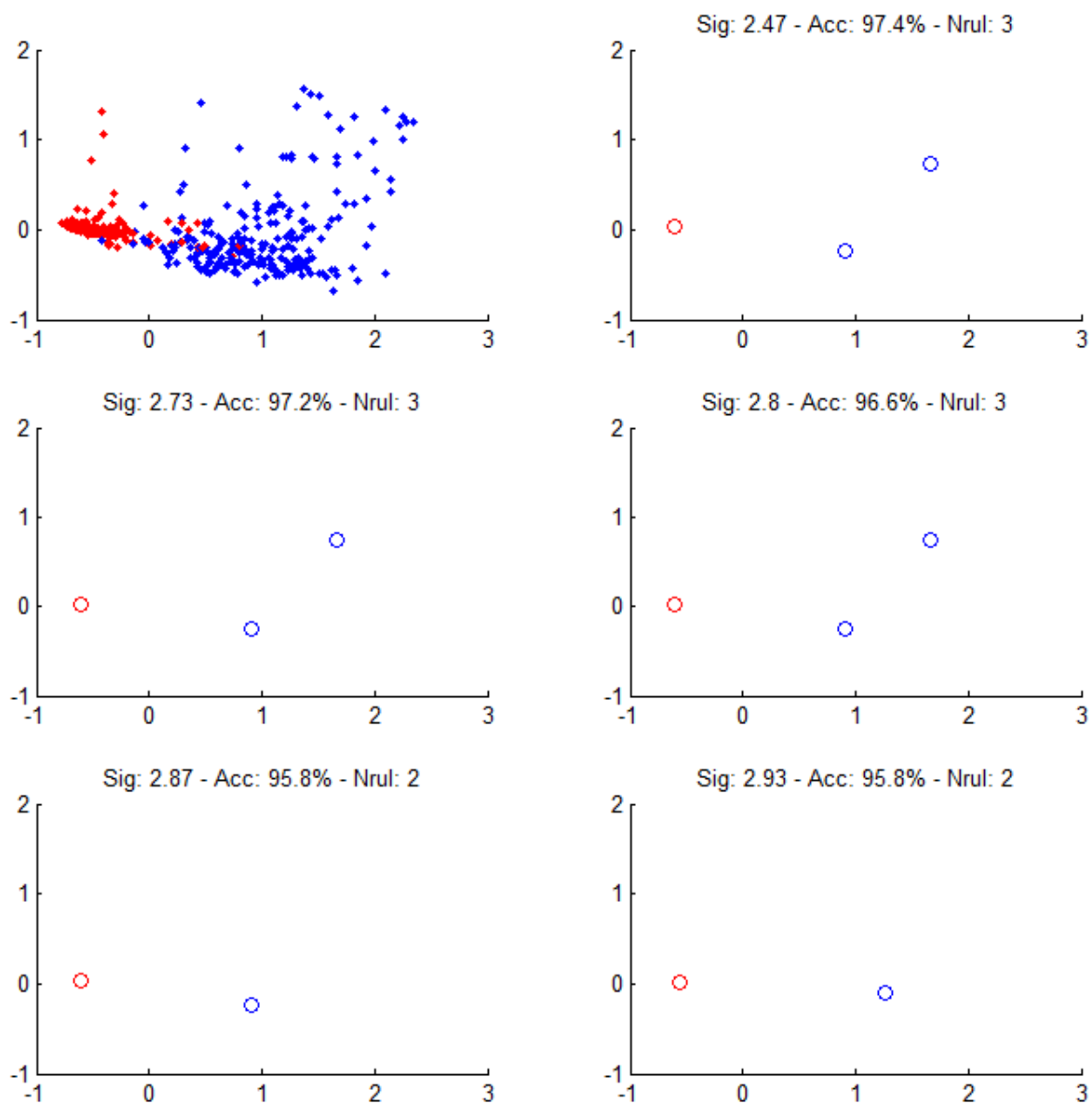


Figura 14 – Visualização completa dos modelos da base Cancer.

4.2.5. Resultados da Base Glass

Com base no gráfico de dispersão x acurácia da base Glass (*Figura 15*), é possível notar o impacto do comportamento de possuir muitas classes misturadas entre si. Quanto menor o número de regras, pior o nível de acurácia alcançado. A visualização completa (*Figura 16*), mostra que para se ter um mínimo de acurácia aceitável (>50%) é necessário um modelo com ao menos 7 regras, uma a mais do que a quantidade real de classes do problema.

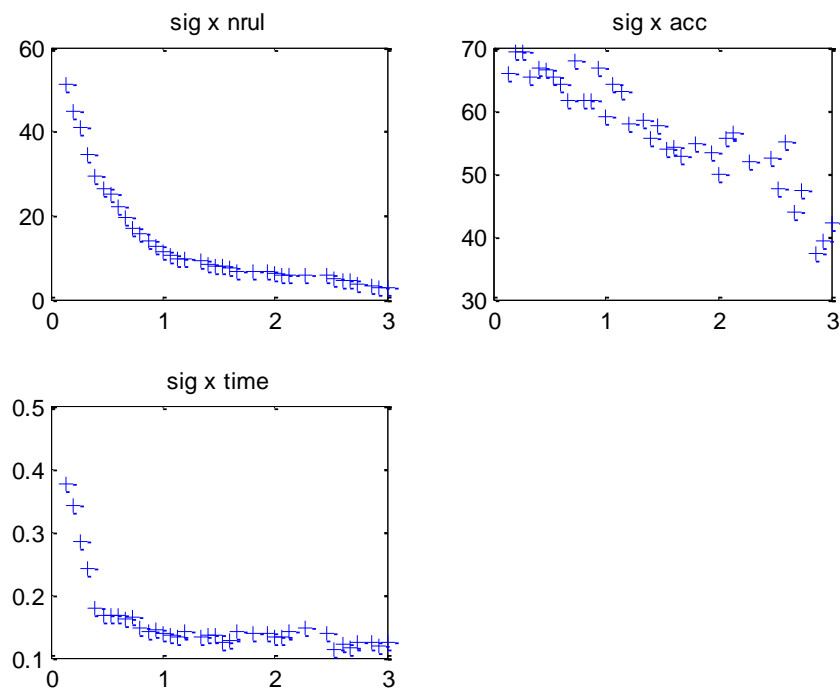


Figura 15 – Relação do parâmetro de dispersão com resultados da base Glass.

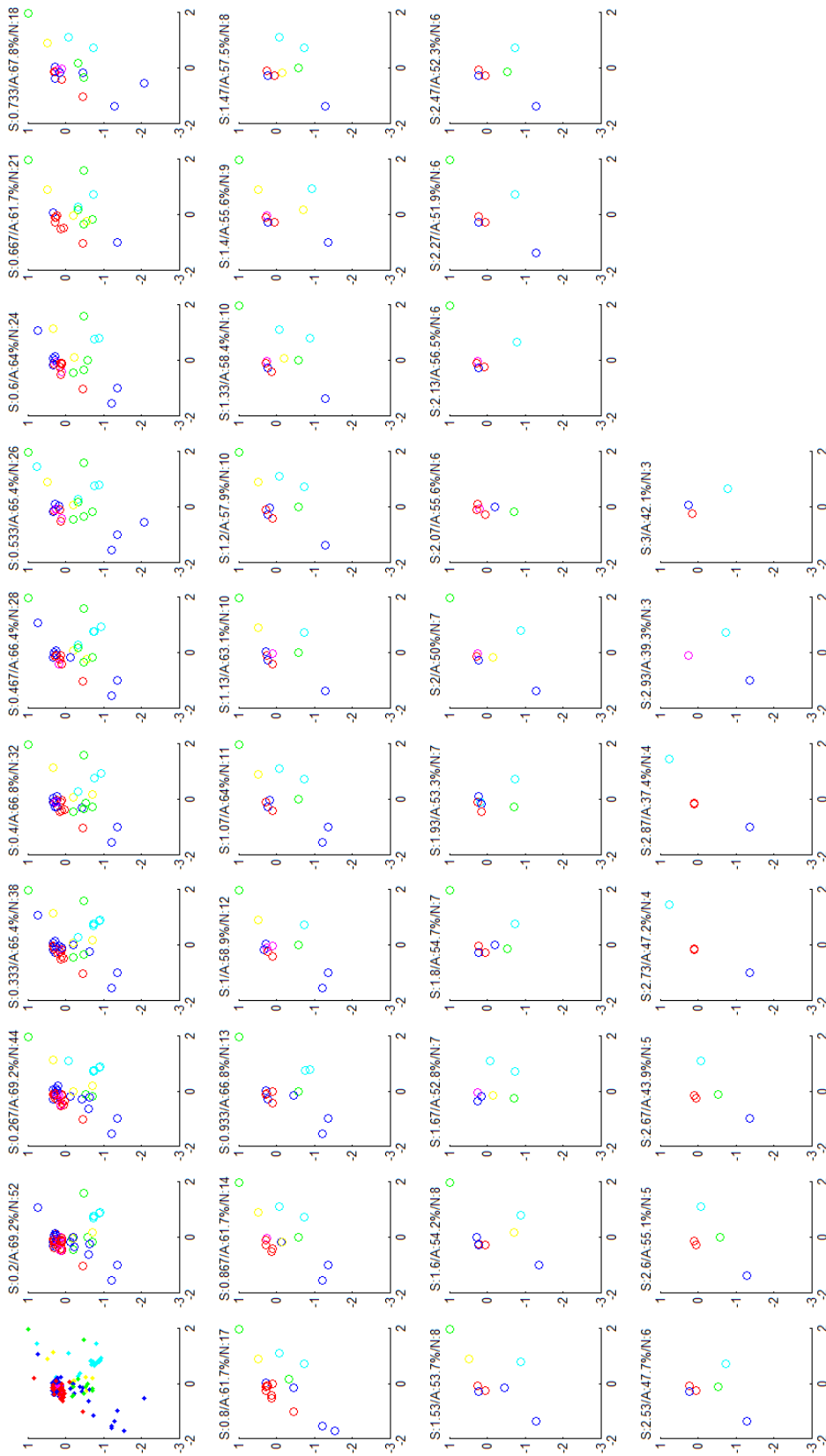


Figura 16 – Visualização completa dos modelos da base Glass.

4.2.6. Resultados da Base Wine

A base Wine caracteriza-se pela facilidade de classificação, como pode ser notado na visualização completa (*Figura 18*). Com o uso de somente 4 regras, já é possível trabalhar ao nível de acurácia de 94,4%. Por não ser uma base com muitos registros, a *Figura 17* mostra o baixo tempo de processamento para todos as dispersões (sempre abaixo de 1s) e a baixa variação no gráfico de acurácia, que conta com somente um resultado com assertividade abaixo dos 90%.

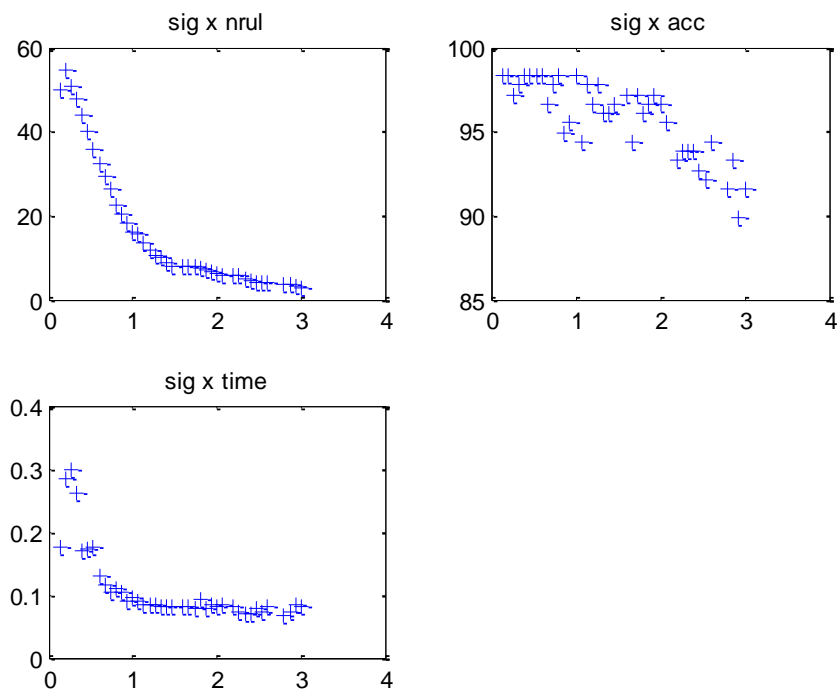


Figura 17 – Relação do parâmetro de dispersão com resultados da base Wine.

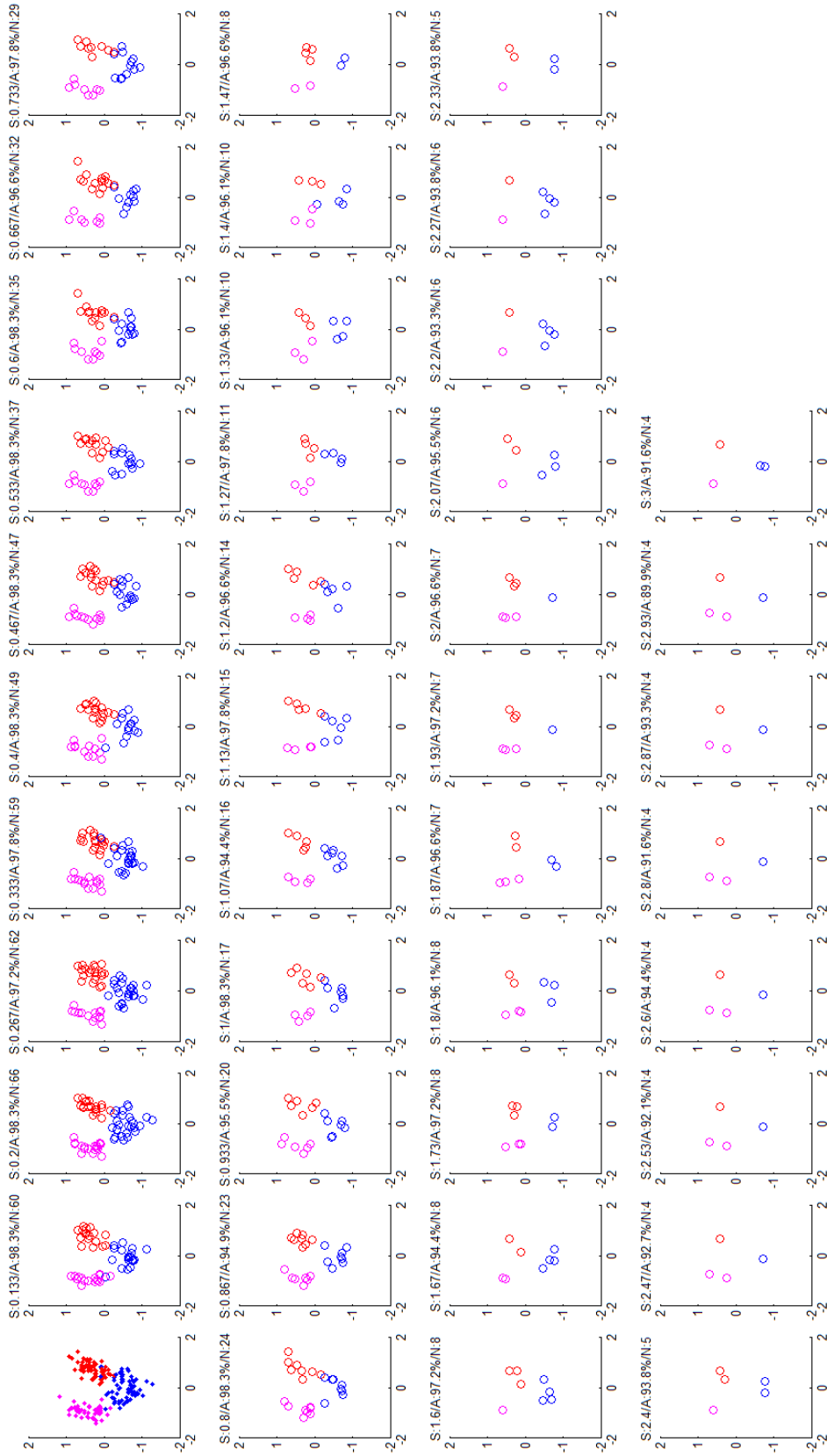


Figura 18 – Visualização completa dos modelos da base Wine.

4.2.7. Resultados da Base Heart

A base Heart, apesar de ser pequena em termos de quantidade de registros, apresenta nível de mistura elevado das 2 classes do problema. Como pode ser visto na visualização completa (*Figura 20*), o melhor modelo já parte de um nível de acurácia de 84,8% com 27 regras. Ao nível de 82,2%, pode-se trabalhar com um modelo bem mais simples, com o uso de somente 7 regras. Por conta do nível de mistura, há maior variabilidade no gráfico de acurácia contra dispersão, conforme pode ser visto na *Figura 19*. Ainda assim, o pior modelo mantém nível de assertividade acima de 75%.

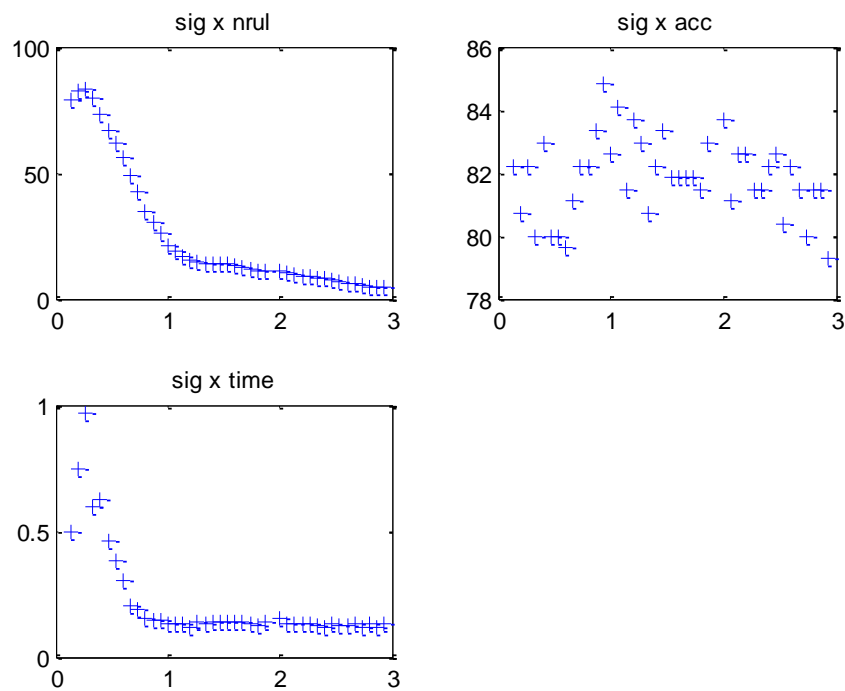


Figura 19 – Relação do parâmetro de dispersão com resultados da base Heart.

