2
Metodologia e Dados

Nas seções 2.1, 2.2 e 2.3 descrevemos as três principais contribuições metodológicas desta tese. O capítulo se encerra na seção 2.4 com a descrição dos bancos de dados utilizados.

Seja X uma amostra de tamanho m composta pelas observações $\{x_1, x_2, ..., x_m;\ x_i \in \Re^n, \forall i = 1,...,m\}$. Considere um ambiente de classificação dicotômico onde cada observação x_i está associada a uma dentre duas classes possíveis, C_1 ou C_2, com rótulos 1 ou 2 respectivamente. Denota-se $y(x_i)$ o rótulo da observação x_i, i.e., $y(x_i) = 1$ se x_i pertence a C_1, e $y(x_i) = 2$ se x_i pertence a C_2.

A motivação para boa parte das contribuições desta tese diz respeito ao enfoque local-global. Métodos locais podem auxiliar na tomada de decisão, pois esta abordagem pode dividir um problema difícil em sub-problemas mais simples. As Figuras 1 e 2 ilustram esta situação com um problema não separável linearmente. Na Figura 1, tem-se uma solução global não linear, enquanto que na Figura 2 tem-se o problema dividido em 8 sub-problemas mais simples. Note que 5 destes 8 sub-problemas possuem soluções triviais, já que possuem observações de uma única classe (duas vezes na classe X e três vezes na classe O). Além disso, nas 3 regiões em que há observações das duas classes, tem-se problemas linearmente separáveis. Sem dúvida, neste caso, a abordagem local tornou o problema mais fácil de ser resolvido.

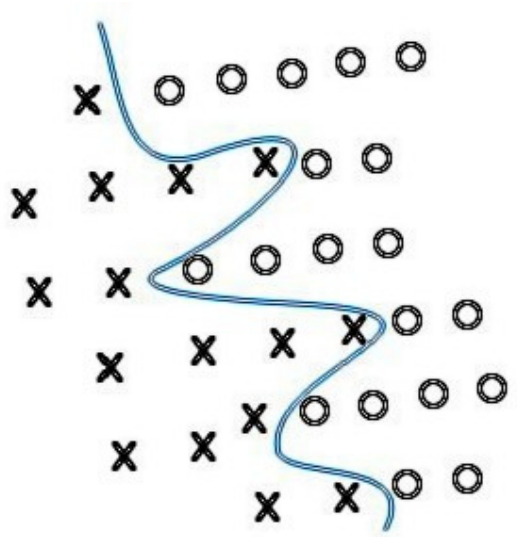

Figura 1: Problema de classificação não separável linearmente com uma solução global.

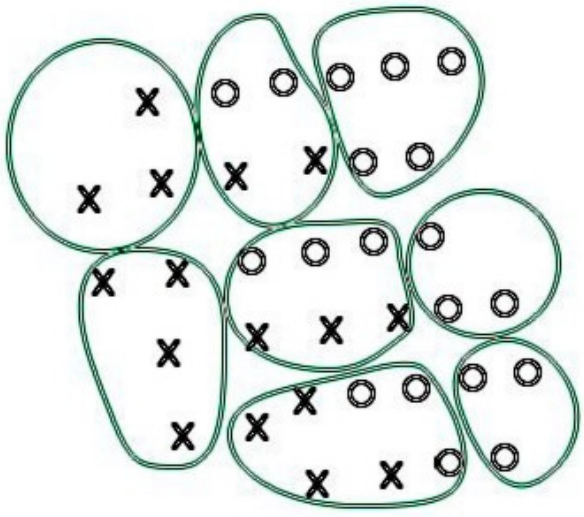

Figura 2: Problema de classificação não separável linearmente com 8 sub-soluções triviais ou separáveis linearmente.

Para realizar a transição entre o problema global e os sub-problemas locais, pode-se tomar uso de métodos globais, aplicados a todas as observações. Por isso, os métodos podem ser chamados de local-global.

2.1.
Classificador por Quantização Vetorial (CQV)

Na primeira contribuição metodológica que denominamos de *Classificador por Quantização Vetorial* (CQV), propomos um algoritmo de classificação que combina os ambientes, supervisionado e não supervisionado. A primeira etapa, não supervisionada, se dá através da aplicação de um procedimento de Quantização Vetorial (QV) com o objetivo de estabelecer uma aproximação quantizada da distribuição da amostra. Desta forma, as observações são divididas em células, associadas a vetores-protótipo (*codebooks*).

Nesta primeira etapa, utiliza-se o algoritmo de QV denominado LBG (Linde-Buzo-Gray) [1] com o objetivo de segmentar a amostra em células. Seja P = $\{p_1, ... , p_r , \quad p_k \in R^n, \ k = 1, ..., r\}$ um conjunto de vetores-protótipo. Cada observação $x_i \in X$, $i = 1,\ldots m$ é associada a um destes protótipos e, conseqüentemente, uma partição de X é obtida através de r células $S_1,..,S_r$. Formalmente, dada uma métrica 'd', define-se uma célula S_h como: $S_h \equiv \{ x_i \in X \mid d(x_i, p_h) \leq d(x_i, p_j), \quad \text{para todo } j \neq h, \ j = 1,...,r \} \quad \text{para } h = 1,..., r, \ i = 1,\ldots m.$ Mais detalhes sobre QV e LBG podem ser encontrados no Apêndice B.

Uma vez definidas as células, passa-se a etapa supervisionada, inicialmente observando que duas situações são possíveis: (i) A célula representada por este protótipo é homogênea, i.e. contém observações de apenas uma das classes; (ii) A célula é heterogênea, contém observações de ambas as classes. O primeiro caso é trivial e associa-se a classe (homogênea) a qualquer observação x que pertença a esta célula. No segundo, executa-se o procedimento, descrito a seguir, para definir qual das classes atribuir à observação x.

Sem perda de generalidade, focaremos em uma observação arbitrária x. Calcula-se então à distância entre x e cada um dos protótipos, escolhendo-se o protótipo mais próximo. Suponha que S_k é a célula associada a este protótipo. Vamos supor que a Figura 3 corresponda à célula S_k, onde m_1 e m_2 são as médias de cada uma das classes dentro de S_k:

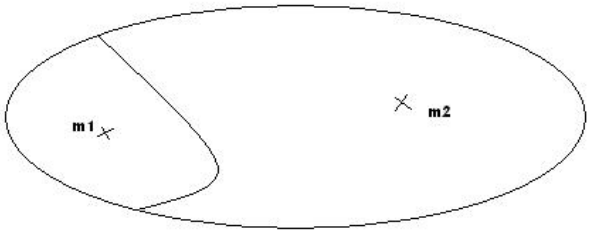

Figura 3: Célula S_k dividida em duas classes com suas médias m_1 e m_2.

Uma maneira de classificar uma observação é tomar o inverso da distância desta observação às médias.

Observe que, na Figura 4, x ∈ C_2 porque $\frac{1}{d(x,m_2)} > \frac{1}{d(x,m_1)}$.

Note que este valor pode ser utilizado como uma estimativa de $p(x \mid C_i)$.

