Universidade Federal do Maranhão
coordenação de Pós-Graduação em Engenharia Elétrica

Dissertação de Mestrado

DIAGNÓSTICO DE DIABETES TIPO II POR CODIFICAÇÃO EFICIENTE E MÁQUINA DE VETOR DE SUPORTE

ÁUREA CELESTE DA COSTA RIBEIRO

São Luís - MA, Brasil

30 de julho de 2009

Sumário

DIAGNÓSTICO DE DIABETES TIPO II POR CODIFICAÇÃO EFICIENTE E MÁQUINAS DE VETOR DE SUPORTE

Dissertação de mestrado submetida à coordenação do Curso de Pós-Graduação de Engenharia de Eletricidade da UFMA como parte dos requisitos para obtenção do título de Mestre em Engenharia de Eletricidade na área de Automação e Controle.

ÁUREA CELESTE DA COSTA RIBEIRO

JUNHO, 2009

Ribeiro, Áurea Celeste da Costa

Diagnóstico de diabetes tipo II por codificação eficiente e máquinas de vetor de suporte/ Áurea Celeste da Costa Ribeiro. – São Luís, 2009.
51 f.

Dissertação (Mestrado em Engenharia Elétrica) – Programa de Pós Graduação em Engenharia de Eletricidade, Universidade Federal do Maranhão, 2009.
1. Diabetes Mellitus (Tipo II) – diagnóstico. 2. Controle automático. I. Título.

CDU 616.379-008.64:62-52

DIAGNÓSTICO DE DIABETES TIPO II POR CODIFICAÇÃO EFICIENTE E MÁQUINA DE VETOR DE SUPORTE

Aurea Celeste da Costa Ribeiro

Dissertação aprovada em 30 de junho de 2009.

Prof. Allan Kardec Duailibe Barros Filho, Ph. D.
(Orientador)

Prof. Guilherme de Alencar Barreto, Dr.
(Membro da Banca Examinadora)

Prof. Sofiane Labidi, Dr.
(Membro da Banca Examinadora)

DIAGNÓSTICO DE DIABETES TIPO II POR CODIFICAÇÃO EFICIENTE E MÁQUINA DE VETOR DE SUPORTE

MESTRADO

Área de Concentração: AUTOMAÇÃO E CONTROLE

ÁUREA CELESTE DA COSTA RIBEIRO

Orientador: Prof. Dr. Allan Kardec Duailibe Barros Filho

Curso de Pós-Graduação

em Engenharia de Eletricidade da

Univesidade Federal do Maranhão

.

A Deus, Neilson e João Gabriel

que alegram minha vida e incentivam-me a ir sempre em frente.

AGRADECIMENTOS

A Deus que esteve comigo em todos os momentos , permitindo tudo ser possível .

Ao meu orientador Allan Kardec Duallibe Barros Filho pela oportunidade, incentivo, crédito, orientação e paciência do início ao final do curso.

Ao meu esposo, melhor amigo e companheiro, Neilson Mendes Mousinho, por sua dedicação, seu companheirismo, apoio, amizade e amor incondicionais.

Aos meus pais João Raimundo de Souza Ribeiro e Júlia de Fátima da Costa Ribeiro e aos meus irmãos, pela incondicional dedicação a minha vida .

A minha amada avó e grande amiga Laila da Silva Ribeiro, por sempre ter me apoiado em cada passo da minha vida com afinco .

Aos meus amigos do PIB: Sid, Cris, Deusdete, Daniel, Anderson, Belinha, Euler, Aryfrance, Ana Lúcia, Henrique, Orlando, Vicente, Enio, Ewaldo, Éder Jr., Flávio, Márcio, André, Denner, Fausto, Mendes, Fábio, pela amizade e comentários oportunos.

Aos demais amigos da UFMA e colegas de outros laboratórios pela companhia e ajuda oportuna.

A todos aqueles que direta e indiretamente apiaram-me e participaram comigo em algum tempo neste período.

À CAPES pela bolsa a mim concedida.

Resumo

Diabetes é uma doença causada pela falência do pâncreas em produzir insulina, é incurável e seu tratamento é baseado em dietas, exercícios e remédios. Os custos com o tratamento, diagnóstico na população e combate à doença tornam-se cada vez mais altos. Sistemas de auxílio ao diagnóstico da doença são uma das soluções para ajudar na diminuição dos custos com a doença.

