Ricardo Ferreira da Rocha

Combinação de Classificadores para Inferência dos Rejeitados

São Carlos, 2012

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Ricardo Ferreira da Rocha

Combinação de Classificadores para Inferência dos Rejeitados

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Orientador: Francisco Louzada-Neto
Co-orientador: Carlos Alberto Ribeiro Diniz

São Carlos, 2012

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

R672cc

Rocha, Ricardo Ferreira da.
Combinação de classificadores para inferência dos rejeitados / Ricardo Ferreira da Rocha. -- São Carlos : UFSCar, 2012.
48 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2012.

1. Estatística. 2. Riscos Financeiros. 3. Combinação de classificadores. 4. *Credit scoring*. 5. Regressão logística. I. Título.

CDD: 519.5 (20ª)

Ricardo Ferreira da Rocha

COMBINAÇÃO DE CLASSIFICADORES PARA INFERÊNCIA DOS REJEITADOS

Dissertação apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovado em 16 de março de 2012.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Francisco Louzada Neto (ICMC-USP/Orientador)

1º Examinador _____
Dr. Eduardo Almeida Prado (Banco PanAmericano)

2º Examinador _____
Prof. Dr. Luis Aparecido Milan (DEs-UFSCar)

Resumo

Em problemas de *credit scoring*, o interesse é associar a um elemento solicitante de algum tipo de crédito, uma probabilidade de inadimplência. No entanto, os modelos tradicionais utilizam amostras viesadas, pois constam apenas de dados obtidos dos proponentes que conseguiram a aprovação de uma solicitação de crédito anterior. Com o intuito de reduzir o vício amostral desses modelos, utilizamos estratégias para extrair informações acerca dos indivíduos rejeitados para que nele seja inferida uma resposta do tipo bom/-mau pagador. Isto é o que chamamos de inferência dos rejeitados. Juntamente com o uso dessas estratégias utilizamos a técnica *bagging* (*bootstrap aggregating*), que é baseada na construção de diversos modelos a partir de réplicas *bootstrap* dos dados de treinamento, de modo que, quando combinados, gera um novo preditor. Nesse trabalho discutiremos sobre alguns dos métodos de combinação presentes na literatura, em especial o método de combinação via regressão logística, que é ainda pouco utilizado, mas com resultados interessantes. Discutiremos também as principais estratégias referentes à inferência dos rejeitados. As análises se dão por meio de um estudo simulação, em conjuntos de dados gerados e em conjuntos de dados reais de domínio público.

Palavras-chave: *Bagging*; Combinação de Modelos; *Credit Scoring* ; Inferência dos Rejeitados, Regressão Logística.

Abstract

In credit scoring problems, the interest is to associate to an element who request some kind of credit, a probability of default. However, traditional models uses samples biased because the data obtained from the tenderers has only clients who won a approval of a request for previous credit. In order to reduce the bias sample of these models, we use strategies to extract information about individuals rejected to be able to infer a response, good or bad payer. This is what we call the reject inference. With the use of these strategies, we also use the bagging technique (bootstrap aggregating), which consist in generate models based in some bootstrap samples of the training data in order to get a new predictor, when these models is combined. In this work we will discuss about some of the combination methods in the literature, especially the method of combination by logistic regression, although little used but with interesting results. We'll also discuss some strategies relating to reject inference. Analyses are given through a simulation study, in data sets generated and real data sets of public domain.

Keywords: Bagging; Credit Scoring; Logistic Regression; Model Combination; Reject Inference.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A análise de crédito é baseada em buscar informações consistentes para uma tomada de decisão num contexto de incertezas e constantes transformações acerca do processo de empréstimo de algum tipo de crédito, com a promessa de pagamento posterior. Essa informação é muito importante para as instituições fornecedoras de crédito, uma vez que parte substancial dos lucros destas instituições provém da concessão de crédito e, sendo assim, é essencial a utilização das melhores ferramentas disponíveis para auxiliar na decisão sobre um solicitante.

A implementação de técnicas de modelagem e aprimoramento contínuo são premissas básicas quando se trata das técnicas para a análise de risco. Os modelos utilizados nessas análises são responsáveis pela movimentação de uma enorme quantia de capital e ganhos relativamente baixos na capacidade de previsão de um modelo são capazes de gerar diferenças significativas no balanço geral de uma instituição.

Os modelos estatísticos são desenvolvidos a partir de amostras de uma população de interesse, de forma que os resultados obtidos refletem as características presentes nos conjuntos de dados utilizados. Então, ao selecionar os dados que serão utilizados num modelo de risco de crédito é fundamental a interpretação sobre a representatividade dessa amostra em relação a população total. Em problemas de *credit scoring* esse fator é predominante, pois existem muitos indivíduos que não foram aprovados num processo de seleção anterior e, consequentemente, não podem ser observado seus comportamentos e suas peculiaridades não são então absorvidas pelos modelos estatísticos. Logo, as amostras usuais não são totalmente representativas da população de interesse, e causam um vício amostral indesejado.

Esse vício pode ser mais ou menos influente no modelo final de acordo com a proporção de rejeitados em relação ao total de proponentes. Quanto maior essa proporção, mais importante é o uso de alguma estratégia para a correção do vício amostral. Para solucionar esse problema, trabalhamos com o que convencionalmente se conhece como inferência dos

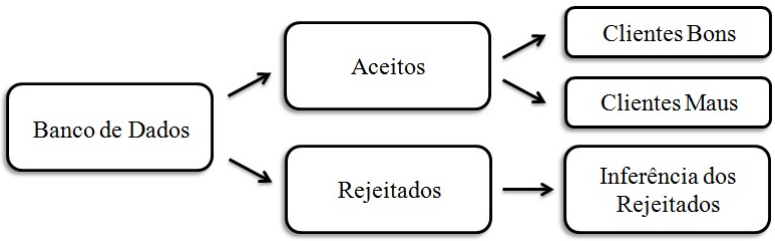

Figura 1.1: *Esquema da distribuição dos dados para um modelo de credit scoring.*

rejeitados, com o papel de estimar o comportamento dos indivíduos que foram rejeitados.

Neste trabalho descrevemos algumas das técnicas de inferência dos rejeitados presentes na literatura, as quais serão implementadas juntamente com os métodos de combinações de modelos, que são descritos nas seções seguintes. No entanto, apenas o uso da inferência dos rejeitados não é suficiente para que haja ganhos significativos na qualidade dos modelos. Hand (1993) conclui que é impossível uma inferência dos rejeitados satisfatória.

Por outro lado, existe o aspecto empresarial, em que a melhoria do modelo se dá na melhor interpretação para os analistas de crédito, ao diminuir os casos em que o escore está próximo do ponto de corte e requer uma nova análise interna (Sabato, 2009).

Com o intuito de aumentar a capacidade preditiva nos modelos de escoragem de crédito com inferência dos rejeitados, propomos a utilização da técnica proposta por Breiman (1996) chamada *bagging*. Esse método combina os escores gerados a partir de diversos modelos, considerando réplicas bootstrap da base de dados. No próximo capítulo apresentaremos a descrição de um problema de *credit scoring*, juntamente com o algoritmo de modelagem de regressão logística e as medidas utilizadas para medir o poder preditivo alcançado nos modelos propostos. No capítulo 3 apresentaremos a técnica *bagging* e as estratégias de combinação de modelos. No capítulo 4 serão discutidos os métodos de inferência dos rejeitados utilizados e nos últimos capítulos apresentaremos os resultados obtidos por meio dos estudos de simulação nos dados gerados e nos dados reais.

Capítulo 2

Preliminares

2.1 Problemas de Credit Scoring

O objetivo de um modelo de *credit scoring* é associar uma probabilidade de inadimplência ao indivíduo que solicita algum tipo de crédito de acordo com diversas características acerca do cliente. Para isso, é necessário buscar e avaliar como que certas características sobre o solicitante, ou do produto solicitado, podem vir a influenciar na variável resposta desejada, que neste caso é a probabilidade de inadimplência. Para isso, utiliza-se o histórico de clientes que já foram analisados como bons ou maus pagadores. A metodologia para o desenvolvimento de um modelo de *credit scoring* segue, basicamente, as seguintes etapas (Alves, 2002):

- Planejamento e definições: tipos de clientes e produtos de crédito para os quais serão desenvolvidos o modelo, finalidades de uso, conceito de inadimplência a ser adotado etc;
- Identificação das variáveis potenciais: caracterização do solicitante de crédito, seleção das variáveis significativas para o modelo etc;
- Coleta dos dados;
- Determinação da fórmula de escoragem via técnicas estatísticas;
- Classificação do solicitante em bom ou mau pagador, a partir de sua escoragem.

As definições de inadimplência podem variar entre as instituições de crédito. Essa determinação pode depender do tipo de crédito concedido, do tempo total de empréstimo etc. Normalmente, após um período de 12 ou 18 meses de observação, já é possível definir se um indíviduo é bom pagador, no entanto, se ocorrer um atraso de 30 dias (ou 90 dias, por exemplo) num certo pagamento, o índivíduo é então classificado como mau pagador.

Por outro lado, quando lidamos com solicitantes rejeitados é porque o elemento obteve uma escoragem muito alta (alto risco de inadimplência) ou alguma restrição de crédito anterior. Muitas instituições adotam políticas de corte para a concessão do crédito, por exemplo, se o solicitante estiver com algum outro crédito pendente, ou com contas atrasadas. Essas informações são acessadas nas chamadas *bureau* de crédito, que possui informações do histórico de vários clientes em diversas operações de risco, como contas de energia, telefone, seguro etc. As principais intituições financeiras consideram modelos estatísticos no auxílio da tomada de decisão. Existem diversas estratégias que podemos assumir para gerar uma classificação ou uma escoragem de crédito. Neste trabalho, utilizaremos a técnica de regressão logística, que é um recurso de modelagem amplamente usado e está descrita na próxima seção.

2.2 Modelo de Regressão Logística

Historicamente, inúmeras estratégias para a elaboração de escores de crédito foram desenvolvidas a partir de diversos algoritmos de modelagem. Dessas estratégias, frequentemente é utilizado o modelo de regressão lógistica para estabelecer a conexão entre as variáveis explanatórias com as respostas observadas. O modelo pode ser interpretado, num contexto geral, como o modelo linear generalizado em que a variável resposta Y tem distribuição de Bernoulli com função de ligação logito.

O modelo de regressão logística possui a vantagem de não precisar supor normalidade nos resíduos e nem variâncias iguais nas observações, além de apresentar uma boa interpretação de seus coeficientes em relação ao problema em questão. Considere $x' = (x_1, \cdots, x_p)$ o vetor de p covariáveis. O logito do modelo de regressão logística é dado pela equação

$$g(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i, \tag{2.1}$$

de tal sorte que a probabilidade condicional desejada é então dada pela equação

$$P(Y = 1) = \pi(x) = \frac{1}{1 + \exp(-g(x))}. \tag{2.2}$$

A estimação de $\beta' = (\beta_0, \cdots, \beta_p)$ é feita por máxima verossimilhança, e sua equação é dada por

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}, \tag{2.3}$$

com (x_i, y_i) representando uma das n observações dos dados.

Trabalhando com o logarítmo de (2.3), deriva-se a expressão em relação a β_0

e β_1 (ou qualquer outro $\beta_i, i = 1, ..., p$) e obtêm-se as equações de verossimilhança:

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \tag{2.4}$$

e

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0. \tag{2.5}$$

Essas equações são usualmente resolvidas numericamente, por meio do algoritmo Newton-Raphson.

Uma vez que os parâmetros do modelo estão estimados, é necessário analisar a adequabilidade do mesmo. Essa análise é feita com o objetivo de identificar quais covariáveis, de fato, contribuem para a explicação da variável resposta. Para isso considera-se o modelo ajustado com todas as covariáveis e compara-se com modelo ajustado sem a varíavel (ou variáveis) que deseja-se verificar, indicando assim a sua significância de cada variável. A estatística

$$G = -2\log\left(\frac{\text{verossimilhança sem as variáveis}}{\text{verossimilhança com as variáveis}}\right) \tag{2.6}$$

é usada na análise da significância de um conjunto de covariáveis e, sob a hipótese de que as variaveis consideradas são iguais a zero, segue uma distribuição χ^2 com p graus de liberdade (Hosmer & Lemeshow, 1989).

2.3 Medidas Preditivas

Com um modelo ajustado, é importante medir a sua capacidade de predição. O modelo de regressão logística nos retorna um resultado no intervalo $[0, 1]$, que chamamos de escoragem. A partir dessa escoragem é tomada a decisão de quais indivíduos serão considerados bons pagadores e quais serão considerados maus pagadores. Essa determinação é feita utilizando um ponto de corte $c \in (0, 1)$ (*cut-off*), de tal forma que os indivíduos com escore maior que c são considerados os potenciais inadimplentes e os com escore menor, os adimplentes. Mais abaixo descreveremos o processo para escolher o valor de c. Logo, podemos associar a escoragem dos clientes à uma classificação e então, calcular diversas métricas que auxiliam na interpretação da adequabilidade do modelo. As medidas são baseadas na comparação das respostas preditas em relação às respostas observadas. Ao comparar, há quatro possíveis situações:

- Verdadeiro Positivo (VP): ocorre quando classificamos um elemento como positivo e, de fato, o elemento foi observado como positivo.
- Verdadeiro Negativo (VN): ocorre quando classificamos um elemento como negativo

e, de fato, o elemento foi observado como negativo.

- Falso Positivo (FP): ocorre quando classificamos um elemento como positivo, mas na verdade o elemento foi observado como negativo.

- Falso Negativo (FN): ocorre quando classificamos um elemento como negativo, mas na verdade o elemento foi observado como positivo.

A partir destes valores, calculamos então as seguintes medidas preditivas:

- **Sensibilidade:** É a proporção entre os classificados como positivos, dentre todos que foram observados como positivos. Representa a probabilidade de um modelo identificar os maus indivíduos da população.

$$SENS = \frac{VP}{VP + FN}. \tag{2.7}$$

- **Especificidade:** É a proporção entre os classificados como negativos, dentre todos que foram observados como negativos. Representa a probabilidade de um modelo identificar os bons indivíduos da população.

$$SPEC = \frac{VN}{VN + FP}. \tag{2.8}$$

- **Valor Preditivo Positivo:** É a proporção entre os classificados como positivos, dentre todos que foram classificados pelo modelo como positivos. Representa a probabilidade de um indivíduo ser positivo dado que o modelo indicou como positivo.

$$VPP = \frac{VP}{VP + FP}. \tag{2.9}$$

- **Valor Preditivo Negativo:** É a proporção entre os classificados como negativos, dentre todos que foram classificados pelo modelo como negativos. Representa a probabilidade de um indivíduo ser negativo dado que o modelo indicou como negativo.

$$VPN = \frac{VN}{VN + FN}. \tag{2.10}$$

- **Acurácia:** É a proporção entre todos os acertos do modelo, em relação a soma de todas as classificações. Representa a capacidade de acerto total de um modelo.

$$ACC = \frac{VP + VN}{VP + VN + FN + FP}. \tag{2.11}$$

- **Coeficiente de Correlação de Matthews:** Representa a correlação entre os valores preditos e observados.

$$MCC = \frac{VP.VN - FP.FN}{\sqrt{(VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)}}. \tag{2.12}$$

Essas medidas permitem uma boa interpretação da adequabilidade do modelo quando usadas em conjunto pois, individualmente, elas podem levar a falsas conclusões. Por exemplo, um modelo com acurácia igual a 95% não é necessariamente um bom modelo, pois não leva em consideração a prevalência amostral. Se a prevalência do evento de interesse for 5% e o modelo indicar todas observações como negativas, então a acurácia seria 95% num modelo que não tem utilidade prática.

Todas essas medidas, exceto o MCC, variam no intervalo $[0, 1]$ e quanto mais próximo seu valor for de 1, melhor é o desempenho avaliado. O MCC varia no intervalo $[-1, 1]$, quanto mais próximo de 1, melhor a predição, enquanto que -1 representa a predição inversa e 0 representa nenhuma correlação entre as observações. Essa medida não sofre impacto da prevalência amostral, o que a torna uma das medidas mais confiáveis na determinação da qualidade de um modelo.

Vamos considerar também uma medida mais próxima do contexto de *credit scoring*, que leva em consideração os erros cometidos. Utilizaremos aqui uma medida baseada na apresentada em Bensic *et al* (2005), chamada **custo relativo**, calculada por

$$CR = \alpha \cdot C_1 \cdot P_1 + (1-\alpha) \cdot C_2 \cdot P_2, \tag{2.13}$$

de tal sorte que α representa a probabilidade de um aplicante ser mau pagador, C_1 é o custo relativo ao cometer o erro de aceitar um mau pagador, C_2 é o custo relativo ao cometer o erro de rejeitar um bom pagador, P_1 é a probabilidade de ocorrer um falso negativo e P_2 é a probabilidade de ocorrer um falso positivo.

Na prática é bem complicado obter as estimativas de C_1 e C_2, pois depende de muitos detalhes da operação em questão, do tipo de crédito etc. Para fazer uma análise sem levar em consideração esses valores, é feito o calculo pra diversas proporções entre C_1 e C_2, fazendo a suposição que $C_1 > C_2$. Essa suposição é equivalente a dizer que o prejuízo em aceitar um mau pagador é maior do que o lucro perdido ao rejeitar um bom pagador. Certamente, essa hipótese se encaixa dentro da maioria dos problemas no contexto de *credit scoring*.

Entretanto, em nossas aplicações estaremos interessados apenas em comparar diversos métodos, sem que o valor dessa medida seja representativa por si própria, sendo o mais relevante a ordem dessas medidas ao comparadas com outras. Sendo assim, ao fixar um

valor para C_1, o custo relativo é crescente em relação a C_2, o que implica que a ordem entre duas análises quaisquer de valores de C_2 é preservada. Então, podemos usar uma equação simplificada, usando $C_1 = C_2 = 1$. Em relação a α, Bensic *et al* (2005) considera como sendo a prevalência amostral, isto é, supõe que a prevalência de maus pagadores dos bancos de dados representa a prevalência real da população de interesse.