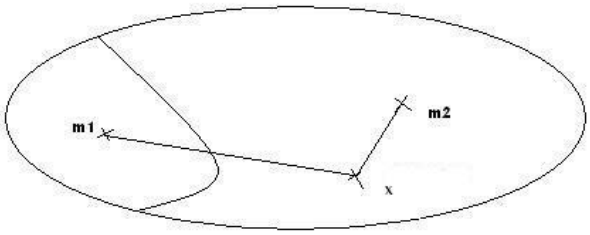

Figura 4: A observação x em S_k classificada de acordo com sua distância às médias.

Entretanto, este critério pode levar a uma tomada de decisão incorreta. Observe a Figura 5. Neste caso, x seria classificada como pertencente a C_1 embora pertença a C_2. Uma possibilidade é levar em consideração o cálculo das

freqüências relativas (estimando as probabilidades a priori). Começa a ser construído um classificador local inspirado no classificador Bayesiano.

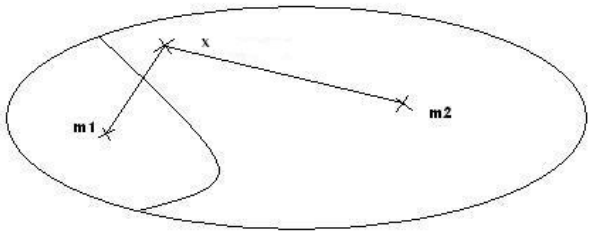

Figura 5: Uma observação classificada erroneamente em virtude do critério utilizado.

Definem-se, então, dois subconjuntos de S_k, ς_k^1 e ς_k^2:

$$\varsigma_k^1 \equiv \{x_i \in S_k \mid y(x_i) = 1\} \quad \text{e} \quad \varsigma_k^2 \equiv \{x_i \in S_k \mid y(x_i) = 2\}. \tag{1}$$

e em seguida, determina-se as freqüências relativas as classes C_1 e C_2 em S_k,

$$f_1^k = \frac{\#\varsigma_k^1}{\#S_k} \quad \text{e} \quad f_2^k = \frac{\#\varsigma_k^2}{\#S_k},$$

(#A representa a cardinalidade de um conjunto A). Note que f_1^k e f_2^k podem ser vistas como estimativas da probabilidade a priori das classes C_1 e C_2 na célula S_k.

Pode-se, assim, definir a razão das probabilidades a priori como $\pi^k \equiv \frac{f_1^k}{f_2^k}$. Iremos utilizar

$$\hat{L}_r = \left(d(x, m_k^1)\right)^{-1} / \left(d(x, m_k^2)\right)^{-1} \tag{2}$$

como estimador da razão de verossimilhança $L_r \equiv p(x \mid C_1) / p(x \mid C_2)$. Aqui, 'd' denota distância, e m_k^1 e m_k^2 são as médias das classes em S_k, explicitadas como:

$$m_k^1 = \frac{1}{\#\varsigma_k^1} \sum_{x \in \varsigma_k^1} x \quad \text{e} \quad m_k^2 = \frac{1}{\#\varsigma_k^2} \sum_{x \in \varsigma_k^2} x.$$

Desta forma, utilizou-se como estimador da verossimilhança dentro da célula S_k, o inverso da distância da observação à respectiva média da classe dentro de S_k.

Baseado no classificador Bayesiano estabelece-se a seguinte regra de decisão na célula S_k:

$$\begin{cases} x \in C_1 & se\ \hat{L}_r \geq (\pi^k)^{-1} \\ x \in C_2 & caso\ contrário \end{cases} \tag{3}$$

$\hat{L}_r$ pode ser considerado uma boa estimativa para a razão de verossimilhança quando as distribuições das classes são unimodais, o que é uma hipótese plausível, considerando-se que estamos trabalhando localmente.

O algoritmo proposto pode ser esquematizado como segue:

1. Execute o algoritmo LBG (Parâmetros: $\varsigma = 10^{-1}$ e $\varepsilon = 10^{-2}$. O algoritmo se encontra no Apêndice B.) com distância Euclidiana para gerar r células $S_1,..., S_r$;

2. Divida cada célula S_k em ς_k^1 e ς_k^2, de acordo com as duas classes (eq. (1));

3. Dada uma observação x, calcule a distância Euclidiana entre x e o conjunto P de protótipos e tome $k = \arg\min_i(d(x, p_i))$. Se S_k é uma célula homogênea, i.e., $\varsigma_k^1 = \varnothing$ ou $\varsigma_k^2 = \varnothing$, vá para o passo 4, caso contrário vá para o passo 5;

4. (célula homogênea) - Atribua a x a mesma classe das demais observações;

5. (célula heterogênea) - Se $d(x,m^1_k) = 0$ e $d(x,m^2_k) = 0$, considere $\hat{L}_r = 1$ e aplique à equação (3) para decidir a classe de x; se apenas $d(x,m^1_k) = 0$, $x \in C_1$; se apenas $d(x,m^2_k) = 0$, $x \in C_2$;

finalmente, se $d(x,m^1_k) \neq 0$ e $d(x,m^2_k) \neq 0$, calcule $\hat{L}_r$ como em (2) e aplique à equação (3) para decidir a classe de x.

Note que, se a distância da observação a ser classificada pelo algoritmo as médias das classes coincidem em S_k, as estimativas das probabilidades a priori f_1^k e f_2^k serão as únicas responsáveis por decidirem pela alocação da classe à observação x. Esta situação é análoga à decisão do classificador Bayesiano quando as probabilidades condicionais as classes são idênticas ($p(x \mid C_1) \equiv p(x \mid C_2)$) e a decisão também é baseada apenas nas probabilidades a priori. No caso da observação coincidir com uma das médias das classes, atribui-se a observação à classe representada por esta média, não havendo, portanto, o cálculo referente à equação (3).

Observe que o número de protótipos escolhido para o algoritmo LBG é fundamental neste processo. Em um extremo, podem-se escolher apenas dois protótipos, o que resultaria em duas células. Neste caso, obviamente, a abordagem local do algoritmo não estaria sendo feita. Em contrapartida, poderíamos estipular cada observação como sendo um protótipo, o que implicaria em uma super parametrização do modelo. Assim, a medida em que o número de protótipos aumenta, a análise do algoritmo é cada vez mais local e o modelo cada vez mais parametrizado. Nesta tese utiliza-se como número de protótipos todas as potências de 2 menores ou iguais a décima parte da quantidade de observações ($2^i \leq \#X/10$, $i \in N$) e seleciona-se a configuração que resulta no melhor desempenho dentro-da-amostra.

Uma possível aplicação do CQV é sua utilização como ferramenta para identificação de observações com rótulo invertido. Assume-se para tal que os rótulos atribuídos às observações podem eventualmente não corresponder ao verdadeiro rótulo. Uma observação classificada erradamente pelo algoritmo CQV pode ser selecionada como pouco representativa de sua classe, e possivelmente suspeita de ter tido seu rótulo invertido por algum mecanismo de ruído. Vale ressaltar que este procedimento é local, dentro de uma célula.

