REGRESSÃO LOGÍSTICA E REDES NEURAIS ARTIFICIAIS:UM PROBLEMA DE ESTRUTURA DE PREFERÊNCIA DO CONSUMIDOR E CLASSIFICAÇÃO DE PERFIS DE CONSUMO

Alexandre Zanini

TD. Mestrado em Economia Aplicada
FEA/UFJF 007/2007

Juiz de Fora
2007

Regressão Logística e Redes Neurais Artificiais: Um Problema de Estrutura de Preferência do Consumidor e Classificação de Perfis de Consumo*

Alexandre Zanini[1]

Resumo: Neste artigo é proposta uma competição entre os métodos de regressão logística multinomial e as redes neurais artificiais em um problema de estrutura de preferência do consumidor, mais especificamente na classificação de perfis de consumo.

Palavras-Chave: preferência do consumidor, regressão logística e redes neurais artificiais.

1. Introdução

Há diversas maneiras de levantar as preferências dos consumidores sobre um produto (ou serviço). No entanto, perguntar diretamente sobre as características de produto (ou serviço) apresenta dificuldades. Ao se decidir sobre a preferência por um determinado produto, não se considera característica a característica, mas o conjunto de características simultâneas que o produto contém. É uma decisão muitas vezes não consciente e difícil de ser manifestada com exatidão pelo decisor.

Uma solução seria construir um conjunto mínimo de possíveis compostos de marketing, apenas com os atributos determinantes de decisão, para que o consumidor manifestasse a preferência pelos produtos (ou serviços). Em seguida, a partir da escolha, seria interessante determinar quantitativamente a utilidade que ele atribuiu, não-consciente ou conscientemente, aos produtos/serviços/conceitos e às suas características específicas no processo de avaliação dos produtos. Dessa forma, poder-se-iam fazer simulações de preferência ou compra (predição) sobre outros produtos/serviços com os mesmos atributos, mas com outras combinações de suas características.

Dentro dos modelos comportamentais, o processo de decisão sobre a escolha de uma alternativa é influenciado por fatores racionais e subjetivos. Os fatores racionais são aqueles explicados a partir de características sócio-econômicas dos indivíduos (ex: renda). Os fatores subjetivos são aqueles que não são expressos diretamente a partir de conceitos econômicos (ex: conforto) ou que são advindos de fatores aleatórios. A combinação de fatores objetivos e subjetivos forma a "preferência do consumidor".

Um dos principais objetivos deste trabalho é justamente medir a preferência do consumidor sobre produtos e serviços competitivos, expressando matematicamente a importância dos diferentes fatores de escolha. A técnica é construída com base no pressuposto de que consumidores tomam decisões complexas baseadas não apenas em cada fator isoladamente, mas na combinação de diversos fatores. Neste processo de escolha, decisões de compra são tomadas não só com base em fatores racionais, mas também são influenciadas por fatores subjetivos que o consumidor não consegue verbalizar

[1] Professor da Faculdade de Economia e Administração da Universidade Federal de Juiz de Fora.
*ARTIGO EM ELABORAÇÃO.

diretamente. Esta subjetividade estaria relacionada aos valores associados às características que compõem um produto, como, por exemplo, sua Marca.

Os produtos ou serviços pesquisados são descritos por meio de "configurações". Cada configuração é uma combinação de "níveis" de diferentes "atributos". Por exemplo, o produto "Carro" pode ser descrito pelos atributos "Fabricante", "Preço" e "Combustível Utilizado", por exemplo. O atributo "Combustível" pode ter dois níveis: "Gasolina" e "Álcool". A combinação dos níveis de cada atributo constrói uma configuração, como por exemplo: "Automóvel: Marca X, Gasolina, R$70.000,00". Através de um questionário estruturado, diversas configurações de um produto (fabricante, preço e combustível, por exemplo) são apresentadas ao consumidor, que por sua vez informa sua preferência para cada uma. Através deste processo, o método consegue captar a subjetividade do processo de escolha do consumidor. Embora não declare sua preferência por esta ou aquela marca ou nível de preço, estas informações estarão representadas pelas suas escolhas, e serão captadas pelo modelo estatístico.

Em linhas gerais, a modelagem permite: (1) determinar a importância de cada característica de um produto na preferência do consumidor e (2) estabelecer um modelo estatístico que represente o julgamento do consumidor, permitindo, desta forma, construir previsões da aceitação para qualquer configuração, mesmo aquelas que não foram testadas diretamente ou que ainda não entraram em produção.

