Detecção de fraudes na distribuição de energia elétrica utilizando support vector machine

Vinícius Dornela Silva †
Rodrigo Arnaldo Scarpel ‡

† Instituto Tecnológico de Aeronáutica/IEMB
donela@gmail.com.br

† Instituto Tecnológico de Aeronáutica/IEMB
rodrigo@ita.br
www.mec.ita.br/~rodrigo

Abstract

Several economics sectors are exposed to fraud made by their own customers. At the Electrician Distribution segment is not different. Several techniques in statistical fields were developed to detect illegal activities, relied in observation classification. The empiric events modeling always become a challenge to get solution in engineering projects. The solution is achieved using a induction process to built a system able to bring up the answer of a previously observed event. The Linear Discriminate Analysis is the quantitative method most used. Recently, an alternative approach arises: The Support Vector Machine (SVM). This paper objectives is train and test a model built using SVM to point out the customers that are performing frauds given the customers company (an Electrician Distribution) data base, doing a efficiency and quality confrontation vis-à-vis the Linear Discriminate Analysis.

Resumo

Os mais variados setores da economia estão sujeitos às fraudes cometidas pelos seus próprios clientes. No setor de distribuição de energia elétrica não é diferente. Muitas técnicas no campo estatístico foram desenvolvidas para detectar atividades fraudulentas, baseando-se em classificações das observações. A solução é obtida utilizando um processo de indução para se construir um sistema capaz de deduzir respostas de fenômenos que já tenham sido observados anteriormente. O método quantitativo mais empregado na classificação de observações é a Análise Discriminante Linear. Recentemente, como alternativa a essa técnica, surgiu o Support Vector Machine (SVM). O objetivo do presente trabalho foi treinar e testar um modelo utilizando o SVM para a classificação de clientes de uma distribuidora de energia elétrica, fazendo uma comparação de eficiência e qualidade vis-à-vis a Análise Discriminante Linear.

Keywords: fraud detection, support vector machine, classification models

Title: Fraud detection in energy distribution using support vector machine

Apdio © *2007 Associação Portuguesa de Investigação Operacional*

1 Introdução

Fraudes cometidas por consumidores tornaram-se tema extremamente comum no cenário atual. Muitos setores da economia são afetados por atividades ilícitas. Exemplos claros acontecem em empresas de telecomunicações – operadoras de celular (aparelhos clonados) e provedores de internet (ligações não autorizadas), empresas de seguros e planos de saúde (falsificação de óbitos e de receitas de medicamentos).

As fraudes representam perdas significativas de receita para as empresas. As perdas são representadas pelo consumo "gratuito" dos produtos e/ou serviços oferecidos, pelos danos causados nos ativos das empresas e, também, por gastos associados a processos judiciais em que os clientes fraudadores devem ser submetidos, onde nem sempre a parte vitoriosa é a empresa lesada.

No setor de distribuição de energia elétrica não é diferente. As ligações clandestinas acontecem por todas as partes e causam perdas relevantes para as companhias distribuidoras. Além das perdas devido ao consumo ilegal, freqüentemente, ocorrem perdas por danificação de ativos (sobrecarga do sistema, queda de tensão e aumento de taxa de manutenção), por realização de inspeções e por acidentes envolvendo o fraudador e o sistema elétrico, que danificam as instalações e, que também, podem provocar a morte de pessoas.

No Brasil, a ocorrência das ligações clandestinas é um fenômeno de intensa "penetração". Segundo a Associação Brasileira de Distribuidores de Energia Elétrica, em amostra com as 18 principais companhias de distribuição de energia do país, detentoras de cerca de 81% da energia elétrica distribuída nacionalmente, o índice nacional de perda comercial (energia requerida / energia consumida), através de ligações ilegais e fraudes, esteve em 5,0%, em 2004, nível ainda considerado alto, já que o padrão mundial é de 1%.

Esses números representam perdas significativas de receita. Em um levantamento realizado pela Agência Nacional de Energia Elétrica, com 26 distribuidoras de energia, as perdas comerciais representavam 4,19% do total de energia produzido em MWh pelas companhias (montante equivalente a R$ 848.965.111,00) e, em algumas empresas, a participação das perdas comerciais atingiram mais de 15% (Figura 1).

Cientes destes números, o problema tornou-se preocupante para as distribuidoras. Assim, programas de detecção e combate a fraudes foram criados e são tidos como prioridade para as empresas de distribuição de energia elétrica.

De acordo com Boccuzzi (2005), são diversas as soluções propostas para reduzir a incidência de fraudes como o desenvolvimento de mecanismos que propiciem aumento/eficiência das inspeções contra fraudes; a conscientização da população e criação de linhas diretas para denúncia; a criação de parcerias entre órgãos, permitindo o cruzamento das bases de dados de departamentos de impostos, polícia, defesa do consumidor, dentre outros, aumentando o rigor dos processos judiciais contra fraudadores, por exemplo, reclamando por compensações financeiras e reduzir a flexibilidade dos acordos judiciais. Ainda segundo o autor a AES Eletropaulo planeja, para 2005, gastos de cerca de R$ 43,5 milhões em ações de combate ao furto e fraude de energia elétrica.