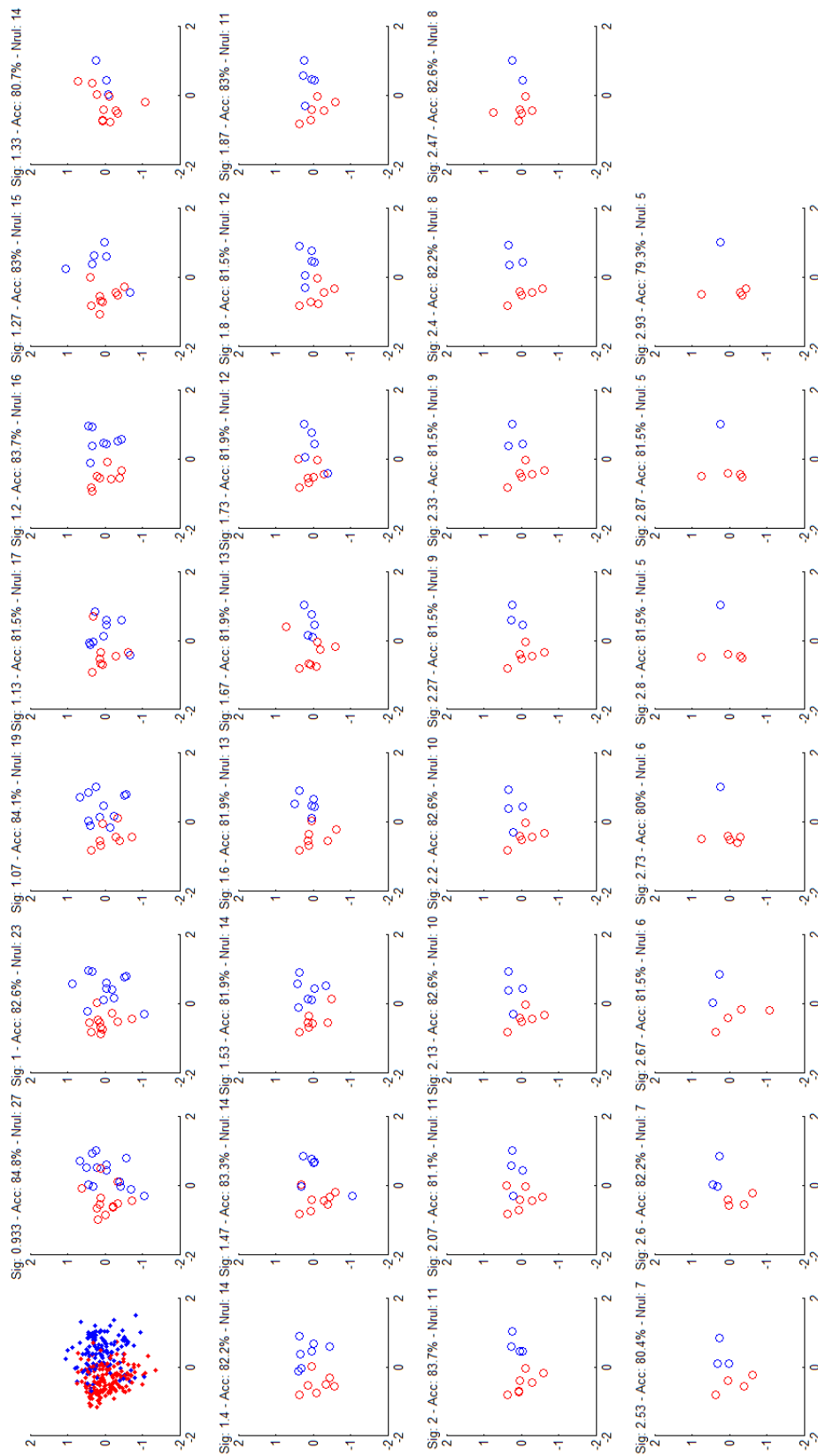


Figura 20 – Visualização completa dos modelos da base Heart.

4.2.8. Resultados da Base Image

Apesar da base Image parecer de difícil classificação, é possível notar na sua visualização completa (*Figura 22*) que o melhor modelo encontrado consegue alcançar 85,7% de acurácia, com o uso de 65 regras *fuzzy*. Ao nível de acurácia de 80%, é possível trabalhar com 10 regras, quantidade perto ao número de classes real do problema. O gráfico de acurácia contra dispersão indica a geração de poucos modelos com assertividade ruim, abaixo dos 50%, conforme pode ser visto na *Figura 21*.

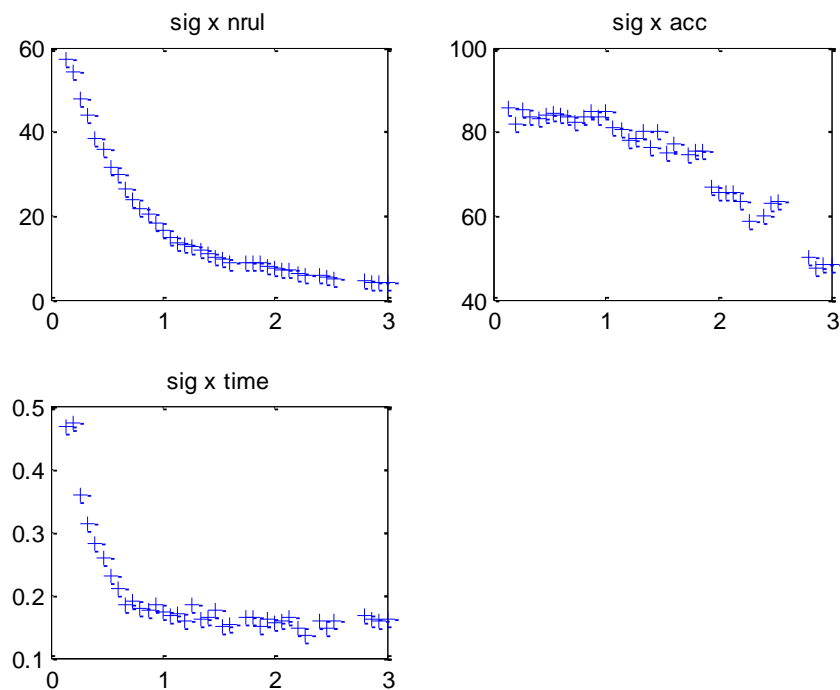


Figura 21 – Relação do parâmetro de dispersão com resultados da base Image.

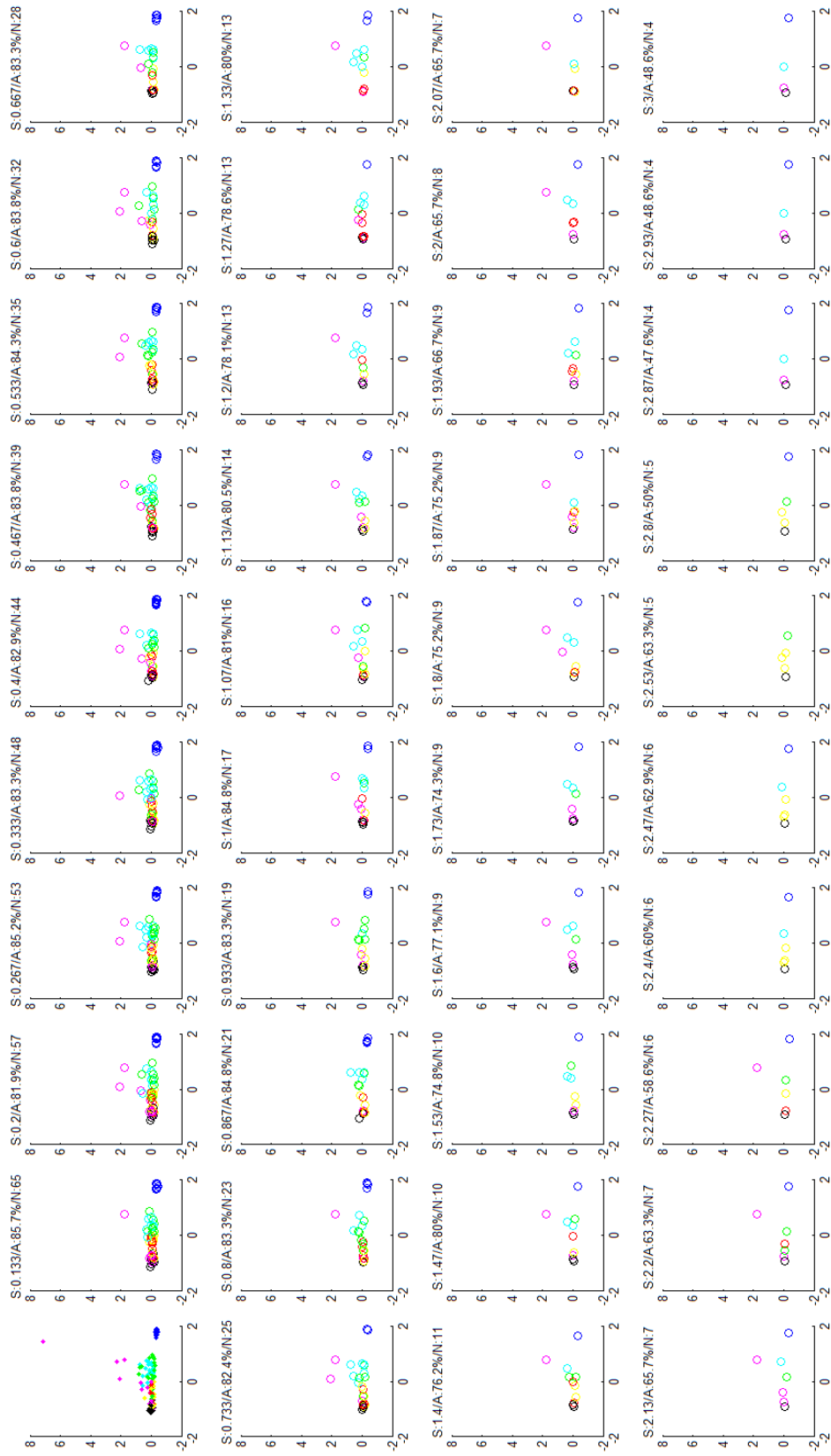


Figura 22 – Visualização completa dos modelos da base Image.

4.2.9. Resultados da Base Ionosphere

A base Ionosphere, assim como outras já citadas, apresenta nível de mistura entre as suas duas classes elevado. Na sua visualização completa (*Figura 24*), é possível notar que com o uso de 8 regras já se tem um nível de acurácia de quase 80%. Mesmo sendo uma base com nível de mistura grande entre classes, os modelos gerados mantêm um bom nível de assertividade, com acurácias sempre acima dos 70%, conforme pode ser visto na *Figura 23*.

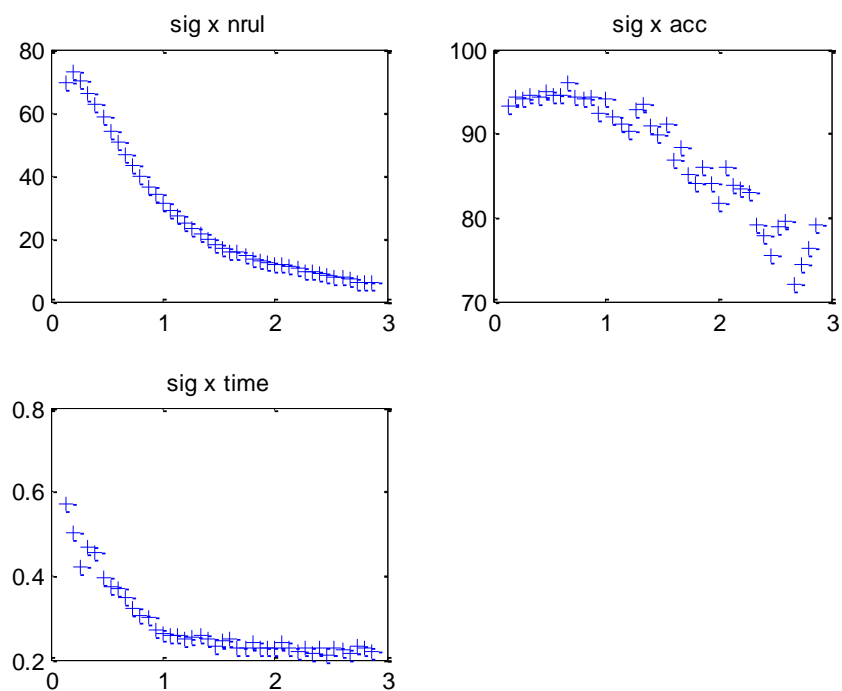


Figura 23 – Relação do parâmetro de dispersão com resultados da base Ionosphere.

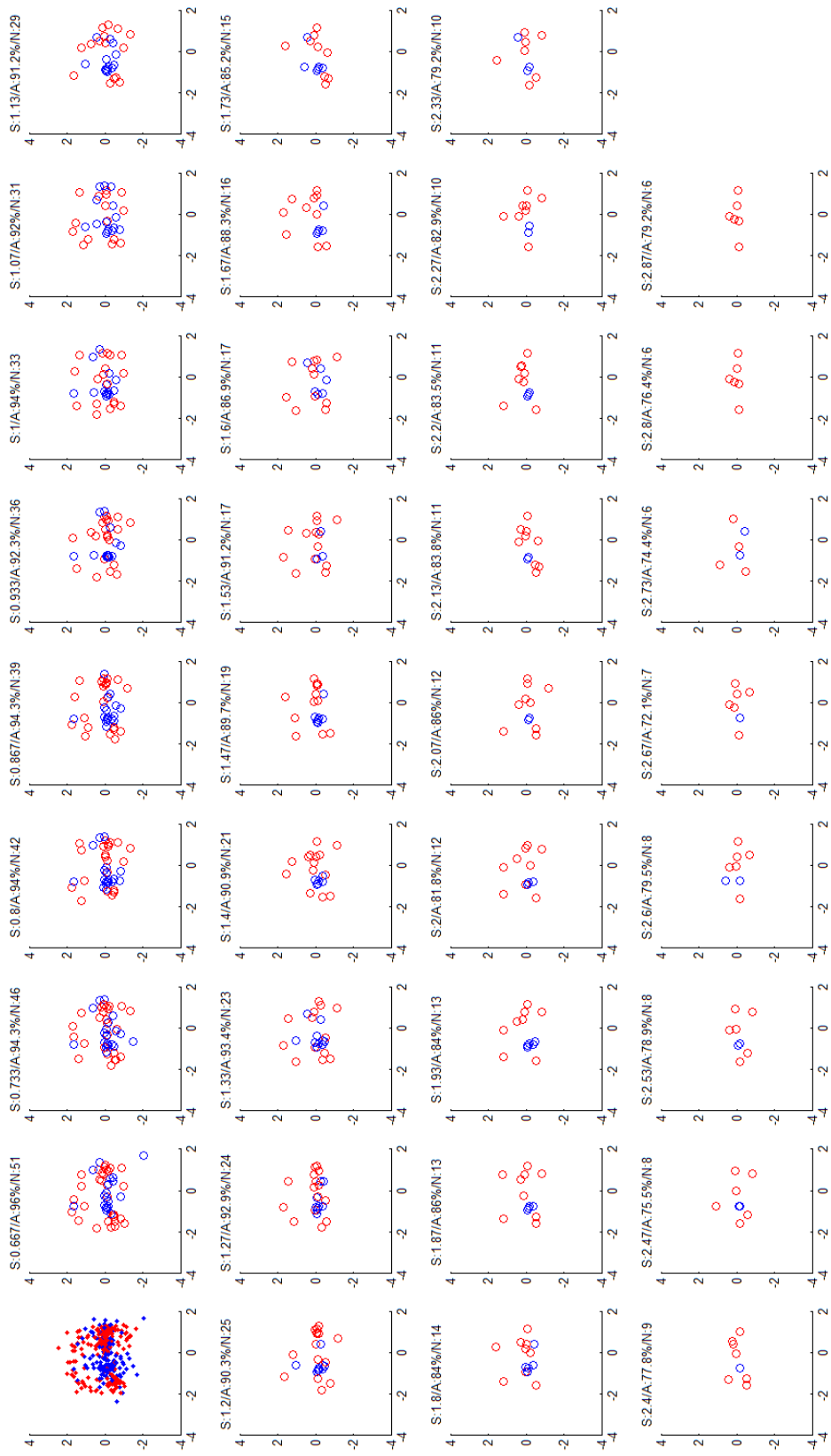


Figura 24 – Visualização completa dos modelos da base Ionosphere.

4.2.10. Resultados da Base Sonar

A base sonar, apesar de aparentemente ter comportamento parecido com a base Ionosphere, apresenta maior dificuldade para o classificador. Na sua visualização completa (*Figura 26*), nota-se que o melhor modelo já parte de 84,6% de acurácia. Ao nível de 70,2%, é possível trabalhar com um modelo com somente 10 regras *fuzzy*. Pelo alto nível de mistura entre classes, é possível notar na *Figura 25* maior variabilidade no gráfico de acurácia contra dispersão, com geração de modelos com assertividade abaixo dos 70%.

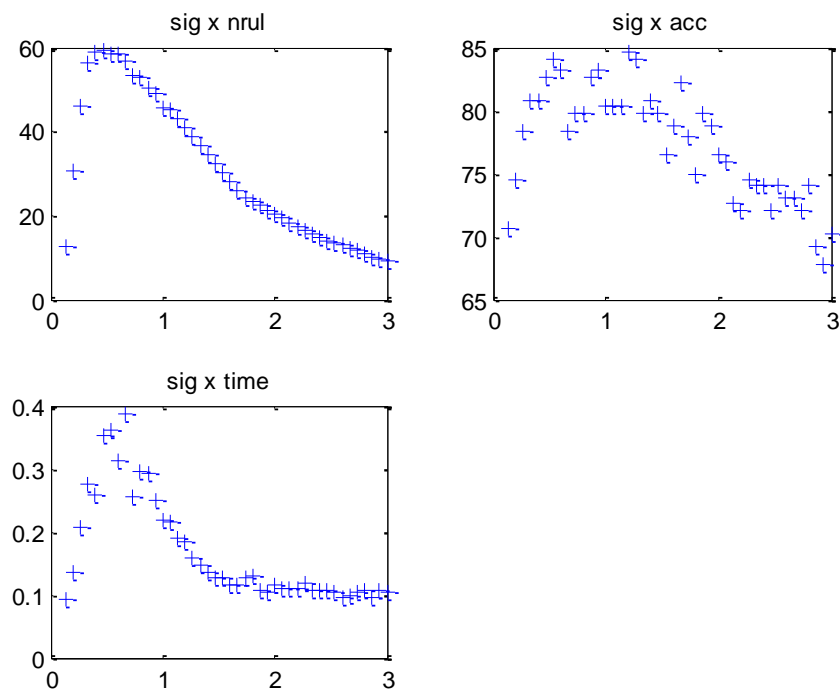


Figura 25 – Relação do parâmetro de dispersão com resultados da base Sonar.