Os métodos de análise selecionados para tratar o problema em questão foram o modelo de regressão logística e uma rede neural artificial (RNA).

2. Regressão Logísticsca

2.1 Definição

Um modelo de regressão pode ser definido como uma equação matemática em que se expressa o relacionamento de variáveis. Nestes modelos, define-se uma variável dependente (Y), ou variável de saída, e procura-se verificar a influência de uma ou mais variáveis ditas variáveis independentes, causais ou explicativas (X's) sobre esta variável dependente. Na equação (1) a seguir, vê-se um exemplo de um modelo de regressão linear.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + \varepsilon_i \qquad (1)$$

Onde:

Y_i: representa a variável dependente;

β_i : são os coeficientes de regressão;

X_i : são as variáveis independentes;

ε_i: erro aleatório[2].

A regressão logística consiste em um tipo de regressão aplicável e preferida quando se tem uma variável dependente categórica dicotômica, ou seja, uma variável nominal ou não métrica que possui apenas dois grupos ou classificações como resultados possíveis como, por exemplo, alto ou baixo, homem e mulher, sim ou não etc[3].

[2] É importante lembrar que todo modelo é uma simplificação da realidade. Desta forma, todos os modelos estatísticos ou probabilísticos apresentam um componente de erro. Isto indica que, mesmo o modelo tendo um bom poder de explicação, ele sempre incorrerá em um que deve ser minimizado.

[3] Existe ainda a regressão logística multinomial, em que a variável dependente é composta por mais de duas categorias.

Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada diretamente. No caso da variável dependente *Y* assumir apenas dois possíveis estados (1 ou 0) e haver um conjunto de *p* variáveis independentes X_1 , X_2 , ... , X_p, o modelo de regressão logística pode ser escrito da seguinte forma:

$$P(Y=1) = \frac{1}{1+e^{-g(x)}} \quad (2)$$

Onde:

$g(x) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}$

Considerando certa combinação de coeficientes β_0, β_1, ..., β_p e variando os valores de *X*, observa-se que a curva logística tem comportamento probabilístico no formato da letra *S*, o que é característica da regressão logística. Esse formato dá à regressão logística alto grau de generalidade, aliada a aspectos muito desejáveis:

a) Quando $g(x) \rightarrow +\infty$, então $P\ (Y = 1) \rightarrow 1$;

b) Quando $g(x) \rightarrow -\infty$, então $P\ (Y = 1) \rightarrow 0$.

Assim, como se pode estimar diretamente a probabilidade de ocorrência de um evento, pode-se estimar a probabilidade de não ocorrência por diferença: $P\ (Y = 0) = 1 - P\ (Y = 1)$. Ao se utilizar a regressão logística, a principal suposição é a de que o logaritmo da razão entre as probabilidades de ocorrência e não ocorrência do evento é linear:

$$\frac{P(Y=1)}{P(Y=0)} = e^{\beta_0+\beta_1 X_i+\beta_2 X_i+...+\beta_p X_{pi}}$$

e, por conseqüência,

$$\ln\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_0 + \beta_1 X_i + \beta_2 X_i + ... + \beta_p X_{pi}$$

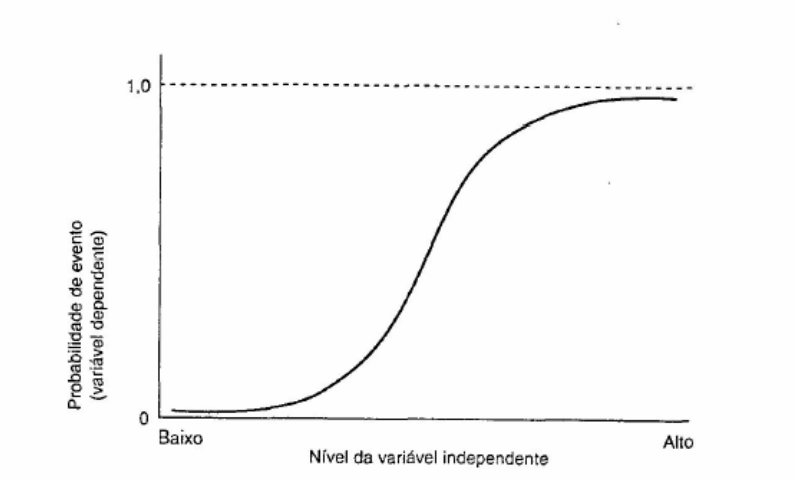
Por essa razão, ao interpretar os coeficientes da regressão logística, opta-se pela interpretação de e^{β} e não diretamente de β. Para utilizar o modelo de regressão logística para discriminação de dois grupos, a regra de classificação é a seguinte:

*Se P(Y=1) > 0,5 então classifica-se Y=1;

* Em caso contrário, classifica-se Y=0.