2.2. A Zona de Risco Generalizada: Seleção de Observações para melhorar a Classificação

Esta segunda contribuição metodológica está relacionada à seleção de observações informativas, dentro do contexto de classificação de padrões e encontra-se publicada em [64]. A *Zona de Risco Generalizada* (ZRG) é um esquema, independente de modelo, que seleciona observações informativas baseado no conceito de dissimilaridade entre densidades de probabilidade [63]. O objetivo é encontrar estas observações e analisar o impacto que elas causam no desempenho de métodos estatísticos de classificação.

Métodos baseados em protótipos atribuem as observações a vetores-protótipo no espaço de observações, em nosso caso, o $\Re^n$. Estes protótipos são alocados apropriadamente utilizando-se as observações dentro-da-amostra (de treinamento) e são associados a uma das classes. As observações do conjunto fora-da-amostra (de teste) são classificadas através da associação à classe do protótipo mais próximo. Em [2], foi proposto um método cuja idéia central era atualizar os protótipos do algoritmo *Learning Vector Quantization* (LVQ) [20] utilizando somente um subconjunto da amostra. O principal objetivo era conduzir os protótipos a convergirem para uma localização mais conveniente, diminuindo assim os erros de classificação. Este subconjunto selecionado da amostra é composto por observações consideradas sob risco de serem capturadas por um protótipo representante de outra classe, e é denominado *Zona de Risco* (ZR).

Uma observação x_i é dita pertencer a ZR se, e somente se, sua distância a qualquer protótipo de outra classe é menor do que à distância entre este protótipo e o protótipo mais próximo da mesma classe [2]. Seja p_c o protótipo mais próximo representando a classe de x_i, e p_r qualquer um dos protótipos representando a outra classe. Diz-se que uma observação x_i pertence à ZR se, e somente se, $d(x_i, p_r) < d(p_c, p_r)$ para qualquer protótipo p_r representante da outra classe. Equivalentemente, x_i pertence à ZR se, e somente se

$$d(x_i, p_r) < d(p_c, p_r).$$

Nesta tese, estende-se o conceito de ZR desenvolvendo-se o que chamamos de *Zona de Risco Generalizada* (ZRG). Esta generalização vem do fato de que a ZRG é uma contribuição independente de modelo de classificação, não estando restrita a métodos com protótipos como a ZR.

Para realizar esta extensão, utiliza-se a divergência de *Cauchy-Schwartz* [3], [8], [12] ($D_{C\text{-}S}$), uma medida da distância entre duas funções de densidade de probabilidade (fdp's). Para evitar o problema de estimação de densidades de probabilidade, que é usualmente difícil, usa-se a abordagem de *Information Theoretic Learning* (ITL) [6] que permite que a informação seja diretamente extraída através das observações disponíveis. ITL é uma metodologia baseada em kernel que está dentro do contexto de teoria da informação.

Os próximos desenvolvimentos irão conduzir à construção da *Zona de Risco Generalizada* (ZRG), que, como afirmado anteriormente, não é restrita a métodos com protótipos ou a qualquer tipo de modelo de classificação. A divergência de Cauchy-Schwartz entre duas funções densidades de probabilidade p e q, definida a seguir, é uma forma de medir a distância entre p e q.

$$D_{C\text{-}S}(p,q) \equiv -\log\left(\frac{\left(\int p(x)q(x)dx\right)^2}{\int p^2(x)dx \int q^2(x)dx}\right) = \log\int p^2(x)dx - 2\log\int p(x)q(x)dx + \log\int q^2(x)dx. \quad (4)$$

Claramente, $D_{C\text{-}S}(p,p) = 0$ e $D_{C\text{-}S}(p,q) = D_{C\text{-}S}(q,p)$ para quaisquer fdp's p e q [8], [12]. Entretanto, $D_{C\text{-}S}$ não é uma métrica porque não satisfaz a desigualdade do triângulo [12].

Sejam p a densidade associada às observações da classe C_1 e q a densidade referente à classe C_2. Segue que a proximidade entre as classes C_1 e C_2 pode ser medida através da $D_{C\text{-}S}(p,q)$. Considerando o contexto de métodos com protótipos, pode-se fazer uma analogia entre a associação de p e q através da $D_{C\text{-}S}$ e uma medida de distância entre protótipos (em $\Re^n$) representativos de classes. Da mesma forma, assim como se pode calcular a distância entre uma observação e um protótipo, pode-se também calcular a divergência entre uma observação e uma fdp (associada a uma classe). Estas analogias são à base da definição de ZRG, que está apresentada adiante, na equação (10), após o desenvolvimento relativo à divergência que se segue.

O cálculo da divergência envolve, em geral, a necessidade de estimação de fdp's [4], [5]. Isto se torna especialmente complicado para variáveis aleatórias contínuas, caso no qual um procedimento de discretização se faz necessário. A dificuldade em estimar as fdp's cresce também de forma importante com o aumento da dimensão das variáveis aleatórias envolvidas. A abordagem de ITL,

proposta em [6], contorna esta dificuldade viabilizando que os cálculos sejam realizados diretamente a partir das observações, como descrito a seguir.

Sejam M e N os tamanhos das amostras geradas pelas fdp's p e q, respectivamente. Estas fdp's podem ser estimadas através de Janelas de Parzen [7]. Seja $G_{\sigma_p^2}:\Re^n \rightarrow \Re$ uma função Gaussiana de média zero e matriz de covariância $\sigma_p^2 I$. Claro que a mesma definição pode ser feita para a fdp q. Uma estimativa $\hat{p}$, da fdp p, utilizando $G_{\sigma_p^2}$ como *kernel*, é dada por:

$$\hat{p} = \frac{1}{M}\sum_{i=1}^{M} G_{\sigma_p^2}(x - x_i)$$

Segue que,

$$\int \hat{p}(x)\hat{q}(x)dx =$$

$$= \int\left(\frac{1}{M}\sum_{i=1}^{M} G_{\sigma_p^2}(x - x_i)\right)\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sigma_q^2}(x - x_j)\right)dx$$

$$= \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\int G_{\sigma_p^2}(x - x_i)G_{\sigma_q^2}(x - x_j)dx$$

$$= \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} G_{\sigma_p^2+\sigma_q^2}(x_i - x_j). \qquad (5)$$

A última igualdade em (5), que resulta de uma simplificação da integral usada no Teorema de Convolução de Gaussianas [8], é um desenvolvimento técnico não trivial [43]♠. Observe que o cálculo de $\int \hat{p}(x)\hat{q}(x)dx$, mostrado na expressão (5), é exato, não são feitas aproximações. É muito interessante notar que o termo final em (5) depende exclusivamente das observações x_i, além da largura de σ na função *kernel*.

De uma maneira similar, obtém-se

$$\int \hat{p}^2(x)dx = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} G_{2\sigma_p^2}(x_i - x_j), \qquad (6)$$

e

$$\int \hat{q}^2(x)dx = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{2\sigma_q^2}(x_i - x_j). \qquad (7)$$

♠ Material adicional pode ser encontrado em http://bioinformatics.musc.edu/~svinga/renyi/.