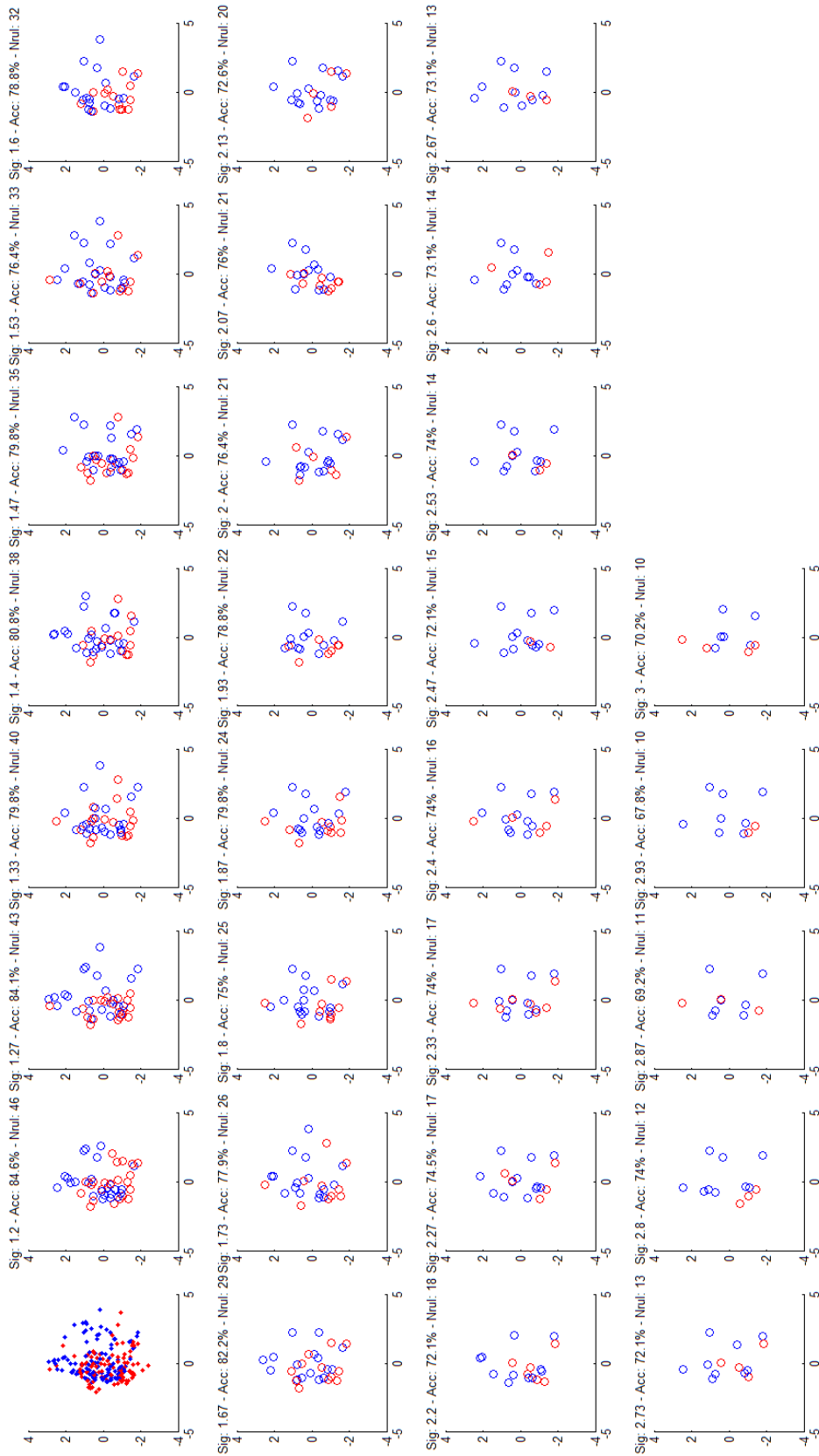


Figura 26 – Visualização completa dos modelos da base Sonar.

4.3 Dendrogramas

A análise da informação do dendrograma² é realizada ao escolher um nível de corte (apresentado no eixo vertical do gráfico) e verificar a quantidade de traços do dendrograma são cortadas por uma linha que seja colocada imediatamente abaixo do nível escolhido. Na tabela apresentada antes de cada dendrograma, se tem a relação entre os níveis, os parâmetros de dispersão e os parâmetros de acurácia. Foram deixadas também as quantidades de regras de cada classe do problema para que se possa verificar os exemplos que serão dados de utilização do dendrograma.

No caso da base Iris, o dendrograma foi apresentado na *Figura 27*, e os parâmetros relacionados a cada nível estão apresentados na *Tabela 3*. No dendrograma, foi feito um corte exemplo imediatamente abaixo do nível 5 e é possível notar que foram cortados dois traços na cor azul, dois na cor vermelha e três na cor verde. Através da tabela de parâmetros, pode-se concluir que o modelo no nível de corte que foi feito está associado ao parâmetro de dispersão igual a 0,60 com acurácia de 90,7%.

Relação entre nível do dendrograma, o parâmetro de dispersão, a acurácia e a quantidade de regras por classe do problema					
Nível	Dispersão	Acurácia	Qtd. Regras Classe 1	Qtd. Regras Classe 2	Qtd. Regras Classe 3
1	0,27	97,3	4	7	8
2	0,33	96,0	4	5	5
3	0,40	94,0	4	5	3
4	0,53	93,3	3	3	3
5	0,60	90,7	2	2	3
6	0,80	87,3	2	2	2
7	0,87	86,7	2	1	2
8	1,07	84,0	2	1	1
9	1,13	83,3	1	1	1

Tabela 3 – Parâmetros Dendrograma Base Iris

² É digno de nota que para a geração do dendrograma no software MATLAB, foi utilizado o algoritmo presente no *Apêndice C*.

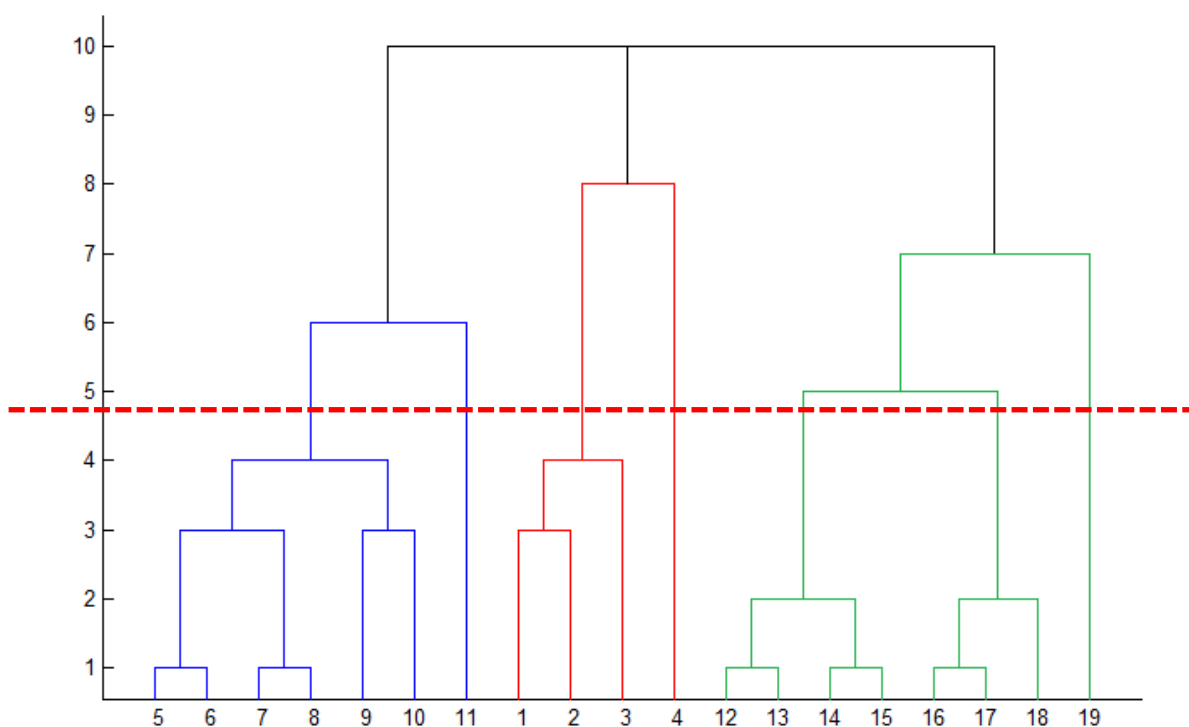


Figura 27 – Dendrograma da Base Iris

No caso da base Balance, o dendrograma foi apresentado na *Figura 28*, e os parâmetros relacionados a cada nível estão apresentados na *Tabela 4*. No dendrograma, foi feito um corte exemplo imediatamente abaixo do nível 2 e é possível notar que foram cortados oito traços na cor azul, seis na cor vermelha e apenas um na cor preta. Através da tabela de parâmetros, pode-se concluir que o modelo no nível de corte que foi feito está associado ao parâmetro de dispersão igual a 0,67 com acurácia de 86,6%.

Relação entre nível do dendrograma, o parâmetro de dispersão, a acurácia e a quantidade de regras por classe do problema

Nível	Dispersão	Acurácia	Qtd. Regras Classe 1	Qtd. Regras Classe 2	Qtd. Regras Classe 3
1	0,47	90,7	17	1	14
2	0,67	86,6	8	1	6
3	0,87	87,5	4	1	6
4	3,00	69,3	2	1	2

Tabela 4 – Parâmetros Dendrograma Base Balance

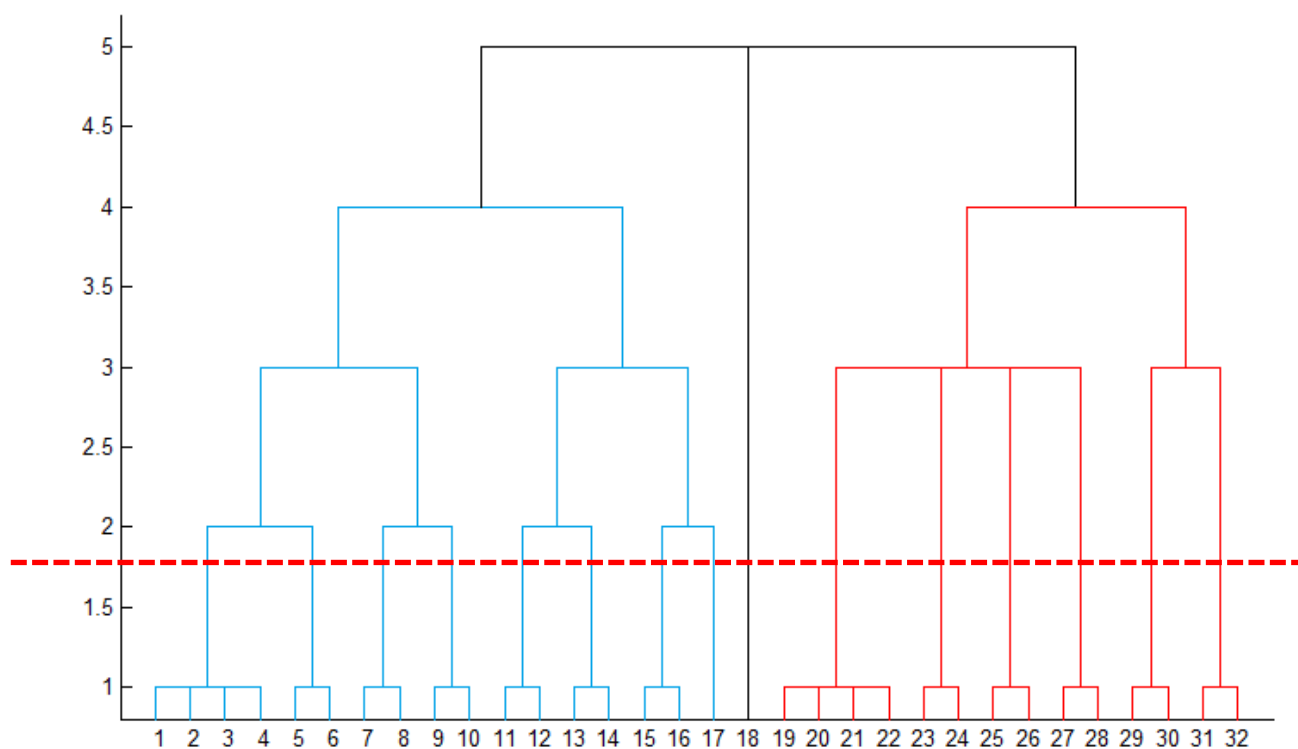


Figura 28 – Dendrograma da Base Balance

Por fim, no caso da base Wine, o dendrograma foi apresentado na *Figura 29*, e os parâmetros relacionados a cada nível estão apresentados na *Tabela 5*. No dendrograma, foi feito um corte exemplo imediatamente abaixo do nível 10 e é possível notar que foram cortados cinco traços na cor verde, sete na cor vermelha e quatro na cor azul. Através da tabela de parâmetros, pode-se concluir que o modelo no nível de corte que foi feito está associado ao parâmetro de dispersão igual a 1,07 com acurácia de 94,4%.

Relação entre nível do dendrograma, o parâmetro de dispersão, a acurácia e a quantidade de regras por classe do problema

Nível	Dispersão	Acurácia	Qtd. Regras Classe 1	Qtd. Regras Classe 2	Qtd. Regras Classe 3
1	0,13	98,3	20	21	19
2	0,40	98,3	18	17	14
3	0,53	98,3	13	14	10
4	0,60	98,3	13	13	9
5	0,73	97,8	9	12	8
6	0,80	98,3	8	8	8
7	0,87	94,9	8	8	7
8	0,93	95,5	6	8	6
9	1,00	98,3	6	7	4
10	1,07	94,4	5	7	4
11	1,13	97,8	5	6	4
12	1,20	96,6	5	5	4
13	1,27	97,8	4	4	3
14	1,33	96,1	3	4	3
15	1,60	97,2	3	4	1
16	2,07	95,5	2	3	1
17	2,33	93,8	2	2	1
18	3,00	91,6	1	2	1

Tabela 5 – Parâmetros Dendrograma Base Wine

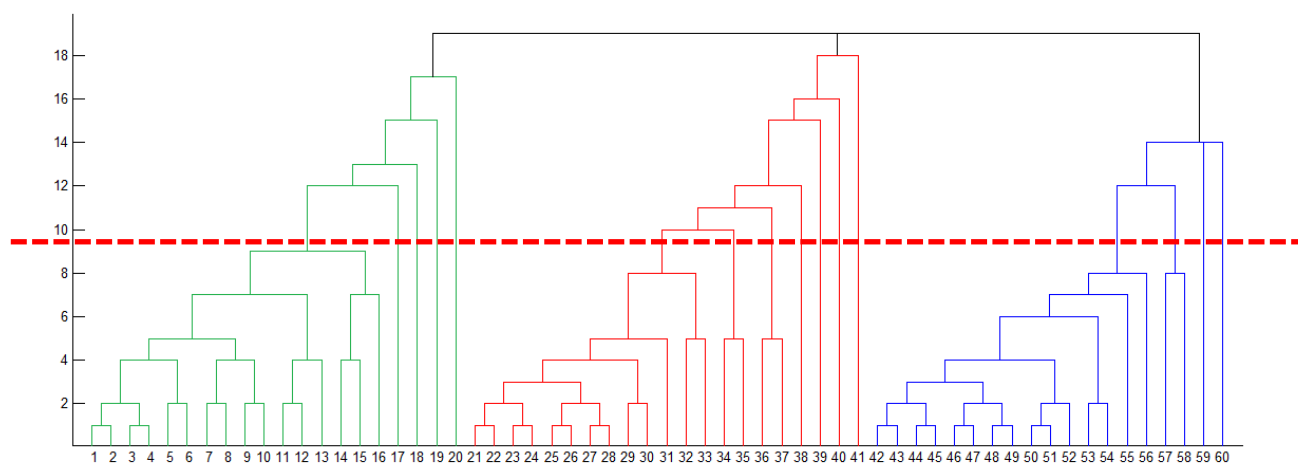


Figura 29 – Dendrograma da Base Wine

CAPÍTULO 5 – Conclusão e Próximos Passos

Com as discussões dos resultados realizadas no capítulo anterior, que envolvem análises com relação ao parâmetro de dispersão do modelo, a acurácia atingida e a interpretabilidade do modelo e dos dados, é possível concluir que tanto os gráficos de visualização completa quanto os dendrogramas auxiliam no entendimento da aplicação da modelagem e dos resultados obtidos.

Com pouco entendimento do comportamento da base de dados gerado através de uma análise descritiva, o usuário da metodologia terá então informações suficientes para escolher o nível com que quer trabalhar, podendo ponderar a relação entre acurácia e a sua capacidade de interpretar o modelo e a aplicação do modelo nos dados.

Entretanto, com o algoritmo gerado para a construção do dendrograma, somente foi possível chegar ao gráfico com as bases que tinham três classes, motivo que levou a apresentação da informação de somente três das dez bases estudadas. A dúvida levantada nesse ponto é se há uma limitação do Matlab na construção de um dendrograma ou se seria o caso de um ajuste do algoritmo gerado. Portanto, é necessário aprofundar a avaliação dessa questão, de modo a contornar os problemas encontrados.

Outro ponto que merece destaque é em relação ao tempo de processamento dos modelos. Para a redução do tempo, é possível abrir dois caminhos. O primeiro de explorar a implementação paralela do algoritmo, visto que o cálculo do modelo a cada iteração não depende do resultado anterior e, portanto, pode acontecer em paralelo. O segundo, seria o de se implementar técnicas que facilitem o cálculo e processamento da análise espectral, que é a parte mais pesada do algoritmo, como por exemplo utilizar a aproximação de Nyström, citada em JIA *et al.* (2011) e SHANG *et al.* (2012).