Fazendo uma síntese, pode-se dizer que um modelo de regressão logística prevê a probabilidade direta de um evento ocorrer. Como se sabe, a probabilidade deve ser um valor limitado entre 0 (zero) e 1 (um) de forma que, se o valor previsto estiver acima de 0,5, aceita-se a hipótese atribuída ao numero 1. Do contrário aceita-se a atribuição dada ao valor 0, qual seja sim ou não, alta ou baixa etc. Esta relação limitada por 0 e 1 caracteriza uma relação não linear que pode ser representada graficamente por uma curva em forma de S, conforme figura 1 a seguir.

Figura 1: Forma da Relação Logística Entre Variáveis Dependente e Independente


1,0
0
Probabilidade de evento
(variável dependente)
Baixo
Alto
Nível da variável independente

2.2. Estimação dos Coeficientes Logísticos

Na equação de regressão logística, para se verificar o efeito ou poder de discriminação de cada uma das variáveis independentes com relação à variável dependente, são calculados os coeficientes de regressão. O cálculo dos coeficientes do modelo é feito através da maximização da função de verossimilhança que calcula a probabilidade de que um evento ocorra (Menard, 1995). Este procedimento é equivalente a minimizar a função logaritmo de verossimilhança (-2LL)[4]. É importante se ter em mente que o que se quer é verificar o poder de ajuste da equação, ou seja, verificar o quanto as variáveis independentes explicam a variável dependente, ou seja, quer se medir o seu poder de influência sobre a variável dependente. Um modelo com bom ajuste terá um valor baixo para -2LL, sendo que o valor mínimo é 0 (zero). Um modelo com ajuste perfeito terá como resposta um valor de verossimilhança igual a 1 (um) e, portanto, -2LL será igual a 0 (zero).

O valor da verossimilhança também pode ser comparado entre equações, onde a diferença representa a mudança no ajuste preditivo de uma equação para outra. Programas estatísticos têm testes estatísticos automáticos para a significância dessas diferenças. O teste qui-quadrado para a redução no valor do logaritmo da verossimilhança fornece uma medida de melhora devido à introdução da(s) variável(eis) independente(s).

É importante ressaltar que, neste trabalho, será estimado um modelo logístico para estimar, a partir do conhecimento de uma série de variáveis, a probabilidade de um cliente comprar um produto específico. A variável dependente (Y) indica se o cliente disse que compraria o produto (=1) ou não (=0) e a série de indicadores (X_1, ..., Xp) constitui o conjunto de variáveis independentes a serem definidas.

Uma das vantagens da regressão logística é que se precisa saber apenas se um evento ocorreu para então usar um valor dicotômico da variável dependente. A partir deste valor dicotômico, o procedimento prevê sua estimativa da probabilidade de que o evento ocorrerá ou não. Se a probabilidade prevista for maior que 0,50, então a previsão será "sim", caso contrário será "não". A regressão deriva seu nome da transformação logística usada com a variável dependente.

O procedimento que calcula o coeficiente logístico compara a probabilidade de um evento ocorrer com a probabilidade de ele não ocorrer. Essa razão de desigualdade pode ser expressa como: $$\frac{\Pr ob(evento\ \ ocorrer)}{\Pr ob(evento\ \ não\ \ ocorrer)} = e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_n X_{ni}}$$

Os coeficientes estimados (β_0, β_1, β_2, ..., β_n) são, na verdade, medidas das variações na proporção das probabilidades, chamada de razão de desigualdade. É importante dizer que um coeficiente positivo indica aumento da probabilidade, ao passo que um valor negativo representa diminuição da probabilidade prevista.