Aplicando (5), (6) e (7) em (4), chega-se à:

$D_{C\text{-}S}(\hat{p}, \hat{q}) =$

$$= \log \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} G_{2\sigma_p^2}(x_i - x_j) \; - 2 \log \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} G_{\sigma_p^2 + \sigma_q^2}(x_i - x_j)$$

$$+ \log \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G_{2\sigma_q^2}(x_i - x_j) \qquad (8)$$

Completa-se, assim, o cálculo da divergência de Cauchy-Schwartz, entre duas fdp's estimadas por Janela de Parzen, com base apenas nas observações.

Por argumentação análoga, pode-se escrever a expressão do divergente entre uma observação, x_k, e uma fdp estimada $\hat{q}$:

$D_{C\text{-}S}(x_k, \hat{q}) =$

$$\log G_{2\sigma_p^2}(0) - 2 \log \frac{1}{N} \sum_{j=1}^{N} G_{\sigma_p^2 + \sigma_q^2}(x_k - x_j) + \log \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G_{2\sigma_q^2}(x_i - x_j). \qquad (9)$$

Posto este desenvolvimento relativo à divergência, estamos agora em condições de definir o que denominamos Zona de Risco Generalizada (ZRG):

Uma observação x_k é dita pertencer a ZRG se, e somente se,

$$D_{C\text{-}S}(x_k, \hat{q}) \; < \; \alpha * D_{C\text{-}S}(\hat{p}, \hat{q}), \text{ se } x_k \in \text{ classe } C_1, \qquad (10.1)$$

$$D_{C\text{-}S}(x_k, \hat{p}) \; < \; \alpha * D_{C\text{-}S}(\hat{p}, \hat{q}), \text{ se } x_k \in \text{ classe } C_2. \qquad (10.2)$$

O parâmetro α nas equações 10.1 e 10.2 é introduzido por razões numéricas; note que o valor do primeiro termo da equação 10.1 (igual argumentação pode ser feita com relação a 10.2) pode se tornar desequilibradamente grande com relação ao valor do segundo termo. Note-se que se trata de dois divergentes, o primeiro entre uma observação e uma fdp e o segundo entre duas fdp's. O divergente entre uma observação e uma fdp é equivalente ao divergente entre duas fdp's, quando em uma destas tem-se apenas uma observação (vide equações 8 e 9). Assim, retornando a (10.1), a menos que uma determinada observação x_k, digamos da classe C_1, seja muito próxima a fdp q, o valor deste termo pode se tornar exageradamente alto. Neste sentido, note-se que o segundo termo do divergente em (9) é o negativo do logaritmo de uma função kernel que pode assumir valores próximos a zero. É plausível considerar que no caso do cálculo do divergente entre duas fdp's, havendo um número razoável de observações, um certo número de parcelas do somatório 'afastará' o segundo termo de (9) do negativo do logaritmo de um valor próximo a zero. Para todos os experimentos, usa-se $\alpha = 3$.

Um procedimento de 'limpeza' do banco é feito antes do cálculo da ZRG. Isto é feito aplicando-se K-NN (K = 3) e removendo-se todas as observações classificadas erradas neste processo. Este procedimento de limpeza é indicado porque, em ambientes ruidosos, pequenos subconjuntos da ZRG (calculadas pelas equações 10.1 e 10.2) podem ser criadas em torno de observações ruidosas. Esta situação, sem que a limpeza seja efetuada, está ilustrada nas Figuras 6 e 7, onde os pontos em vermelho na Figura 6 são as observações ruidosas (com os rótulos invertidos artificialmente) criando indesejáveis pequenos subconjuntos de ZRG como ilustrado na Figura 7. Uma vez eliminadas estas observações ruidosas, os subconjuntos da ZRG criados em torno delas também desaparecem (veja Figura 8).

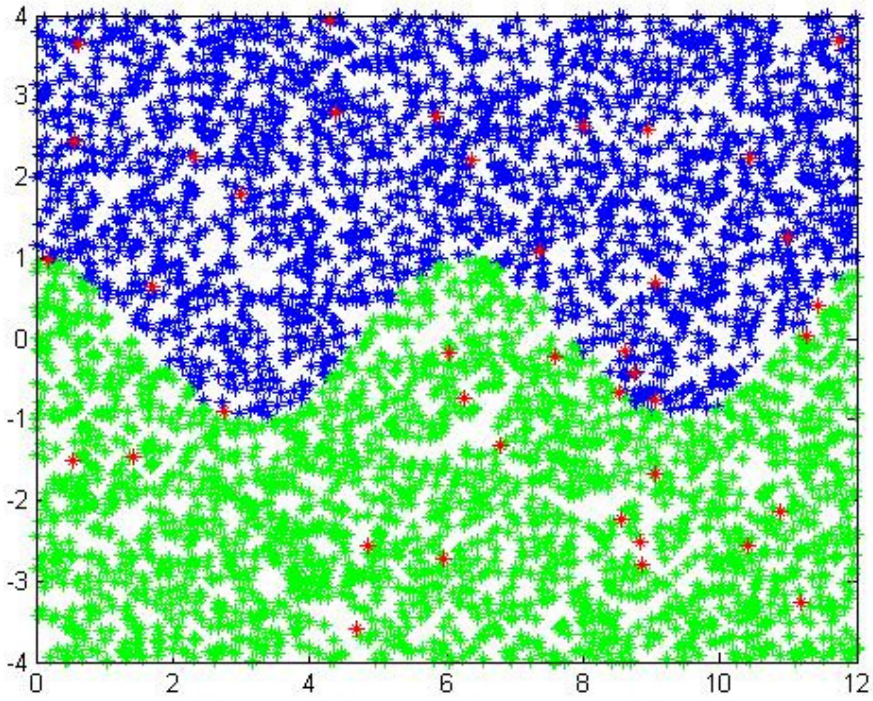

Figura 6: Duas classes divididas pela função cosseno. Observações em vermelho são aquelas cujo rótulo esta invertido artificialmente.

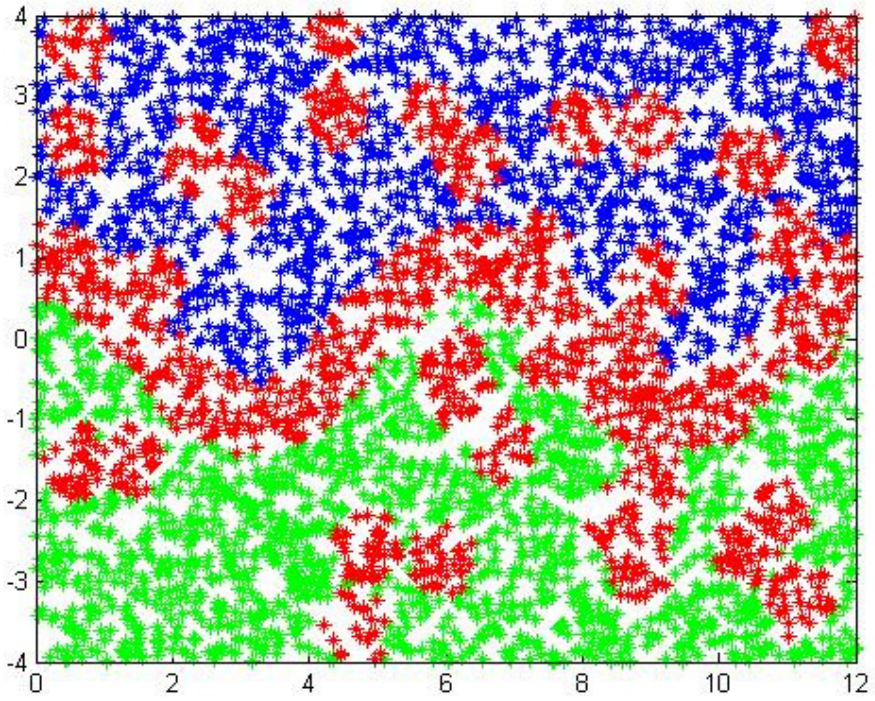

Figura 7: ZRG com regiões ruidosas. Sigma = 0.1.