Propõe-se ainda, como continuação da pesquisa, um estudo que compare a abordagem utilizada com outras que também explorem a relação entre acurácia e interpretabilidade, a fim de proporcionar possibilidades metodológicas alternativas que permitam ao usuário o melhor entendimento dos dados com que trabalha.

REFERÊNCIAS BIBLIOGRÁFICAS

ABREU, N. M. M. “Teoria Espectral dos Grafos: Um Híbrido entre a Álgebra Linear e a Matemática Discreta e Combinatória com Origens na Química Quântica”, **Tema Tendências em Matemática Aplicada e Computacional**, v. 6, n. 1, p. 1-10, 2005.

ALCALÁ, R., CANO, J.R., CORDÓN, O., HERRERA, F., VILLAR, P., ZWIR, I. “Linguistic modeling with hierarchical systems of weighted linguistic rules”, **Journal of Approximate Reasoning**, v. 32 n. 2-3, p. 187–215, 2003.

CASTRO, V. E. “Why so many clustering algorithms: a position paper”, **SIGKDD Explorations Newsletter**, v. 4, n. 1, p. 65-75, jun. 2002.

CHEN, Y. X., WANG, J. Z. “Support vector learning for fuzzy rule based classification systems”, **IEEE Transactions of Fuzzy Systems**, v. 11, p. 716-728, dez. 2003.

EVSUKOFF, A. G., BRANCO, A. C. S., GALICHET, S. “Intelligent Data Analysis and Model Interpretation with Spectral Analysis Fuzzy Symbolic Modeling”, **International Journal of Approximate Reasoning**, v. 52, p. 728-750, 2011.

EVSUKOFF, A. G., GALICHET, S., DE LIMA, B. S. L. P., EBECKEN, N. F. F. “Design of interpretable fuzzy rule-based classifiers using spectral analysis with structure and parameters optimization”, **Fuzzy Sets Syst.**, v. 160, p. 857–881, 2009.

FILIPPONE, M., CAMASTRA, F., MASULLI, F., ROVETTA, S. “A survey of kernel and spectral methods for clustering”, **Pattern Recognition**, v. 41, n. 1, p. 176-190, jan. 2008.

FRED, A. L. N, JAIN, A. K. “Combining multiple clusterings using evidence accumulation”, **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, p. 835-850, 2005.

HAN, J., KAMBER, M. **Data Mining: Concepts and Techniques**. 2 ed. San Francisco, Morgan Kauffman Publishers, 2006.

JAIN, A. K. “Data clustering: 50 years beyond K-means”, **Pattern Recognition Letters**, v. 31, n. 8, p. 651-666, jun. 2010.

JAIN, A. K., MURTY, M. N., FLYNN, P. J. “Data Clustering: A Review,” **ACM Computing Surveys**, v. 31, n. 3, p. 264-323, 1999.

JIA, J., XIAO, X., LUI, B, JIAO, L. “Bagging-based spectral clustering ensemble selection”, **Pattern Recognition Letters**, v. 32, n. 10, p. 1456-1467, 2011.

KOHAVI, R. “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”, **International Joint Conference on Artificial Intelligence**, v. 2, p. 1137–1145, 1995.

LI, W., NG, W. K., N. G., LUI, Y., ONG, K. L. “Enhancing the Effectiveness of Clustering with Spectra Analysis”, **IEEE Trans. Knowledge and Data Eng.**, vol. 19, n. 7, p. 887-902, jul. 2007.

LUXBURG, U. “A tutorial on spectral clustering”, **Statistics and Computing**, v. 17, n. 4, p. 395-416, dez. 2007.

LUXBURG, U. V., BOUSQUET, O., BELKIN M. “Limits of spectral clustering”. In: SAUL, L. K., WEISS Y., BOTTOU L. (Eds). **Advances in Neural Information Processing Systems (NIPS)**, v. 17. MIT Press, Cambridge, MA, 2005.

MIRZAEI, A., RAHMATI, M. “A novel hierarchical-clustering-combination scheme based on fuzzy-similarity relations”, **IEEE Trans. Fuzzy Syst.**, v. 18, n. 1, p. 27- 39, 2010.

NG, A. Y., JORDAN, M. I., WEISS, Y. “On spectral clustering: analysis and an algorithm”, in: DIETTERICH, T. G., BECKER, S., GHAMRANI, Z. (Eds.), **Advances in Neural Information Processing Systems**, v. 14, MIT Press, Cambridge, MA, 2002.

SANTOS, C. K., EVSUKOFF, A. G., DE LIMA, B. S. L. P., EBECKEN, N. F. F. “Identificação de Relações Potenciais em Redes Sociais através da Detecção Espectral da Estrutura de Comunidades”. In: **30º Congresso Ibero-Latino-Americano de Métodos Computacionais em Engenharia (CILAMCE'09)**, Buzios, 2009.

SHANG, F. JIAO, L. C., SHI, J., WANG, F., GONG, M. “Fast affinity propagation clustering: a multilevel approach Pattern Recognition”, **Pattern Recognition**, v. 45, n. 1, p. 474–486, jan. 2012.

XU, R., WUNDSCH, D. “Survey of clustering algorithms”, **IEEE Transactions on Neural Networks**, v. 16, n. 3, p. 645-678, 2005.

APÊNDICE A – Descrição Detalhada das Bases de *Benchmark*

Para descrever detalhadamente o comportamento de cada base, foram gerados no *software* MATLAB:

- O gráfico de projeção em duas dimensões baseado em ACP, que nos problemas de classificação supervisionada permite visualizar a distribuição dos registros de cada classe em cada projeção. Sabe-se que em problemas de múltiplas variáveis é impossível visualizar os dados em mais do que três dimensões e que a visualização 3D já se torna confusa;
- O *data image*, ferramenta que permite visualizar a presença de grupos e *outliers* nos dados gerada a partir da matriz de distâncias entre os registros;
- A matriz de correlação, apresentada também de forma visual, permite a análise de correlação linear entre as variáveis do problema. A presença de forte correlação entre duas variáveis pode representar redundância de informação;
- *Scatter Plot*, ou gráficos de pontos, foram gerados para algumas das bases utilizadas, com o objetivo de evidenciar o comportamento relacional de algumas das variáveis dos problemas.

A.1. Descrição da Base Iris

A base de dados Iris contém 150 registros e três classes com 50 registros associados cada, e cada classe refere-se a um tipo de planta Íris. As variáveis representam duas medidas (comprimento e largura) das pétalas e sépalas das plantas analisadas.

Base Iris	
Quantidade de Variáveis	4
Quantidade de Registros	150
Quantidade de Grupos	3

Tabela 6 – Características da base Iris.

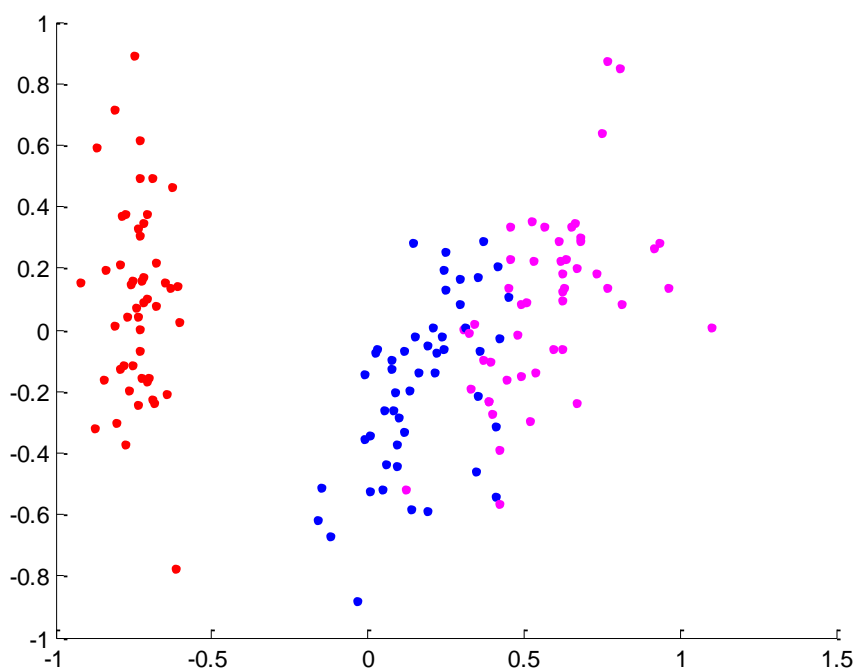


Figura 30 – Gráfico de projeção da base Iris.

Através do gráfico de projeção, representado na *Figura 30*, é possível observar que uma das três classes é linearmente separável das outras duas, que não são linearmente separáveis umas das outras.

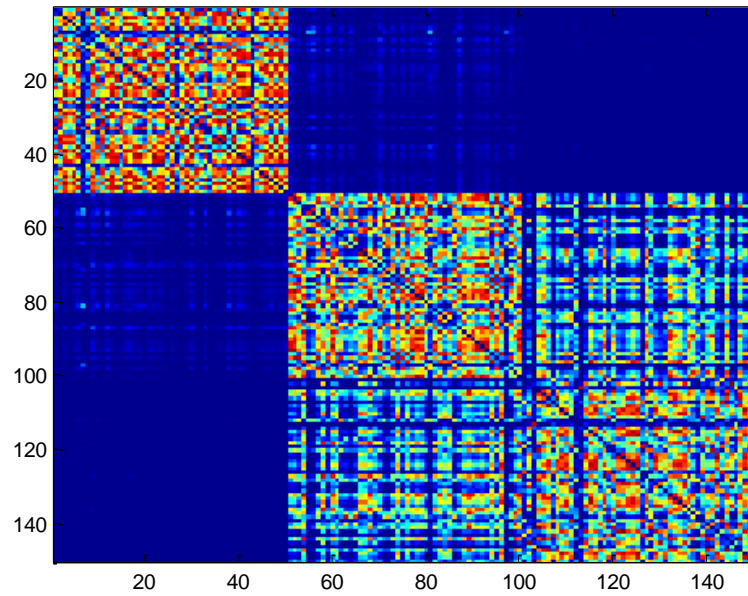


Figura 31 – Data Image da base Iris.

O *Data Image* (Figura 31) evidencia a presença de três classes na base Iris e também que a ordem dos dados está influenciada pelas classes. A utilização da validação cruzada para o cálculo da acurácia fará com que essa ordenação não interfira no resultado.

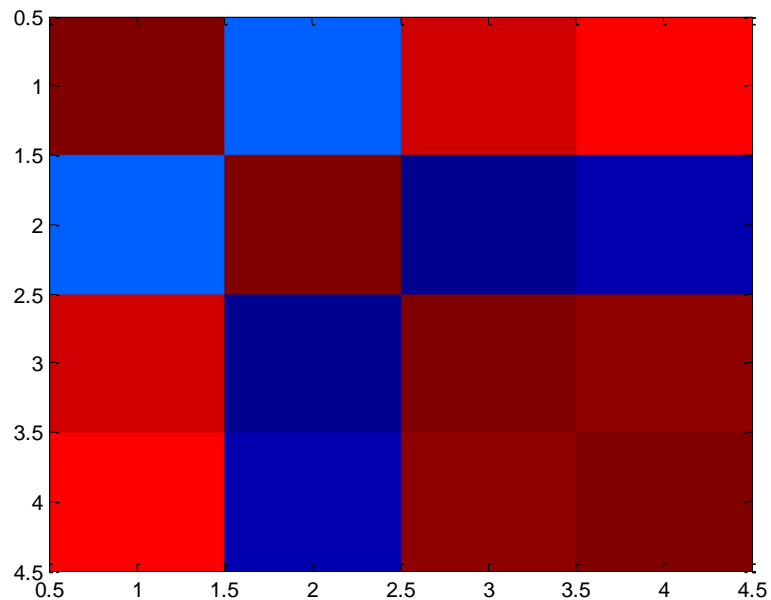


Figura 32 – Matriz de Correlação da base Iris.

A matriz de correlação (*Figura 32*) mostra forte correlação linear positiva (o acréscimo de uma implica em acréscimo da outra) entre as variáveis 3 e 4 e moderada entre 1 e 2, e 1 e 3. A variável 2 está correlacionada de modo negativo (o acréscimo de uma implica em decréscimo da outra) com as demais variáveis do problema.

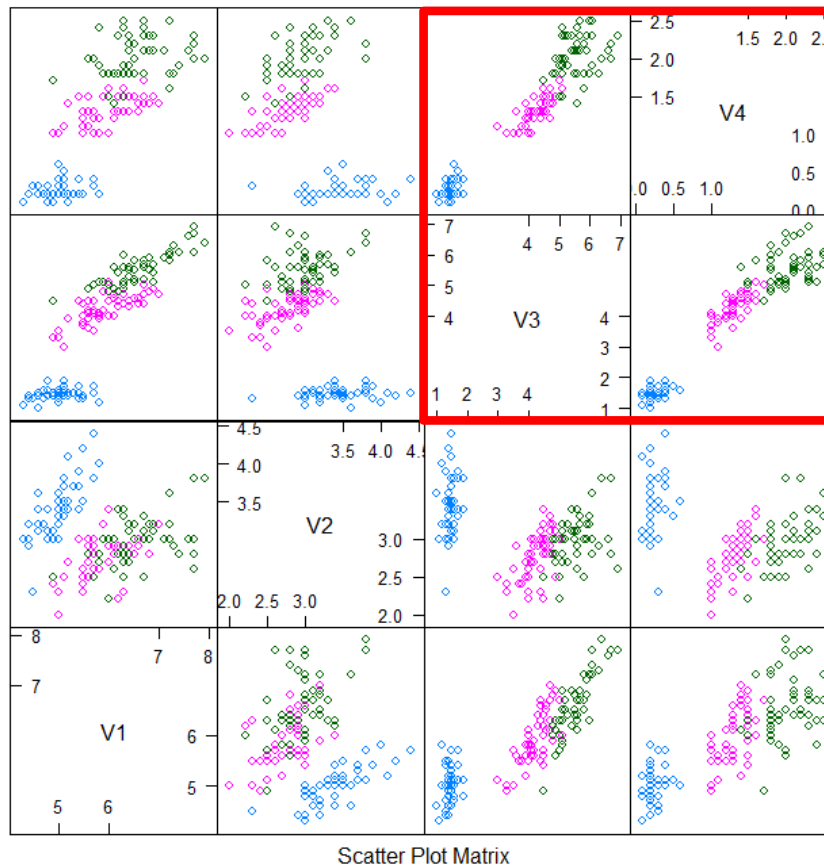


Figura 33 – Scatter Plot da base Iris.

Para ilustrar melhor a incidência de correlação linear positiva entre as variáveis da base Iris, foi gerado no *software* R um *scatter plot*, apresentado na *Figura 33*. Foram destacados em vermelho os gráficos de pontos entre as variáveis 3 e 4 e é nítido o comportamento linear entre elas.

A.2. Descrição da Base Balance

A base de dados Balance contém 625 registros com 4 variáveis e também apresenta 3 classes. Cada registro representa o resultado da escala de uma balança, que pode ser esquerda, direita ou balanceado. As variáveis representam duas medidas (peso e distância) a esquerda e a direita.

Base Balance	
Quantidade de Variáveis	4
Quantidade de Registros	625
Quantidade de Grupos	3

Tabela 7 – Características da base Balance.

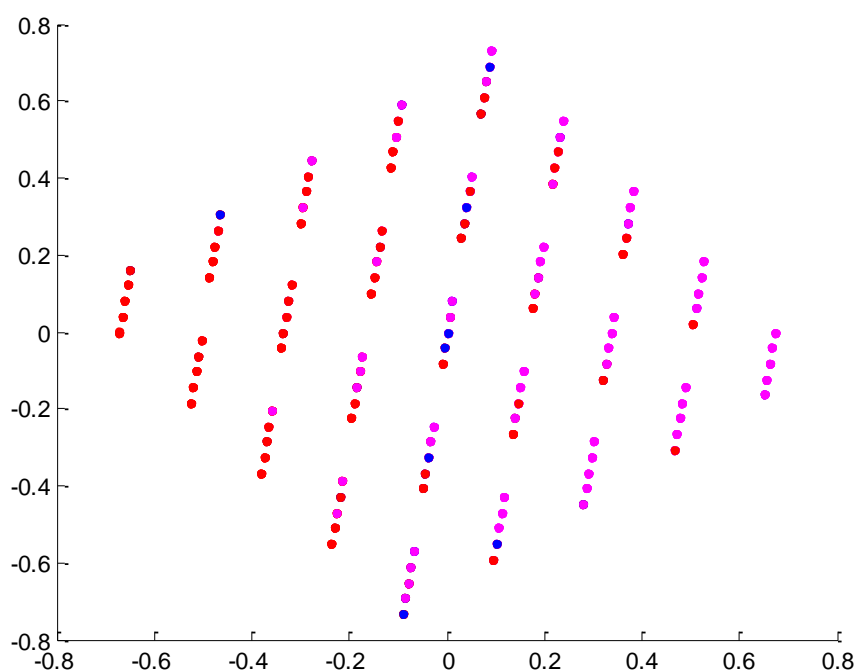


Figura 34 – Gráfico de projeção da base Balance.

Através do gráfico de projeção, representado na *Figura 34*, é possível observar uma das características dessa base que é a fraca presença da classe “balanceado” representada em azul na imagem. Geralmente essa situação é problemática para os algoritmos de classificação supervisionada, pois a tendência é que essa classe seja desprezada.

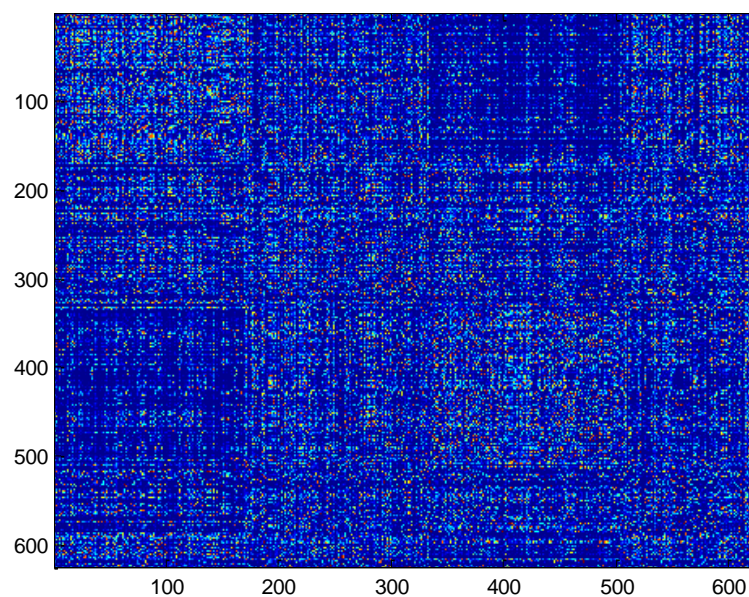


Figura 35 – Data Image da base Balance.