O uso destas medidas em conjunto é suficiente para a nossa análise de modelos, uma vez que o objetivo principal deste trabalho é comparar as técnicas propostas. No entanto, sabemos que a realidade numa instituição financeira pode ser bem diferente. Normalmente, os solicitantes de crédito são avaliados de acordo com a faixa de escore em que se encontram, isto é, uma vez que um escore é determinado pelas técnicas de modelagem da instituição, ele é então analisado juntamente com outros solicitantes que obtiveram crédito semelhante. Cada uma destas faixas podem ter suas próprias políticas para determinar a concessão ou não do crédito. Na prática, medidas como as taxas de inadimplência e lucratividade por faixa de escore são as mais utilizadas.

2.4 Bancos de Dados

Nos modelos aqui desenvolvidos, além de dados simulados, utilizaremos dois bancos de *credit scoring* de livre domínio, disponíveis na internet. O primeiro deles é o *German Credit Data*. Esse banco de dados consiste de 20 variáveis cadastrais, 13 categóricas e 7 numéricas, e 1000 observações de utilizadores de crédito, dos quais 700 são correspondentes a bons pagadores e 300 (prevalência de 30% de positivos) a maus pagadores.

O segundo é o *Australian Credit Data*, que consiste de 14 variáveis, 8 categóricas e 6 contínuas, e 690 observações, das quais 307 (prevalência de 44,5% de positivos) são elementos inadimplentes e 383 são adimplentes.

Ambos bancos de dados, e muitos outros, podem ser acessados no website do *UCI Machine Learning Repository*.

Capítulo 3

Combinação de Modelos

3.1 Bagging de Modelos

Proposto por Breiman (1996), o *bagging* (*bootstrap aggregating*) é uma técnica onde constroem-se diversos modelos baseados nas réplicas *bootstrap* de um banco de dados de treinamento. Todos os modelos são então combinados, de forma a encontrar um preditor que represente a informação de todos modelos gerados. Breiman propõe uma combinação feita por votos, em que a resposta final é a resposta dada pela resposta da maioria dos modelos. Seja B o número de réplicas utilizadas, o procedimento *bagging* pode então ser descrito nos seguintes passos:

- Gerar $L_1^*, \cdots, L_B^*$ réplicas *bootstrap* da amostra treinamento L;
- Para cada réplica gera-se o modelo com preditor $S_i^*, i = 1, \cdots, B$;
- Combinando os modelos, chega-se no preditor *bagging* S^*.

A característica que deve estar presente no conjunto de dados para que este procedimento apresente bons resultados é a instabilidade. Breiman (1996) define, informalmente, um modelo como instável se pequenas variações nos dados de treinamento leva a grandes alterações nos modelos ajustados.

Quanto mais instável é o classificador básico, mais sortidos serão os modelos ajustados pelas réplicas *bootstrap*, fazendo com que cada modelo possua informações diferentes, de forma que a contribuição para o preditor combinado seja maior. Se o classificador básico for estável, as réplicas gerariam praticamente os mesmos modelos e não haveriam aumentos significantes no preditor combinado final. Algoritmos de modelagem como redes neurais e árvores de decisão são exemplos de classificadores usualmente instáveis (Kuncheva, 2004). Em Büllmahn & Yu (2002), é feita uma análise do impacto da utilização do *bagging* no

erro quadrático médio e na variância do preditor final, utilizando uma definição algébrica de instabilidade.

Desde que esta técnica foi apresentada, diversas variantes foram desenvolvidas. Ainda em Büllmahn & Yu (2002), é proposto a variante *subagging* (*subsample aggregating*), que consiste em retirar amostras aleatórias simples dos dados de treinamento, mas de tamanho menor. A combinação é feita usualmente por voto majoritário, mas no entanto, também admite uso de outras técnicas. Essa estratégia apresenta resultados interessantes, com resultados ótimos quando o tamanho das amostras é metade do tamanho do conjunto de dados de treinamento (*half-subagging*). No artigo é mostrado que os resultados com *half-subagging* são praticamente iguais aos do *bagging*, principalmente em amostras pequenas.

Em Louzada-Neto *et al* (2011) é proposto um procedimento que generaliza a ideia de reamostragem do *bagging*, chamado *poly-bagging*. A estratégia é fazer reamostras sucessivas nas próprias amostras *bagging* originais. Cada reamostragem aumenta um nível na estrutura e complexidade da implementação. Os resultados em dados simulados foram expressivos, mostrando que é possível reduzir ainda mais a taxa de erro de um modelo. A técnica se mostra poderosa em diversas configurações de tamanhos amostrais e prevalências. É feito uma aplicação em dados de *credit scoring* como exemplo em um conjunto de dados reais.

É importante ressaltar que quando utilizamos a ferramenta *bagging* é enfraquecido o entendimento da explicação do modelo em relação a interpretação dos coeficientes da regressão, pois o escore final de um novo cliente se dá por meio da escoragem obtida em diversos modelos da estrutura de *bagging*, que são então depois combinados. Uma alternativa para a interpretação do modelo é o uso da meta-análise, que explora o uso de ferramentas estatísticas para analisar uma coleção de estudos independentes. Para mais informações consulte DerSimonian e Laird (1986).

Na modelagem via *bagging*, a aplicação dos novos clientes deve passar por todos os modelos construídos na estrutura. Cada elemento será escorado, em nosso caso, por 25 modelos de forma a obter 25 escoragens diferentes. Com essas informações, um novo escore será obtido por meio da aplicação dos escores anteriores numa determinada função combinação.

Uma vez implementado a estrutura de *bagging*, é necessário o uso de alguma ferramenta para agregar todos os preditores em um único, representando assim o modelo combinado. Na próxima seção discutiremos as abordagens propostas para tais combinações. Para tanto considere os preditores S_i e a função combinação $c(S_1, \cdots, S_B) = S^*$.

3.2 Combinações via Médias

A combinação via média é uma das mais presentes na literatura. É uma combinação de simples implementação e não necessita a estimação de novos parâmetros na estrutura da modelagem. Em equações temos

$$S^* = c(S_1, \cdots, S_B, \alpha) = \frac{1}{B}\sum_{i=1}^{B} S_i^*. \tag{3.1}$$

Em termos gerais, como proposto em Kuncheva (2004), podemos escrever a equação (3.1) como caso particular da equação

$$S^* = \left(\frac{1}{B}\sum_{i=1}^{B}(S_i^*)^\alpha\right)^{\frac{1}{\alpha}}, \tag{3.2}$$

quando $\alpha = 1$.

Essa formulação permite a dedução de outros tipos menos comuns de combinação, que podem ser utilizadas em situações mais específicas. Ao variar o valor de α pode-se controlar a modelagem de acordo com a relevância do problema em questão.

Além do caso $\alpha = 1$ gerando a combinação por média, temos outros casos particulares interessantes. Se $\alpha = -1$, a equação (3.2) representa uma combinação via média harmônica, isto é,

$$S^* = \left(\frac{1}{B}\sum_{i=1}^{B}\frac{1}{S_i^*}\right)^{-1}. \tag{3.3}$$

Se $\alpha \to 0$, a equação (3.2) representa uma combinação via média geométrica:

$$S^* = \left(\prod_{i=1}^{B} S_i^*\right)^{\frac{1}{B}}. \tag{3.4}$$

Se $\alpha \to -\infty$, a equação (3.2) representa uma combinação pelo mínimo:

$$S^* = \min_{i=1}^{B} S_i^*. \tag{3.5}$$

Se $\alpha \to \infty$, a equação (3.2) representa uma combinação pelo máximo:

$$S^* = \max_{i=1}^{B} S_i^*. \tag{3.6}$$

Estas estratégias podem ser usadas de acordo com o conservadorismo, ou otimismo, que deseja-se exercer sobre a modelagem em questão. Quanto menor o valor de α, a combinação

fica mais próxima da combinação via mínimo, que é otimista por tomar o menor escore dentre os modelos gerados. Se escolhermos valores altos para α, o valor do escore tenderá a aumentar, representando assim uma combinação com tendências conservadoras. Em nossas modelagens consideraremos a análise para diversos valores de α, de modo que seja possível encontrar seu melhor valor no contexto de *credit scoring*, de acordo com as medidas calculadas.

3.3 Combinações via Votos

A combinação por voto é também uma estratégia muito simples e comumente usada. No contexto que estamos trabalhando, é necessário antes fazer a associação da escoragem com a classificação final dos elementos. Consideremos a variável C_i^*, que corresponde a classificação associada a escoragem S_i^*, definida a partir do ponto de corte c escolhido.

$$C_i^* = 1 \text{ se } S_i^* > c \text{ e } C_i^* = 0 \text{ no caso contrário.} \tag{3.7}$$

Lembrando que, na nossa notação, S_i^* corresponde a um vetor de escores e a equação (3.7) refere-se a classificação elemento a elemento da escoragem. A partir dos classificadores C_i^*, define-se a combinação por voto majoritário da seguinte maneira:

$$C^* = 1 \text{ se } \sum_{i=1}^{B} C_i^* \geq \left[\frac{B}{2}\right] \text{ e } C^* = 0 \text{ no caso contrário,} \tag{3.8}$$

com [.] representando a função maior inteiro.

Nos casos em que B é ímpar, temos uma maioria absoluta dos classificadores, no entanto, quando B é par pode ocorrer casos de empate e, segundo a definição em (3.8), será classificado como 1.

Neste trabalho, analisaremos a combinação via voto de uma maneira geral, variando todos os possíveis números de votos. A equação geral pode ser descrita por

$$C_i^* = 1 \text{ se } \sum_{i=1}^{B} C_i^* \geq k \text{ e } C_i^* = 0 \text{ no caso contrário,} \tag{3.9}$$

com $k = 0, \cdots, B$.

Da mesma maneira que nas variações pela combinação por média, em razão de se obter um modelo mais conservador ou otimista, é possível aumentar ou diminuir a quantidade de votos necessários para uma classificação. Essas variações podem ajudar em casos específicos em que deseja-se uma sensibilidade ou especificidade maior, de acordo com o problema em questão. Em medicina, é comum em alguns testes usar a regra pelo voto unânime, isto

é, o classificador final é considerado positivo se, e somente se, todos os classificadores individuais assim apontarem (Kuncheva, 2004).

3.4 Combinações via Regressão Logística

Essa é uma estratégia apresentada em (Zhu *et al*, 2002), que consiste em combinar os preditores considerando-os como covariáveis em um modelo de regressão logística explicando a variável resposta. Em equações temos:

$$S_0 = \log\left(\frac{P(Y=1|S_1,\cdots,S_B)}{1-P(Y=1|S_1,\cdots,S_B)}\right) = \beta_0 + \sum_{i=1}^{B}\beta_i S_i, \tag{3.10}$$

de sorte que P(Y=1) representa a probabilidade do evento de interesse.

Essa combinação pode ser interpretada como uma espécie de combinação linear ponderada, de forma que o modelo de regressão logística aponta os modelos mais influentes na explicação da variável resposta por meio de seus coeficientes.

Para tornar os valores de S_0 no intervalo $[0,1]$ pode-se usar qualquer função monotônica que respeite os domínios desejados. A monotonicidade da função garante a ordenação correta da escoragem final. Neste trabalho usaremos a função logito, que já é comumente usada nesses casos.

A combinação linear ponderada é a combinação

$$S^* = \sum_{i=1}^{B} w_i S_i^*, \text{ tal que } \sum_{i=1}^{B} w_i = 1. \tag{3.11}$$

Quando escolhemos os valores de w_i de forma que maximize uma ou mais medidas preditivas, acaba-se tornando muito custoso computacionalmente. Para pequenos valores de B o processo já é bastante ineficaz, inviabilizando uma escolha livre para este parâmetro, que normalmente não é tão baixo. Nesse sentido, a combinação via regressão logística apresenta uma boa alternativa e é computacionalmente eficaz. De modo a tornar a interpretação ainda mais próxima, consideraremos também o caso em que a combinação é feita via regressão logística sem intercepto.

Ainda no trabalho de Zhu *et al* (2002), é verificada uma propriedade interessante acerca desta combinação. São definidos conceitos como suficiência e irrelevância entre dois escores, que são utilizados para estudar em que condições um escore pode ser influente numa combinação. Ela pode gerar ganhos preditivos na combinação de dois escores mesmo quando um dos escores é dominante em relação ao outro. Foi feito um estudo com 600.000 consumidores de crédito, com prevalência de aproximadamente 10%, utili-

zando duas escoragem diferentes, uma obtida a partir de um *bureau* de crédito e outro por modelagem dos dados cadastrais dos clientes. É mostrado que o escore obtidos pelos dados cadastrais é suficiente em relação ao outro e, apesar disso, o escore via *bureau* não é irrelevante para o modelo combinado via regressão logística.

Capítulo 4

Inferência dos Rejeitados

Uma premissa fundamental na modelagem estatística é que a amostra selecionada para o modelo represente a população total da qual se deseja estimar. Nos problemas de *credit scoring* geralmente essa premissa é violada, pois são utilizados apenas os proponentes aceitos, que puderam ter o comportamento observado. Os rejeitados, por sua vez, não são observados e são usualmente descartados do processo de modelagem.

A inferência dos rejeitados é a associação de uma resposta para o indivíduo não observado, de forma que seja possível utilizar suas informações em um novo modelo. Os principais métodos podem ser verificados em Ash & Meester (2002), Banasik & Crook (2005), Crook & Banasik (2004), Crook & Banasik (2007), Feelders (2003), Hand (2001) e Parnitzke (2005).

Por mais que seja simples a definição do problema que estamos trabalhando, é um trabalho complexo construir técnicas realmente eficientes em inferência dos rejeitados. As técnicas, por sua vez, possuem a característica de serem mais ineficazes de acordo com que a proporção de rejeitados aumenta. E quanto maior a proporção de rejeitados, mais é necessário alguma estratégia para reduzir o vício amostral (Ash & Meester, 2002).

Neste texto consideramos as técnicas da reclassificação, ponderação e parcelamento, descritas a seguir.

4.1 Método da Reclassificação

Uma das estratégias mais simples para aderir os proponentes rejeitados ao modelo é simplesmente considerar toda população dos rejeitados como sendo maus pagadores. Essa estratégia procura reduzir o viés amostral baseado na ideia de que na população dos rejeitados espera-se que a maioria sejam maus pagadores, embora certamente possa haver bons pagadores em meio aos rejeitados. Adotado esse método, toda população dos bons que foram rejeitados são classificados erroneamente e, consequentemente, os proponentes

com perfis similares são prejudicados (Thomas *et al*, 2002).

No entanto, pela característica desta técnica, é de se esperar um modelo mais sensível, em que os elementos positivos serão identificados melhores, o que é de grande importância no contexto de escoragem de crédito.

4.2 Método da Ponderação

Provavelmente esta é a estratégia mais presente na literatura. Como proposto em Crook & Banasik (2005), esse método consiste em assumir que a probabilidade de um cliente ser mau pagador independe do fato de ter sido aceito ou não. A representação dos rejeitados é feita pelos proponentes que possuem escore semelhante, mas que foram aceitos. Estes carregam a informação dos rejeitados para o modelo através de pesos atribuídos, que são calculados de acordo com seu escore. Os indivíduos que foram aceitos, ainda que com escores altos, serão os que vão carregar os maiores pesos, representando assim a população dos rejeitados. No entanto, para implementar esse método, é necessário ter em mãos um modelo inicial que separe a população de todos os proponentes em aceitos e rejeitados e que fornece a probabilidade de inadimplência de cada indivíduo.

Em Parnitzke (2005), é alcançado um aumento de 1,03% nas capacidade de acerto total em dados simulados, e nenhum aumento quando baseado num conjunto de dados reais. Em Alves (2008), os resultados são bem similares aos do modelo logístico usual.

4.3 Método do Parcelamento

De acordo com Parnitzke (2005), para desenvolver essa estratégia, deve-se considerar o modelo proposto quando utilizado somente os proponentes aceitos. O próximo passo é dispor os solicitantes utilizados no modelo em faixas de escore. É razoável determinar essas faixas de forma que os elementos escorados se distribuam de modo uniforme. Em cada faixa de escore, verifica-se a taxa de inadimplência e então escora-se os rejeitados, para distribuídos nas mesmas faixas de escore antes construídas. Para cada rejeitado é associado uma resposta do tipo bom ou mau pagador, de forma aleatoria e de acordo com as taxas de inadimplência observadas nos proponentes aceitos. E então é construído um modelo com os clientes aceitos e rejeitados com sua devida resposta inferida.

De acordo com que os escores aumentam, a concentração de maus fica maior em relação a de bons pagadores, essa proporção é então utilizada para distribuir os rejeitados, que pertencem a tais faixas de escores, conforme indicado nas duas últimas colunas da Tabela (4.1).

Tabela 4.1: *Esquema da distribuição dos rejeitados no método do parcelamento*

Faixa de Escore	Bons	Maus	% Maus	Rejeitados	Bons	Maus
0-200	285	20	0,0656	25	23	2
200-400	450	85	0,1589	35	29	6
400-600	295	135	0,314	95	65	30
600-800	180	200	0,5263	260	123	137
800-1000	90	260	0,7429	375	96	279

Os resultados apresentados por essa técnica também são similares aos usuais, em alguns casos, levando a pequenos melhoramentos.

4.4 Outros Métodos

Uma estratégia não muito conveniente já proposta é a de aceitar todos os solicitantes por um certo período de tempo para que seja possível criar um modelo completamente não viciado. No entanto, essa ideia não é bem vista pois o risco envolvido em aceitar proponentes nos escores mais baixos pode não compensar o aumento de qualidade que o modelo pode vir a gerar. Outra ideia seria aceitar apenas uma pequena parcela dos que seriam rejeitados, entretanto, pelos mesmos motivos citados anteriormente, essa prática tem difícil justificativa (Hand, 2001).