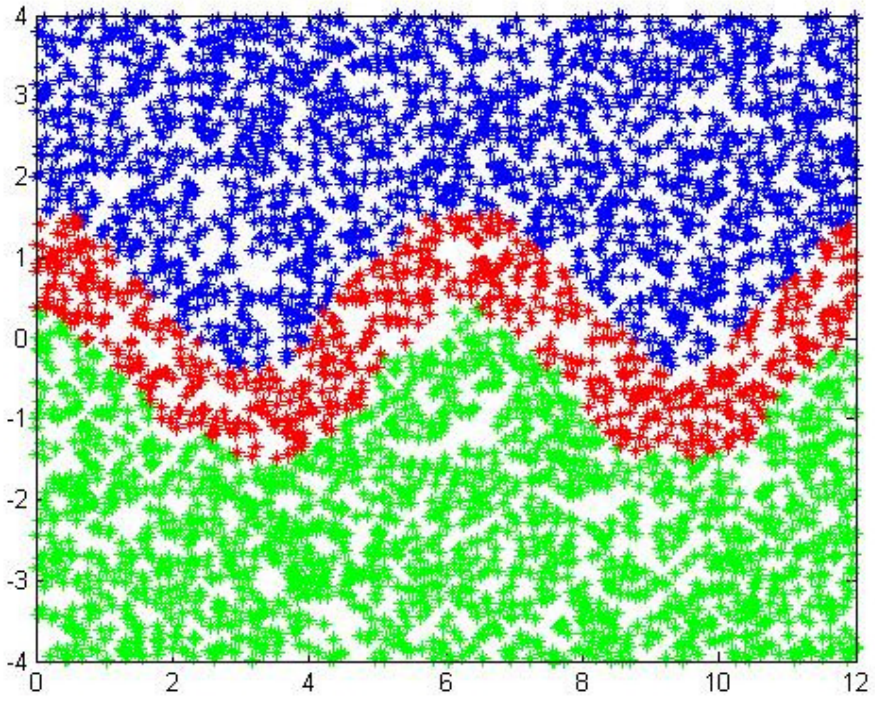

Figura 8: ZRG depois do procedimento de limpeza. Sigma = 0.1.

Em relação à largura de sigma, utilizamos quatro abordagens diferentes: as primeiras duas baseadas na abordagem proposta de [41]; (i) validação cruzada; (ii) largura ótima supondo que as observações têm distribuição normal com variância unitária e o kernel é normal (equação 4.14 de [41]); (iii) valor médio de (i) e (ii) e (iv) Valor fixo (0.1). Considerando os experimentos propostos nesta tese, utilizamos o sigma que forneceu o melhor equilíbrio entre desempenho e número de observações selecionadas.

Assim, para os diversos métodos de classificação, utilizam-se, na fase dentro-da-amostra, apenas as observações pertencentes a ZRG.

O algoritmo de ZRG pode ser esquematizado como segue:

1. Dada à amostra X, utilize o procedimento de limpeza descrito nesta subseção, através de K-NN (K = 3), removendo-se todas as observações classificadas erradas neste processo;
2. Calcule os diferentes valores para as larguras do sigma (i) validação cruzada; (ii) largura ótima supondo que as observações têm distribuição normal com variância unitária e o kernel é normal (equação 4.14 de [41]); (iii) valor médio de (i) e (ii) e (iv) Valor fixo (0.1);
3. Para cada largura de sigma, execute 10.1 e 10.2 com $\alpha = 3$, encontrando as observações que pertencem a cada uma das ZRG's;

Utilize somente as observações de cada uma das ZRG's para a fase dentro-da-amostra. O conjunto que obtiver o melhor equilíbrio entre desempenho e número de observações selecionadas é o escolhido.

2.2.1. Uma Aplicação da ZRG a Seleção de Observações

Nesta seção, apresentamos um método para analisar a representatividade das observações em relação a sua classe, através de um procedimento de categorização da amostra. Neste sentido, após o término do procedimento, cada uma das observações é categorizada em um dos seguintes grupos:

- Típico (T): Observações claramente representativas de sua classe, possivelmente localizadas em regiões distantes da fronteira de decisão;

- Risco (R): Observações de risco, aquelas que se encontram vizinhas à fronteira de decisão. Se corretamente identificadas estas observações podem ser as únicas utilizadas na fase de ajustes dentro-da-amostra, como uma estratégia para melhorar a performance de classificação;
- Inversão de Rótulo (Inv): Observações não-representativas de sua classe que possivelmente são resultado de mecanismo de inversão de rótulo por ruído.

Em [9], encontra-se uma contribuição anterior que considera a abordagem de categorização de observações em três grupos: observações típicas, críticas e ruídos, onde críticas seriam as observações vizinhas às fronteiras (as que categorizamos como Risco) e ruídos aquelas que possuem evidência de inversão de rótulo. Os algoritmos DROP3 [57] e IB3 [56] também possuem a característica de identificar observações ruidosas e excluí-las para que não degradem o desempenho fora-da-amostra.

Considere-se uma observação com rótulo invertido. Trata-se de uma observação com o rótulo de uma classe, porém imersa na outra. Eventualmente, esta observação pode ser considerada estar sob risco e, portanto, pertencer a ZRG. Na Figura 7 exemplifica-se uma situação, onde duas classes divididas pela função cosseno possuem 20 observações em cada uma delas com o rótulo invertido deliberadamente. Na Figura 8, tem-se o estabelecimento da ZRG com sigma igual a 0.1. Note que, além da fronteira de decisão, aparecem vários pequenos subconjuntos da ZRG em torno das observações com rótulo invertido.