O *Data Image* (Figura 35) da Balance não evidencia a presença das classes do problema e também não dá indícios de viés na ordem dos dados.

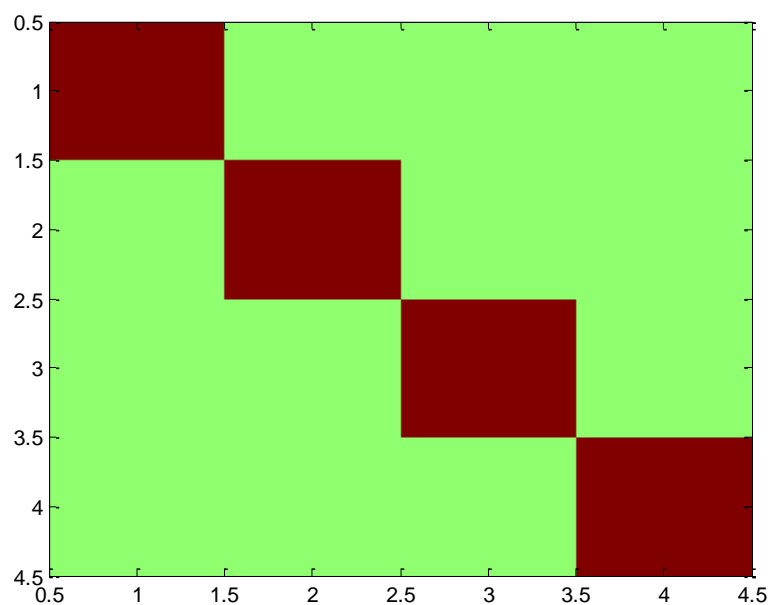


Figura 36 – Matriz de Correlação da base Balance.

A matriz de correlação (*Figura 36*) mostra ausência de correlação linear entre todos os pares de variáveis do problema.

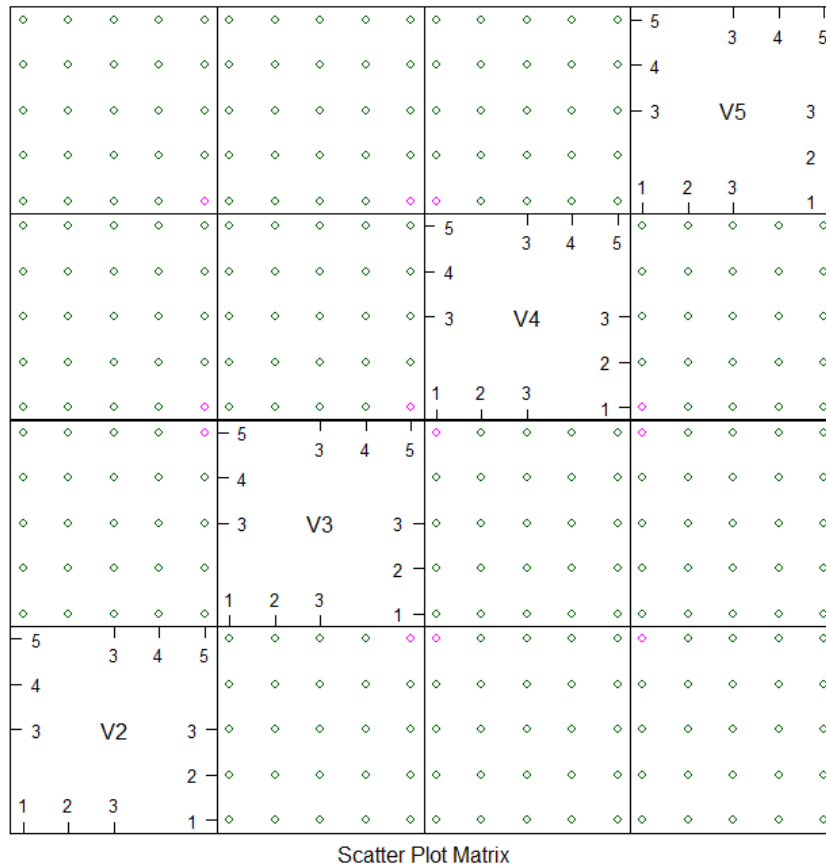


Figura 37 – Scatter Plot da base Balance.

Com ajuda do *scatter plot* (*Figura 37*) é possível confirmar a informação visualizada na matriz de correlação, de ausência de correlação linear entre as variáveis do problema.

A.3. Descrição da Base Diabetes

A base Diabetes contém 768 registros com 8 variáveis e apenas 2 classes. Cada registro representa uma paciente, do sexo feminino e com mais de 21 anos de idade com herança de uma tribo indígena norte-americana, que tem como característica ter a maior

prevalência de diabetes tipo 2 no mundo. As variáveis representam características das pacientes, como por exemplo, o número de vezes grávidas, pressão arterial, índice de massa corporal e idade.

Base Diabetes	
Quantidade de Variáveis	8
Quantidade de Registros	768
Quantidade de Grupos	2

Tabela 8 – Características da base Diabetes.

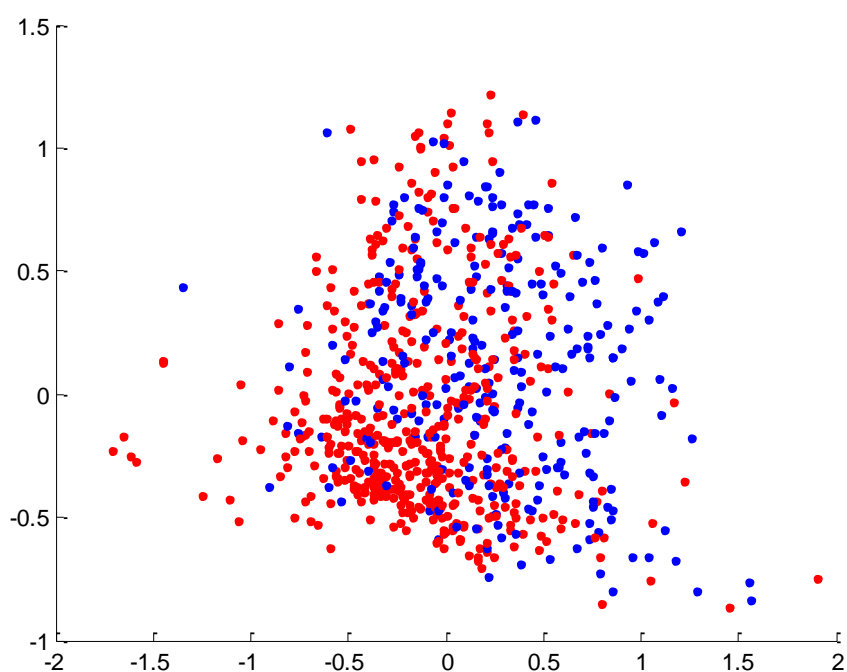


Figura 38 – Gráfico de projeção da base Diabetes.

Pelo gráfico de projeção (*Figura 38*), é possível observar que as duas classes não são linearmente separáveis. Pelo grau de interseção entre as classes, é possível que os algoritmos de classificação supervisionada tenham dificuldades em termos de acurácia.

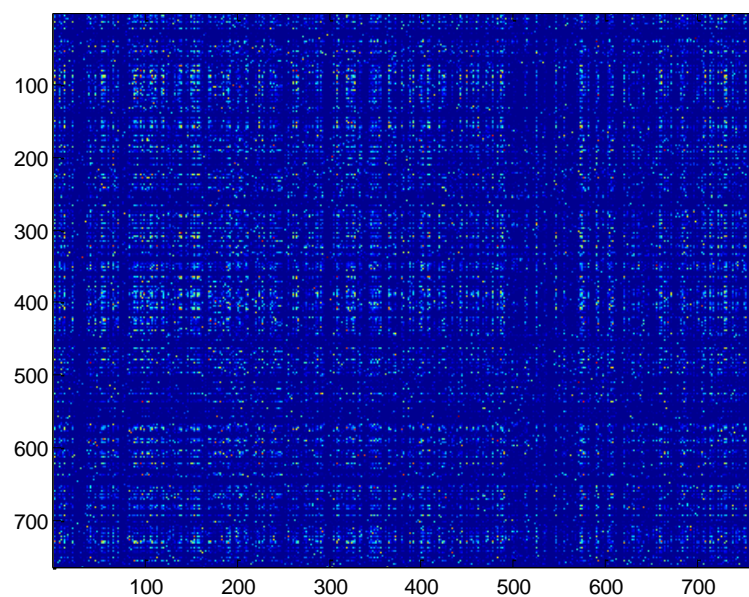


Figura 39 – Data Image da base Diabetes.

Através do *Data Image* (Figura 39) da base Diabetes, não é possível observar evidências da presença das duas classes do problema assim como não há indícios de viés na ordem dos dados.

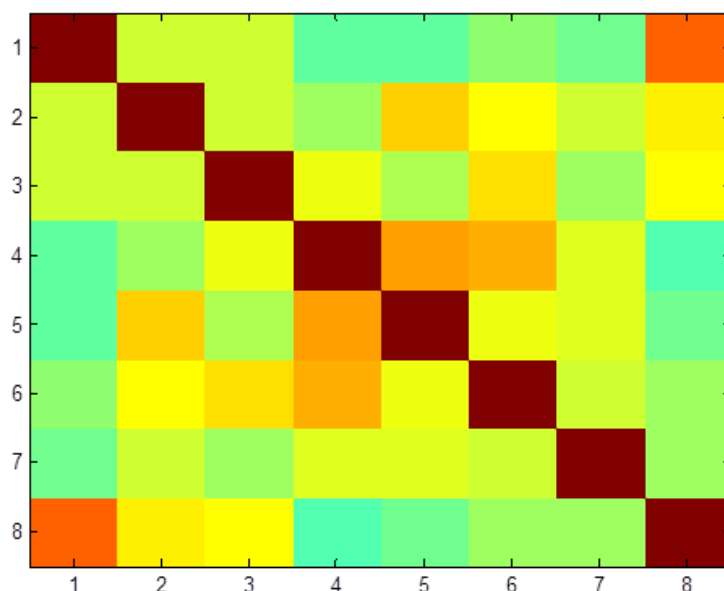


Figura 40 – Matriz de Correlação da base Diabetes.

A matriz de correlação (*Figura 40*) mostra fraca correlação linear entre os pares de variáveis do problema. O maior indício de correlação está presente entre as variáveis 1 e 8.

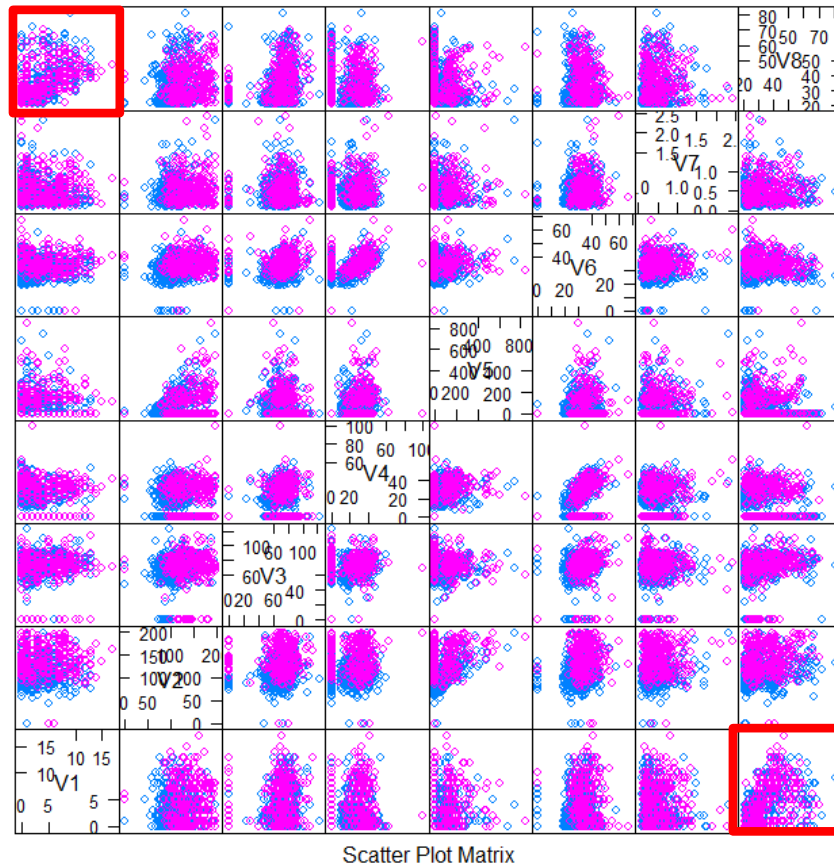


Figura 41 – Scatter Plot da base Diabetes.

É possível confirmar a informação obtida com a matriz de correlação através do *scatter plot* ilustrado na *Figura 41*. Foram destacados em vermelho os gráficos de pontos entre as variáveis 1 e 8, que na matriz de correlação foram apontadas com correlação moderada.

A.4. Descrição da Base Cancer

A base Cancer, assim como a base Diabetes e a base Balance, tem como característica a grande quantidade de registros em comparação às demais bases utilizadas. Será visto na

apresentação de resultados que essa característica é problemática para a metodologia proposta neste trabalho, pela complexidade dos cálculos envolvidos. A base Cancer conta com 9 variáveis e 683 registros, separados em somente 2 classes.

Base Cancer	
Quantidade de Variáveis	9
Quantidade de Registros	683
Quantidade de Grupos	2

Tabela 9 – Características da base Cancer.

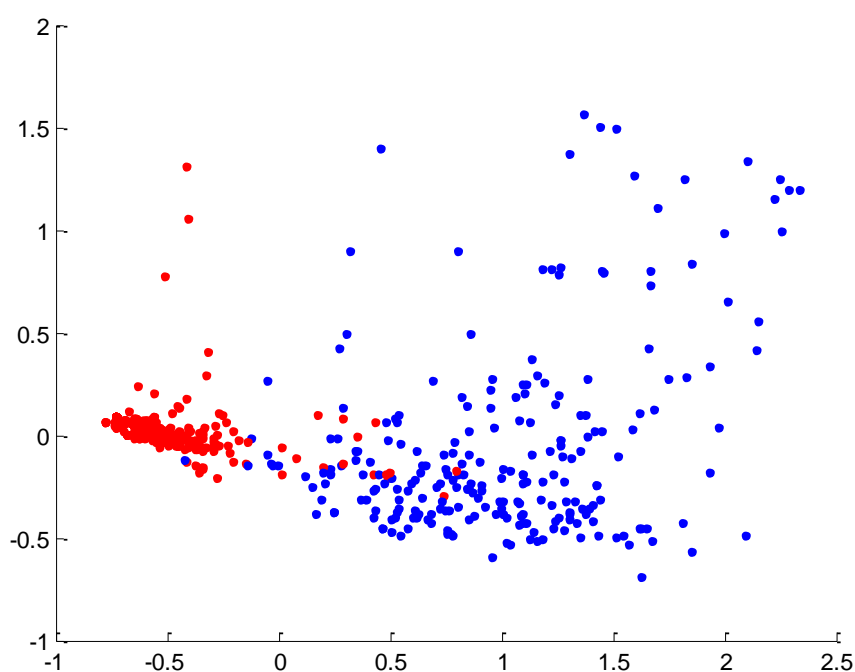


Figura 42 – Gráfico de projeção da base Cancer.

O gráfico de projeção apresentado na *Figura 42* indica que as duas classes do problema são quase linearmente separáveis, ao contrário da base Diabetes, por exemplo. Esse comportamento facilita a atuação dos algoritmos de classificação supervisionada.

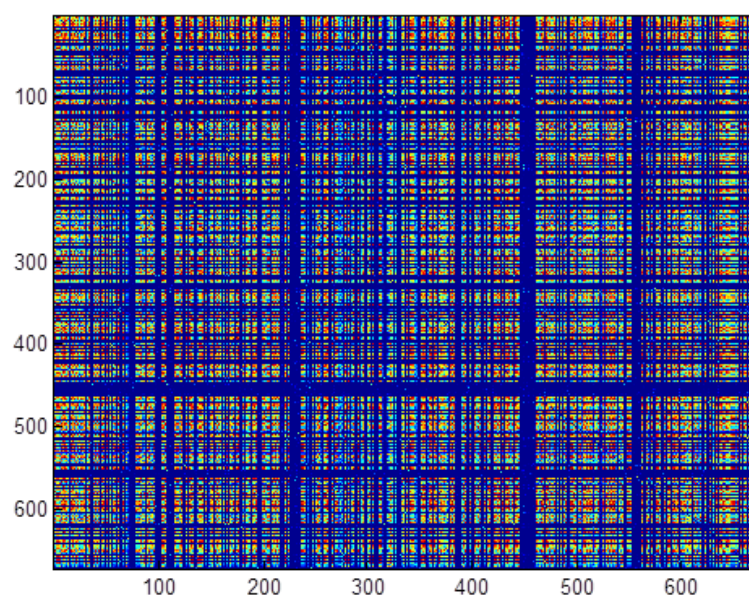


Figura 43 –Data Image da base Cancer.

Com o *Data Image* (Figura 43) da base Cancer, não é possível observar evidências da presença das duas classes do problema assim como não há indícios de viés na ordem dos dados na base.

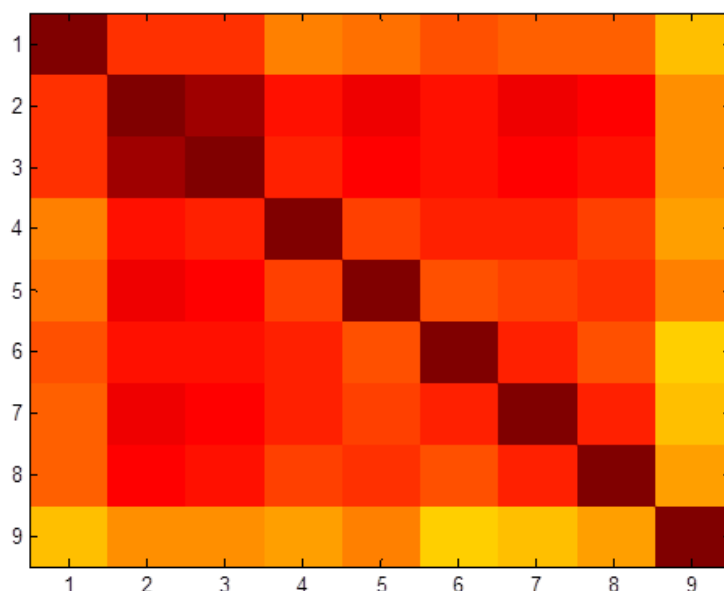


Figura 44 – Matriz de Correlação da base Cancer.