No que diz respeito ao desenvolvimento de mecanismos que propiciem aumento / eficiência das inspeções contra fraudes é muito comum o desenvolvimento de modelos de classificação que auxiliem na busca por prováveis fraudadores. A Figura 2 ilustra o funcionamento do mecanismo de detecção de fraudes, em que um modelo aponta, dentre os clientes de uma empresa, os prováveis fraudadores.

No passado, os modelos de classificação mais utilizados eram os baseados em regras e apresentavam, dessa forma, uma grande limitação, já que as regras utilizadas nem sempre eram geradas, empiricamente, a partir de bases de dados existentes, mas sim, por meio de regras julgamentais, diminuindo a confiabilidade desses mecanismos.

Segundo Scarpel (2005), recentemente, novos mecanismos foram desenvolvidos como as técnicas de mineração de dados, de reconhecimento de padrões, estatísticas e de inteligência artificial e, estão sendo utilizadas de forma crescente na formação de modelos de detecção de fraudes, onde as regras são geradas a partir dos dados históricos de casos comprovados de fraudes.

		Perdas técnicas (MWh)	Perdas comerciais (MWh)	Perdas consumidores cativos (MWh)	Perdas consumidores livres (MWh)	Perdas Totais (cativos+livres) (MWh)	Perdas técnica (%)	Perdas comercial (%)	Perdas Comerciais (R$)
1	LIGHT	1,203,170	3,120,497	4,323,667	494,373	4,818,040	6.06%	15.73%	229,387,752
2	ELETROPAULO	2,062,543	2,510,985	4,573,528	78,053	4,651,581	6.33%	7.71%	218,455,695
3	AMPLA	974,281	1,114,714	2,088,995	70,096	2,159,091	13.29%	15.21%	88,151,583
4	COELBA	1,086,974	589,825	1,679,799	-	1,676,799	12.06%	6.54%	37,141,280
5	CPFL PIRATININGA	589,316	402,736	992,053	103,994	1,096,047	5.76%	3.93%	35,485,094
6	CEMIG	2,405,806	448,885	2,854,691	87,048	2,941,738	6.98%	1.30%	29,837,395
7	CEEE	561,286	289,034	850,320	23,858	874,178	8.82%	4.54%	25,622,864
8	CPFL	1,334,627	249,441	1,584,068	49,610	1,633,678	6.98%	1.30%	21,282,310
9	ELEKTRO	500,867	269,967	770,835	72,254	843,088	4.82%	2.60%	20,279,959
10	ESCELSA	465,917	241,562	707,479	171,442	878,921	7.84%	4.06%	19,612,442
11	COPEL	1,123,477	218,481	1,341,958	179,804	1,521,762	6.22%	1.21%	18,105,553
12	COELCE	678,812	253,433	932,245	-	932,245	11.35%	4.27%	17,588,268
13	CELESC	783,621	185,119	968,740	26,993	995,733	5.64%	1.33%	17,138,354
14	CELPA	746,218	281,368	1,027,586	-	1,027,586	17.77%	6.70%	14,397,597
15	BANDEIRANTE	664,793	113,002	777,795	201,067	978,862	7.19%	1.22%	9,431,125
16	AES SUL	367,507	89,456	456,963	-	456,963	4.98%	1.21%	7,898,093
17	CEB	262,007	89,224	351,231	22,377	373,608	7.23%	2.46%	7,625,975
18	COSERN	327,198	124,467	451,666	-	451,666	11.13%	4.23%	6,784,716
19	RGE	583,838	66,314	650,152	-	650,152	9.24%	1.05%	5,887,357
20	ENERSUL	434,613	81,644	516,257	8,443	524,700	15.40%	2.89%	5,621,176
21	ENERGIPE	207,292	86,908	294,200	5,824	300,024	10.96%	4.59%	4,718,216
22	CEMAT	439,966	32,165	472,131	-	472,131	12.68%	0.93%	2,562,287
23	CELTINS	139,983	33,374	173,357	-	173,357	16.24%	3.87%	2,499,372
24	CELB	18,167	25,298	43,464	3,248	46,712	3.82%	5.32%	1,941,857
25	CFLCL	109,589	6,799	116,388	-	116,388	11.28%	0.70%	829,120
26	SULGIPE	24,791	8,321	33,113	-	33,113	11.67%	3.92%	679,674
	TOTAL	18,096,659	10,933,019	29,032,681	-	30,628,163	9.30%	4.19%	848,965,111

Figura 1: Distribuidoras de energia elétrica com maiores perdas comerciais e técnicas, em MWH e R$, entre janeiro/2003 e abril/2005