A fim de estabelecer a separação entre observações pertencentes a ZRG e as com inversão de rótulo, considera-se a divergência entre a observação e sua própria densidade estimada. Assim, calculamos as divergências $D_{C\text{-}S}(x_i, \hat{p})$ se $x_i \in$ classe C_1 e $D_{C\text{-}S}(x_i, \hat{q})$ se $x_i \in$ classe C_2. Define-se a seguir, para cada uma das duas classes, os seguintes conjuntos: T_i, observações típicas de cada banco de dados. São aquelas que não satisfazem à equação (10), ou seja, não pertencem a ZRG para sua classe; Inv_i, observações com inversão de rótulo; e R_i, observações de risco:

$$T_1 \equiv \{x_i \notin ZRG \wedge x_i \in C_1\}$$
$$T_2 \equiv \{x_i \notin ZRG \wedge x_i \in C_2\} \quad (11)$$

$$Inv_1 \equiv \{x_i \notin T_1 \wedge x_i \in C_1 \mid \frac{D_{C\text{-}S}(x_i, \hat{q})}{D_{C\text{-}S}(x_i, \hat{p})} < 1\}$$
$$Inv_2 \equiv \{x_i \notin T_2 \wedge x_i \in C_2 \mid \frac{D_{C\text{-}S}(x_i, \hat{p})}{D_{C\text{-}S}(x_i, \hat{q})} < 1\}; \quad (12)$$

$$R_1 \equiv \{x_i \notin T_1 \wedge x_i \in C_1 \mid \frac{D_{C\text{-}S}(x_i, \hat{q})}{D_{C\text{-}S}(x_i, \hat{p})} \geq 1\}$$
$$R_2 \equiv \{x_i \notin T_2 \wedge x_i \in C_2 \mid \frac{D_{C\text{-}S}(x_i, \hat{p})}{D_{C\text{-}S}(x_i, \hat{q})} \geq 1\}; \quad (13)$$

Assim, categorizamos as observações através de (11), (12) e (13) como desejado.

O algoritmo pode ser caracterizado como:

1. Dada à amostra X, calcule os diferentes valores para as larguras do sigma: (i) validação cruzada; (ii) largura ótima supondo que as observações têm distribuição normal com variância unitária e o kernel é normal (equação 4.14 de [41]); (iii) valor médio de (i) e (ii) e (iv) Valor fixo (0.1);

Para cada largura de sigma, encontre as observações pertencentes aos conjuntos (11), (12) e (13) utilizando $\alpha = 3$ para o cálculo da divergência (eqs. 10.1 e 10.2), categorizando as observações da amostra X.

2.3. Quantizador Vetorial das Fronteiras de Decisão (QVFD)

Nesta terceira seção metodológica, desenvolve-se um procedimento para estabelecer protótipos representativos das observações na região das fronteiras de decisão.

No algoritmo LVQ clássico estabelece-se previamente a quantidade de protótipos e em seguida realiza-se um procedimento para ajustar a localização dos mesmos no espaço das observações. Em contraste, no procedimento proposto

nesta seção, não é necessário o estabelecimento a priori do número de protótipos, estes são adicionados iterativamente, isto é, a cada iteração, um novo protótipo é incluído em função da minimização do erro de classificação no conjunto dentro-da-amostra. A idéia central é adicionar protótipos em regiões em que o valor de uma função de custo, baseada na razão entre a distância entre o protótipo mais próximo da mesma classe e o mais próximo da outra classe, é alto. Após o término da quantização, é realizado um procedimento de enxugamento do conjunto de protótipos, retirando-se aqueles que não pioram a taxa de acerto dentro-da-amostra. Tipicamente, os protótipos obtidos pelo método estarão em regiões vizinhas às fronteiras de decisão, já que foram adicionados em regiões onde o erro era alto e ao mesmo tempo, permaneceram influenciando no desempenho de classificação após a redução do conjunto. Por isso, dizemos que o algoritmo quantiza as fronteiras, no sentido de que produz uma aproximação da distribuição na região da fronteira através dos protótipos. Com este procedimento, não só se delimita a fronteira, mas como se estabelecem às margens nesta fronteira (vide Figura 9). Uma vez estabelecidos os protótipos com as observações dentro-da-amostra pode-se delinear a fronteira fora-da-amostra utilizando estes protótipos como referências para alocação das observações através da regra de vizinhos mais próximos. Um algoritmo de quantização vetorial para classificação similar ao proposto nesta tese foi apresentado em [13]. A semelhança é o processo de adicionar protótipos de forma iterativa baseado na minimização de uma função de custo. Entretanto, existem diferenças na função de custo, no critério de parada e não há o procedimento de redução no conjunto de protótipos, o que leva o método proposto em [13] a quantizar todo o espaço e não focar na fronteira de decisão.

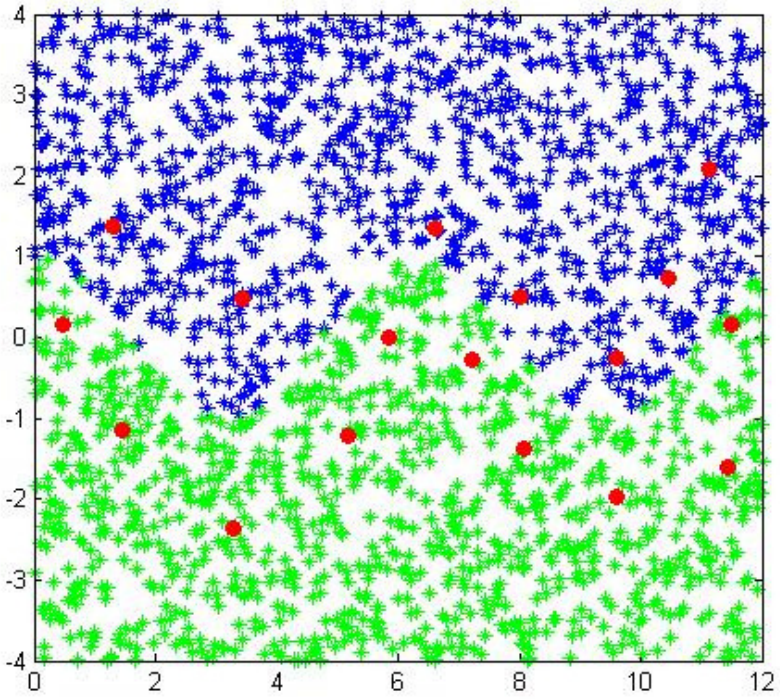

Figura 9: Resultado do QVFD para o banco de dados apresentado na seção 2.4.1.

As seguintes etapas devem ser percorridas para implementação do Quantizador Vetorial das Fronteiras de Decisão (QVFD):

Notação: C(x) = 1 se x ∈ classe C_1 e C(x) = 2 se x ∈ classe C_2.

Etapa 1 (Inicialização): Estabelecer os centróides de cada classe como protótipos iniciais:

$$p_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_j \qquad i = 1, 2.$$

N_i é a cardinalidade da classe i, x_j são os elementos da classe i.

Etapa 2 (a partir da minimização da função de custo baseada no erro de classificação, incrementar o número de protótipos e determinar a localização do novo protótipo):

Considere a função de custo:

$$J(x) = \frac{\min(d(x, p_z))}{\min(d(x, p_r))}$$

onde p_z é o protótipo mais próximo da mesma classe de x e p_r é o protótipo mais próximo da outra classe. 'd' representa a distância Euclidiana.