Ao contrário da base Diabetes, a matriz de correlação (*Figura 44*) da base Cancer mostra forte correlação linear positiva entre os pares de variáveis do problema, com exceção da última variável ou variável 9, o que representa redundância de informação. O maior índice de correlação está presente entre as variáveis 2 e 3.

A.5. Descrição da Base Glass

A base Glass contém 214 registros com 8 variáveis e 6 classes. Cada registro representa a análise de uma amostra de vidro, com o objetivo de identificar de qual tipo é o vidro. A base foi gerada em um estudo motivado por investigação criminológica. Na cena de um crime, o vidro deixado pode ser usado como prova, e para isso é necessário que ele seja corretamente identificado. As variáveis representam níveis de materiais encontrados na composição da amostra analisada, como magnésio, alumínio, ferro, etc.

Base Glass	
Quantidade de Variáveis	9
Quantidade de Registros	214
Quantidade de Grupos	6

Tabela 10 – Características da base Glass.

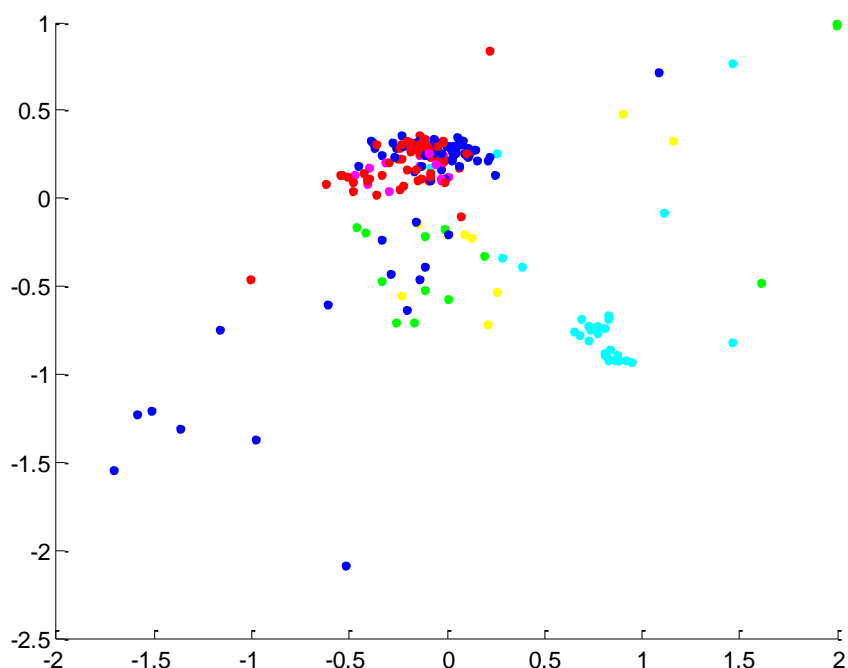


Figura 45 – Gráfico de projeção da base Glass.

O gráfico de projeção da base Glass, apresentado na *Figura 45*, indica há múltiplos comportamentos em termos de classes. É fácil de perceber, por exemplo, que as classes representadas em magenta e azul escuro não são linearmente separáveis mas encontram-se afastadas das classes representadas em verde e amarelo. O comportamento de classes misturadas que pode ser observado, geralmente impacta de modo negativo no desempenho do classificador.

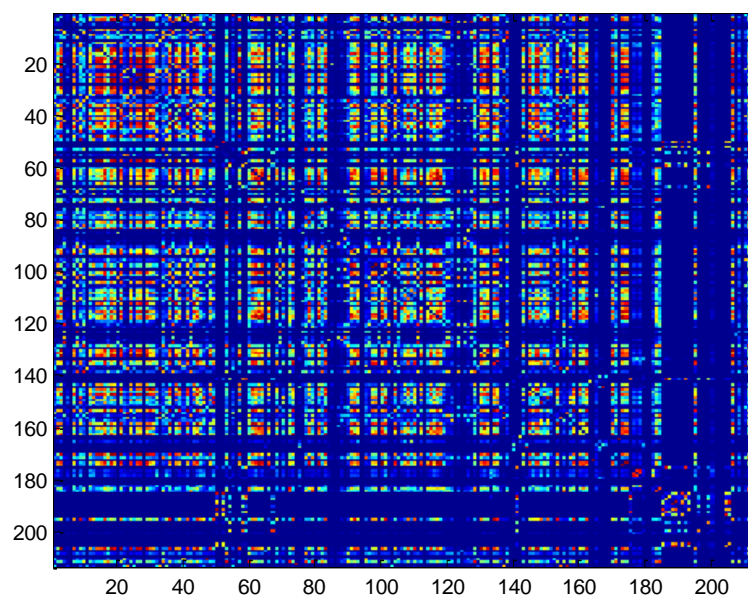


Figura 46 –Data Image da base Glass.

Com o *Data Image* (Figura 46) da base Glass, não é possível observar evidências da presença das classes do problema assim como não há indícios de viés na ordem dos dados na base.

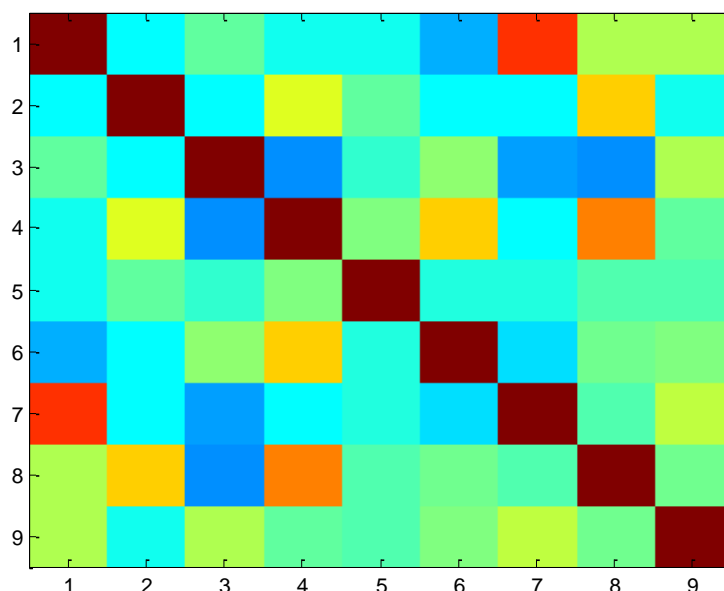


Figura 47 – Matriz de Correlação da base Glass.

Entretanto, a matriz de correlação (*Figura 47*) mostra existência de correlação linear positiva entre as variáveis 1 e 7 e também correlação linear negativa entre as variáveis 3 e 8, e 3 e 4.

A.6. Descrição da Base Wine

A base Wine representa o resultado de uma análise química de vinhos cultivados na mesma região da Itália, mas derivados de 3 cultivos diferentes. A análise determinou as quantidades de 13 constituintes encontrados em cada um dos três tipos de vinhos. Portanto, a base é composta por 13 variáveis e 3 classes, que representam os 3 tipos de vinho. São encontrados 178 resultados de análise, que constituem os registros da base.

Base Wine	
Quantidade de Variáveis	13
Quantidade de Registros	178
Quantidade de Grupos	3

Tabela 11 – Características da base Wine.

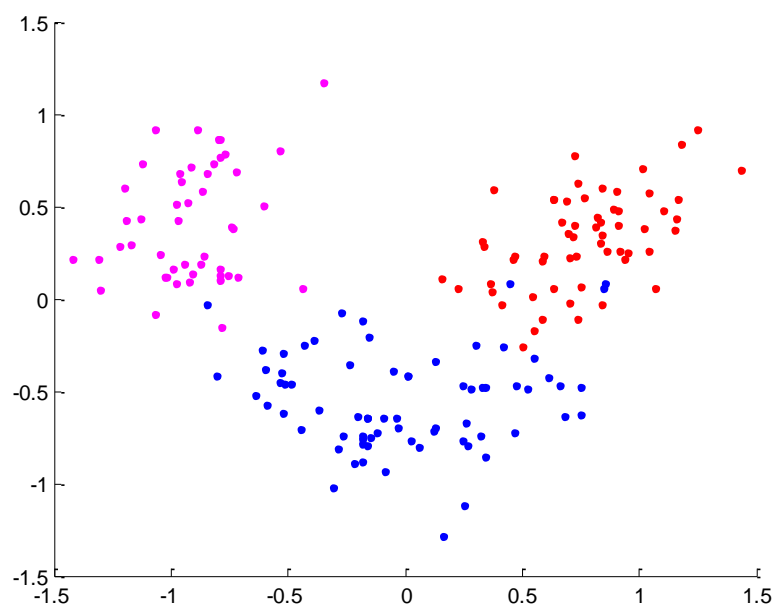


Figura 48 – Gráfico de projeção da base Wine.

O gráfico de projeção apresentado na *Figura 48*, indica que trata-se de uma base de fácil classificação, uma vez que é composta de 3 classes linearmente separáveis com poucos registros de interseção.

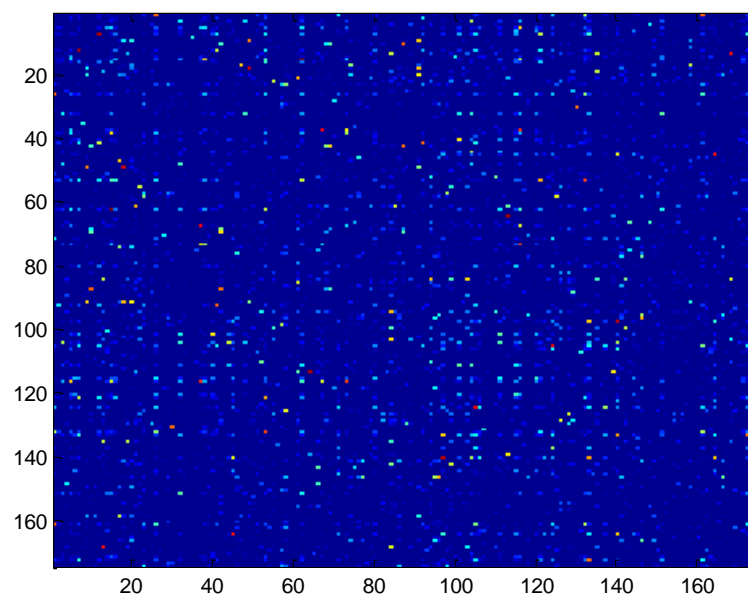


Figura 49 – Data Image da base Wine.

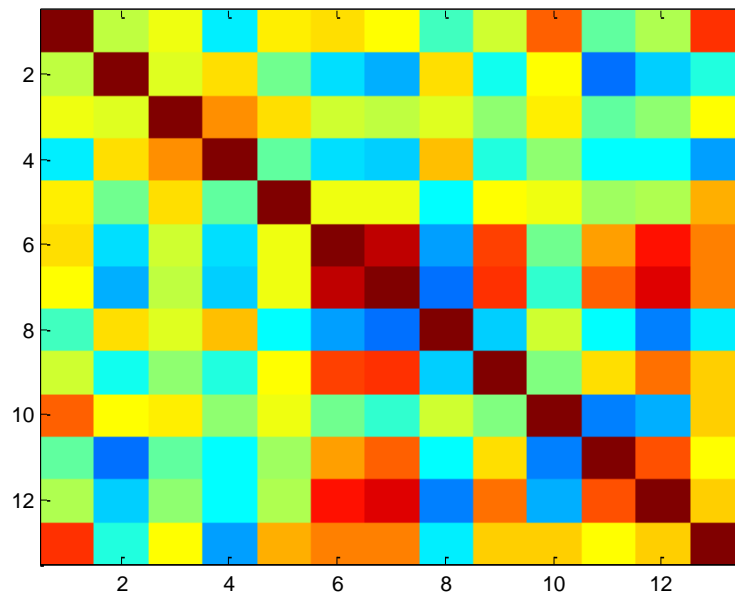


Figura 50 –Matriz de Correlação da base Wine.

Com o *Data Image* da base Wine, representado na *Figura 49*, não é possível observar evidências da presença das classes do problema assim como não há indícios de viés na ordem dos dados na base. Entretanto, a matriz de correlação (*Figura 50*) mostra existência de variáveis bastante correlacionadas positivamente, como é o caso das variáveis 6 e 7. As variáveis 2 e 11 apresentam forte correlação linear negativa.

A.7. Descrição da Base Heart

A base Diabetes contém 270 registros com 13 variáveis e apenas 2 classes. Cada registro representa uma paciente, que foi avaliado no sentido de identificar se ele possui ou não doenças cardíacas (as duas classes do problema). As variáveis representam características das pacientes, como por exemplo, idade, sexo, nível de colesterol, tipo de dor no peito, etc.

Base Heart	
Quantidade de Variáveis	13
Quantidade de Registros	270
Quantidade de Grupos	2

Tabela 12 – Características da base Heart.

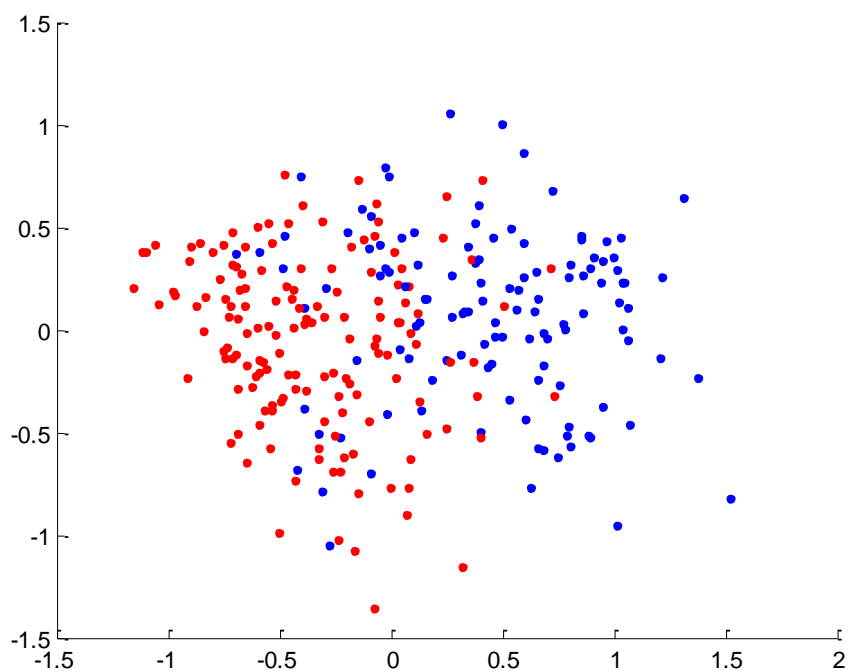


Figura 51 – Gráfico de projeção da base Heart.

Com o gráfico de projeção da base Heart, representado na *Figura 51*, é possível notar que essa base tem características semelhantes à base Diabetes, pois é composta de duas classes que não são linearmente separáveis.

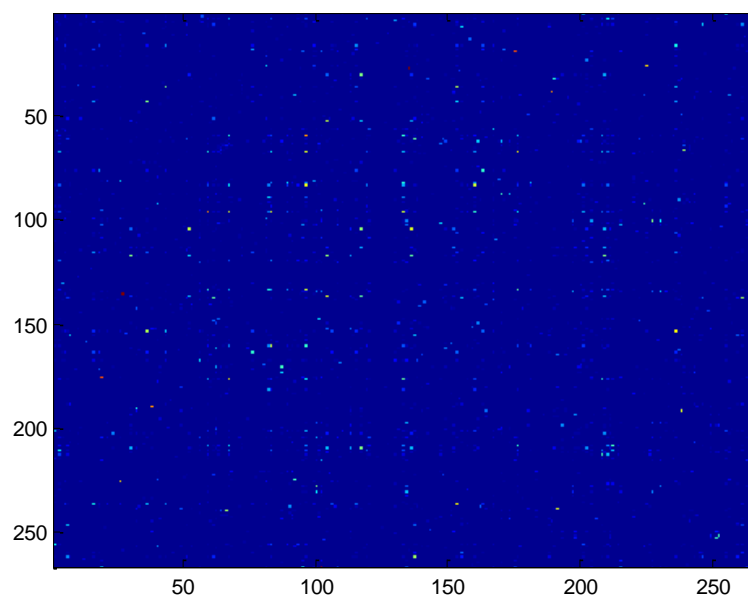


Figura 52 – Data Image da base Heart.

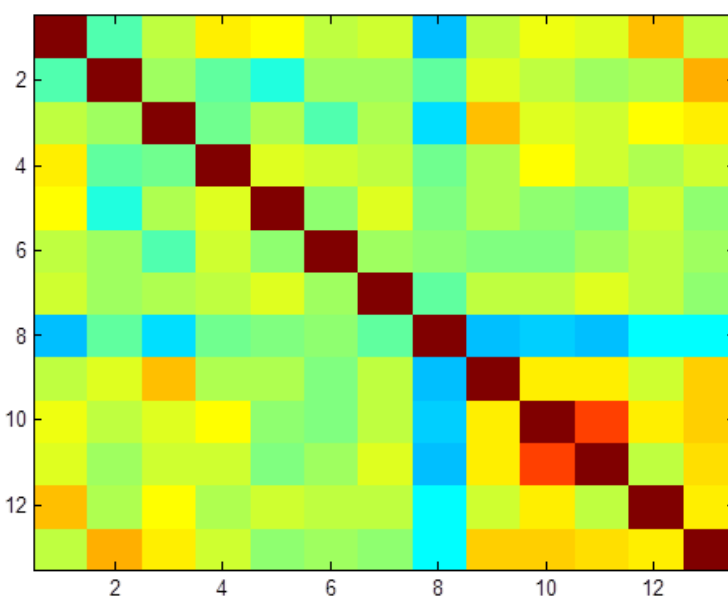


Figura 53 – Matriz de Correlação da base Heart.