Nosso método propõe um sistema de auxílio de diagnóstico baseado nas máquinas de vetor de suporte para uma classe e na codificação eficiente através da análise de componentes independentes para classificar uma base de dados de pacientes em diabéticos e não-diabéticos.

Primeiramente, foram feitos testes de classificação com as características não-invasivas e invasivas da base de dados juntas. Em seguida, fizemos um teste sem as características invasivas da base de dados, que são glicose e insulina em jejum, que são feitas com a coleta sanguínea. Obteve-se uma taxa de acurácia de 99,84% e 99,28%, respectivamente. Outros testes foram feitos sem as características invasivas, tirando uma característica não-invasiva por vez, com o fim de observar a influência de cada uma no resultado final.

Palavras-chaves: Diabetes, Diagnóstico, Redundância, Codificação Eficiente, Máquinas de vetor de suporte para uma classe.

Abstract

Diabetes is a disease caused by the pancreas failing to produce insulin. It is incurable and its treatment is based on a diet, exercise and drugs. The costs for diagnosis and human resources for it have become high and inefficient. Computer-aided design (CAD) systems are essential to solve this problem.

Our study proposes a CAD system based on the one-class support vector machine (SVM) method and the efficient coding with independent component analysis (ICA) to classify a patient's data set in diabetics or non-diabetics.

First, the classification tests were done using both non-invasive and invasive characteristics of the disease. Then, we made one test without the invasive characteristics: plasma glucose concentration and 2-Hour serum insulin (mu U/ml), which use blood samples. We have obtained an accuracy of 99.84% and 99.28%, respectively. Other tests were made without the invasive characteristics, also excluding one non-invasive characteristic at a time, to observe the influence of each one in the final results.

Keywords: Diabetes. classification. Redundancy. Efficient Coding and One-Class SVM.

LISTA DE ABREVIATURAS E SIGLAS

- AC Algoritmo do projeto StatLog que oferece interação com o usuário via gráficos.
- ANFIS Adaptive-Network-Based Fuzzy (*Redes Adaptativas baseadas em lógica Fuzzy*)
- ARTMAP-IC Um tipo de arquitetura de redes neurais.
- ALLOC80 Algoritmo estatístico baseado em estimação da densidade.
- BNN Bayesian neural nets (*Redes neurais baysianas*)
- Baytree Algoritmo do projeto StatLog baseado em árvores de decisão na abordagem Baysiana.
- Backprop Algoritmo Backpropagation.
- CAEP Classification by aggregating emerging patterns (*Surgimento de padrões por classificação através de agregação*)
- CE Codificação eficiente.
- CI Componentes independentes.
- C4.5 Um dos algoritmos do projeto StatLog baseado em árvores de decisão.
- Cal5 Algoritmo do projeto StatLog projetado para atributos contínuos e discretos, acrescentando um sub-algoritmo capaz de manipular os atributos discretos desordenados.
- CAR Class association rules (*Regra de associação de classes*)

- CART Um dos algoritmos do projeto StatLog que significa: Classification and Regression Tree (Árvore de classificação e regressão).
- CASTLE Algoritmo de abordagem baysiana.
- CN2 Algoritmo do projeto StatLog de regras de aprendizado batizado de CN por causa de seus criadores Clark e Nilblett.
- CWN Continuous-weight networks (*Redes de pesos contínuos*)
- DMSK Differential minimum shift keying (*Deslocamento mínimo diferencial keying*)
- Datamind Software baseado em regras de decisão.
- DIPOL92 Algoritmo híbrido que usa discriminação logística e métodos estatísticos não-paramétricos.
- Discrim Algoritmo LDA (Linear discriminant analysis) em português *análise discriminante linear*.
- DWN Discrete weight networks (*Redes de pesos discretos*)
- GRNN Generalized regression neural network (*Redes neurais de regressão generalizada*)
- Gnosis Ferramenta para desenvolvimento de redes polinomiais.
- ITI Incremental Decision Tree Induction (*Indução de árvore de decisão incrementais*)
- ICA Independent component analysis (*Análise de componentes independentes*)
- ITrule Algoritmo do projeto StatLog baseado em árvores de decisão.