Outro método é o uso de informações de mercado (*bureau* de crédito), que baseia-se em utilizar informações obtidas de alguma central de crédito que consta de registros de atividade de crédito dos proponentes, o que permite verificar como que os proponentes duvidosos se comportam em relação aos outros tipos de compromissos, como contas de energia, de telefone, seguros etc.

Os proponentes rejeitados são avaliados em dois momentos, o primeiro é quando solicitam o crédito e o segundo momento é algum tempo depois, de tal forma que esse tempo é um período de avaliação pré-determinado. No primeiro momento, pode ser que os proponentes não possuam nenhuma irregularidade e isso pode permanecer como está ou adquirir irregularidades durante o período de avaliação, assim como os que possuíam irregularidades, podem ou não possuir no segundo momento. Assim, após feita uma comparação entre as informações obtidas e as informações da proposta de crédito, classifica-se o indivíduo com bom ou mau pagador.

Logo, constrói-se um novo modelo considerando o banco de dados com os clientes aceitos (classificados como bom ou mau pagador segundo a própria instituição) acrescido dos clientes rejeitados com resposta definida a partir de suas informações de mercado. Para construir um modelo com esta estratégia, deve-se considerar que, obviamente, há mais informações acerca dos proponentes do que nas demais aqui abordadas, e, portanto,

espera-se um modelo melhor, no entanto, o acesso a essas informações pode requerer um investimento financeiro que não deve ser desconsiderado (Rocha & Andrade, 2002).

Capítulo 5

Estudo de Simulação dos Métodos de Combinação

5.1 Especificações da Estrutura de Bagging e Combinação de Modelos

A implementação é feita de maneira similar nos diversos modelos aqui desenvolvidos. Uma vez que o conjunto de dados está definido, conforme proposto em Hosmer & Lemeshow (1989), separa-se 70% dos dados disponíveis como amostra de treinamento e os 30% restantes ficam reservados para o cálculo das medidas de desempenho dos modelos. Com a amostra de treinamento em mãos, são retiradas 25 réplicas *bootstrap* e então constroem-se os modelos da estrutura do *bagging*. O valor de 25 réplicas foi escolhido baseado no trabalho de Breiman (1996), que é mostrado que as medidas preditivas analisadas convergem rapidamente em relação ao número de modelos. A diferença entre a modelagem com 25 e 50 réplicas foram mínimas.

Com os modelos construídos nas amostras *bootstrap*, é feita a escoragem da amostra teste por cada modelo e os escores são então sujeitos aos métodos de combinação e determinam um preditor final. A escolha dos pontos de corte é feita de tal forma que maximize o MCC do preditor final, analisando numericamente seu valor em cada incremento de 0,01 no intervalo $[0, 1]$. Para resultados estáveis, foram simuladas 1000 vezes cada modelagem. O software utilizado nos ajustes foi o SAS (versão 9.0) e o processo de seleção de variáveis utilizado nas regressões foi o *stepwise*.

Em todos os modelos foram utilizadas subamostras estratificadas em relação a variável resposta, isto é, cada subamostra gerada preservou a prevalência da resposta observada.

As combinações via médias são fácilmente implementadas, dependendo apenas da álgebra a ser aplicada nas escoragens das amostras testes. A equação geradora de com-

binações será analisada nos valores de $\alpha = -11, -10, \cdots, 10, 11$, e nos valores extremos, isto é, combinando via mínimo e via máximo.

Nas combinações via votos, é necessário classificar cada escoragem gerada pelas réplicas *bootstrap*. A classificação do escore é feita buscando o valor do ponto de corte, por todo intervalo $[0, 1]$, que maximiza a medida de desempenho MCC. Analisaremos os modelos em todas as possíveis contagens de votos, isto é, para todo $k = 1, 2, \cdots, 25$.

A combinação por regressão logística também é de fácil implementação. Inicialmente, considera-se os modelos *bagging* da amostra de treinamento escoradas na própria amostra de treinamento. Com essa escoragem é gerado o banco de dados para uma regressão logística, em que cada escoragem corresponde a uma covariável da regressão. Então, usa-se os coeficientes gerados na última regressão para gerar o escore combinado da amostra teste e sua classificação se dá da mesma maneira da combinação por média, maximizando o MCC. Consideraremos o caso da regressão logística sem intercepto, que é o que mais se aproxima, num contexto de interpretação, de uma combinação ponderada e o caso da regressão logística com intercepto, a fim de verificar seu impacto como parâmetro extra na combinação.

Nas próximas seções descreveremos a experimentação nos dados reais e a nossa estrutura de estudo de simulação. As figuras resumem os resultados obtidos pelas médias em cada método de combinação, além da aplicação usual do modelo lógistico, sem combinação. No estudo foram feitas 1000 simulações, variando a distribuição da amostra teste e treinamento.

5.2 Dados Gerados via Breiman

Usaremos como base da geração de dados a proposta de Breiman (1998), em que a população dos negativos é dada por k covariáveis com distribuição normal com média zero e variância k. A população dos positivos é gerada a partir de uma normal com média $\frac{1}{\sqrt{k}}$ e variância k. Nessa simulação usaremos $k = 4$, utilizando 2000 observações geradas e examinadas nas prevalências 40%, 20%, 10% , 5% e 2,5% de elementos positivos. A estrutura de modelagem é similar a feita nos dados reais.

Nas combinações via votos, na prevalência 40% (Figura (5.1)), pode-se perceber a têndencia no aumento da acurácia e da diminuição do risco relativo, a medida em que k aumenta. Juntamente, quando os valores de k começam a se tornar mais elevados, a especificidade aumenta e a sensibilidade diminui. Isso indica que o modelo esta escolhendo praticamente todos os casos como negativo. Sendo assim, a acurácia torna-se uma medida menos representativa, mesmo quando possui as maiores medidas para os maiores valores de k. O custo relativo tem variação pequena e seu menor valor se dá quando $k = 20$.

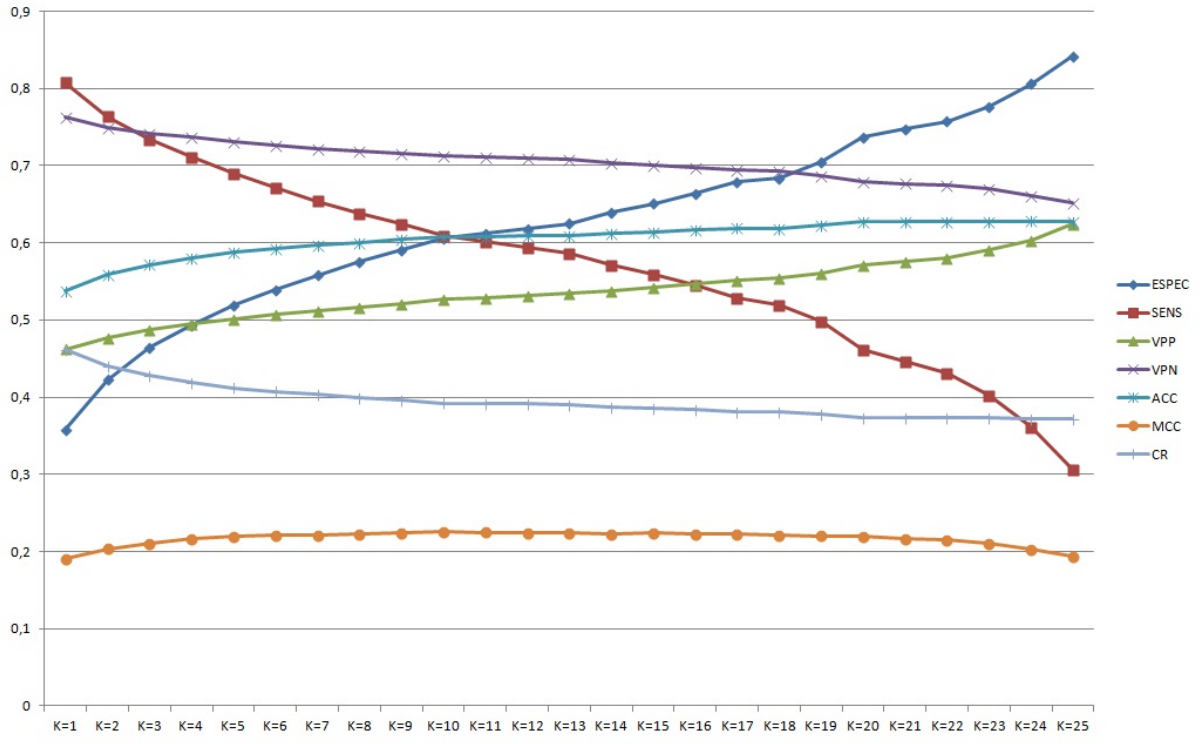

Figura 5.1: *Combinações via votos nos dados Breiman 40%.*

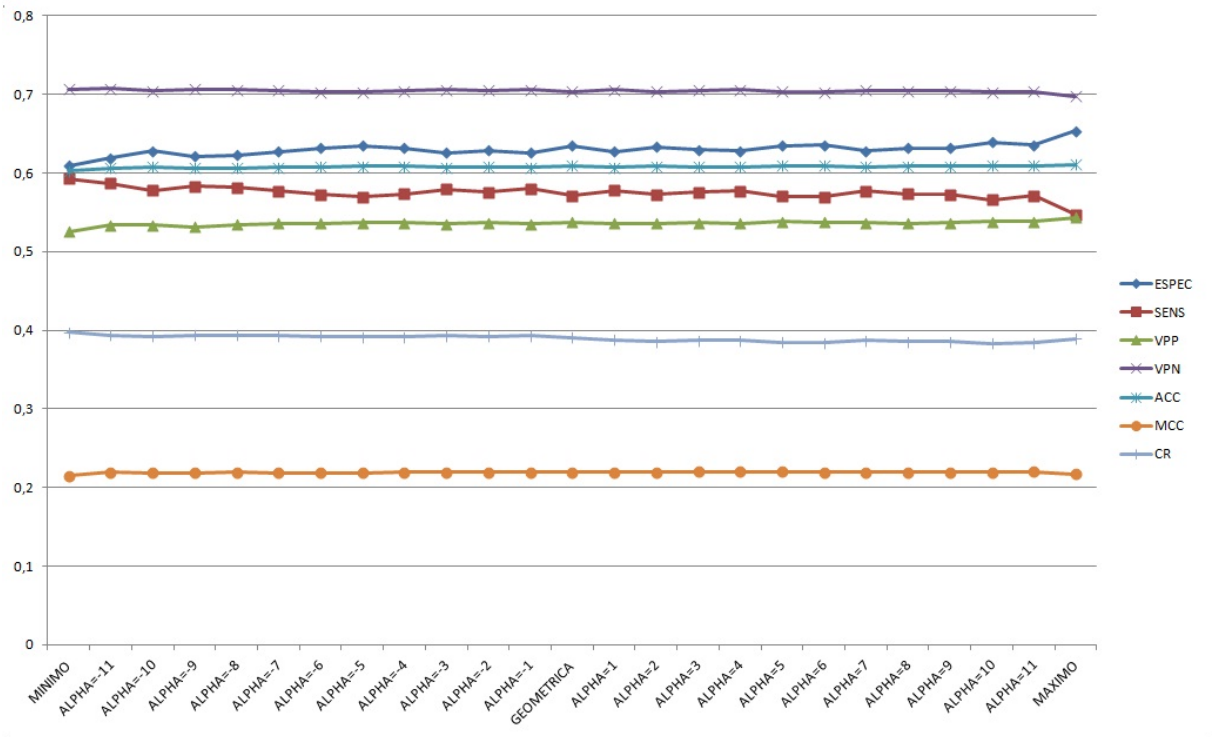

Figura 5.2: *Combinações via médias nos dados Breiman 40%.*

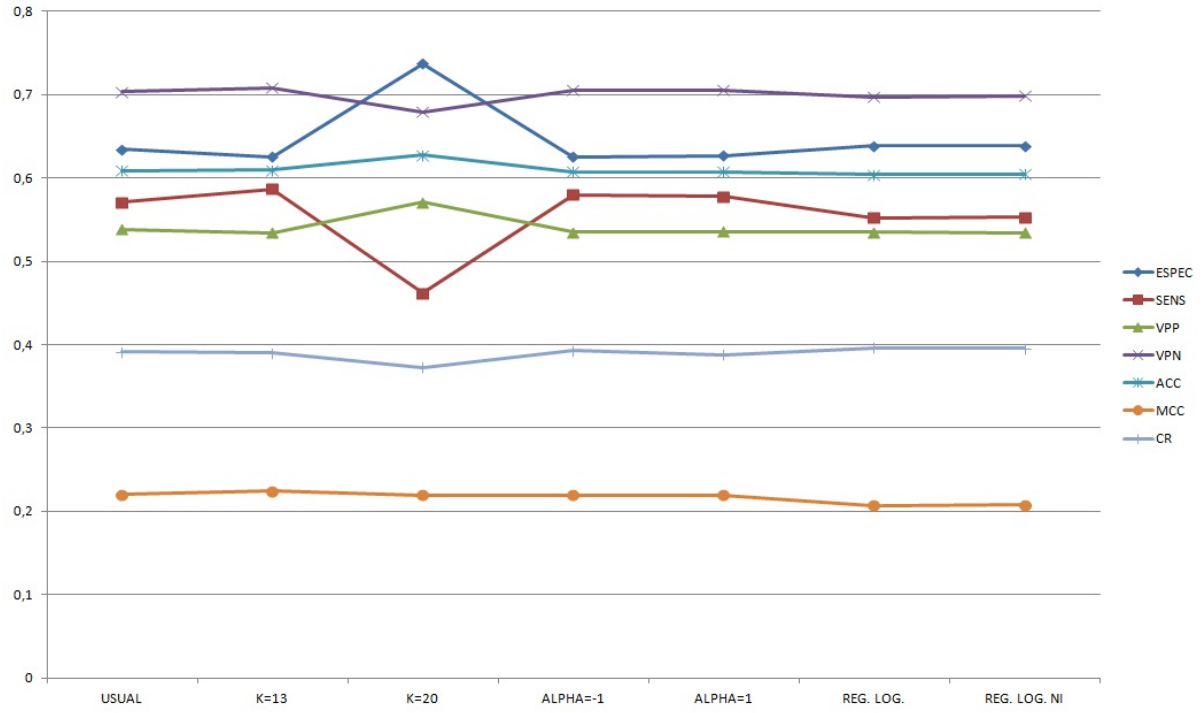

Figura 5.3: *Comparação entre os melhores modelos obtidos para os dados Breiman 40%.*

O coeficiente de correlação começa crescente, se estabiliza, e decresce nos valores finais de k. Isso indica um ajuste global mais adequado, tornando mais equilibrado a relação entre a sensibilidade e a especificidade. A acurácia é pouco menor do máximo atingido.

Na análise das combinações via médias (Figura(5.2)), os resultados se mostraram bastante estáveis, variando muito pouco, de modo que todos os modelos apresentaram valores praticamente iguais.

Na Figura (5.3) estão reunidos o modelo usual, os modelos combinados via votos com $k = 13$ e $k = 20$, os modelos via médias com $\alpha = -1$ e $\alpha = 1$ (média harmônica e aritmética, respectivamente), juntamente com os modelos via regressão logística.

Nesse cenário, os resultados também variaram pouco. A combinação por regressão logística apresentou desempenho um pouco inferior aos demais nas medidas MCC, acurácia e sensibilidade. Os outros modelos foram muito similares, inclusive em relação o modelo usual.

Na prevalência de 20% e 10%, Figuras (5.4), (5.5), (5.6), (5.7), (5.8) e (5.9), as combinações por votos e médias apresentaram resultados muito semelhantes aos anteriores, mudando apenas em valor absoluto, mas seguindo as mesmas tendências. Ao comparar com o modelo usual e aos combinados via regressao logística, os resultados também se mantiveram. Utilizamos os mesmos $k = 13$ e $k = 20$, junto com $\alpha = -1$ e $\alpha = 1$.