Com o *Data Image* da base Heart, representado na *Figura 52*, não é possível observar evidências da presença das classes do problema assim como não há indícios de viés na ordem dos dados na base. Com a matriz de correlação (*Figura 53*), é possível observar somente a

existência de correlação linear positiva entre as variáveis 10 e 11 e que a variável 8 se correlaciona de modo negativo com algumas outras variáveis, como a 1 e a 11, por exemplo.

A.8. Descrição da Base Image

A base Image é caracterizada pela distribuição de seus 210 registros nas 7 classes presentes e 18 variáveis. É possível observar através do gráfico de projeção, ilustrado na *Figura 54*, que há ao mesmo tempo, classes bastante separadas como a de cor azul escuro, por exemplo, e classes bastante misturadas como as de cor preto e magenta. Essa característica, geralmente, dificulta o desempenho do classificador.

Base Image	
Quantidade de Variáveis	18
Quantidade de Registros	210
Quantidade de Grupos	7

Tabela 13 – Características da base Image.

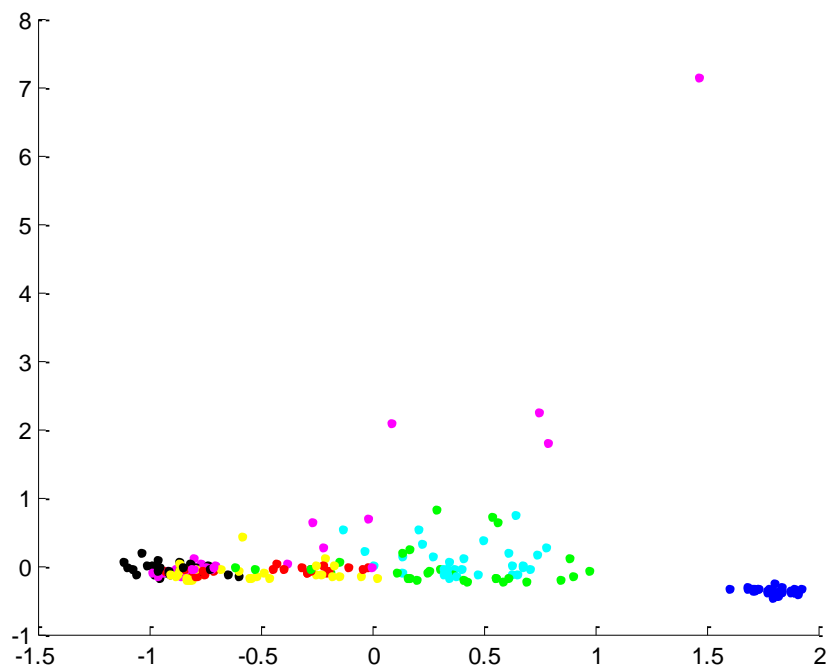


Figura 54 – Gráfico de projeção da base Image.

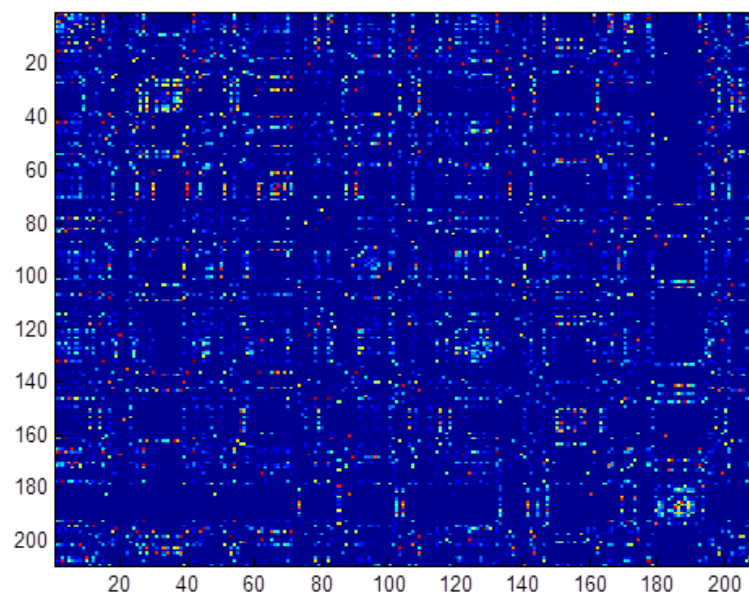


Figura 55 – *Data Image* da base Image.

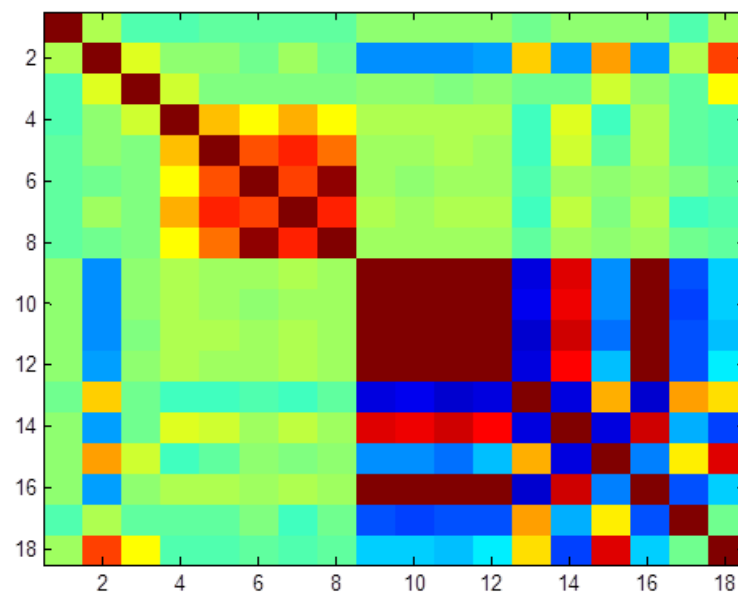


Figura 56 – Matriz de Correlação da base Image.

Com a matriz de correlação (*Figura 53*), é possível notar presença de fortes correlações positivas e negativas entre as variáveis da base. Essa característica implica em redundância de informação presente, que geralmente prejudica o desempenho do classificador sem apresentar vantagem em termos de acurácia. Vale o comentário de que, com a grande quantidade de variáveis presentes, a leitura de um gráfico do tipo *scatter plot* já seria bastante prejudicada.

A.9. Descrição da Base Ionosphere

A base Ionosphere tem 32 variáveis e 351 registros, classificados entre 2 classes somente. Estes dados de radar foram coletados por um sistema cujo objetivo era a identificação de elétrons livres na ionosfera. "Bons" retornos de radar são aqueles que mostram evidências de algum tipo de estrutura na ionosfera. "Maus" retornos são aqueles que não o fazem. A separação dos dados de radar entre "bons" e "maus" caracterizam as classes do problema.

Base Ionosphere	
Quantidade de Variáveis	32
Quantidade de Registros	351
Quantidade de Grupos	2

Tabela 14 – Características da base Ionosphere.

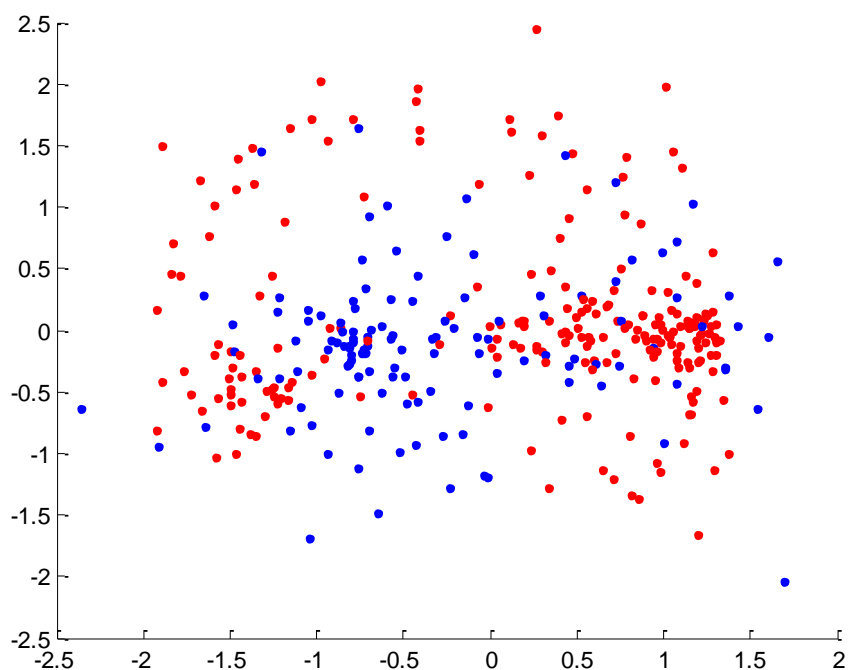


Figura 57 – Gráfico de projeção da base Ionosphere.

É possível observar no gráfico de projeção (*Figura 57*) que as duas classes são bastante misturadas, como também é o caso da base Diabetes.

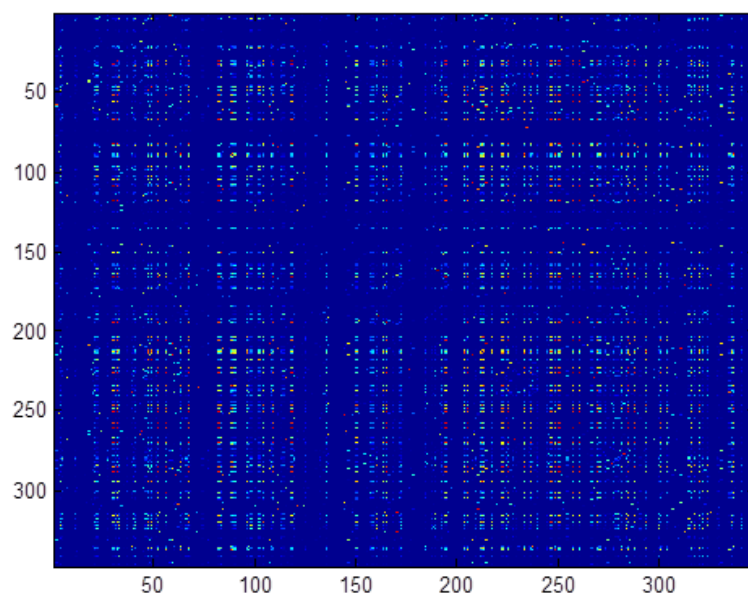


Figura 58 – Data Image da base Ionosphere.

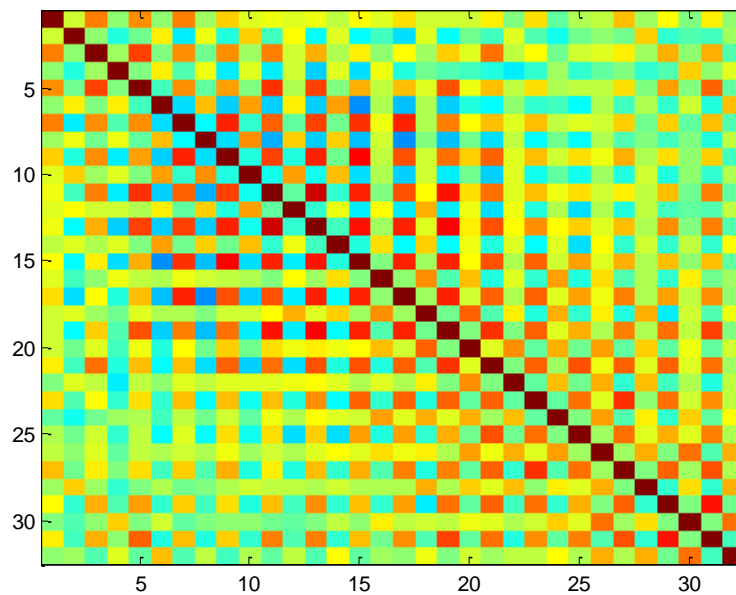


Figura 59 – Matriz de Correlação da base Ionosphere.

Com o *Data Image* da base Ionosphere, representado na *Erro! Fonte de referência não encontrada.*, não é possível observar evidências da presença das classes do problema assim como não há indícios de viés na ordem dos dados na base. Já com a matriz de correlação (*Figura 59*), é possível observar presença de correlação linear moderada entre algumas variáveis, provavelmente pela redundância comum em arrecadação de informações de radar.

A.10. Descrição da Base Sonar

A base Sonar é caracterizada pela grande quantidade de variáveis presentes (60), apesar de contar somente com 208 registros e a separação em duas classes. Os dados foram gerados em um estudo de classificação de sinais de sonar que retornam de cilindros de metal e cilindros de rochas aproximadamente cilíndricas. Cada variável representa a energia dentro de uma faixa de frequência específica, avaliada em um período de tempo. Os registros são classificados como “M” se o objeto é metal ou “R” se é rocha.

Base Sonar	
Quantidade de Variáveis	60
Quantidade de Registros	208
Quantidade de Grupos	2

Tabela 15 – Características da base Sonar.

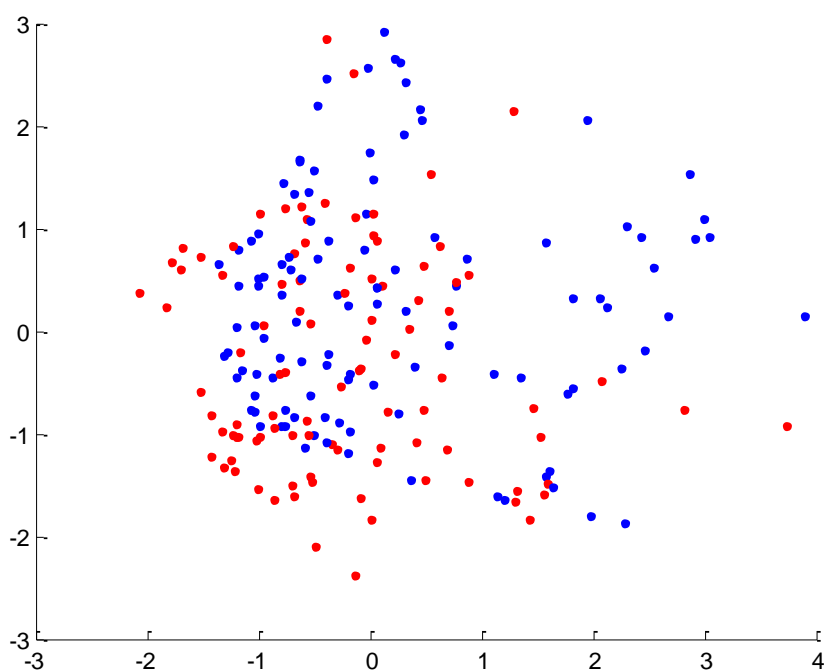


Figura 60 – Gráfico de projeção da base Sonar.

É possível observar no gráfico de projeção (*Figura 60*) que os dados têm comportamento semelhante ao da base Ionosphere, uma vez que são duas classes com bastante interseção.

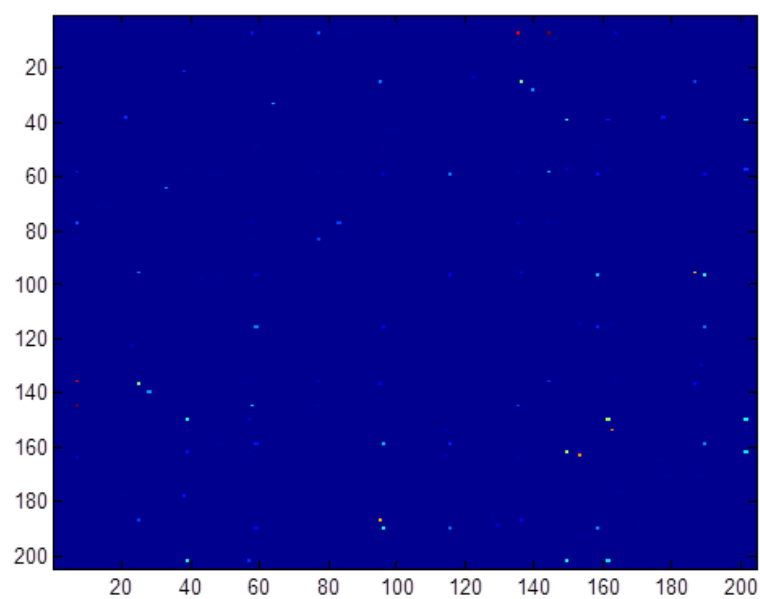


Figura 61 –Data Image da base Sonar.

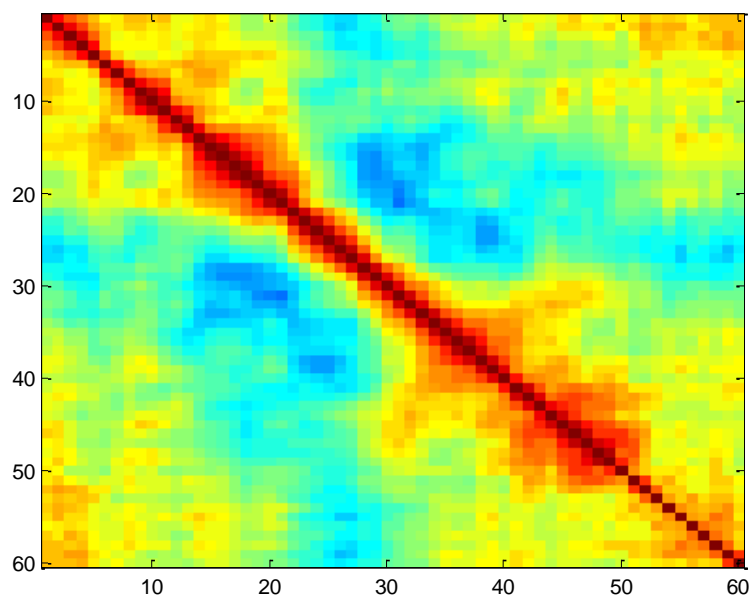


Figura 62 – Matriz de Correlação da base Sonar.

Com o *Data Image* da base Sonar, representado na *Figura 61*, não é possível observar evidências da presença das classes do problema assim como não há indícios de viés na ordem

dos dados na base. Já com a matriz de correlação (*Figura 62*), é possível observar presença de várias variáveis correlacionadas tanto positivamente quanto negativamente. Pela quantidade de variáveis, é possível observar uma regionalização do gráfico, com áreas de correlação moderada negativa, áreas de correlação moderada positiva (perto da diagonal) e áreas que representam ausência de correlação entre as variáveis associadas.

APÊNDICE B – Apresentação dos Gráficos de Melhor Modelo

Serão apresentados os gráficos de projeção por ACP das bases de *benchmark* utilizadas com o posicionamento dos centros de regras do melhor modelo gerado em termos de acurácia.