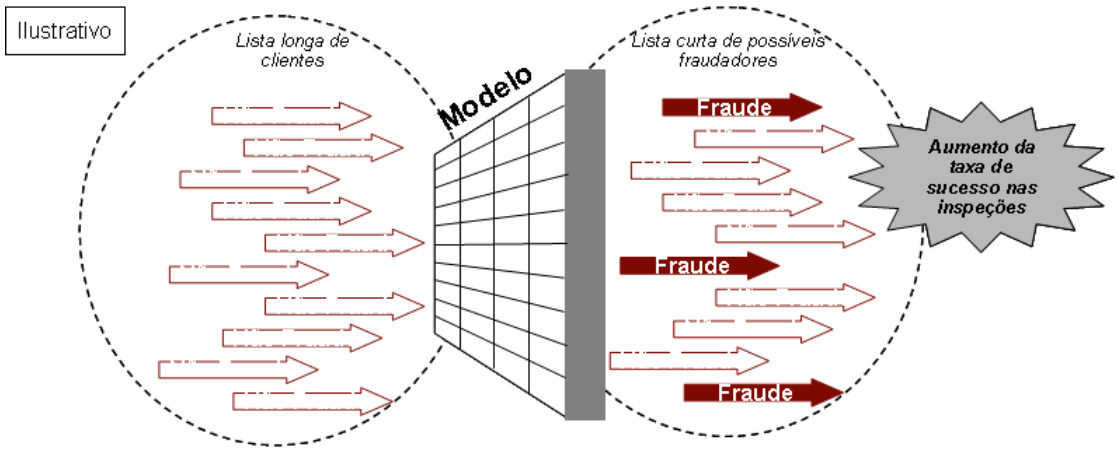


Figura 2: Esquema ilustrativo do mecanismo de detecção de fraudes

O objetivo deste trabalho é o desenvolvimento de um modelo para detecção de fraudes em sistemas de distribuição de energia elétrica utilizando support vector machine (SVM),

para indicar, dentre a lista de clientes, os mais prováveis de estarem cometendo fraude e, conseqüentemente, quais devem ser inspecionados.
O desempenho do modelo desenvolvido será comparado ao desempenho da técnica análise discriminante linear que, historicamente, é o método quantitativo mais utilizado na criação de modelos de classificação.

2 O Método Empregado

A modelagem de fenômenos empíricos sempre representou desafios para soluções de problemas em engenharia. Nesse tipo de modelagem, um processo de indução é utilizado para se construir um sistema capaz de deduzir respostas de fenômenos que já tenham sido observados anteriormente. A análise discriminante é o método quantitativo muito empregado para este propósito, por ser um método de fácil implementação e que apresenta uma performance aceitável.
Uma alternativa a este método é o uso do support vector machine (SVM). O SVM é uma técnica de programação matemática desenvolvida por Vladimir Vapnik [Vapnik, 1998] e que vem sendo utilizada de forma crescente, devido aos muitos recursos atrativos e promissora performance empírica. Sua formulação incorpora o princípio da Minimização de Risco Estrutural (SRM), diferentemente, de técnicas mais tradicionais que utilizam o principio da Minimização de Risco Empírico (ERM), o que promove ao SVM maior generalização [Gunn, 1998].
A utilização de SVM é crescente, principalmente, na área de reconhecimento de padrões. Alguns trabalhos desenvolvidos neste área são os de Guyon et al. (2002), Bin et al. (2000), Bradley e Mangasarian (2000) e Byvatov e Schneider (2003).
Segundo Vapnik (1999), o SVM é um procedimento construtivo universal de aprendizagem baseado em "statistical learning theory". O termo universal significa que o SVM pode ser utilizado para o aprendizado de várias representações como as funções de base radial, "splines" e funções polinomiais.
O problema de classificação pode ser apresentado considerando um problema de classificação binária (duas classes apenas) sem perda de generalidade. Assim, o objetivo é separar as observações em dois grupos, utilizando uma função que tenha sido deduzida de exemplos disponíveis e que seja capaz de separar futuras observações com uma certa precisão. Exemplificando, para uma amostra de dados (Figura 3) que possa ser separada por hiperplanos lineares, existem inúmeras possibilidades de resposta, mas apenas uma que maximiza a margem (maximiza a distância entre o hiperplano e o ponto mais próximo de cada classe). É importante observar que podemos utilizar hiperplanos não lineares também para fazer a separação da amostra.

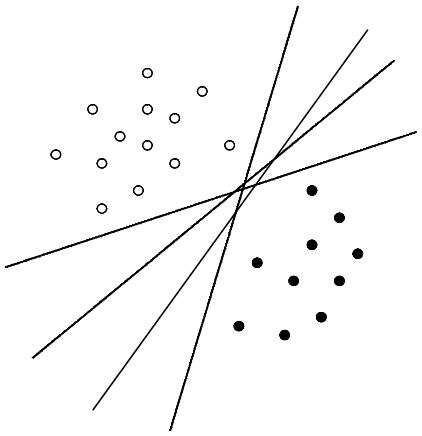


Figura 3: Possíveis hiperplanos de separação para uma amostra de dados

2.1 Caso 1: Separação Linear

Tratando-se de classificação binária, o problema é achar uma função paramétrica, linear ou não, para um hiperplano de separação dos pontos em dois conjuntos no R_m, em que m é o número de dimensões existentes. No caso onde o problema seja separável por um hiperplano linear com um conjunto de N observações $x_i = (x_{i1},..., x_{iN})$ e respostas binárias $yi \in -1,1$ têm-se três hiperplanos:

1. Hiperplano de Separação: H_0: $y = w^t x + b = 0$ que separa as observações.
2. Hiperplano Superior: H_1: $y = w^t x + b = +1$ que é definido por pelo menos 1 ponto pertencente ao grupo com y = +1.
3. Hiperplano Inferior: H_2: $y = w^t x + b = -1$ que é definido por pelo menos 1 ponto pertencente ao grupo com y = -1.