Nas prevalências menores, 5% e 2,5%, ao combinar por votos os melhores resultados ficam nos valores intermediários de k, seguindo ainda a tendência observada nas pre-

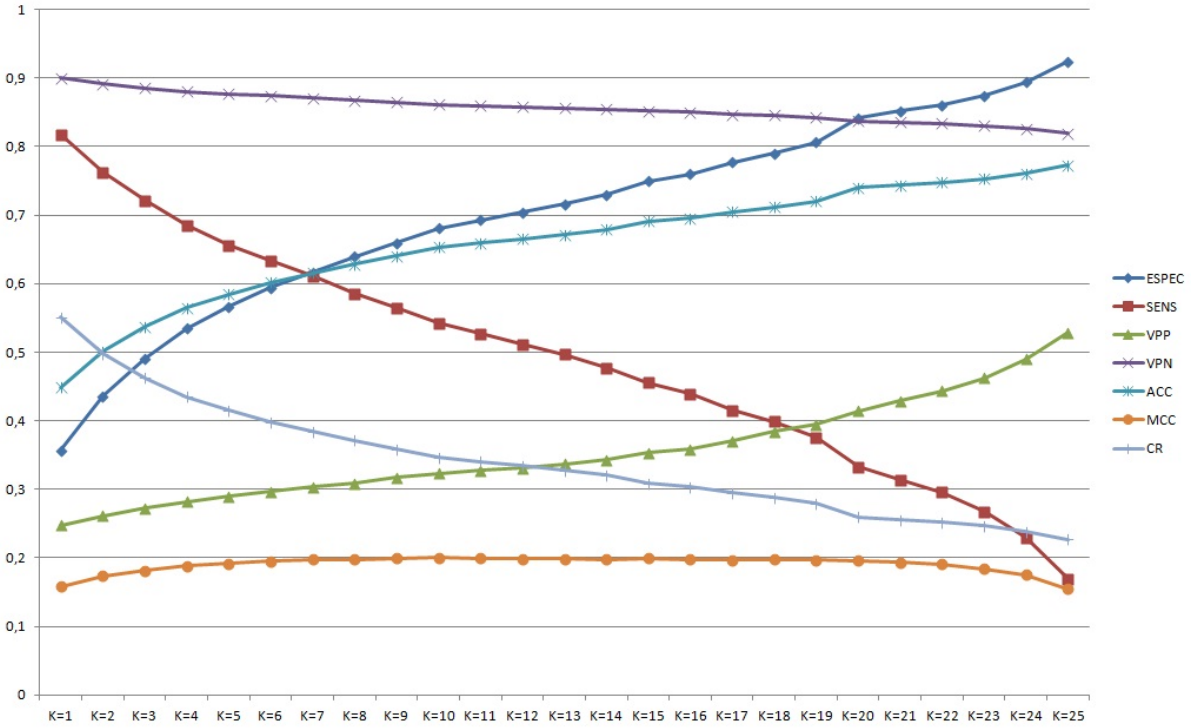

Figura 5.4: *Combinações via votos nos dados Breiman 20%.*

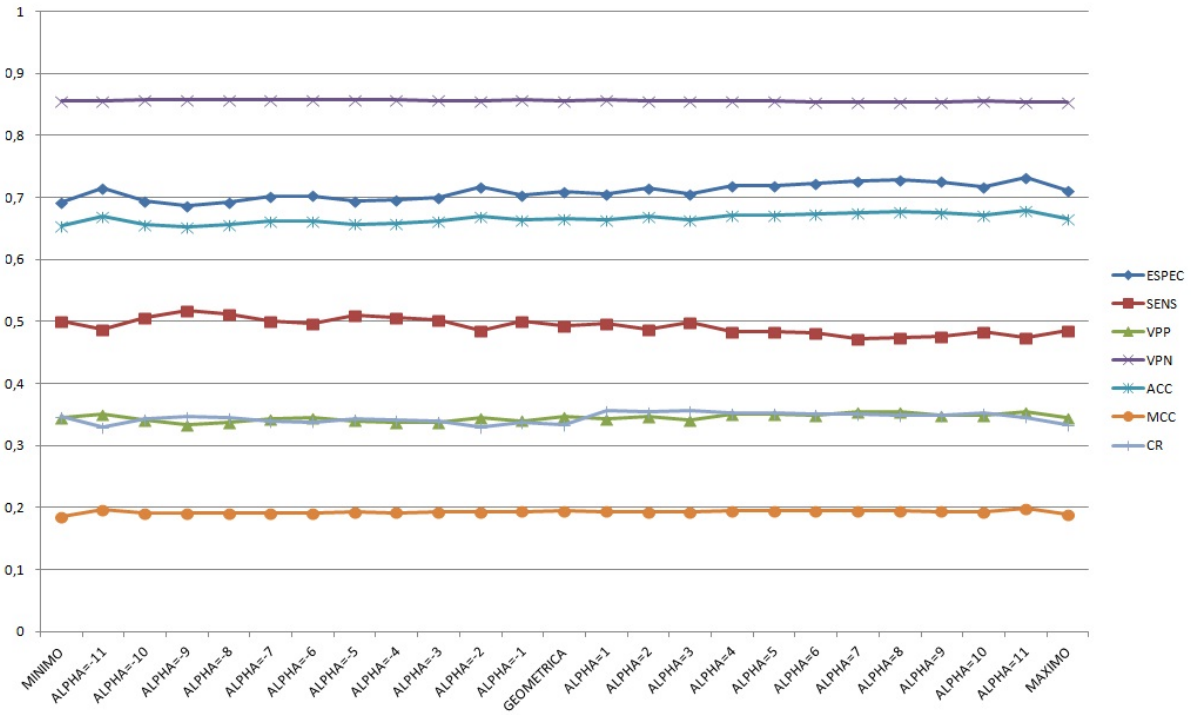

Figura 5.5: *Combinações via médias nos dados Breiman 20%.*

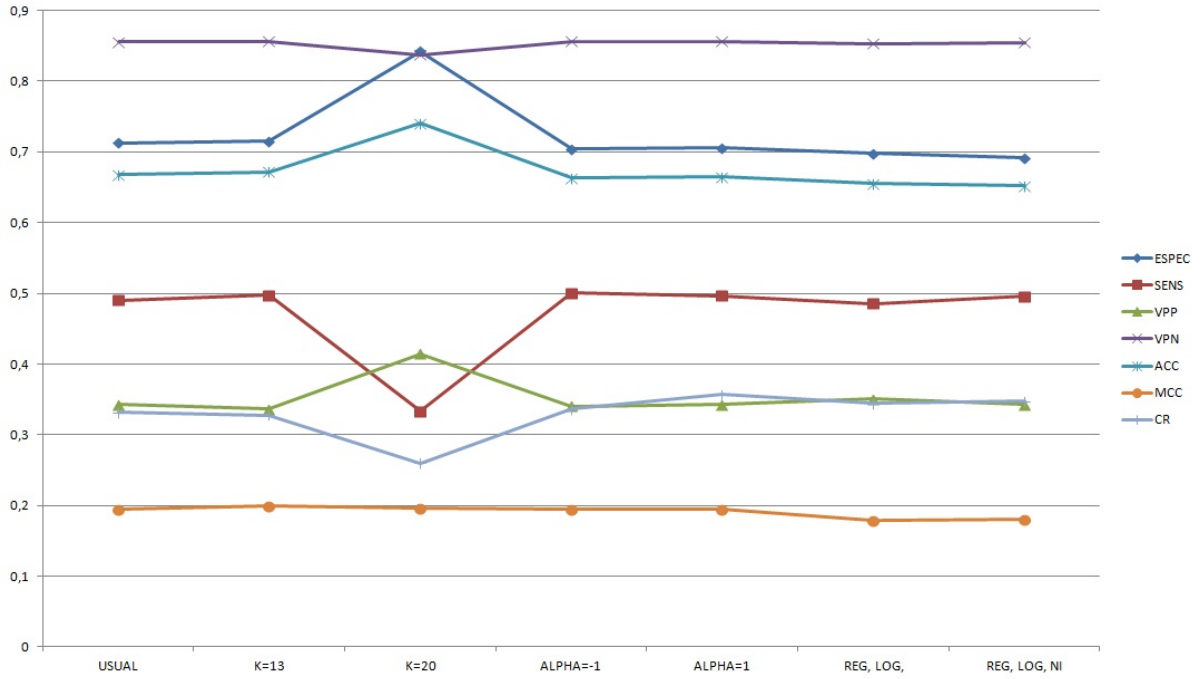

Figura 5.6: *Comparação entre os melhores modelos obtidos para os dados Breiman 20%.*

valências maiores. Na combinação por médias, o resultado muda um pouco. Os valores negativos de α obtém os menores valores do custo relativo e possui sensibilidades maiores, ainda que com pouca diferença. A correlação muda muito pouco e, sendo assim, consideraremos os melhores modelos aqueles com o menor risco relativo, a saber, $\alpha = -7$ e $\alpha \to 0$ (combinação por média geométrica) na prevalência 5% e $\alpha = -1$ e $\alpha \to 0$ na prevalência 2,5%.

Na comparação de todos os modelos, Figura (5.9), a melhor combinação se dá pelo voto majoritário, em termos de correlação e acurácia, ainda que pouco diferentes dos demais.

Nesses dados gerados, o comportamento das combinações mostraram-se similares em relação a prevalência amostral, se diferenciando mais nos casos quando a prevalência é menor. Nas combinações por votos, ao utilizar $k = 13$, isto é, o voto majoritário, temos as melhores medidas no geral, com a melhor correlação. Nas combinações por médias, há poucas diferenças na prevalências maiores, mas nas menores, a combinação via média geométrica mostrou-se capaz de reduzir o risco relativo, mantendo a acurácia e o MCC similar aos demais.

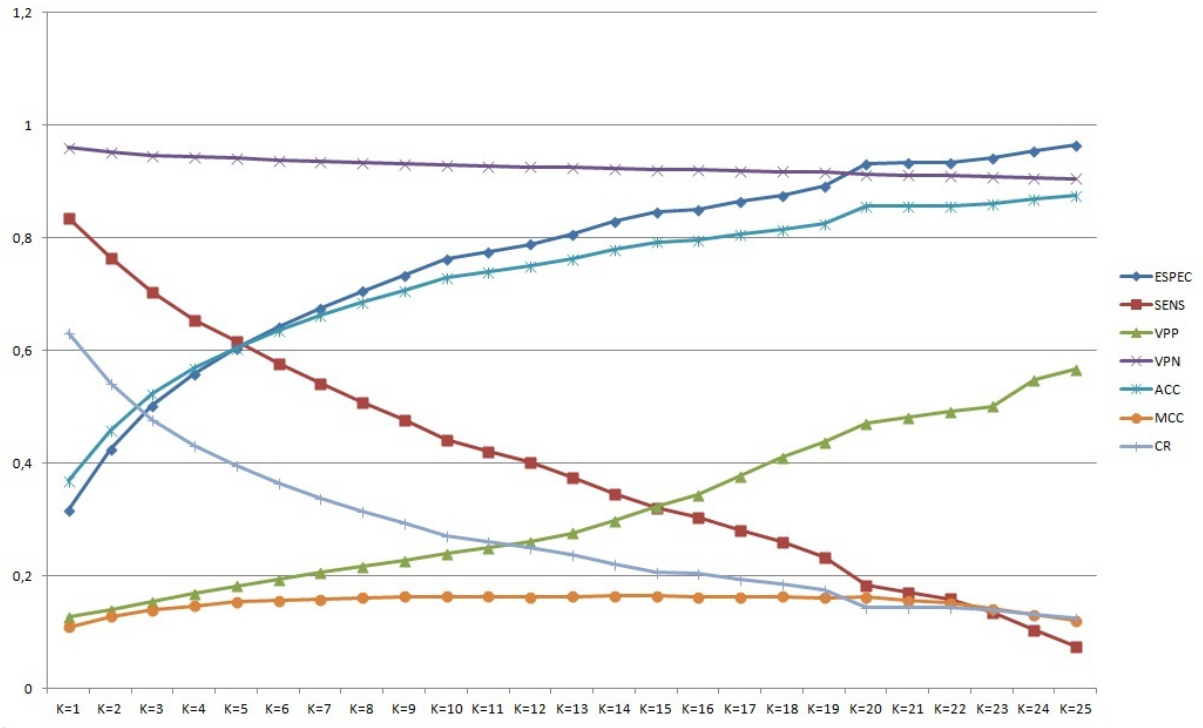

Figura 5.7: *Combinações via votos nos dados Breiman 10%.*

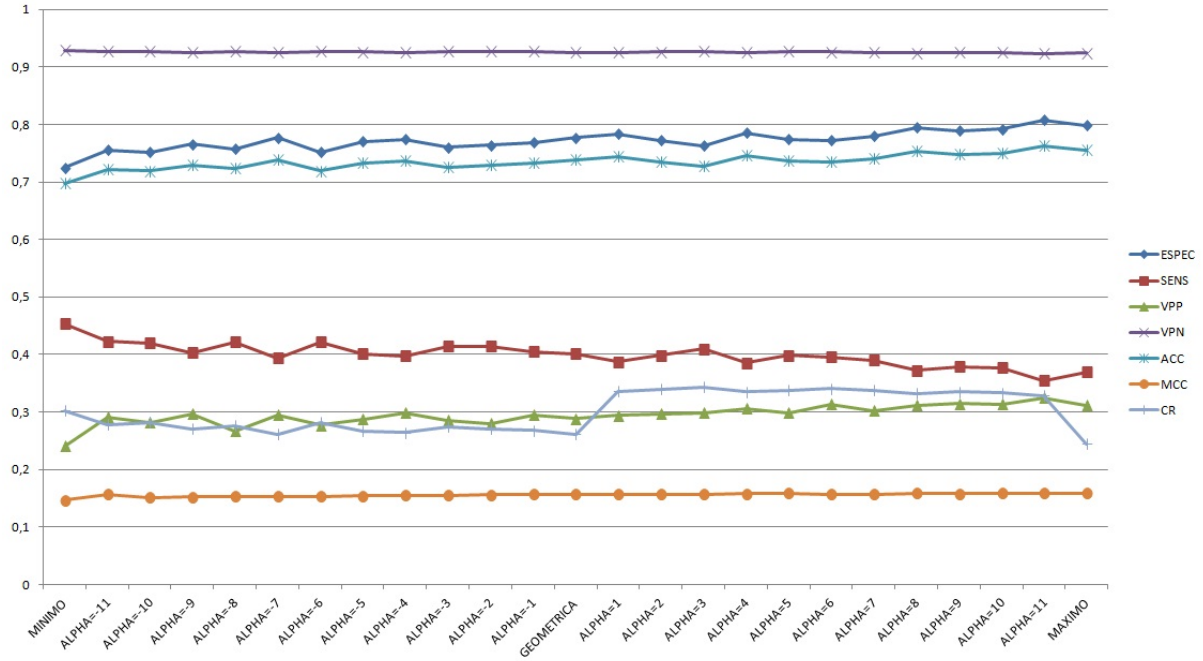

Figura 5.8: *Combinações via médias nos dados Breiman 10%.*

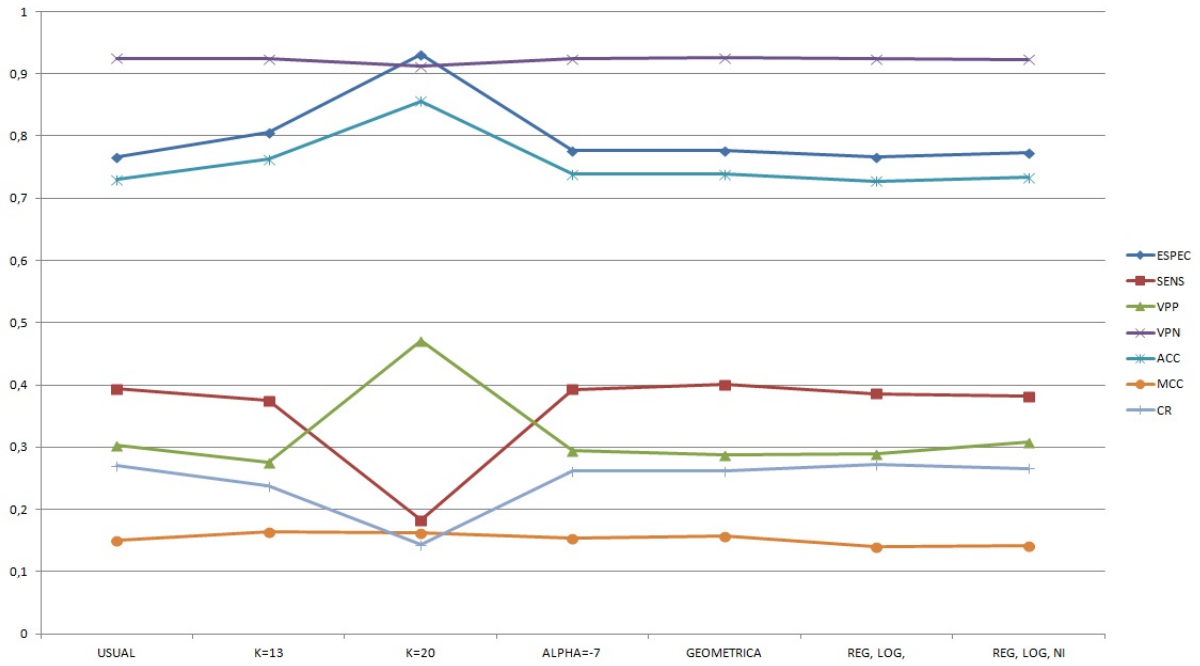

Figura 5.9: *Comparação entre os melhores modelos obtidos para os dados Breiman 10%.*