[4] Nos modelos de regressão linear comumente vistos na literatura (ver GUJARATI, 2005), vê-se que os coeficientes de regressão são calculados através da minimização da função de erro quadrático, procedimento conhecido como Mínimos Quadrados Ordinários (MQO). Já na regressão logística, o cálculo é feito através da minimização da função de verossimilhança (na verdade, é -2 *vezes* o logaritmo do valor da verossimilhança e é chamada de -2LL, ou -2logverossimilhança). Um modelo bem ajustado terá um valor pequeno para -2LL.

2.3 Cálculo do Pseudo R^2

Para determinar o ajuste geral do modelo, ou seja, medir seu poder de explicação, são utilizados na regressão logística métricas similares às usadas na análise de regressão tradicional, em que se calcula os chamados coeficientes de explicação, determinação ou simplesmente R^2. Na regressão logística, estes coeficientes recebem a denominação de "Pseudo R^2" e são calculados através da equação (3) a seguir[5]:

$$R^2_{\log it} = \frac{-2LL_{nulo} - (-2LL_{\mathrm{mod}\,elo})}{-2LL_{nulo}} \qquad (3)$$

2.4 Teste de Hosmer e Lemeshow

O teste Hosmer e Lemeshow ou teste HL tem a finalidade de avaliar a validade preditiva do modelo de Regressão Logística. É baseado não no valor de verossimilhança, mas na visão real da variável dependente (Hosmer e Lemeshow, 1989).

Considerando-se Y como o valor real da variável e $\hat{Y}$ como o valor previsto, o teste é feito com intuito de medir a proximidade de ambos. A hipótese nula (hipótese de teste) é que não existe diferença significativa entre o valor real e o valor previsto, ou seja, equivale a dizer que o modelo tem bom poder de ajuste. Quanto menor é o valor da diferença entre Y e $\hat{Y}$, mais os valores previstos se aproximam dos reais e, portanto, melhor desempenho preditivo tem o modelo. Desta forma, um fator positivo a favor do modelo é quando se aceita a seguinte hipótese nula: $H_0 : Y = \hat{Y}$ *ou* $H_0 : Y - \hat{Y} = 0$.

3 - Redes Neurais

3.1 - Conceituação

A motivação original desta metodologia foi a tentativa de modelar a rede de neurônios humanos visando compreender o funcionamento do cérebro. Portanto, como o próprio nome da metodologia revela, sua motivação inicial foi a de realizar tarefas complexas que o cérebro executa com elevada efetividade (por exemplo: reconhecimento de padrões, percepção e controle motor) através da simulação de seu funcionamento.

Hoje, as redes neurais artificiais são uma metodologia estatística eficiente e capaz de resolver uma gama de problemas importantes. Trata-se, portanto, de um processador capaz de extrair conhecimento experimental disponibilizando-o para uso prático (tomada de decisões por exemplo).

Uma grande vantagem de usar uma rede neural é a capacidade de resolver problemas sem a necessidade de definição de listas de regras (como nos sistemas especialistas e nos modelos estatísticos clássicos) ou de modelos explícitos. Isto possibilita tratar de situações onde é difícil criar modelos adequados da realidade ou situações com freqüentes mudanças no ambiente. Atenta-se que, grande parte desta sua adequabilidade funcional deve-se à sua capacidade em inferir relações não lineares complexas. Frente a

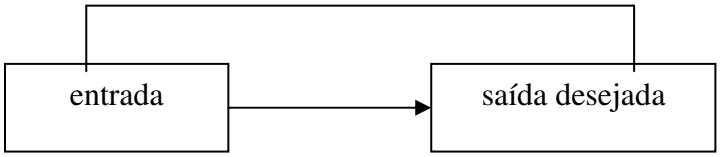
[5] Existem várias métricas para este "pseudo R^2". A que será utilizada neste trabalho é o coeficiente de Nagelkerke.

estas suas propriedades, hoje, podemos observar sua aplicabilidade principalmente nas áreas de classificação de padrões (Haykin, 1995 e Bishop, 1995) e de previsão (Zanini, 2000).

3.2 - Modelagem e aprendizado em redes neurais

A modelagem em redes neurais pode ser sintetizada basicamente nos seguintes passos:
1) Escolher os exemplos (observações) a serem usados para o treinamento, de modo que o ambiente seja bem representado. Salienta-se que os exemplos consistem em pares do tipo (entrada, saída desejada).