Note que, se J(x) > 1, a observação x está sendo classificada de forma errada. Define-se, então, o custo associado a cada protótipo. Considere as células definidas por cada protótipo:

$$S_i = \{x \in X \mid d(x, p_i) < d(x, p_j), \forall j, j \neq i\}.$$

Define-se o custo de um protótipo p_i da seguinte forma: (i) considere todos os elementos de sua célula que possuam custo maior que 1; (ii) O somatório de todos estes valores será o custo do protótipo. Assim,

$$Custo(p_i) = \sum J(x), \quad (x \in S_i) \wedge (J(x) > 1)$$

Seja

$$c = \arg\max_i (Custo(p_i)) \tag{14}$$

O novo protótipo $p_{(\#P+1)}$ será a observação mais próxima ao centróide das observações classificadas erroneamente pelo protótipo p_c, onde c foi obtido em (14). Note que, se existir no máximo uma observação classificada de forma errada em cada célula, o novo protótipo teria que coincidir com esta observação. Neste caso, o procedimento é interrompido para evitar uma super-parametrização do modelo. Após a inserção do novo protótipo, o conjunto de protótipos P é ajustado através do algoritmo de quantização vetorial LBG, considerando o critério: "Um protótipo jamais é atualizado se a nova célula contém menos de 50% das observações da mesma classe do protótipo".

Etapa 3 (Estipular um critério de parada para o número de protótipos): Critério de parada:

(1) Erro zero; e
(2) Custo zero; e
(3) A diferença entre o percentual de erro na classe do protótipo adicionado antes e depois da adição do protótipo é menor que um parâmetro β. Nos experimentos referentes a esta tese, utiliza-se $\beta = 0.5$.

Etapa 4 (Reduzir o conjunto de protótipos P, retirando todos aqueles que não pioram a taxa de classificação dentro-da-amostra):

Para cada i = 1,...,#P, a classificação é calculada para P\{pi}. Caso a classificação não piore, P = P – {pi}. Assim, protótipos que não influenciam na classificação são retirados.

Após o processo de redução, o algoritmo é concluído com o conjunto final de protótipos P. A metodologia para o conjunto fora-da-amostra utiliza este conjunto reduzido e classifica uma nova observação através da regra do vizinho mais próximo.

2.4.
Sobre os Dados

Os dez bancos de dados, descritos na seqüência, foram utilizados nos experimentos que serão descritos no próximo capítulo. Além de quatro conjuntos de dados sintéticos (três deles gerados pelos autores e um deles disponível na *internet*), usados para experimentos controlados, foram usados seis bancos oriundos de problemas reais disponíveis na *internet*[♠].

2.4.1.
Classes divididas através de função cosseno

O primeiro conjunto sintético foi construído gerando-se a divisão de duas classes através de uma função cosseno (Figura 10). Conjunto 1: foram geradas 1030 observações da classe C_1 e 1027 observações da classe C_2 para utilização dentro-da-amostra e 1060 observações C_1 e 1041 observações C_2, fora-da-amostra.

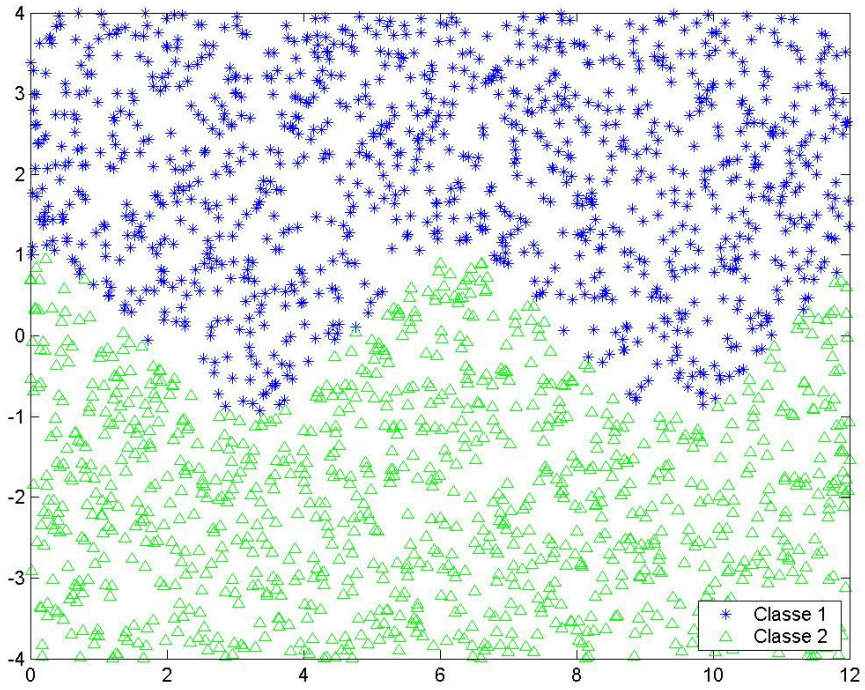

Figura 10: Classes C_1 e C_2 divididas através de função cosseno.

♠ http://www.ics.uci.edu/~mlearn/MLSummary.html.

2.4.2.
Quadrados Sobrepostos

Este conjunto foi construído gerando-se duas classes determinadas por dois quadrados com 25% da área em comum (Figura 11). Aplicado para metodologias de seleção de observações. Foram geradas 1000 observações para cada classe.

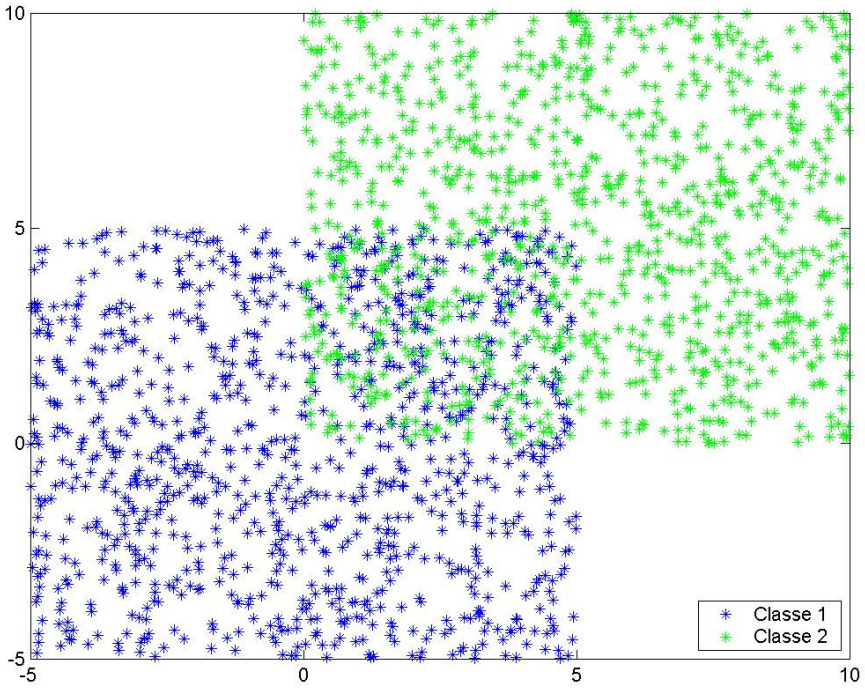

Figura 11: Classes C_1 e C_2 – Quadrados Sobrepostos.