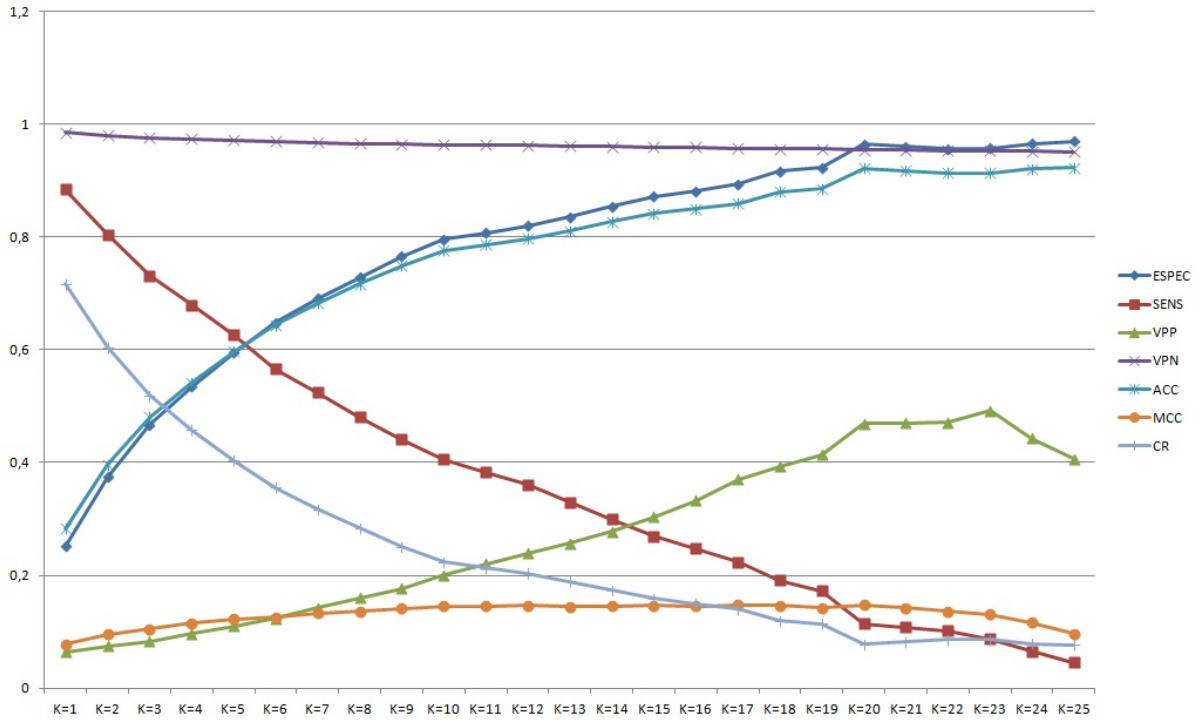

Figura 5.10: *Combinações via votos nos dados Breiman 5%.*

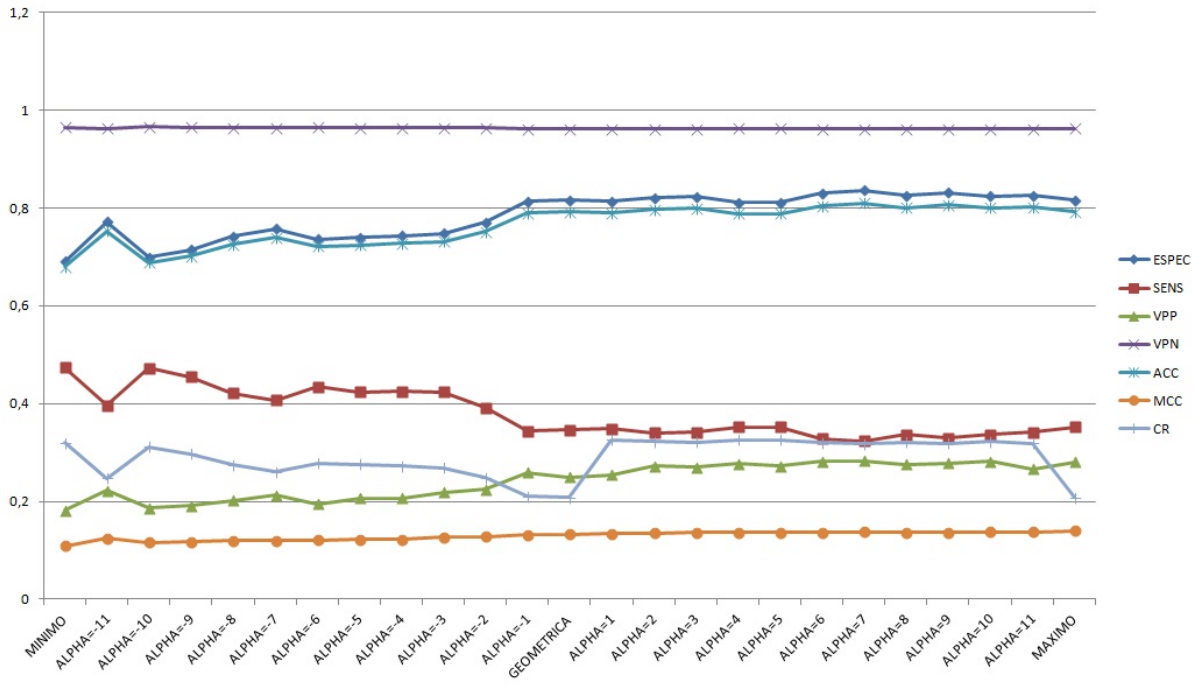

Figura 5.11: *Combinações via médias nos dados Breiman 5%.*

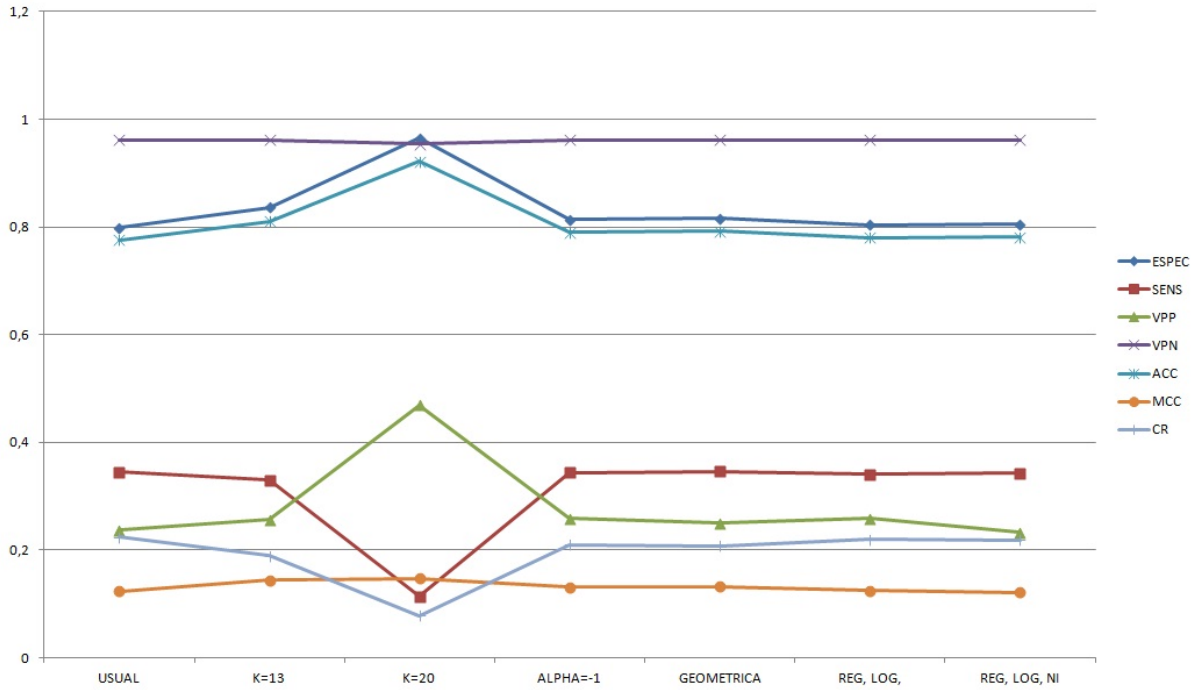

Figura 5.12: *Comparação entre os melhores modelos obtidos para os dados Breiman 5%.*

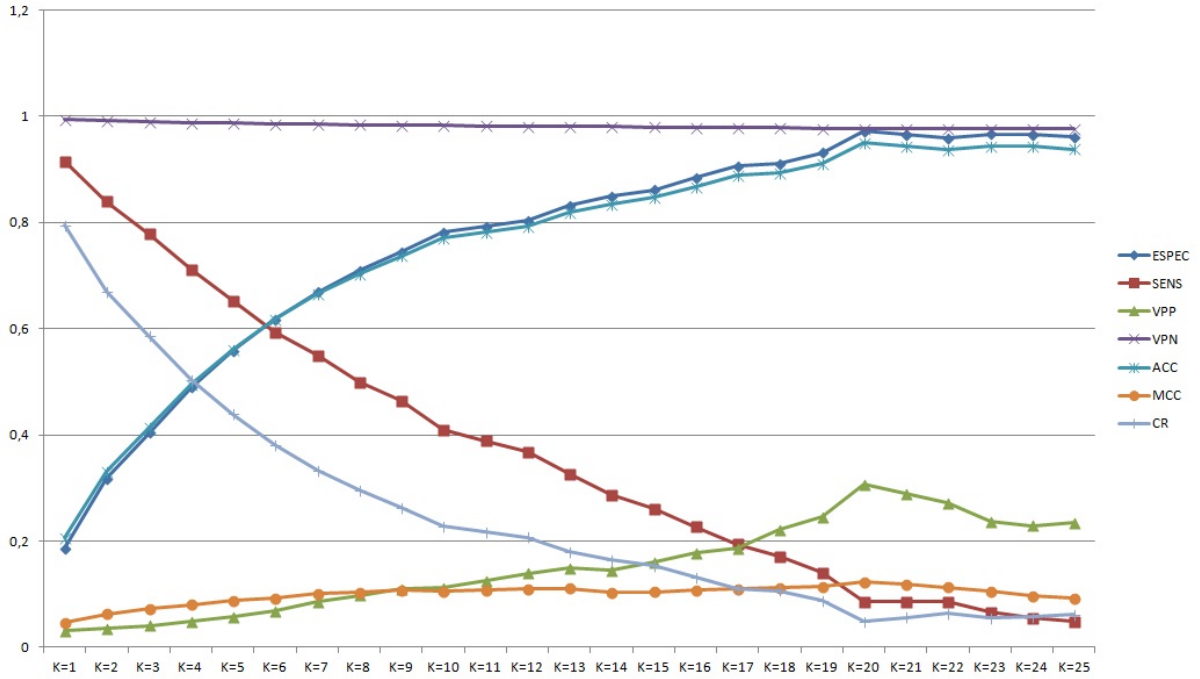

Figura 5.13: *Combinações via votos nos dados Breiman 2,5%.*

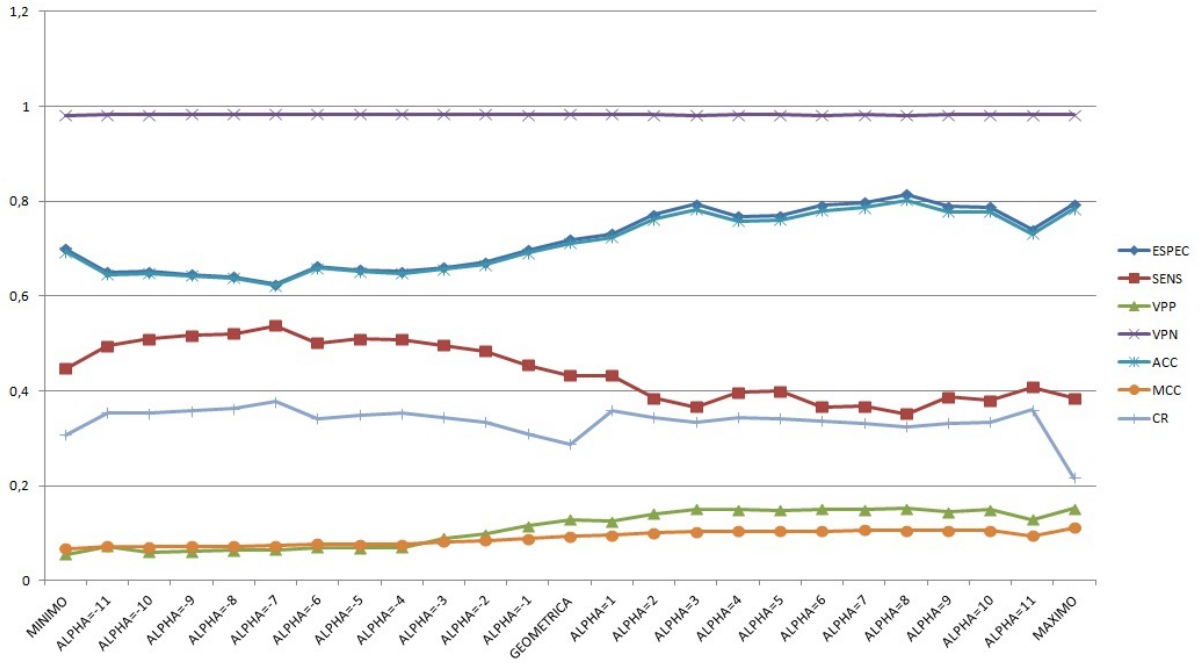

Figura 5.14: *Combinações via médias nos dados Breiman 2,5%.*

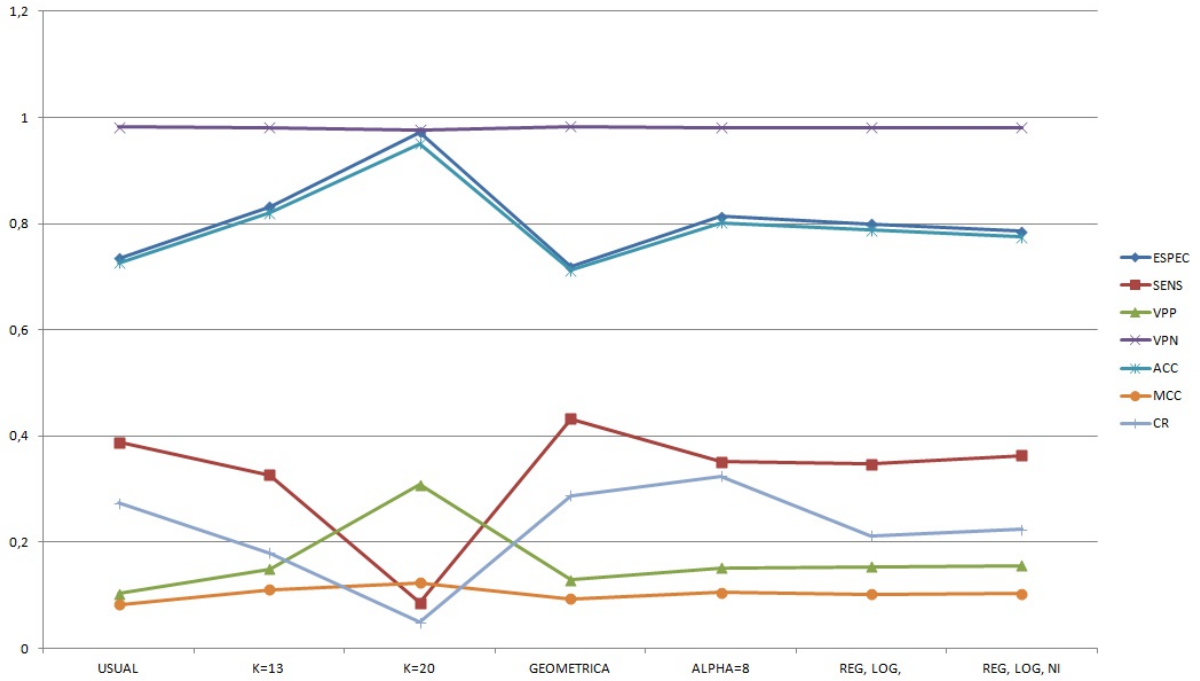

Figura 5.15: *Comparação entre os melhores modelos obtidos para os dados Breiman 2,5%.*

5.3 Aplicação em Dados Reais

No *german credit data* fizemos a análise em relação as combinações via média, votos e regressão logística, e também o modelo usual. A Figura (5.16) resume os resultados obtidos pelas combinações por votos.

A medida em que os valores de k aumentam, o modelo torna-se menos conservador. A sensibilidade e o valor preditivo negativo são maiores quando $k = 1$ é decresce em cada valor de $k > 1$. A situação contrária ocorre na especificidade e no valor preditivo positivo, pois os maiores valores estão associados aos maiores valores de k.

A maior acurácia e menor custo relativo estão em $k = 20$, num modelo com alta especificidade e baixa sensibilidade. O coeficiente de correlação atinge seu pico em $k = 9$ e é inferior ao encontrado na combinação com $k = 20$.

Note que a curva do custo relativo segue descrescente, ao passo que a acurácia é crescente, e tendem a se estabilizar depois de $k = 13$, aproximadamente.

Na Figura (5.17) está representado as combinações via médias, em que obtivemos resultados relativamente mais estáveis. A sensibilidade aumentou junto com α e a especificidade diminiu. As demais medidas ficaram relativamente estáveis, com pouca variação. O menor custo relativo é apontado pela combinação via mínimo, no entanto, possui o menor MCC e sensibilidade.

Nos valores positivos de α encontramos os melhores valores para o MCC, sendo seu máximo em $\alpha = 4$, juntamente com a melhor sensibilidade.