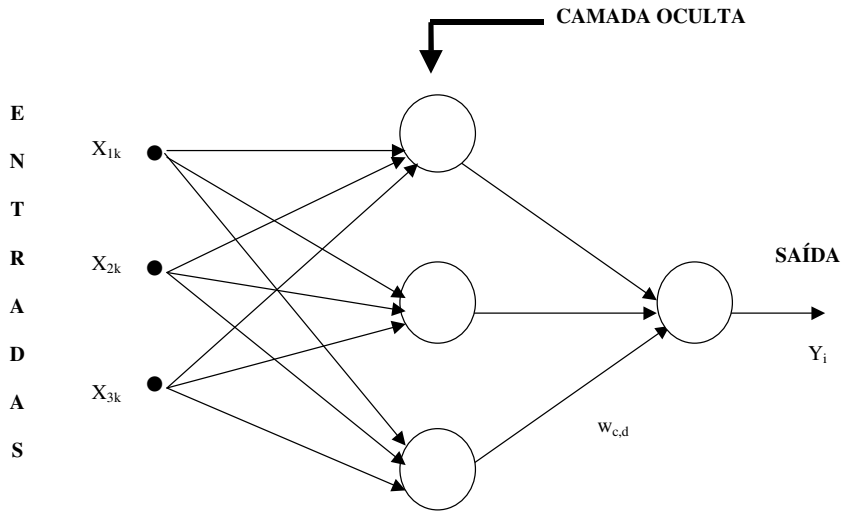
EXEMPLOS



2) Escolher os exemplos que serão utilizados para validação (avaliação da capacidade de generalização do modelo).
3) Escolher a arquitetura apropriada.
4) Escolher o algoritmo para treinamento da rede.

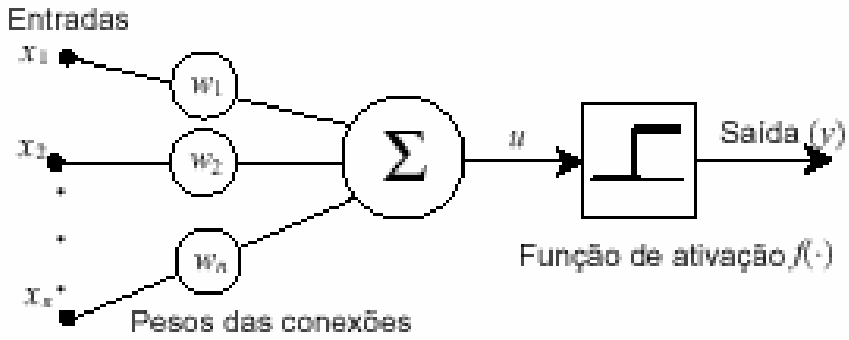
A figura 2 mostra um exemplo de uma rede *Feedforward* (propagação direta) com três entradas, uma camada escondida e uma camada de saída (são representados 4 neurônios). Ela está incorporando informações através do vetor $X_k=[X_{1k}\ X_{2k}\ X_{3k}]^T$ para produzir como resultado Y_i. O peso $w_{c,d}$ conecta o neurônio "c" ao neurônio "d" da camada imediatamente posterior.

Figura 02: : Rede *Feedforward*



Já na figura 03 a seguir vê-se a representação funcional de apenas um neurônio artificial. Considerando "p" entradas ligadas a um somador por "p" pesos, então $u = \sum_{k=1}^{p} w_k x_k$ e $y = f(u)$.

Figura 03: Representação funcional de um neurônio artificial



A função "f" é chamada de função de ativação e caracteriza o aspecto "não-linear" de uma rede neural. É importante ressaltar que esta função precisa ser diferenciável e não decrescente devido ao mecanismo de ajuste dos pesos. Em geral, as funções sigmoidal e tangente hiperbólica são as ativações mais usadas. Neste trabalho optou-se pela primeira, até por ser a mesma utilizada no modelo de regressão logística.

Especificada a metodologia de modelagem quando do uso de redes neurais, é importante que se entenda como a rede "adquire" conhecimento. O conhecimento em redes neurais é adquirido através de um processo de aprendizado ou de treinamento. A informação é transformada e armazenada nas "densidades de conexão" ou pesos sinápticos.

O conhecimento é então produzido utilizando estes pesos para combinar as não linearidades (o que implica numa distribuição da não linearidade na rede). Salienta-se que este conhecimento é incorporado sem que se tenha que explicitar modelos *a priori*, ou seja, através de um aprendizado por exemplos, o modelo é construído implicitamente a partir dos dados.