A Figura 4 ilustra os hiperplanos de separação, superior e inferior no espaço m=2.
Os pontos que definem os hiperplanos H_1 e H_2 são chamados de "support vectors" e a orientação do plano de separação (H_0) é feita de forma que a distância entre H_1 e H_2 seja máxima.

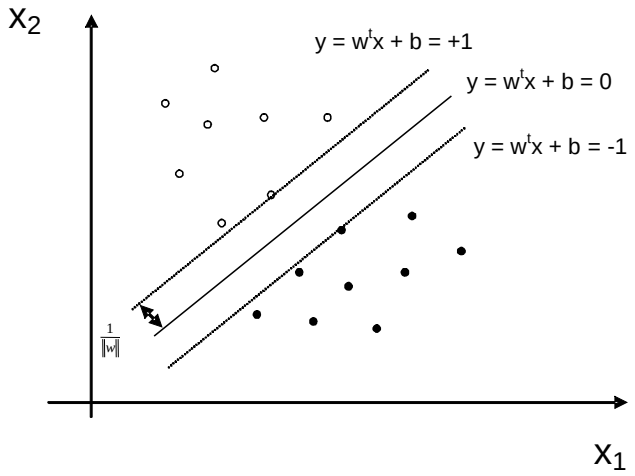


Figura 4: Hiperplanos de separação para o caso m=2

Uma vez que a distância entre os hiperplanos H_1 e H_2 é calculada por $D(H_1, H_2) = \frac{2}{\| w \|} = \frac{2}{w^t w}$ e o objetivo é encontrar os parâmetros w que maximizem essa distância. Desta forma, define-se a função objetivo do problema como

$$Minimizar\ Z = \frac{1}{2} w^t w \tag{1}$$

Como restrições, ao problema separável por um hiperplano linear, para que não hajam pontos entre H_1 e H_2, têm-se $w^t x - b \geq +1$ para y = +1 e $w^t x - b \leq -1$ para y = -1. Essas duas restrições podem ser combinadas fazendo-se com que a formulação matemática desse problema tenha como restrição

$$y\left(w^t x - b\right) \geq +1 \tag{2}$$

Assim, o problema de separação tem m+1 incógnitas ($w_1,...,w_M$,b). A estimativa dos parâmetros é definida pelos pontos sobre H_1 e H_2, chamados de "support vectors", de forma que os outros pontos podem ser movidos livremente sem alterar o resultado da otimização.

2.2 Caso 2: Não Linearmente Separáveis

Na maioria das situações, a base de dados não pode ser linearmente separada (Figura 5). Existem duas maneiras de abordar o novo problema, que dependem do conhecimento anterior e da estimativa de ruído da base de dados.
Se for esperado (ou previamente conhecido) que um hiperplano pode separar as classes corretamente, introduz-se um custo adicional na função objetivo para penalizar os erros de classificação existentes. Outra alternativa é utilizar uma função mais complexa para descrever os limites do modelo.

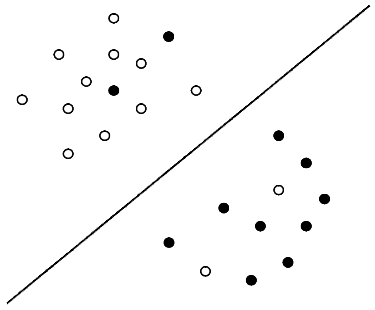


Figura 5: Situação não separável por hiperplano linear

Nesse caso, introduz-se N variáveis de folga ($\xi_i \geq 0$, i=1,…,N), de forma a criar uma penalidade na função objetivo e uma folga nas restrições. Portanto, a formulação do problema de separação no caso inseparável por um hiperplano linear é

$$Minimizar \;\; Z = \frac{1}{2} w^t w + C\left(\sum_{i=1}^{N} \xi_i\right) \tag{3}$$

$$Sujeito\ a \;\; y_i\left(w^t x_i - b\right) \geq +1 - \xi_i \quad ,\mathrm{i} = 1,...,\mathrm{N} \tag{4}$$

$$\xi_i \geq 0 \quad ,\mathrm{i} = 1,...,\mathrm{N} \tag{5}$$

em que C é uma constante de penalização (C > 0). Esse problema tem N+m+1 incógnitas ($\xi_1,\ldots,\xi_N,w_1,\ldots,w_M,b$). É possível sofisticar o modelo utilizando outros hiperplanos de separação como funções polinomiais, splines ou funções de base radial, por exemplo.