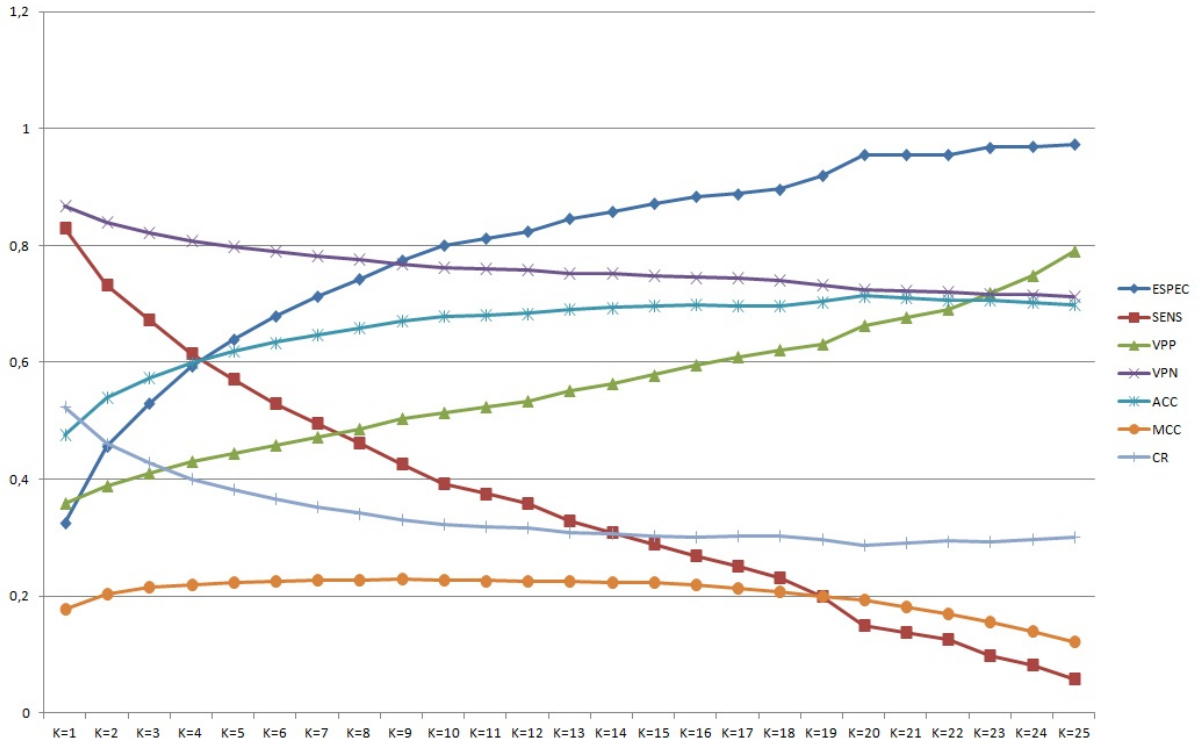

Figura 5.16: *Combinações via voto no german credit data.*

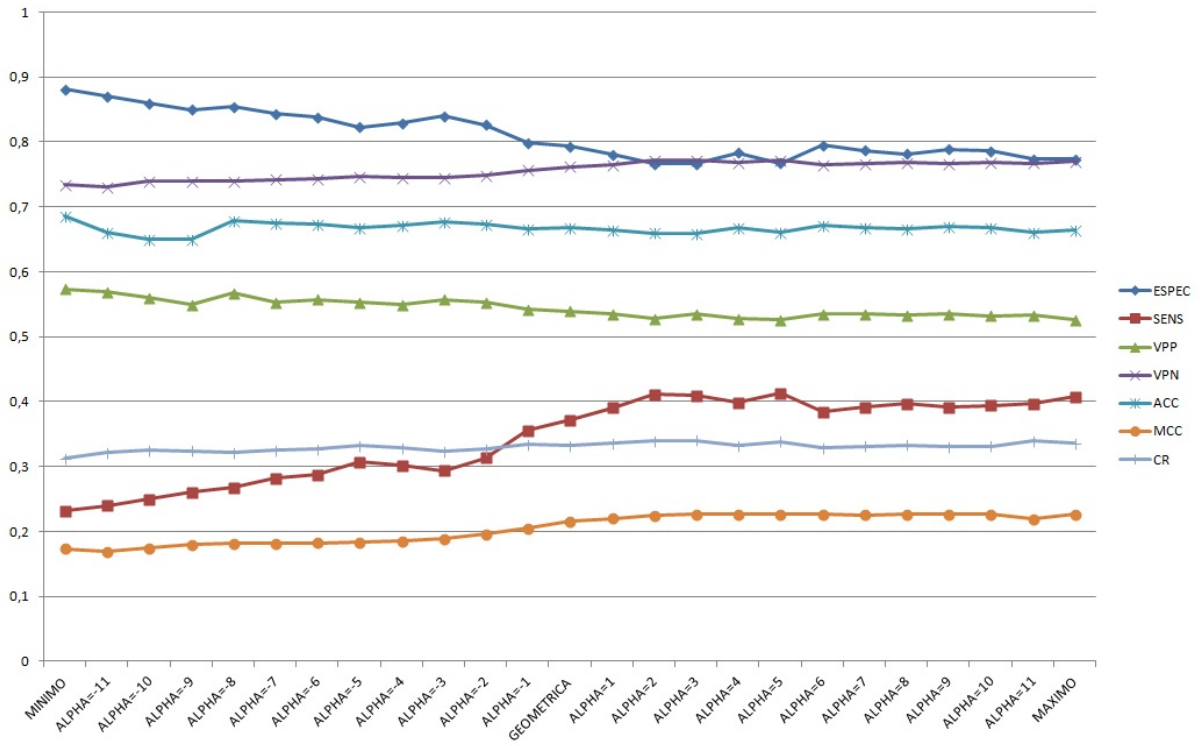

Figura 5.17: *Combinações via médias no german credit data.*

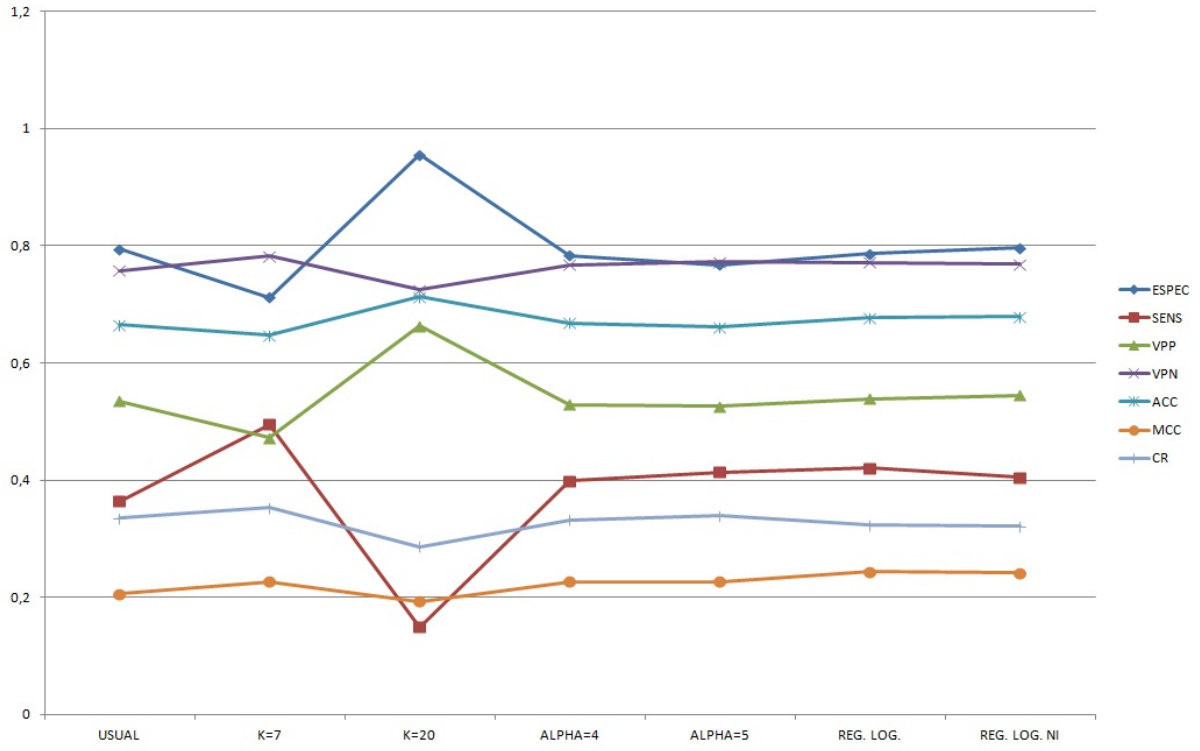

Figura 5.18: *Comparação entre os melhores modelos obtidos para o german credit data.*

Diante desses resultados, tomaremos os dois melhores valores de k e de α e compararemos juntamente com o modelo usual e as combinações via regressão logística. Consideramos os valores obtidos em $k = 7$ e $k = 20$, e $\alpha = 4$ e $\alpha = 5$. A Figura (5.18) ilustra os valores obtidos.

As combinações via regressão logística apresentaram bastantes semelhantes. Claro, a influência do intercepto apenas translada os escores, de forma que não afeta a classificação final pois o que importa realmente é a ordem dos escores. No entanto, o fato de não usar intercepto pode levar a alterações nos outros parâmetros estimados na combinação, o que justifica as pequenas diferenças entre os modelos.

O menor custo relativo está na combinação por voto com $k = 20$, entretanto, simultaneamente apresenta os menores valores de MCC e sensibilidade (menores também que o modelo sem combinação alguma).

As combinações via regressão logística apresentaram os melhores valores para a correlação e o segundo melhor resultado em relação ao custo relativo, acurácia e especificidade.

As combinações via votos no *australian credit data* estão descritas na Figura (5.19). A sensibilidade decai rapidamente com os valores de k, atingindo valores muito baixos nos últimos valores. O mesmo ocorre com a especificidade, que sobre rapidamente até $k = 10$ e depois se estabiliza. O custo relativo apresentou uma curva cujo menor valor ocorre em $k = 5$, assim como a acurácia e MCC, mantendo ainda bons resultados na especificidade

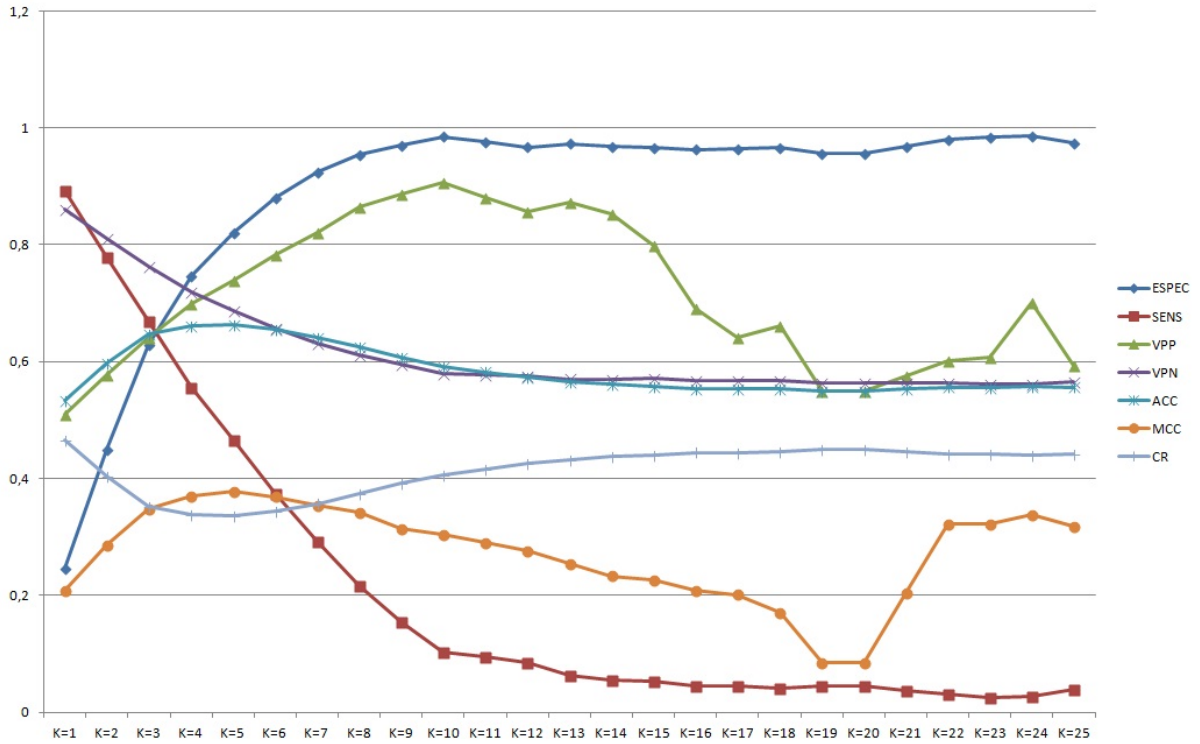

Figura 5.19: *Combinações via votos no australian credit data.*

e sensibilidade.

Nas combinações via médias, os melhores desempenhos foram obtidos em $\alpha = 2$ e $\alpha = 3$, com o maior coeficiente de correlação e acurácia e também o menor custo relativo, sem se desestabilizar nas outras medidas. Nos valores negativos de α os resultados foram em geral inferiores dos obtidos nos valores positivos.

Na Figura (5.21) estão representados as combinações via votos com $k = 4$ e $k = 5$, as combinações via médias com $\alpha = 2$ e $\alpha = 3$, juntamente com o modelo usual e os modelos combinados por regressão logística.

Notemos o desempenho da combinação via regressão logística, que obteve os melhores valores em todas as medidas, exceto a especifidade por uma pequena diferença. A acurácia, coeficiente de correlação e custo relativo obtiveram resultados com diferenças grandes em relação aos demais modelos, indicando assim um bom adequamento neste conjunto de dados.

Na implementação dos dados reais pudemos observar um potencial na utilização de α e k como parâmetros de calibração de uma combinação. Em ambos bancos de dados os melhores modelos foram encontrados utilizando valores diferentes dos que representam as combinações mais comuns, geralmente mais abordados na literatura. Na próxima seção utilizaremos a combinação por regressão logística nos dados reais e combinação por votos nos dados gerados, utilizando $k = 7$, de forma a geral um modelo mais conservador e mais

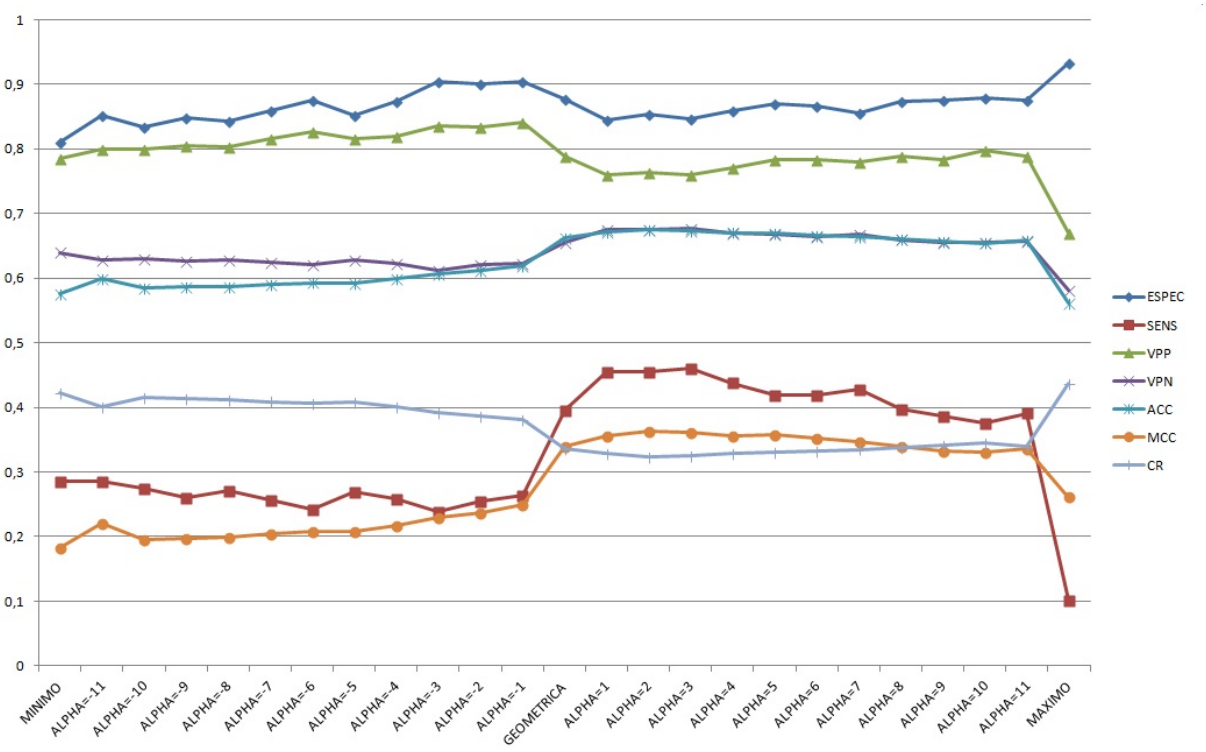

Figura 5.20: *Combinações via médias no australian credit data.*

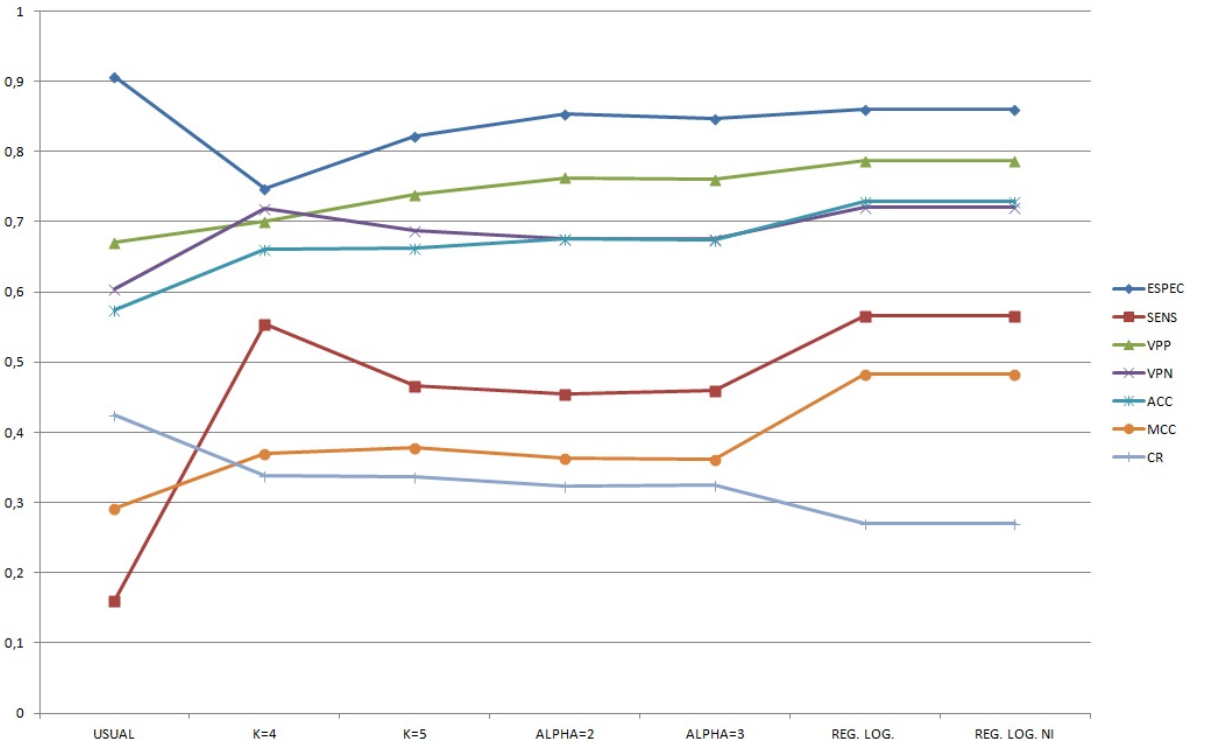

Figura 5.21: *Comparação entre os melhores modelos obtidos para o australian credit data.*

eficaz no aumento da sensibilidade.