Base Iris:

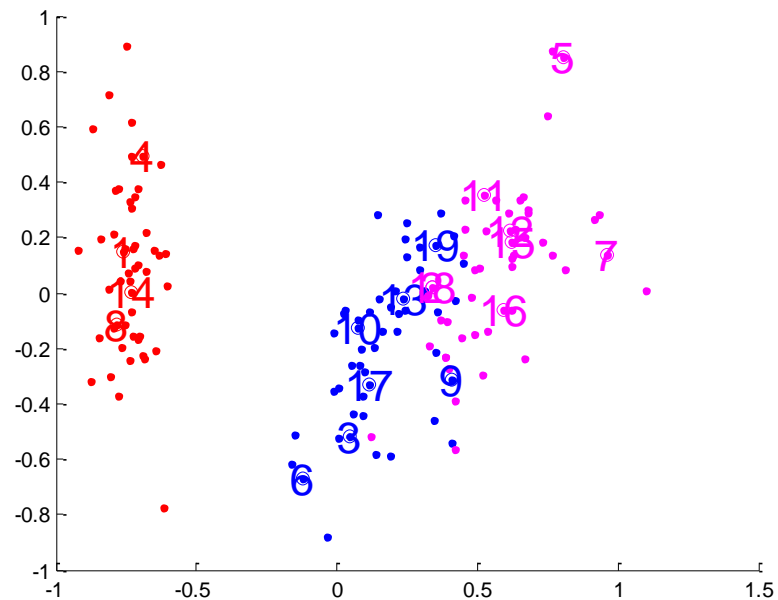


Figura 63 – Projeção da base Iris com os centros de regra do melhor modelo.

Base Balance:

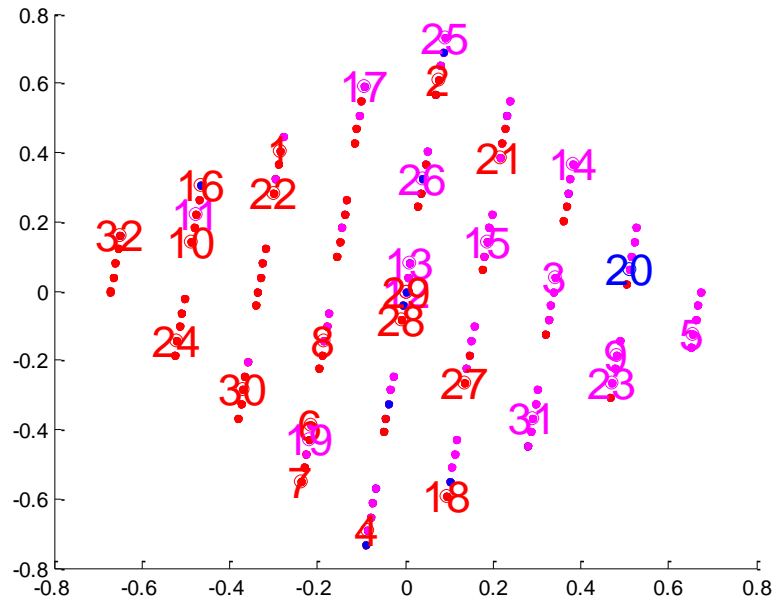


Figura 64 – Projeção da base Balance com os centros de regra do melhor modelo.

Base Diabetes:

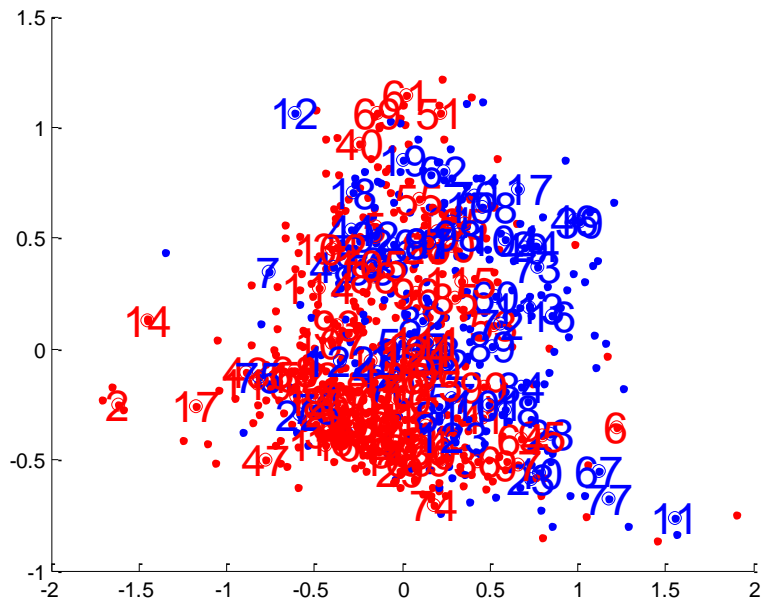


Figura 65 – Projeção da base Diabetes com os centros de regra do melhor modelo.

Base Cancer:

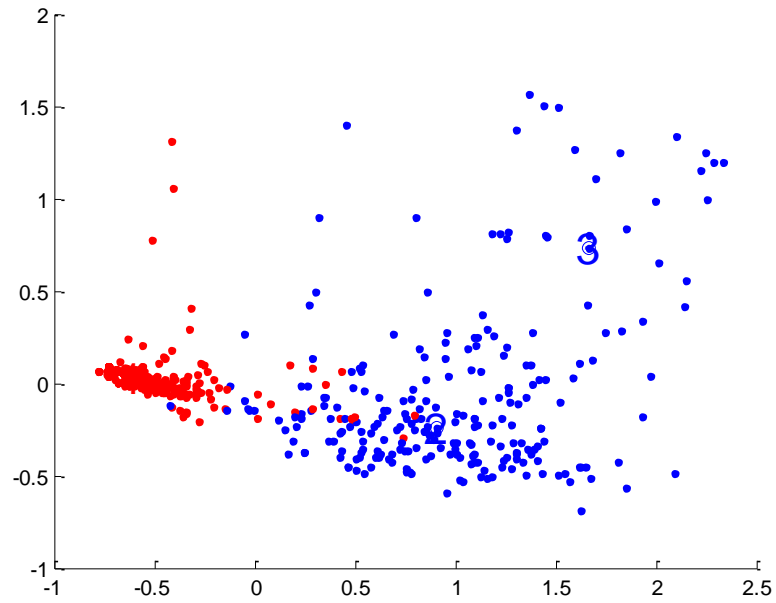


Figura 66 – Projeção da base Cancer com os centros de regra do melhor modelo.

Base Glass:

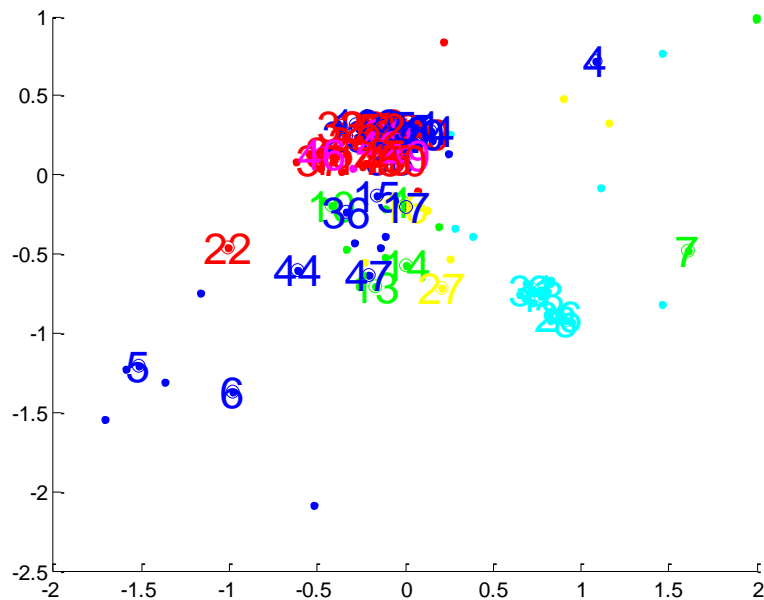


Figura 67 – Projeção da base Glass com os centros de regra do melhor modelo.

Base Wine:

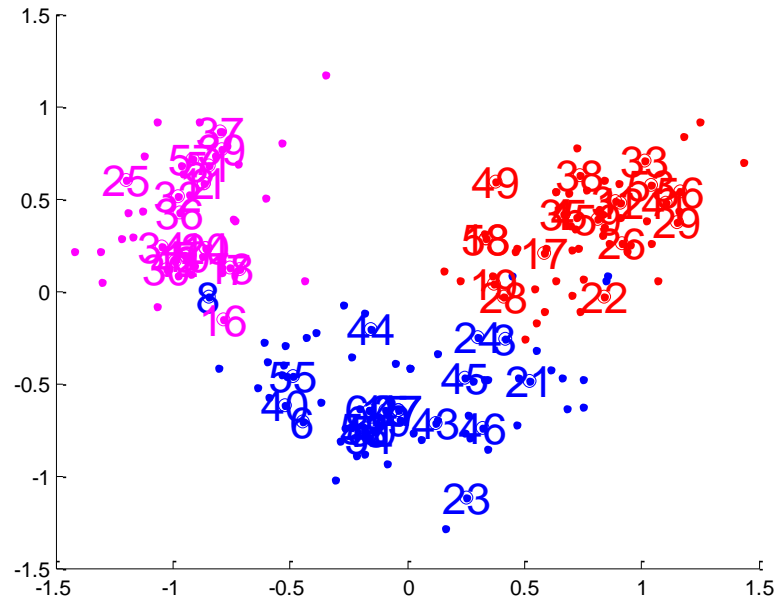


Figura 68 – Projeção da base Wine com os centros de regra do melhor modelo.

Base Heart:

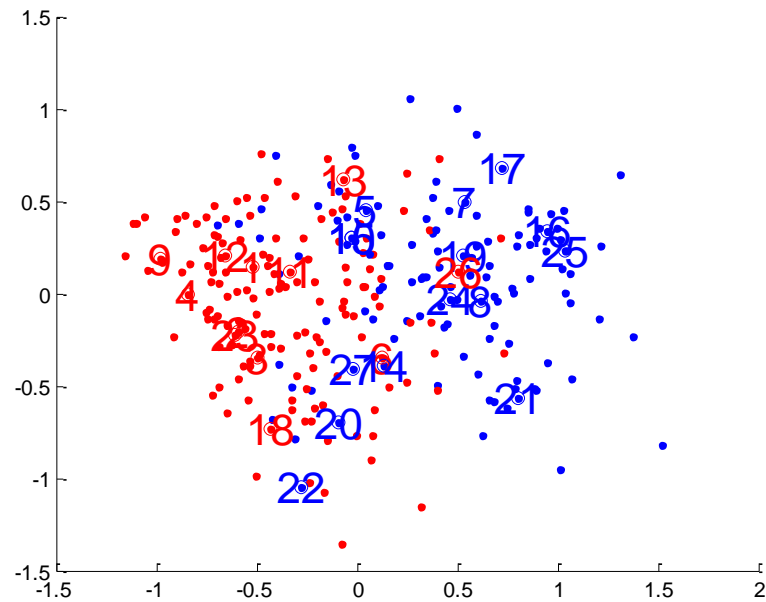


Figura 69 – Projeção da base Heart com os centros de regra do melhor modelo.

Base Image:

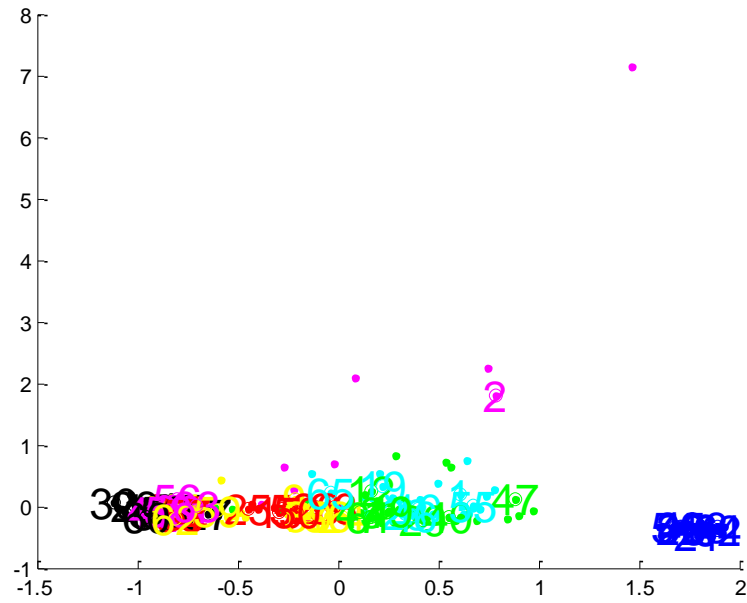


Figura 70 – Projeção da base Image com os centros de regra do melhor modelo.

Base Ionosphere:

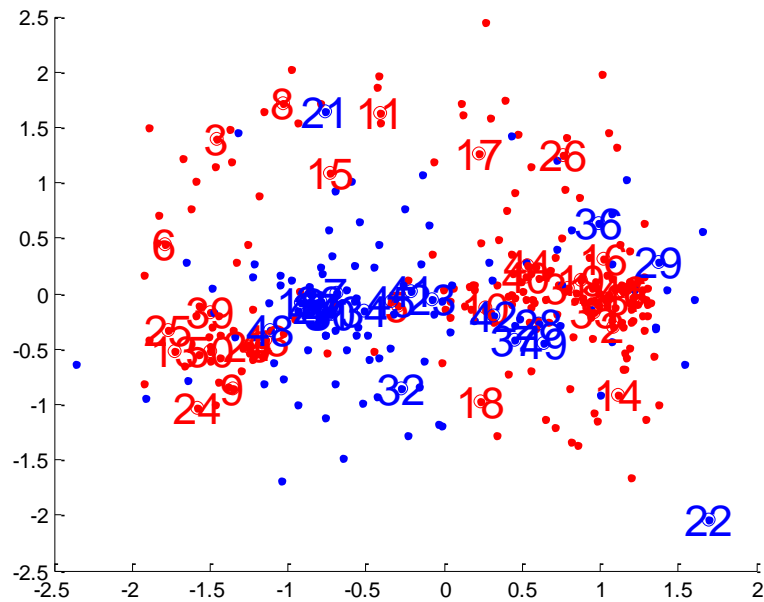


Figura 71 – Projeção da base Ionosphere com os centros de regra do melhor modelo.

Base Sonar:

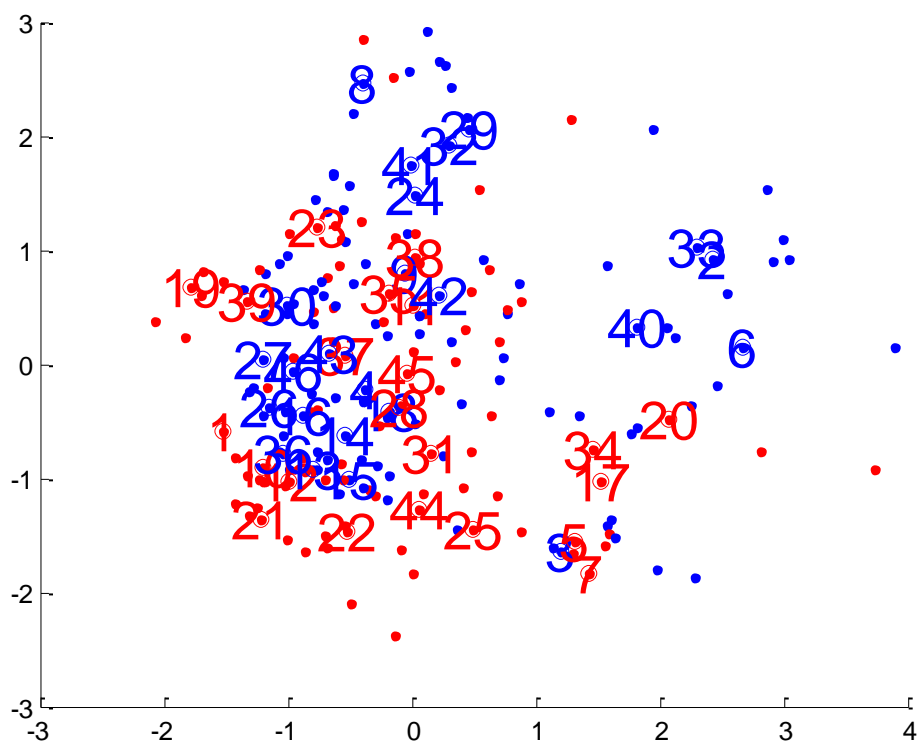


Figura 72 – Projeção da base Sonar com os centros de regra do melhor modelo.

APÊNDICE C – Algoritmo Para o Dendrograma no MATLAB

Por conta de algumas particularidades encontradas durante a confecção dos dendrogramas no MATLAB, foram necessários ajustes no conjunto de resultados gerados a partir da escolha dos modelos a serem hierarquizados. Esses ajustes estão descritos a seguir:

Algoritmo: Geração de Dendrograma no MATLAB

Entrada: Conjunto de modelos com informações de parâmetros completas.

Saída: Dendrograma.

- 1 INICIO
- 2 Avaliar cada modelo do conjunto de modelos selecionados no sentido da quantidade de regras de cada classe do problema.
- 3 Eliminar os modelos em que tenha ocorrido aumento de regras em ao menos uma classe ou que não tenha havido diminuição da quantidade de regras em ao menos uma classe, em relação ao modelo anterior.
 - O próximo passo é importante para que haja fechamento na utilização da função dendrograma no MATLAB:
- 4 Verificar se o último modelo do conjunto de modelos selecionados é composto de uma regra por classe. Em caso negativo, inserir “resultado final” com essa característica.
 - Como a função dendrograma do MATLAB funciona com agrupamento de pares de grupos, então o salto na quantidade de regras de uma classe não pode ser grande em relação ao próximo modelo:
- 5 Avaliar os modelos resultantes com relação ao salto de um modelo para o próximo na quantidade de regras de cada classe.
- 6 Se o número inteiro superior à metade da quantidade de regras em uma classe supera a quantidade de regras da mesma classe do próximo modelo, então inserir “resultado parcial”, associado ao mesmo valor de dispersão do próximo modelo.

7 Fazer agrupamento a cada duas regras até que a quantidade de regras do próximo modelo seja obtido. Alocar os agrupamentos realizados junto ao parâmetro de dispersão associado ao modelo correspondente.

-- Unificação da parte superior do dendrograma:

8 Alocar nos agrupamentos gerados o fechamento do dendrograma.

9 Gerar dendrograma.