A rede neural aprende, então, o ambiente através de um processo iterativo de modificação dos pesos de interligação, a partir de estímulos fornecidos pelo ambiente. O tipo de aprendizado é determinado pelo modo com que se promove a adaptação dos parâmetros e isso pode ser feito de dois modos:

1) Aprendizado Supervisionado – usa-se um conjunto de pares, entrada e saída, previamente conhecidos que representam a realidade;

2) Aprendizado Não Supervisionado – não se usa um conjunto de exemplos previamente conhecidos. Uma medida da qualidade da representação do ambiente pela rede é estabelecida e os parâmetros são modificados de modo a otimizar esta medida. Este tipo de aprendizado é muito utilizado na área de reconhecimento de padrões (Kohonen, 1995).

3.3 - Aprendizado supervisionado e o algoritmo de *Backpropagation*

Neste trabalho, utilizou-se apenas o aprendizado supervisionado. Este é baseado no erro de saída da rede. Obviamente, para que seja possível calcular um erro na saída, é

preciso que haja um conjunto de pares (entrada – saída desejada) conhecidos. Desta forma, os pesos são ajustados visando minimizar o erro de saída.

O algoritmo mais freqüentemente empregado no aprendizado supervisionado de redes neurais é conhecido como Retropropagação do Erro (*Backpropagation)*. Destaca-se que, este algoritmo tem fundamentos matemáticos baseados na regra da cadeia e no método do gradiente descendente [Haykin, 1998].

O algoritmo de *Backpropagation* consiste basicamente em:

1) Calcular o erro na saída da rede.

2) Retropropagar o erro na rede calculando de que forma modificações nos pesos afetam o erro.

3) Modificar os pesos de modo a minimizar o erro.

É importante ressaltar que o objetivo é sempre minimizar o erro médio. Para isto, faz-se modificações nos pesos padrão a padrão. A média destas modificações (sobre todos os exemplos) é uma estimativa da modificação que se obteria em se minimizando o erro sobre todo o conjunto.

Desta forma, o mecanismo de aprendizado pode ser então sintetizado em três passos:

Passo 1) a rede é estimulada pelo ambiente através de exemplos extraídos do mesmo.

Passo 2) os pesos são modificados através de um procedimento iterativo, como resultado do passo 1.

Passo 3) a rede passa a responder ao ambiente de uma nova forma em decorrência das mudanças efetuadas no passo 2.

3.4. - Generalização

Uma rede neural opera em duas fases distintas: treinamento (*in sample)* e testes (*out of sample)*, quando se pretende atingir a generalização. Na primeira fase, a rede retira do ambiente o conhecimento. Na segunda, a qualidade da primeira fase é posta em cheque, para que possa ser utilizada para os fins almejados quais sejam: projeção, classificação, segmentação, entre outros. Desta forma, o que se faz, na prática, é escolher um conjunto de dados para o treinamento e separar um outro para a generalização.

É importante salientar que uma rede neural generaliza com sucesso quando a relação entrada-saída aprendida no treinamento é representativa do ambiente gerador. Outro aspecto interessante é que quanto mais simples a arquitetura, maior a probabilidade de uma generalização adequada. A justificativa consiste no fato de que, ao adicionarmos complexidade ao modelo, se está sujeito a uma generalização fraca pois, por ser excessivamente complexo o modelo capta a maioria dos detalhes dos dados de treinamento em detrimento de propriedades mais gerais.

No caso das redes neurais, dado o processo de estimação e ajuste dos pesos, quanto mais camadas e neurônios nas camadas ocultas existirem, mais complexo será o modelo e, portanto, mais "pobre" será a generalização (Haykin, 1998). Isto implica no fato de que, quanto menor a quantidade de dados, mais simples deve ser o modelo.

4 – Análise de Dados e Resultados

Para a implementação das análises, foram selecionados dados obtidos através de uma pesquisa de mercado. O questionário foi aplicado para uma amostra de 461 indivíduos e investigava aspectos relativos à compra de um carro da marca Y. Trabalhou-se apenas com variáveis qualitativas com apenas duas categorias. Estas variáveis podem ser vistas no quadro 01 a seguir:

Quadro 01: Variáveis no modelo

COMPARIA O CARRO Y?	SIM (1)
	NÃO (0)
SEXO	Feminino (0)
	Masculino (1)
Importância: VISUAL	Alta (1)
	Baixa (0)
Importância: POTÊNCIA DO MOTOR	Alta (1)
	Baixa (0)
Importância: COR	Alta (1)
	Baixa (0)
Importância: SEGURANÇA	Alta (1)
	Baixa (0)
Importância: PREÇO	Alta (1)
	Baixa (0)
Importância: COMBUSTÍVEL	Alta (1)
	Baixa (0)