Capítulo 6

Estudo de Simulação dos Métodos de Inferência dos Rejeitados

6.1 Especificações da Implementação da Inferência dos Rejeitados

Para simular a situação em que temos rejeitados na amostra foram separados os indivíduos do banco de dados de ajustes que obtiveram escore mais alto segundo um modelo proposto com uma metade aleatória de observações do banco de dados para teste.

Implementar o método da reclassificação é muito simples, em cada indivíduo da população dos rejeitados é inferida a resposta mau pagador e então, com os aceitos mais os rejeitados é construído o modelo de regressão logística e o *bagging*.

Na estratégia da ponderação devemos ter inicialmente um modelo aceita - rejeita, que fornece a probabilidade de inadimplência de todos os proponentes. Para simular esse modelo, usamos um modelo auxiliar, gerado a partir da outra metade da amostra teste. Com esse modelo escoramos cada indivíduo e associamos um peso em cada elemento da população dos aceitos. O peso para o indivíduo i é dado por

$$P_i = \frac{1}{1 - E_i}, \tag{6.1}$$

sendo E_i o seu escore. A ideia é que o peso seja inversamente proporcional ao escore obtido, fazendo com que os indivíduos aceitos mais próximos do ponto de corte obtenham peso maior, representando assim a população dos rejeitados. Usamos $1 - E_i$ porque estamos sempre modelando a probabilidade da ocorrência do evento de interesse.

No método do parcelamento devemos inferir o comportamento dos rejeitados a partir das taxas de inadimplência observadas na população dos aceitos. O procedimento é gerar

um modelo a partir dos aceitos e dividir os proponentes em faixas de escores homogêneas. Consideramos 7 faixas de escore. Escolhemos esse número devido a divisibilidade que é necessária em relação ao tamanho das amostras de treinamento. Em cada faixa é calculada a taxa de inadimplência, verificando quantos são maus pagadores em relação ao total. Essa proporção aumenta na medida em que os escores aumentam, nas faixas mais altas esperam-se altas taxas de inadimplência enquanto que nos escores menores esperamos taxas de inadimplência menores.

Ainda com o modelo dos aceitos, escora-se a população dos rejeitados e distribuem-se nessas mesmas faixas de escore e, então, é atribuído a resposta bom/mau pagador aleatoriamente, na mesma proporção das taxas obtidas nos aceitos. Assim a inferência está completa e o modelo final é gerado com os aceitos acrescidos dos rejeitados.

A análise é feita considerando 10%, 30% e 50% de rejeitados simulados. Cada modelo foi simulado 200 vezes, variando a distribuição da amostra teste no caso dos dados reais e variando a semente da simulação nos dados de Breiman. Nestas análises, considere as tabelas que constam dos valores médios dos resultados.

6.2 Dados Gerados via Breiman

Nessa seção, a implementação foi feita considerando amostras de tamanho 10000, de modo que seja sempre ocorrente os eventos positivos, mesmo nas prevalências menores.

Na prevalência de 40% (Tabela(6.1)), os dados de Breiman ficaram pouco sensíveis em relação aos métodos implementados. Todos os modelos obtiveram resultados muito parecidos em todas as concentrações de rejeitados. O modelo que apresenta um melhor desempenho é o do método da reclassificação.

Os modelos combinados não foram capazes de aumentar o MCC ou a taxa de acerto em relação aos modelos usuais. Os resultados foram apenas pouco maiores, com ganhos apenas nos décimos de centésimos. No entanto, quando há estrutura de combinação ocorre uma troca entre a sensibilidade e a especificidade, fazendo com que os novos modelos obtenham sensibilidade aumentada, em troca da especificidade diminuida. Isso reflete no custo relativo, que na maioria das situações foi diminuido ao usar o *bagging*.

Os resultados para a prevalência 20% (Tabela(6.2)) foram parecidos com os anteriores. A estratégia da ponderação e parcelamento apresentaram resultados muito similares ao usual, enquanto que o método da reclassificação foi o melhor. Em relação ao custo relativo, também foi diminuido no geral. Cada método apresentou melhorias independente da prevalência utilizada. Nos modelos com *bagging*, a estratégia do parcelamento apresentou os melhores resultados para a sensibilidade, sendo a maior obtida em todas concentrações de rejeitados. O mesmo ocorreu com o método da reclassificação em relação a acurácia.

Tabela 6.1: *Inferência dos rejeitados nos dados Breiman com prevalência 40%.*

		Sem Bagging			Com Bagging		
		10%	30%	50%	10%	30%	50%
USUAL	SPEC	0,63626	0,63724	0,63256	0,59945	0,59074	0,57324
	SENS	0,56503	0,56413	0,56693	0,60529	0,61403	0,63081
	VPP	0,51809	0,51691	0,51643	0,50739	0,50529	0,50140
	VPN	0,69257	0,69252	0,69306	0,69955	0,70108	0,70434
	ACC	0,60777	0,60800	0,60631	0,60179	0,60006	0,59627
	MCC	0,20560	0,20511	0,20407	0,20572	0,20546	0,20478
	CR	0,34943	0,36118	0,39734	0,32550	0,33493	0,35687
RECLASS.	SPEC	0,63288	0,61869	0,62383	0,60460	0,59581	0,58874
	SENS	0,57027	0,58577	0,58064	0,60111	0,61037	0,61659
	VPP	0,51781	0,51373	0,51422	0,50983	0,50750	0,50566
	VPN	0,69444	0,69686	0,69554	0,69991	0,70154	0,70242
	ACC	0,60783	0,60552	0,60655	0,60320	0,60163	0,59988
	MCC	0,20730	0,20724	0,20683	0,20758	0,20748	0,20659
	CR	0,32863	0,34000	0,37153	0,32443	0,34140	0,36713
POND.	SPEC	0,63107	0,63840	0,64423	0,60017	0,59398	0,57311
	SENS	0,57075	0,56308	0,55491	0,60480	0,61001	0,63120
	VPP	0,51676	0,51712	0,51881	0,50760	0,50575	0,50127
	VPN	0,69361	0,69215	0,69029	0,69953	0,70025	0,70445
	ACC	0,60694	0,60827	0,60850	0,60202	0,60039	0,59634
	MCC	0,20574	0,20508	0,20369	0,20592	0,20487	0,20489
	CR	0,33313	0,36057	0,38400	0,32507	0,32837	0,34713
PARC.	SPEC	0,63332	0,63495	0,62721	0,59127	0,58098	0,55995
	SENS	0,56748	0,56432	0,56910	0,61206	0,62201	0,63848
	VPP	0,51600	0,51571	0,51141	0,50547	0,50180	0,49529
	VPN	0,69243	0,69117	0,69099	0,70084	0,70152	0,70340
	ACC	0,60699	0,60670	0,60396	0,59958	0,59739	0,59136
	MCC	0,20429	0,20275	0,19910	0,20466	0,20306	0,19845
	CR	0,33760	0,37040	0,39180	0,32473	0,34063	0,36737

Na prevalência 10% (Tabela(6.3)), temos resultados iguais em todos os métodos nas quantidades 10% e 30% de rejeitados. Com 50%, os métodos de inferência dos rejeitados se destacam em relação ao MCC.

A combinação de modelos aponta o modelo da reclassificação, no geral, como melhor do que os demais, apresentando a melhor taxa de acerto e MCC. Entretanto, o método do parcelamento possui a maior sensibilidade quando os rejeitados são 50%.

Nas prevalências de 5% e 2,5% (Tabelas (6.4) e (6.5)), os resultados são equivalentes nas prevalências de rejeitados menores e vão diminuindo a medida que os rejeitados aumentam, mas o método da reclassificação diminui menos, apresentando um desempenho melhor quando são 50% de rejeitados. Nos modelos combinados temos que o modelo com reclassificação tem taxa de acerto melhor nas prevalências de rejeitados menores, entretanto, fica com a maior sensibilidade.

O custo relativo nas prevalências menores obtiveram resultados ainda mais expressivos, com diferenças bastantes significativas. Em todas as análises, os valores para o valor preditivo positivo e negativo foram os que menos variaram, sendo praticamente iguais em

Tabela 6.2: *Inferência dos rejeitados nos dados Breiman com prevalência 20%.*

		Sem Bagging			Com Bagging		
		10%	30%	50%	10%	30%	50%
USUAL	SPEC	0,70370	0,68587	0,69786	0,65413	0,64170	0,61440
	SENS	0,48728	0,50867	0,48382	0,54810	0,56095	0,58340
	VPP	0,30563	0,29926	0,30285	0,29131	0,28820	0,28160
	VPN	0,84850	0,85068	0,84710	0,85482	0,85587	0,85723
	ACC	0,66042	0,65043	0,65505	0,63292	0,62555	0,60820
	MCC	0,16992	0,16988	0,16437	0,17144	0,17050	0,16532
	CR	0,27332	0,27332	0,39734	0,21433	0,21433	0,35687
RECLASS.	SPEC	0,71860	0,71317	0,68984	0,68944	0,69870	0,67121
	SENS	0,47532	0,48153	0,50815	0,50905	0,49943	0,53047
	VPP	0,31231	0,31115	0,30162	0,30461	0,30655	0,29617
	VPN	0,84811	0,84879	0,85122	0,85161	0,85070	0,85336
	ACC	0,66994	0,66684	0,65350	0,65336	0,65885	0,64306
	MCC	0,17471	0,17487	0,17306	0,17477	0,17518	0,17305
	CR	0,21630	0,21630	0,37153	0,20710	0,20710	0,36713
POND.	SPEC	0,70769	0,68931	0,69596	0,65424	0,63953	0,61804
	SENS	0,48262	0,50518	0,48733	0,54788	0,56358	0,58108
	VPP	0,30728	0,29972	0,30274	0,29123	0,28750	0,28266
	VPN	0,84802	0,85020	0,84772	0,85473	0,85617	0,85721
	ACC	0,66267	0,65248	0,65424	0,63297	0,62434	0,61065
	MCC	0,17014	0,16975	0,16560	0,17128	0,17049	0,16651
	CR	0,24433	0,24433	0,38400	0,21307	0,21307	0,34713
PARC.	SPEC	0,68938	0,70845	0,67969	0,64210	0,62724	0,58186
	SENS	0,50373	0,47780	0,49442	0,55870	0,57188	0,60367
	VPP	0,30012	0,30277	0,29057	0,28790	0,28187	0,27207
	VPN	0,84948	0,84648	0,84574	0,85522	0,85601	0,85674
	ACC	0,65225	0,66232	0,64264	0,62542	0,61617	0,58622
	MCC	0,16867	0,16525	0,15319	0,16905	0,16544	0,15431
	CR	0,22297	0,22297	0,39180	0,19787	0,19787	0,36737

todas situações de comparação propostas.

Nesses dados, é notável a melhor adequabilidade da estratégia da reclassificação. Em alguns casos, o método do parcelamento ou ponderação pode apresentar uma qualidade específica melhorada, que pode ser aproveitada de acordo com o problema. A combinação de modelos mostra-se também eficaz, não apenas pelo aumento, ainda que ligeiro, nas medidas, mas na maneira geral de como elas retornam as medidas preditivas como um todo. Como já falado, não só no contexto de *credit scoring* mas em tantos outros, muitas vezes é muito mais importante obter maior sensibilidade do que especificidade. Nesse ponto, os modelos combinados demostraram ser bastantes eficientes.

6.3 Aplicação em Dados Reais

No *australian credit data* podemos observar que as medidas caem, de acordo com o aumento da concentração de rejeitados, o que era de se esperar. O método da reclassificação foi o que melhor se adequou, apresentando um ganho pequeno, mas significativo em

Tabela 6.3: *Inferência dos rejeitados nos dados Breiman com prevalência 10%.*

		Sem Bagging			Com Bagging		
		10%	30%	50%	10%	30%	50%
USUAL	SPEC	0,73066	0,74867	0,69899	0,67345	0,65430	0,59031
	SENS	0,45163	0,42400	0,42453	0,52520	0,54250	0,58660
	VPP	0,18128	0,18736	0,17228	0,16205	0,15679	0,14774
	VPN	0,92533	0,92365	0,92015	0,92936	0,92974	0,92992
	ACC	0,70276	0,71621	0,67154	0,65863	0,64312	0,58994
	MCC	0,13185	0,12961	0,09797	0,13269	0,12939	0,11527
	CR	0,27332	0,26519	0,29453	0,21433	0,22050	0,21539
RECLASS.	SPEC	0,73657	0,70967	0,64088	0,70725	0,68163	0,61753
	SENS	0,45207	0,48040	0,56033	0,48543	0,51447	0,58603
	VPP	0,17944	0,16999	0,15277	0,17332	0,16427	0,14953
	VPN	0,92542	0,92684	0,93049	0,92721	0,92872	0,93197
	ACC	0,70812	0,68675	0,63283	0,68507	0,66492	0,61438
	MCC	0,13560	0,13327	0,12914	0,13455	0,13319	0,12860
	CR	0,21630	0,24190	0,39269	0,20710	0,24923	0,38661
POND.	SPEC	0,73437	0,75426	0,71109	0,67402	0,65707	0,59430
	SENS	0,44793	0,41993	0,41490	0,52440	0,53973	0,58513
	VPP	0,18420	0,18479	0,18327	0,16243	0,15757	0,14821
	VPN	0,92498	0,92330	0,92039	0,92934	0,92968	0,93006
	ACC	0,70573	0,72083	0,68147	0,65906	0,64534	0,59339
	MCC	0,13203	0,12965	0,10207	0,13277	0,12986	0,11671
	CR	0,24433	0,22970	0,27336	0,21307	0,20153	0,23163
PARC.	SPEC	0,74555	0,74737	0,67774	0,67579	0,66995	0,55019
	SENS	0,43467	0,42697	0,46083	0,51947	0,52210	0,61870
	VPP	0,18192	0,19149	0,16090	0,16194	0,15820	0,13699
	VPN	0,92396	0,92334	0,92122	0,92888	0,92831	0,93056
	ACC	0,71446	0,71533	0,65605	0,66015	0,65517	0,55704
	MCC	0,13068	0,12924	0,09649	0,13106	0,12717	0,10634
	CR	0,22297	0,23350	0,28749	0,19787	0,21193	0,27221

relação ao modelo usual. Foi também o modelo com a maior sensibilidade e que menos foi afetado pelo aumento dos rejeitados.

O uso do *bagging* junto com a combinação via regressão logística (com intercepto) apresentou excelentes resultados. Em todos os casos obteve-se um aumento expressivo nas medidas, mostrando o melhor resultado ainda no método da reclassificação. Vale notar que, nos modelos com *bagging*, enquanto que no modelo usual a sensibilidade diminui a medida que os rejeitados aumentam, no método da reclassificação acontece o contrário, o que é uma propriedade muito interessante.

Em relação ao custo relativo, obteve-se melhorias na maioria ds situações. E pode-se notar também uma tendência de maior diminuição no caso em que as prevalências são maiores, como no caso do método da reclassificação.

Os métodos do parcelamento e da reclassificação não mostraram uma boa adequação nesse conjunto de dados, com resultado inferior aos usuais, salvo em relação ao custo relativo.

No *german credit data*, prevalência 10%, obtivemos resultados semelhantes nos méto-

Tabela 6.4: *Inferência dos rejeitados nos dados Breiman com prevalência 5%.*

		Sem Bagging			Com Bagging		
		10%	30%	50%	10%	30%	50%
USUAL	SPEC	0,77366	0,76544	0,79206	0,69140	0,66627	0,67886
	SENS	0,39459	0,36610	0,29681	0,50473	0,51287	0,47120
	VPP	0,13935	0,15290	0,15654	0,09200	0,09284	0,10467
	VPN	0,96187	0,95962	0,95705	0,96497	0,96491	0,96268
	ACC	0,75471	0,74547	0,76729	0,68206	0,65860	0,66848
	MCC	0,10248	0,08457	0,07111	0,10138	0,09510	0,08532
	CR	0,23298	0,31697	0,22915	0,17931	0,18221	0,18039
RECLASS.	SPEC	0,79292	0,73941	0,60205	0,76414	0,70092	0,58520
	SENS	0,37187	0,44460	0,59260	0,41047	0,49073	0,61220
	VPP	0,13712	0,09053	0,07705	0,11522	0,08643	0,07493
	VPN	0,96100	0,96280	0,96642	0,96206	0,96399	0,96695
	ACC	0,77187	0,72467	0,60157	0,74646	0,69041	0,58655
	MCC	0,10238	0,09748	0,09119	0,10210	0,09728	0,09045
	CR	0,28952	0,30827	0,35936	0,25171	0,32376	0,36912
POND.	SPEC	0,77489	0,77727	0,78584	0,69024	0,66757	0,66629
	SENS	0,38980	0,35400	0,30840	0,50500	0,51340	0,48660
	VPP	0,14259	0,15324	0,15934	0,09269	0,09232	0,10415
	VPN	0,96174	0,95954	0,95749	0,96496	0,96486	0,96318
	ACC	0,75564	0,75610	0,76196	0,68098	0,65986	0,65731
	MCC	0,10159	0,08518	0,07307	0,10077	0,09562	0,08584
	CR	0,24483	0,27789	0,23888	0,19464	0,18875	0,18869
PARC.	SPEC	0,74359	0,77369	0,77594	0,70519	0,67887	0,65451
	SENS	0,42913	0,36293	0,31913	0,48587	0,49780	0,48747
	VPP	0,13689	0,19933	0,15242	0,10340	0,10007	0,08722
	VPN	0,96292	0,96050	0,95779	0,96452	0,96406	0,96250
	ACC	0,72786	0,75315	0,75310	0,69423	0,66982	0,64616
	MCC	0,10149	0,09278	0,06958	0,10162	0,09306	0,07439
	CR	0,25816	0,24917	0,23435	0,20344	0,19381	0,16214

dos da reclassificação, usual e ponderação, sendo o primeiro com MCC pouco maior do que os demais, ao passo que, o método da ponderação obteve a maior acurácia. A medida que os rejeitados aumentam, o modelo usual torna-se o com a maior taxa de acerto, junto ainda com o método da ponderação, mas o maior MCC fica ainda com o método da reclassicação, que tem a maior sensibilidade também. O custo relativo não apresentou um desempenho satisfatório, pois em vários casos ele é aumentado com o uso do *bagging* e nas outras não diminui muito. No geral, o método da ponderação obteve os menores custos relativos.