- IndCART Algoritmo derivado do algoritmo CART.
- NewID Um dos algoritmos do projeto StatLog baseado em árvores de decisão.
- KnowledgerMiner Ferramenta de desenvolvimento para redes polinomiais.
- Kohonen Kohonen network algorithm (*Algoritmo redes de kohonen*).
- K-NN K-NEAREST NEIGHBOUR (*K-vizinho mais proxímo*).
- LMDT Linear Machine Decision Tree (*Árvore de decisão de máquina linear*)
- LVQ Learning Vector Quantizers (*Quantizadores de vetor de aprendizado*).
- Logdisc Logistic discrimination (Discriminação Logítica)
- MQ Expert Ferramenta para construção de redes polinomiais.
- MARS Multivariate Adaptive Regression Splines (*Regressão adaptativa multivariada splines*)
- MSDD Multi-Stream Dependency Detection (Detecção de dependência por Multi-stream).
- MFN Multiplier-free feedforward networks (Redes de transmissão de multippicação livre)
- NewID Algoritmo do projeto Statlog baseado em árvores de decisão.
- NaiveBay Algoritmo do projeto StatLog baseado em árvores de decisão.
- NeuroShell2 Software experimental com arquitetura das redes neurais com interface windows.
- OMS Organização mundial de saúde
- OLQV Algoritmo derivado do LQV.

- PCA Principal component analysis (*Análise de componentes principais*)
- PolyNet Ferramenta para desenvolvimento de redes polinomiais.
- PcOLPARS Software com vários algoritmos utilizado para a construção de uma rede neural.
- PRW Software com vários algoritmos utilizado para a construçào de uma rede neural.
- PPR Projection Pursuit Regression(*Regressão de busca da projeção*).
- PSA Penalized log likelihood smoothing spline analysis of variance (Análise da variância pela probabilidade logaritmica com o método de suavização spline)
- Quaddisc Algoritmo QDA (Quadratic discriminant analysis) em português *análise discriminante quadrática.*
- Regres. Logística Regresão Logística
- RBF Radial bases function (Função de base radial)
- RNA ADAP hebbian Redes neurais artificiais adaptativas usando aprendizado Hebbiano
- RNA Redes neurais artificiais
- BNN Bayesian neural net (*Rede neural baysiana*).
- S-plus Pacote estatístico comercializado pela empresa Insightful Corporation.
- See5 Algoritmo baseado em árvores de decisão.
- Scenario Algoritmo baseado em árvore de decisão.

- SVM -One class Support vector machine- One class(*Máquinas de vetor de suporte para uma classe*)
- SMART Algoritmo estatístico.
- StatLog Statistical log analyzer (*Análise estatística por algoritmo*).
- WizWhy É uma ferramenta de mineração de dados que análisa os dados e faz previsões sobre estes, é baseada em regras de decisão.

Capítulo 1

Introdução

O diabetes é uma doença causada pela falência do pâncreas em produzir insulina, ou alternativamente, quando o corpo não pode usar efetivamente este hormônio. É incurável e seu tratamento é baseado em dietas, exercícios físicos e remédios. A Organização mundial de saúde (OMS) estima que mais de 180 milhões de pessoas no mundo tem a doença e este número é projetado perto dos 366 milhões de pessoas para 2030 [1].

Existem vários tipos de manifestações do diabetes melitus no ser humano. Por isso há diferentes classificações da doença, entre as mais comuns estão o diabetes tipo 1 (DM1), diabetes tipo 2 (DM2) e o diabetes gestacional.

O DM1 surge quando o organismo deixa de produzir a insulina ou produz uma pequena quantidade deste hormônio, é mais comum em pessoas com menos de 35 anos, mas pode surgir em qualquer idade. O DM2 tem como peculiaridade a contínua produção de insulina pelo pâncreas, que por consequência ocorre uma anomalia denominada "resistência Insulínica", esta impede que as células metabolizem glicose suficiente da corrente sangüínea, além disso o DM2 possui um fator hereditário maior do que no tipo 1 e uma grande relação com a obesidade e o sedentarismo, sua incidência é maior após os 40 anos. O diabetes gestacional é a alteração das taxas

de açúcar no sangue que aparece ou é detectada pela primeira vez na gravidez. Pode persistir ou desaparecer após o parto [3].

Apesar do crescimento de dados sobre esta doença em bases de dados de hospitais e de institutos médicos no mundo todo, os métodos tradicionais de análise de dados existentes tornam-se ineficientes. Geralmente, estes dados são utilizados somente para chegar à uma quantidade aproximada de doentes com o fim de prever os gastos dos governos com medicamentos, programas de prevenção e relatórios a organizações mundiais, este último, como faz a OMS. Métodos computacionais podem transformar eficientemente estes dados em informação útil para o diagnóstico da doença, prevenção e controle.