A variável de saída (Y) é a resposta que representa a "intenção" de comprar o carro. Tem-se ainda uma variável de gênero (masculino e feminino) para caracterizar o consumidor. As demais foram obtidas através da codificação das respostas quanto à importância (escala de 0 a 10) atribuída ao se avaliar um carro. Os resultados do modelo de regressão logística podem ser vistos nos quadros 02 a 04 a seguir:

Quadro 02: Betas e Significância das Variáveis – Modelo Final

Variáveis	Beta	Exp(Beta)	p-valor
Constante	0,364	1,439	0,024
Sexo	1,219	3,384	0,025
Segurança	1,099	3,001	0,052
Combustível	0,714	2,042	0,002
Marca	1,419	4,133	0,003

Quadro 03: Valores para Pseudo R^2 e Teste Hosmer e Lemeshow (HL) – Modelo Final

Desempenho	Estatísticas
Nalgerkerke R^2	74,3%
Teste Hosmer e Lemeshow	$\chi^2 = 2,956$ (p-valor = 0,889)

Quadro 04: Valores Observados e Valores Previstos – Modelo Final

Observado		Previsto		
		Compraria o carro Y?		
		Sim	Não	Percentual de acerto
Compraria o carro Y?	Sim	139	103	92,9%
	Não	102	117	89,5%
				91,8%

Como visto anteriormente, uma das vantagens do modelo de regressão logística é que os coefientes de regressão (na verdade *exp* do coeficiente) representam uma "razão de chances", ou seja, a chance do evento acontecer em relação a não acontecer. Desta forma, vê-se que os homens têm 3,4 mais chances de comprarem o carro Y do que as mulheres. Já pessoas que deram importância alta para os atributos segurança, combustível e marca têm respectivamente 3, 2 e 4 vezes mais chances de comprar o carro Y do que aqueles que deram importância baixa. Isto é um fator positivo a favor do carro em questão e estes são atributos que poderiam ser melhor explorados numa abordagem de marketing direto, por exemplo.

O coeficiente de Nagelkerke (74%) indica um bom poder de previsão do modelo, o que é corroborado pelo teste HS que indica boa aderência entre valores observados e previstos. Esta constatação é corroborada pelo percentual de acerto do modelo que chega a 92% no total.

Para a RNA as variáveis foram codificadas numa forma vetorial como pode ser visto no quadro 05 a seguir:

Quadro 05: Codificação das variáveis para a RNA

Faixas	Codificação
Variáveis Independentes	
Baixo (0)	[0;1]
Alto (1)	[1;0]
Variável Dependente	
Sim	[1;0]
Não	[0;1]

Esta formatação é importante, pois pode-se calcular, assim como na regressão logística, os percentuais de acerto e erro da RNA (bem como o percentual de "dúvida" na classificação). Para isto, é importante se estabelecer alguns pontos de corte nos vetores para se avaliar a acertabilidade.

Durante a fase do treinamento da RNA, vários tipos de arquitetura foram testadas. Optou-se ao final, por uma topologia onde tem-se as quatro entradas descritas anteriormente, apenas uma camada oculta e a camada de saída com dois neurônios. Para este tipo de arquitetura, diferentes parâmetros como o número de épocas, o objetivo de erro e o gradiente mínimo foram também testados. Percebeu-se, entretanto, que os resultados eram muito similares. Desta forma, comentar-se-á sobre as duas tipologias de RNA que melhores resultados geraram para um objetivo de erro de 0,001. Os resultados podem ser vistos no quadro 06 a seguir:

Quadro 06: % Acerto da RNA

Corte	Categoria	Número de Neurônios na Camada Oculta	
		3	5
[0,8 0,2]	Não	71,4	71,4
	Sim	85,7	100,0
[0,7 0,3]	Não	85,7	85,7
	Sim	85,7	100,0
[0,6 0,4]	Não	85,7	85,7
	Sim	100,0	100,0
Erro Quadrático Médio		18,1	13,8