3 Análises e Resultados

Os dados utilizados foram fornecidos por uma distribuidora de energia elétrica que será chamada de EMPRESA. Serão consideradas informações relacionadas a questões geográficas, perfil do cliente (tipo de residência e forma de pagamento) e dos padrões de consumo (valor da conta, consumo médio e variação da conta no último período).
A base de dados fornecida pela EMPRESA possui informações de 596 clientes de um determinado período, sendo que 298 deles representam clientes fraudulentos (realizaram ligações clandestinas) e 298 representando clientes honestos.

As informações disponíveis para cada cliente e que servirão como parâmetros de avaliação são:

- Código do Cliente: cada cliente possui um registro único para evitar dados duplicados.
- Tempo de residência: representa o número de anos que o cliente reside no local. Toda a vez que é solicitada mudança de endereço e/ou mudança de nome do titular, inicia-se a contagem.
- Valor da última conta: representa, em R$, o valor da última conta do cliente em questão.
- Valor da conta média: representa, em R$, o valor médio da conta do cliente dos últimos 6 meses. Esse foi o período de tempo estipulado como ótimo e viável para a implementação deste modelo. Períodos menores poderiam estar sujeitos a ruídos de sazonalidade e períodos maiores demandariam um banco de dados com capacidades de armazenamento inviáveis.
- Região: representa a região de habitação do cliente, podendo ser Leste, Oeste, Norte e Sul.
- Forma de pagamento: representa se o cliente opta por pagamento da conta em débito automático (Sim/Não).
- Variação de valor da última conta: representa qual foi a variação em relação ao valor da última conta $\left(\frac{valor\ da\ conta\ no\ mês\ atual}{valor\ da\ conta\ no\ mês\ anterior}\right)$.
- Fraude: representa a informação de que o respectivo cliente é honesto ou se comete fraude (Sim/Não).

A base de dados foi dividida em duas partes (uma utilizada para treinar o modelo e a outra utilizada para testar a performance do modelo). Essa divisão foi feita de forma aleatória.

Na atividade de formulação do problema e implementação, utilizou-se a relação de variáveis, conforme a Tabela 1.

Tabela 1: Relação de variáveis utilizadas

X1	Região Leste (binária)
X2	Região Oeste (binária)
X3	Região Norte (binária)
X4	Região Sul (binária)
X5	Forma de Pagamento: Débito automático (binária)
X6	Tempo de Residência
X7	Valor da última conta
X8	Valor de conta média
X9	Variação do valor da última conta
Y	Classificação do cliente (Fraude/Honesto)

A validação dos modelos foi feita utilizando as matrizes de confusão geradas a partir dos conjuntos de treino e de validação. Essa análise permitiu uma comparação clara e objetiva de performance e eficiência dos métodos.

Na comparação de eficiência e performance do SVM, utilizou-se duas funções: Linear e Polinomial de grau 2. Em ambas as análises utilizou-se o custo adicional C (constante de penalização), para que o modelo fosse mais suscetível a possíveis ruídos na base de dados e, também, para que se pudesse controlar o grau de tolerância em relação a erros.

3.1 Modelo 1: Função Linear

A função objetiva (minimizar) utilizando função linear pode ser representada por

$$Z = \frac{1}{2}(w_1^{\,2} + w_2^{\,2} + w_3^{\,2} + w_4^{\,2} + w_5^{\,2} + w_6^{\,2} + w_7^{\,2} + w_8^{\,2} + w_9^{\,2}) + C\left(\sum_{i=1}^{N} \xi_i\right) \quad (6)$$

Para os testes, fez-se uso de dois valores para a constante de penalização: C_1=0,01 e C_2=10. Assumindo valor 0,01 para a constante de penalização C_1 o valor da função objetivo foi igual a 2 e para C_2 =10 a função objetivo apresentou valor igual a 1750.
Nas Tabelas 2 e 4 são apresentados os parâmetros estimados pelos modelos usando C_1=0,01 e C_2=10, para cada variável estudada e as matrizes de confusão para C_1=0,01 e C_2=10 são mostradas nas Tabelas 3 e 5, respectivamente.