O uso de combinação de modelos foi mais eficaz nas prevalências menores de rejeitados. Em todos os métodos é possível verificar um aumento significante. Novamente, o método da reclassificação foi o que obteve os melhores resultados e manteve a propriedade observada de, nos modelos com *bagging*, aumentar a sensibilidade junto com a quantidade de rejeitados (o que não ocorre nos demais métodos nem nos modelos sem *bagging*).

Os modelos, em geral, apresentaram bons resultados em relação ao aumento das medidas preditivas quando trabalhando com a combinação de modelos. Percebe-se que seu uso,

Tabela 6.5: *Inferência dos rejeitados nos dados Breiman com prevalência 2,5%.*

		Sem Bagging			Com Bagging		
		10%	30%	50%	10%	30%	50%
USUAL	SPEC	0,80665	0,77664	0,80892	0,74821	0,71052	0,71111
	SENS	0,30750	0,32089	0,26895	0,40627	0,43627	0,42187
	VPP	0,17631	0,09700	0,08506	0,12505	0,07499	0,06920
	VPN	0,97986	0,97942	0,97823	0,98131	0,98123	0,98089
	ACC	0,79418	0,76524	0,79542	0,73966	0,70366	0,70388
	MCC	0,07350	0,05985	0,05252	0,07927	0,06971	0,06216
	CR	0,12694	0,12694	0,16959	0,09983	0,09983	0,11837
RECLASS.	SPEC	0,82220	0,74805	0,62025	0,77511	0,69971	0,57501
	SENS	0,32760	0,43267	0,55987	0,38973	0,49653	0,61440
	VPP	0,09442	0,04661	0,04097	0,07917	0,04553	0,03983
	VPN	0,98017	0,98160	0,98330	0,98093	0,98266	0,98403
	ACC	0,80983	0,74017	0,61874	0,76548	0,69463	0,57599
	MCC	0,08023	0,07449	0,06741	0,07924	0,07214	0,06456
	CR	0,23140	0,23140	0,29603	0,23150	0,23150	0,30093
POND.	SPEC	0,80752	0,76585	0,79437	0,70529	0,68788	0,69967
	SENS	0,30253	0,33520	0,28493	0,44600	0,45813	0,42773
	VPP	0,17309	0,09744	0,09769	0,10316	0,07036	0,07013
	VPN	0,97972	0,97972	0,97867	0,98173	0,98170	0,98098
	ACC	0,79489	0,75508	0,78163	0,69881	0,68213	0,69287
	MCC	0,07231	0,06106	0,05433	0,07419	0,06592	0,06006
	CR	0,22880	0,22880	0,27167	0,09070	0,09070	0,11470
PARC.	SPEC	0,79025	0,83434	0,80454	0,71511	0,73827	0,71460
	SENS	0,33093	0,25213	0,27453	0,43360	0,39653	0,39893
	VPP	0,16614	0,13405	0,10278	0,11125	0,08210	0,07539
	VPN	0,97988	0,97844	0,97851	0,98149	0,98071	0,98033
	ACC	0,77876	0,81978	0,79129	0,70808	0,72973	0,70671
	MCC	0,07514	0,06580	0,05449	0,07591	0,07029	0,05708
	CR	0,13460	0,13460	0,10693	0,09940	0,09940	0,06257

concomitante com as técnicas de inferência dos rejeitados, pode trazer melhores benefícios na modelagem de *credit scoring*.

Tabela 6.6: *Inferência dos rejeitados no australian credit data.*

		Sem Bagging			Com Bagging		
		10%	30%	50%	10%	30%	50%
USUAL	SPEC	0,81577	0,83640	0,93270	0,90910	0,91973	0,95685
	SENS	0,38247	0,34607	0,18663	0,49944	0,40461	0,21449
	VPP	0,78290	0,80213	0,86486	0,82863	0,81716	0,81951
	VPN	0,67840	0,66734	0,61080	0,71249	0,67469	0,61265
	ACC	0,62295	0,61820	0,60070	0,72680	0,69050	0,62650
	MCC	0,31492	0,29732	0,24889	0,54816	0,49113	0,41194
	CR	0,40297	0,41317	0,42637	0,39377	0,40017	0,40197
RECLASS.	SPEC	0,80279	0,82423	0,81820	0,91865	0,90198	0,88315
	SENS	0,42888	0,38146	0,33517	0,49697	0,52438	0,53966
	VPP	0,77510	0,78079	0,76942	0,84437	0,82208	0,80088
	VPN	0,68922	0,66869	0,66220	0,71359	0,72027	0,72296
	ACC	0,63640	0,62720	0,60325	0,73100	0,73395	0,73030
	MCC	0,33108	0,32485	0,28626	0,56103	0,54796	0,52834
	CR	0,39480	0,39617	0,40257	0,39120	0,39293	0,39963
POND.	SPEC	0,93117	0,93523	0,94360	0,92892	0,92486	0,93225
	SENS	0,13090	0,14416	0,12112	0,16045	0,16404	0,14292
	VPP	0,84548	0,83496	0,86095	0,81497	0,79751	0,82764
	VPN	0,58867	0,59705	0,58899	0,60442	0,60280	0,59642
	ACC	0,57505	0,58320	0,57760	0,58695	0,58630	0,58100
	MCC	0,21532	0,23893	0,22877	0,25146	0,23340	0,22158
	CR	0,41397	0,39660	0,41310	0,39690	0,39610	0,40093
PARC.	SPEC	0,82414	0,87541	0,87757	0,89568	0,86162	0,89243
	SENS	0,30371	0,22180	0,21011	0,52539	0,54831	0,38820
	VPP	0,74920	0,74826	0,66729	0,81885	0,77650	0,77370
	VPN	0,65100	0,62264	0,60909	0,72074	0,71256	0,65735
	ACC	0,59255	0,58455	0,58055	0,73090	0,72220	0,66805
	MCC	0,24761	0,26282	0,21562	0,53848	0,46546	0,35887
	CR	0,41027	0,41890	0,42543	0,39457	0,40907	0,41797

Tabela 6.7: *Inferência dos rejeitados no german credit data.*

		Sem Bagging			Com Bagging		
		10%	30%	50%	10%	30%	50%
USUAL	SPEC	0,76371	0,84486	0,89352	0,76600	0,81257	0,83819
	SENS	0,39300	0,28011	0,18656	0,40811	0,32367	0,28133
	VPP	0,51568	0,57179	0,59152	0,52274	0,55837	0,58415
	VPN	0,76446	0,74514	0,72698	0,76423	0,74733	0,73959
	ACC	0,65250	0,67543	0,68143	0,65863	0,66590	0,67113
	MCC	0,20374	0,19771	0,18054	0,21096	0,18750	0,18189
	CR	0,33000	0,37270	0,39910	0,26731	0,31740	0,38280
RECLASS.	SPEC	0,71762	0,73714	0,66505	0,76700	0,64457	0,64381
	SENS	0,45767	0,43722	0,47122	0,44878	0,57156	0,56033
	VPP	0,50502	0,50442	0,43902	0,51357	0,46591	0,45966
	VPN	0,77615	0,77430	0,78279	0,77422	0,79998	0,79852
	ACC	0,63963	0,64717	0,60690	0,67153	0,62267	0,61877
	MCC	0,21600	0,22211	0,19386	0,24223	0,23349	0,22457
	CR	0,29077	0,36420	0,39980	0,21231	0,24920	0,26620
POND.	SPEC	0,78681	0,84310	0,88362	0,81333	0,86248	0,88190
	SENS	0,36522	0,27944	0,19611	0,34756	0,29711	0,22822
	VPP	0,52650	0,56567	0,59888	0,55878	0,60485	0,60349
	VPN	0,75836	0,74274	0,73139	0,75367	0,74847	0,73450
	ACC	0,66033	0,67400	0,67737	0,67360	0,69287	0,68580
	MCC	0,19947	0,18609	0,17576	0,20676	0,22109	0,18185
	CR	0,44000	0,42090	0,42990	0,39192	0,41350	0,41140
PARC.	SPEC	0,75219	0,79490	0,86848	0,77043	0,80252	0,82557
	SENS	0,36256	0,29944	0,20733	0,38500	0,32433	0,25467
	VPP	0,49688	0,51014	0,59824	0,52506	0,57702	0,60598
	VPN	0,75472	0,74218	0,73017	0,76081	0,74978	0,73634
	ACC	0,63530	0,64627	0,67013	0,65480	0,65907	0,65430
	MCC	0,17135	0,16359	0,17002	0,19480	0,17944	0,14649
	CR	0,33538	0,41120	0,41470	0,22462	0,27180	0,33540

Capítulo 7

Consideraçoes Finais

7.1 Conclusões

Inicialmente propomos o estudo acerca dos métodos de combinação, analisando as diversas medidas preditivas alcançadas no modelo logístico usual em relação aos modelos com os diversos tipos de combinações abordadas. Implementamos a combinação gerada por médias, analisando em diversos pontos relevantes. Também foram analisadas as combinações por votos, que usam a classificação da escoragem na combinação, de forma a gerar escores combinados de acordo com a quantidade de votos que deseja-se para determinar um classificador final. E também analisamos um método de combinação ainda pouco estudado, via regressão logística, que utiliza os coeficientes da regressão como parâmetros na combinação dos escores, nas versões com intercepto e sem intercepto.

Em relação aos nossos resultados, nos dados reais obtivemos um bom aumento no desempenho dos modelos cuja combinação foi feita por regressão logística. Essa combinação obteve os melhores resultados para a acurácia, MCC e custo relativo, se destacando mais no *australian credit data*.

Nessas simulações pode-se notar ainda que os melhores resultados para a combinação via médias não se dá nos casos particulares (combinação é via média, média harmônica, média geométrica, máximo e mínimo). No *german credit data* encontramos a melhor combinação por média utilizando $\alpha = 2$ e no *australian credit data* usando $\alpha = 4$. No caso da combinação por votos, os melhores resultados se dá ao utilizar $k = 7$ e $k = 5$, respectivamente, que ficou bem distante da combinação usual majoritária.

Nos dados gerados segundo Breiman os métodos de combinações apresentam resultados com ganhos pequenos por alguns, e percas pequenas por outros. O uso do voto majoritário foi o que trouxe as maiores taxas de acerto e MCC em todas as prevalências. Utilizando $k = 20$ também trouxe bons resultados, geralmente com custo relativo menor, mas com baixa sensibilidade e alta especifidade. A combinação por regressão logística apenas se

deu melhor que no modelo usual na prevalência de 2,5%, que representa a modelagem de algum evento raro e ainda não foi melhor que a estratégia por votos. Nas combinações por médias os resultados foram muito semelhantes nas maiores prevalências. Nas prevalências baixas, valores de α negativos foram capazes de diminuir o custo relativo e aumentar a sensibilidade.

Em uma segunda parte, analisamos os métodos de inferência dos rejeitados, que por si só não são tão eficientes no aumento taxa de acerto total e MCC dos modelos. Para tanto propomos o uso conjunto com combinação de modelos. Consideramos três dos métodos mais presentes na literatura e construímos seus modelos usuais e com *bagging*, usando a estratégia de combinações conforme os resultados da simulação anterior. Consideramos também o impacto da prevalência da população dos rejeitados, levando em conta os casos com 10%, 30% e 50%.

As técnicas de inferência dos rejeitados apresentaram todas resultados bem simulares nos dados de Breiman. A estratégia que se ressaltou foi a da reclassificação, com ganhos pequenos, mas em quase todas as situações. Os métodos do parcelamento e ponderação apresentaram bons resultados apenas em alguns casos isolados. No geral, as combinações foram eficientes para melhorar o desempenho nos modelos com inferência dos rejeitados, assim como no usual, de tal forma que incorpora as características obtidas pelo uso da inferência dos rejeitados.

Em praticamente todas as prevalências de rejeitados, o método da reclassificação apresentou um bom desempenho, tanto nos modelos usuais quanto nos com *bagging*. Esse método foi capaz de manter o MCC e taxa de acerto equivalentes ao modelo usual, mas aumentando a sensibilidade do modelo e com custo relativo bastante reduzido. Essa característica é fundamental e deve ser notada, pois quando se trata de *credit scoring*, é mais interessante ter sensibilidade alta do que especificidade alta. Em alguns casos, o uso do *bagging* ainda foi capaz de potencializar essa característica.

Embora as medidas não tenham se diferenciado muito nos dados gerados, o uso do *bagging* foi capaz de reduzir significantemente o custo relativo na maioria dos cenários.

Em síntese, de acordo com nossos resultados podemos dizer que a melhor implementação para um modelo de *credit scoring* seria o uso da reclassificação juntamente com a estrutura de *bagging*.

Em relação as combinações, a ideia de variar os valores de k e α como parâmetros de calibração da combinação é bastante eficaz e podem trazer melhorias em relações as combinações usuais. Sua utilização pode gerar uma combinação mais adequada, com a possibilidade de se ajustar ao contexto de interesse. Foi possível verificar que no contexto da modelagem de crédito, os melhores modelos encontrados não foram provenientes de nenhum caso particular de combinação, mostrando assim sua eficiência como parâmetro

de calibração.

7.2 Propostas Futuras

Nesse trabalho, podemos notar diversos ramos que ainda podem ser analisados e usados na continuaçao do trabalho. O estudo da adequabilidade das combinações pode extender-se para outros algoritmos de modelagem, como redes neurais, que é também bastante usado no contexto de *credit scoring*.

Ainda nas combinações, a utilização dessas técnicas em outros problemas também é de interessante análise, em modelagens de medicina e biologia, por exemplo.

Na área de inferência de rejeitados, uma nova simulação com uma geração de dados diferentes também é interessante, concomitante com o uso de outras técnicas de modelagem.

Referências Bibliográficas

1. Alves, M. C., 2008. Estratégias para o desenvolvimento de modelos de credit score com inferência dos rejeitados - Dissertação de Mestrado - Instituto de Matemática e Estatística, USP São Paulo.

2. Ash, D., Meesters, S., 2002. Best Practices in Reject Inferencing. Apresentação na credit risk modelling and decisioning conference. Wharton Financial Institution Center, Philadelphia.

3. Banasik, J.L., Crook, J.N., 2005. Credit Scoring, augmentation and lean models. Journal of the Operational Research Society 56, 1072-1091.

4. Bensic, M., Sarlija, N. and Zekic-Susac, M., 2005. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. Intelligent Systems in Accounting, Finance and Management, 13: 133?150.

5. Breiman, L., 1996. Bagging Predictors. Machine Learning. 24(2): 123-140p.

6. Breiman, L., 1998. Arcing Classifiers. The Annals of Statistics, v.26, n. 3, 801-849.

7. Buhlmann, P., Yu, B., 2002. Analyzing bagging. The Annals of Statistics, vol. 30, pp. 927-961.

8. Crook, J.N., Banasik, J.L., 2004. Does reject inference really improve the performance of application scoring models?. Journal of Banking and Finance 28, 857-874.

9. Crook, J.N., Banasik, J.L., 2007. Reject inference, augmentation, and sample selection. European Journal of Operational Research 183, 1582-1594.

10. DerSimonian, R., Laird, N., 1986. Meta-analysis in clinical trials. Controlled Clinical Trials, vol 7, 177?188.

11. Feelders, A.J., 2003., An overview of model based reject inference for credit scoring. Technical report, Utrecht University, Institute for Information and Computing Sciences.

12. Feelders, A.J., 2000. Credit scoring and reject inference with mixture models. International Journal of Intelligent Systems in Accounting, Finance and Management 9, 1-8.

13. Hand, D.J., 2001. Reject inference in credit operations: theory and methods. The Handbook of Credit Scoring. Ed. Elizabeth Mays, Glenlake Publishing Company, 225-240.

14. Hosmer, D., Lemeshow S., 1989. Applied Logistic Regression. New York: Wiley.

15. Kuncheva L.I., 2004. Combining Pattern Classifiers. Methods and Algorithms, Wiley.

16. Louzada-Neto, F., Amaral, G. J. A., Abreu, H. J., Guirado, L., Ferreira, M. R. P., Silva, P. H. F., 2009. Medidas Estatísticas da Capacidade Preditiva de Modelos de Classificação em Credit Scoring. Revista Tecnologia de Crédito, 68, 7-27.

17. Louzada-Neto, F., Anacleto, O., Candolo, C., Mazucheli, J., 2011. Poly-Bagging Predictors for Classification Modelling for Credit Scoring. Expert Systems with Applications, v. 38, p. 12717-12720.

18. Parnitzke, T., 2005. Credit scoring and the sample selection bias. Institute of Insurance Economics. University of St. Gallen. Gallen, Switzerland.

19. Rocha, C. A., Andrade F. W. M., 2002. Metodologia para Inferência de Rejeitados no Desenvolvimento de Credit Scoring Utilizando Informações de Mercado. Revista Tecnologia de Crédito, 31, 46-55.

20. Sabato, G., 2009. Solucionando o Viés em Modelos de Credit Scoring: A Inferência de Rejeição. Revista Tecnologia de Crédito, 63, 56-64.

21. Thomas, L. C., Edelman, D. E. & Crook, J. N., 2002. Credit Scoring and its Applications. Monographs on Mathematical Modelling and Computation, Philadelphia: Society for Industrial and Applied Mathematics.

22. Zhu, H., Beling, P. A., Overstreet, G. A., 2001. A study in the combination of two consumer credit scores. Journal of Operational Research Sociaty, 52, 974-980.