2.4.3.
Círculos Concêntricos

Determinaram-se duas classes C_1 e C_2 respectivamente por um círculo e uma rosca concêntricos sem superposição (Figura 12). Foram geradas 123 observações da classe C_1 e 2611 observações da classe C_2 para utilização dentro-da-amostra e 127 observações C_1 e 2646 observações C_2, fora-da-amostra.

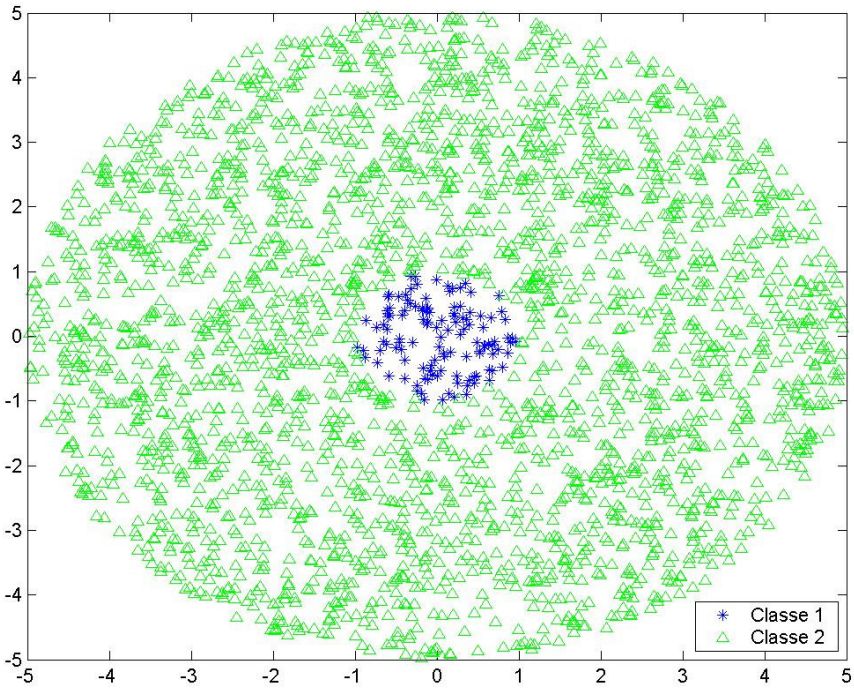

Figura 12: Classes C_1 e C_2 – Círculos Concêntricos.

2.4.4. Waveform

O banco de dados waveform, disponível em [59] foi gerado sinteticamente. Possui três classes de *waves* (nesta tese usa-se classe C_1 contra as demais), 5000 observações (3000 para fase dentro-da-amostra e 2000 para a fase fora-da-amostra), e 40 atributos de entrada, todos incluindo ruído. Os últimos 19 atributos de entrada são ruídos, todos com média zero e variância um.

2.4.5. Reconhecimento da Letra B

O objetivo é identificar um grande número de *pixels* retangulares em preto e branco como uma das 26 letras maiúsculas do alfabeto inglês. Como estamos em um ambiente de classificação binário, escolhemos a letra B para representar a classe C_1 e todas as outras 25 letras formam C_2. O banco consiste de 20000 observações (10600 para fase dentro-da-amostra e 9400 para a fase fora-da-amostra) e 16 atributos de entrada.

2.4.6.
Statlog – Landsat Satellite

O banco consiste em valores multi-espectrais de pixels em vizinhanças 3x3 em uma imagem de satélite e da classificação associada ao pixel central de cada vizinhança. O objetivo é prever esta classificação dados os valores multi-espectrais. Novamente, como tratamos de problemas de classificação binários, as 6 classes originais foram reduzidas a duas (classe C_1 contra as demais). Existem 6435 observações (4435 para a fase dentro-da-amostra e 2000 para a fase fora-da-amostra) e 36 atributos de entrada.

2.4.7.
Diagnóstico de Doenças Cardíacas

Este banco foi construído através da composição de quatro bancos de dados [14] referentes a um problema relacionado com diagnóstico de doenças cardíacas. O objetivo é prever o estágio de doença vascular através da estimação no estreitamento de vasos de grande calibre. Cada um destes quatro bancos que formam a composição usada nesta tese teve as observações coletadas, respectivamente, nos seguintes hospitais: Cleveland Clinic Foundation; Hungarian Institute of Cardiology; V.A. Medical Center, e do Zurich University Hospital. Todos os bancos tinham originalmente 76 atributos, mas somente 13 destes atributos são considerados relevantes e utilizados, em geral, na literatura [15]. Depois da eliminação de observações com atributos faltantes, o banco de dados resultou em 740 observações (pacientes) e 10 atributos de entrada. Neste banco de dados, utilizou-se validação cruzada (k-fold) para classificação com k = 10.

2.4.8.
Classificação de Punção Aspirativa por Agulha Fina de Nódulo Mamário para Diagnóstico

Freqüentemente usado como *benchmark* [2], [51], [53], [60], este banco foi provido pela University of Wisconsin Hospitals, Madison. Estes dados são provenientes de características citológicas de massa tumoral de mama. O conjunto de atributos de entrada é composto por 9 características citológicas de aspirados com agulha fina de massa tumoral mamária benigna e maligna. Entre estes

atributos de entrada estão: uniformidade do tamanho e da forma das células, nucléolos normais, adesividade das junções, espessura das junções, presença de mitoses e cromatina nuclear. Todos estes atributos assumem valores entre 1 e 10, e as classes são 'benigno' e 'maligno'. Após a eliminação de observações com atributos faltantes, restaram-se 683 pacientes. Neste banco de dados, utilizou-se validação cruzada (k-fold) para classificação com k = 10.

2.4.9. Ionosfera

Trata-se de um banco referente a dados coletados pelo sistema de radar na baía de Goose, em Labrador. Este sistema consiste de um "*phased array"* de 16 antenas de alta freqüência com um total de potência transmitido da ordem de 6.4 kilowatts. O alvo são elétrons livres na ionosfera. O retorno considerado 'Bom' dos radares são aqueles que mostram evidência de algum tipo de estrutura na ionosfera. O retorno 'Ruim' é o oposto; o sinal deles passou pela ionosfera. Os sinais recebidos foram processados usando uma função de auto-correlação cujos argumentos são o tempo de pulso e o número de pulso. Existem 17 números de pulsos descritos por duas características por pulso, o que resulta em 34 atributos de entrada. Existem 351 observações. Novamente, k-fold com k = 10 foi utilizado para classificação.

2.4.10. Câncer de Pulmão

Este banco foi utilizado em [58] para ilustrar o poder do plano discriminante ótimo em conjuntos "mal-postos". As observações descrevem três tipos de cânceres de pulmão usando 56 atributos de entrada. Aqui, usa-se a patologia 1 contra as demais no problema de classificação binária. O banco possui originalmente 32 observações, que resultaram em 27, 8 de C_1 e 19 de C_2, após remover observações com atributos faltantes.