Observando o quadro 06, salienta-se que, definindo um objetivo de erro em 0,001, apesar do erro quadrático médio ter ficado em 18,1 e 13,8 para três e cinco neurônios "escondidos" respectivamente, conseguiu-se boa performance de classificação. Neste sentido, chama a atenção os altos percentuais de acerto que a RNA conseguiu. Começando com um corte de [0,8 0,2], observa-se que com 5 neurônios na camada escondida consegue-se melhorar a classificação na categoria "Sim" em relação à estrutura com 3 neurônios (100 contra 85,7%). Quando altera-se o corte para [0,7 0,3], obtém-se melhoras em ambas as arquiteturas na categoria "Não" (85,7 contra 71,4%). Por fim, modificando o corte para [0,6 0,4], com 3 neurônios "escondidos" aufere-se melhora na classificação da categoria "Sim" (100 contra 85,7%), ficando inalterado a performance com 5 neurônios na camada escondida neste caso.

A comparação do resultado de classificação das duas metodologias pode ser visto no quadro 07:

Quadro 07: Comparação das duas metodologias: Percentual de Acerto

Metodologias	Sim	Não
Regressão Logística	92,9	89,5
Rede Neural Artificial	100,0	85,7

Primeiramente é importante observar que ambas as técnicas têm menor acertividade na categoria "Não". Vê-se ainda que a RNA, quando comparada com a regressão logística, teve um melhor percentual de acerto na categoria "Sim", o que não acontece na categoria "Não".

5 – Conclusões

Este trabalho abordou um problema de classificação importante na área de marketing e que diz respeito à modelagem da estrutura de preferência do consumidor. Classificou-se os clientes através da possibilidade ou não destes comprarem um determinado produto (carro). Entretanto, é importante ressaltar que a classificação é importante em várias outras áreas e pode ser utilizada, por exemplo, para calcular a probabilidade de um cliente não cumprir com os prazos de pagamento (risco de crédito), ou seja, classificando-se os clientes em adimplentes ou inadimplentes. Na área médica, por exemplo, pode-se calcular a chance de um paciente desenvolver determinado desfecho (doença, por exemplo) a partir da exposição a determinados fatores de risco.

A regressão logística possibilitou selecionar exatamente aquelas variáveis (neste caso, atributos) relevantes para a compra do produto em questão. Foram estes o "sexo" e a importância atribuída para os atributos "segurança", "combustível" e "marca". Uma das vantagens desta metodologia é que os coeficientes de regressão são lidos na forma de uma "razão de chances", indicando o efeito de cada atributo sobre a variável dependente, ou seja, sobre a compra ou não do produto. Este fator não se verifica, por exemplo, no uso das redes neurais artificiais, onde os pesos sinápticos não têm uma interpretação lógica. Por outro lado, se o foco estiver apenas na classificação em "certo ou errado", "sim ou não", "alto ou baixo", "adimplente ou inadimplente" viu-se que, neste problema, a RNA apresentou melhor desempenho nesta classificação dos objetos.

A aplicação prática destas metodologias são muitas. Pode-se, por exemplo, através de um modelo bem ajustado, gerar informações importantes para uma estratégia de marketing direto, abordando-se com mais ênfase aqueles fatores (atributos) que maximizariam a chance do cliente adquirir determinado produto/serviço. Como dito anteriormente, estes modelos também podem ser utilizados para minimizar o risco de crédito na medida que é possível classificar os clientes em adimplentes e inadimplentes identificando-se seus perfis. Em síntese, os modelos matemáticos aqui utilizados geram informações importantes para a tomada de decisões dos agentes em diversas áreas do conhecimento.

6- Referências

BISHOP, C.M.. *Neural Networks for Pattern Recognition*. Oxford : Clarendon Press, 1995.

GUJARATI, D. *Econometria Básica*. 4ª. ed. Ed. Campus. 2006.

HAYKIN, S.. *Neural networks*: *a comprehensive foundation*. 2nd ed. New Jersey: Prentice Hall. 1998.

HOSMER, D.; LEMESHOW, S. *Applied logistic regression*. New York: John Wiley & Sons, 1989.

KOHONEN, T.. Self-organizing maps. New York: Springer Verlag. 1995.

MENARD, SCOTT.. *Applied logistic regression analysis*. 1995.

ZANINI, A.. *Redes neurais e regressão dinâmica: um modelo híbrido para previsão de curto prazo da demanda de gasolina automotiva no Brasil*. Dissertação de Mestrado, DEE, PUC-Rio, abril, 2000.