Tabela 2: Coeficientes obtidos para C=0,01, SVM Linear

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	b
-0,0273	0,0005	0,1768	-0,15	-0,0159	-0,0426	0,0037	-0,013	0,0345	2,4488

Tabela 3: Matrizes de confusão obtidas para C=0,01, SVM Linear

Treino		*PARA*	
		Fraude	Honesto
DE	Fraude	79,90%	38,30%
	Honesto	20,10%	61,70%

Validação		*PARA*	
		Fraude	Honesto
DE	Fraude	78,50%	38,90%
	Honesto	21,50%	61,10%

Tabela 4: Coeficientes obtidos para C=10, SVM Linear

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	b
1,1068	0,7865	1,1256	-3,019	-0,0938	-0,0875	-0,0092	-0,001	0,5261	2,5926

Tabela 5: Matrizes de confusão obtidas para C=10, SVM Linear

Treino		*PARA*	
		Fraude	Honesto
DE	Fraude	74,50%	26,20%
	Honesto	25,50%	73,80%

Validação		*PARA*	
		Fraude	Honesto
DE	Fraude	68,50%	24,20%
	Honesto	31,50%	75,80%

Observa-se na Tabela 3, para os dados de treinamento, que dos classificados como fraudadores 79,9% eram, realmente, fraudadores e dos classificados como honestos 61,7% eram honestos, obtendo-se, portanto, uma eficiência total de 70,8%. Já para os dados de validação 78,5% eram fraudadores e 61,1% eram honestos, com eficiência total de 69,8%.
É possível verificar na Tabela 5, para os dados de treinamento, que 74,5% foram classificados como fraudadores e 73,8% como honestos, obtendo-se, uma eficiência total de 74,2%. Para os dados de validação 68,5% foram considerados fraudadores e 75,8% foram considerados honestos, com eficiência total de 72,1%.

3.2 Modelo 2: Polinômio de 2º grau

A função objetivo (minimizar), utilizando função polinomial de 2º grau, foi definida por

$$Z = \frac{1}{2}(w_1^2 + w_2^2 + w_3^2 + w_4^2 + w_5^2 + w_6^2 + w_7^2 + w_8^2 + w_9^2 + w_{10}^2 + w_{11}^2 + w_{12}^2 + w_{13}^2) + C\left(\sum_{i=1}^{N} \xi_i\right) \quad (7)$$

Em relação aos parâmetros w_{10}, w_{11}, w_{12} e w_{13}, eles são aplicados às variáveis: $x_{10} = x_6^2$, $x_{11} = x_7^2$, $x_{12} = x_8^2$ e $x_{13} = x_9^2$. Para os testes, utilizou-se uso, também dois valores para a constante de penalização: C_1=0,01 e C_2=10. Assumindo valor 0,01 para a constante de penalização C_1 o valor da função objetivo foi igual a 2 e para C_2 =10 a função objetivo apresentou valor igual a 1613.

Nas Tabelas 6 e 8 são apresentados os parâmetros estimados pelos modelos usando C_1=0,01 e C_2=10, para cada variável estudada e as matrizes de confusão para C_1=0,01 e C_2=10 são mostradas nas Tabelas 7 e 9, respectivamente.

Tabela 6: Coeficientes obtidos para C=0,01, SVM Polinomial

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	w_{13}	b
0,044	-0,02	0,134	-0,07	-0,015	0,056	0,012	-0,056	-0,004	-0,001	-0,001	0,002	0,041	1,807

Tabela 7: Matrizes de confusão obtidas para C=0,01, SVM Polinomial

Treino		*PARA*	
		Fraude	Honesto
DE	Fraude	74,50%	33,60%
	Honesto	25,50%	66,40%

Validação		*PARA*	
		Fraude	Honesto
DE	Fraude	67,10%	33,60%
	Honesto	32,90%	66,40%

Tabela 8: Coeficientes obtidos para C=10, SVM Polinomial

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	w_{13}	b
1,695	1,471	1,73	-4,896	-0,031	-0,255	0,053	-0,027	-1,507	0,002	-0,001	0	0,515	6,035

Tabela 9: Matrizes de confusão obtidas para C=10, SVM Polinomial

Treino		*PARA*	
		Fraude	Honesto
DE	Fraude	73,20%	22,10%
	Honesto	26,80%	77,90%

Validação		*PARA*	
		Fraude	Honesto
DE	Fraude	66,40%	24,20%
	Honesto	33,60%	75,80%

Verifica-se na Tabela 7, para os dados de treinamento, que dos classificados como fraudadores 74,5% eram, realmente, fraudadores e dos classificados como honestos 66,4% eram honestos, obtendo-se, portanto, uma eficiência total de 70,5%. Para os dados de validação, observa-se que 67,1% eram fraudadores e 66,4% eram honestos, com eficiência total de 66,8%.

Na Tabela 9, para os dados de treinamento, observa-se que 73,2% foram classificados como fraudadores e 77,9% como honestos, obtendo-se, uma eficiência total de 75,5%. Para os dados de validação 66,4% foram considerados fraudadores e 75,8% foram considerados honestos, com eficiência total de 71,1%.

3.3 Modelo 3: Análise Discriminante

Para a avaliação do desempenho do SVM fez-se sua comparação com o desempenho da análise discriminante linear. Por essa técnica, cria-se um hiperplano de separação linear, sendo seus parâmetros estimados pelo método dos mínimos quadrados ordinários. Assim, tem-se

$$\tilde{Y} = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + a_6X_6 + a_7X_7 + a_8X_8 + a_9X_9 + \beta \quad (8)$$

sendo a função objetivo (minimizar) obtida por:

$$Minimizar\ Z = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2 \quad (9)$$

em que, $Y_i = 1$ se a observação é um caso de fraude e $Y_i = 0$ caso contrário. Em relação ao ponto de corte, estabeleceu-se um valor de $\tilde{Y}i=0,5$, assim se o valor de $\tilde{Y}i$ for maior que 0,5 a observação será classificada como fraudulenta.

A Tabela 10 mostra os parâmetros estimados pelo modelo, usando análise discriminante linear, para cada variável estudada e na Tabela 11 é apresentada a matriz de confusão usando-se análise discriminante linear. Verifica-se, para os dados de treinamento, que 73,2% foram classificados como fraudadores e 73,2 como honestos 66,4%, apresentando eficiência total de 73,2%. Para os dados de validação, observa-se que 67,8% eram fraudadores e 72,5%, com eficiência total de 70,1%.

Tabela 10: Coeficientes obtidos para a Análise Discriminante Linear

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	b
0,9247	0,8932	1,0349	0,3905	-0,0147	-0,0124	-0,0037	0,0006	0,2548	0

Tabela 11: Matrizes de confusão obtidas utilizando Análise Discriminante Linear

Treino		*PARA*	
		Fraude	Honesto
DE	Fraude	73,20%	26,80%
	Honesto	26,80%	73,20%

Validação		*PARA*	
		Fraude	Honesto
DE	Fraude	67,80%	27,50%
	Honesto	32,20%	72,50%

3.4 Comparação entre as técnicas

A partir dos resultados obtidos fez-se a comparação entre a performance das técnicas analisando-se, conjuntamente, as matrizes de confusão obtidas.

Percebe-se que ao utilizar uma constante de penalização de baixo valor, os modelos estudados, utilizando-se o SVM, mostraram uma maior eficiência para classificação de clientes fraudulentos do que àquela apresentada na classificação de clientes honestos, tanto no conjunto de treino como no conjunto de validação (por exemplo, 79,9% vs 61,7%; 74,5% vs 66,4%). Já o modelo utilizando a técnica de Discriminante Linear mostrou uma mesma eficiência na classificação do conjunto de testes, sendo mais eficiente na classificação de clientes honestos (72,5% vs 67,8%) no conjunto de validação.

Com relação à eficiência total, tanto no conjunto de treino, como no conjunto de validação, o modelo utilizando a Análise Discriminante mostrou-se mais eficaz. Entre os modelos utilizando a técnica do SVM, em ambos os conjuntos, o modelo Linear apresentou maior eficiência global.

Apesar dos modelos utilizando o SVM apresentarem menores eficiências globais, mostraram-se mais adequados ao presente estudo, pois o principal objetivo de uma

distribuidora de energia elétrica é identificar, dentro de seu banco de clientes, quais são os clientes que comentem fraudes e, uma vez identificado o fraudador, a empresa apresenta gastos com fiscalização e processos de acusação. Assim, os melhores modelos são aqueles que apresentam uma alta taxa de acertos ao identificar clientes fraudadores.
Em relação à constante de penalização, aumentando o seu valor os resultados tornaram-se piores, diminuindo-se a eficiência na identificação dos clientes fraudulentos. Em contrapartida, melhorou-se a performance na identificação de clientes honestos, dentro da base de dados. Assim, os resultados sugerem o grande poder de generalização dessa técnica.
A Figura 6 mostra, para todos os modelos estudados, a comparação de eficiência global para a identificação de fraudes. Verifica-se que o modelo utilizando o SVM, com função linear e constante de penalização C=0,01, provavelmente, apresenta-se como o mais indicado para a classificação de clientes de uma distribuidora de energia elétrica e a Figura 7 mostra, a classificação dos modelos estudados, considerando-se apenas a eficiência global.

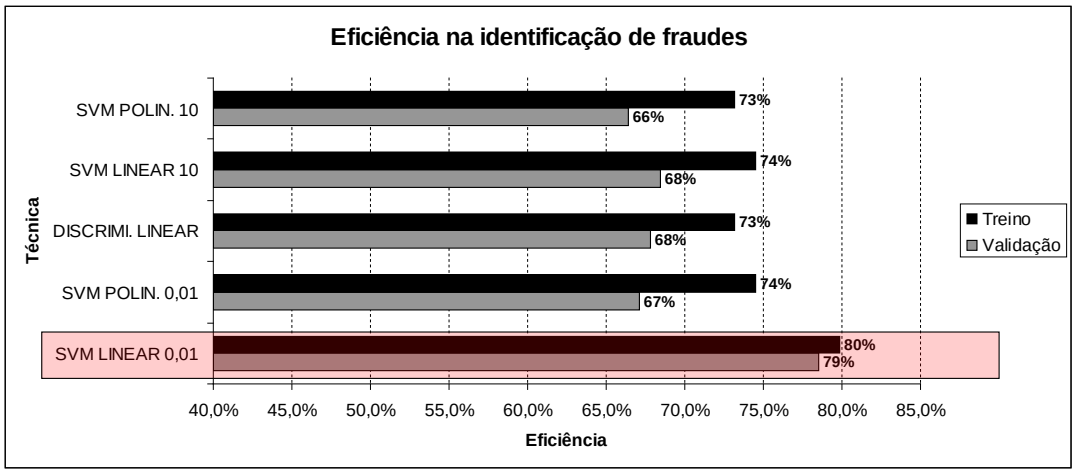


Figura 6: Comparação de eficiência para identificação de fraudes

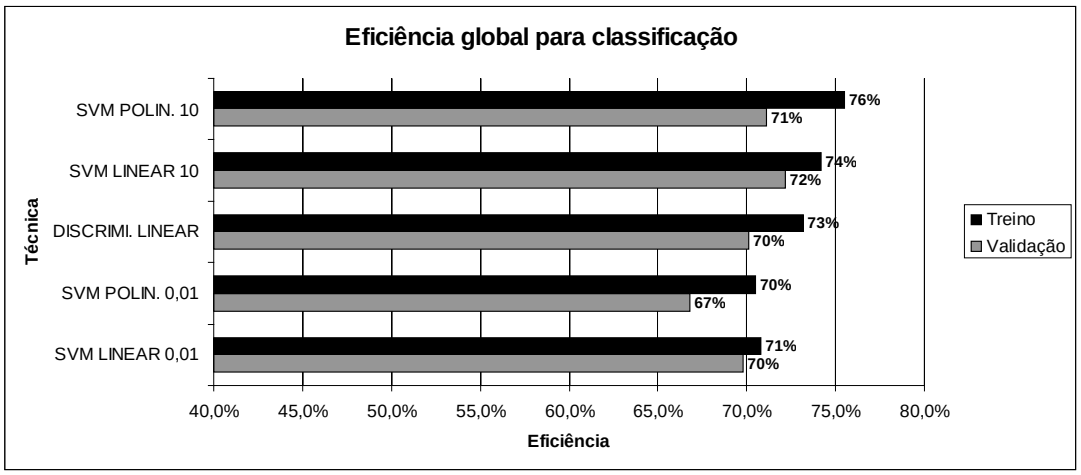


Figura 7: Eficiência global dos modelos (treino e validação)

4 Conclusões

O presente trabalho mostrou-se de grande valor ao sugerir que a técnica do SVM é uma abordagem bastante atrativa para ser utilizada na modelagem de fenômenos onde o objetivo é de reconhecimento de padrões. Os resultados obtidos demonstraram que essa técnica possui uma boa performance, sendo uma alternativa viável à Análise Discriminante Linear que é a técnica de classificação mais utilizada.
É importante, também, mencionar a relação existente entre a qualidade dos resultados obtidos e a natureza da base de dados. Modelos baseados em técnicas de aprendizado são totalmente dependentes da qualidade da base de dados, que é utilizada para treino. A utilização de uma formulação extremamente complexa e correta, em uma base de dados apresentando incoerências e ruídos, proporcionaria resultados não eficientes. Assim, devido a bom desempenho dos modelos, pode-se concluir que a base de dados fornecida pela EMPRESA foi satisfatória.
Como sugestão de trabalhos futuros, pretende-se aplicar o SVM utilizando funções mais complexas, como a "Multi-Layer Perceptron" e "Exponential Radial Basis". Esses modelos podem apresentar respostas mais eficientes, tanto relacionadas à performance global quanto à performance na identificação de clientes fraudadores. Pretende-se, também, investigar a relação entre a performance dos modelos utilizando o SVM e a constante de penalização, C, buscando determinar o melhor valor a ser utilizado no modelo.

5 Referências

Bin Z., Yong L. e Shao-Wei, X. (2000) Support Vector Machine and its Application in Handwritten Numeral Recognition, 15th International Conference on Pattern Recognition (ICPR'00), Vol 2, pp. 2720-2734.

Boccuzzi, C. V. (2005) Combate a fraude de energia, Anais do I Workshop Contra Furto e Fraude de Energia e Roubo de Condutores e Equipamentos, Curitiba-PR.

Bradley, P. S. e Mangasarian, O. L. (2000) Massive data discrimination via linear support vector machines, Optimization Methods and Software, Vol 13, pp. 1–10.

Byvatov E. e Schneider G. (2003) Support vector machine applications in bioinformatics. Applied Bioinformatics, Vol 2, No 2, pp. 67-77.

Gunn, S. R. (1998) Support Vector Machine for classification and regression, University of Southhampton, VA.

Guyon, I., Weston, J., Barnhill, S. e Vapnik, V. N. (2002) Gene selection for cancer classification using support vector machines, Machine Learning, Vol 46, pp. 389–422.

Scarpel, R. A. (2005) Utilização de Support Vector Machine em previsão de insolvência de empresas, Anais do XXXVII Simpósio Brasileiro de Pesquisa Operacional, pp. 671-677.

Vapnik, V. N. (1998) Statistical learning theory, John Wiley & Sons, NY.

Vapnik, V. N. (1999) An overview of Statistical learning theory, IEEE transactions on neural networks, Vol 10, No 5, pp. 988-999.