Este trabalho tem o foco sobre a classificação de pacientes em diabéticos e não-diabéticos do tipo 2 por meio de dados clínicos relevantes à classificação, estes estão em uma base de dados pública. Para isto, foram utilizados dois métodos computacionais; um para a extração das características, no qual foi utilizado o método de codificação eficiente através da análise das componentes independentes e outro para a classificação dos pacientes em diabéticos e não diabéticos, que foi classificador de máquinas de vetor de suporte para uma classe, poderemos visualisar melhor esta metodologia no diagrama de blocos da figura 1.1.

A base de dados utilizada para teste do método já foi utilizada por alguns pesquisadores com seus respectivos métodos de classificação utilizados para fazer a separação de classes: Diabéticos e não-diabéticos, como pode-se observar na tabela 1.1. Todos os métodos desta tabela foram aplicados utilizando a base de dados completa, a base de índios Pima.

1.1 Complicações do diabetes melitus

Há algumas décadas as doenças do aparelho circulatório são a primeira causa de morte em vários países do mundo. Segundo registros oficiais só no Brasil é a primeira em casos de morte. Dimensionando, em 2000, correspondeu a mais de 27 %

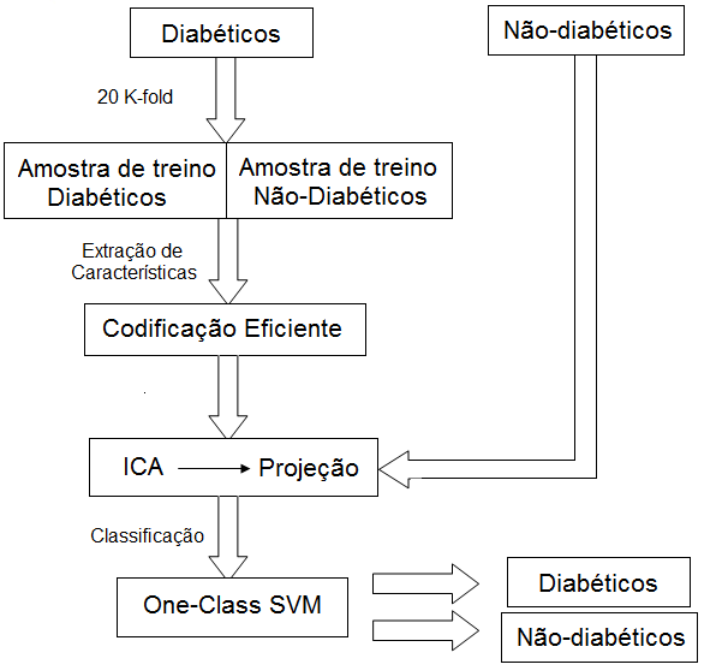

Figura 1.1: Fluxograma da metodologia proposta para a classificação dos pacientes: Na primeira fase usa-se um classificador para uma classe, separa-se as amostras de diabéticos em treino e teste pelo método de validação cruzada 20k-fold, extraem-se as características pela codificação eficiente por meio de ICA, as funções bases obtidas do conjunto de treino são projetadas no conjunto de treino e teste de diabéticos e não-diabéticos. A última fase diz respeito à classificação das amostras em diabéticos e não-diabéticos.

Tabela 1.1: tabela contendo pesquisadores e seus respectivos métodos utilizados na base de dados de índios Pima completa

Pesquisador (ano)	Acurácia	Método
Smith , Everhart et al. (1998) [27]	76,00%	RNA ADAP hebbian
Quinlan (1993)	71,10%	C4.5
Wahba, Gu et al. (1992)	72,00%	PSA
Michie, Spiegelhalter et al. (1994)	77,50%	Algoritmo Discrim
	73,80%	Algoritmo Quaddisc
	77,70%	Algoritmo Logdisc
	76,80%	Algoritmo SMART
	69,90%	Algoritmo ALLOC80
	67,60%	Algoritmo k-NN
	74,20%	Algoritmo CASTLE
	74,50%	Algoritmo CART
	72,90%	Algoritmo IndCART
	71,10%	NewID
	72,40%	AC
	72,90%	Baytree
	73,80%	NaiveBay
	71,10%	CN2
	73,00%	C4.5
	75,50%	Itrule
	75,00%	Cal5
	72,70%	Kohonen
	77,60%	Dipol92
	75,20%	Backprop
	75,70%	RBF
	72,80%	LVQ
Oates (1994)	71,33%	(MSDD)
Bioch, van der Meer et al. (1996)	79,50%	BNN
Ripley (1996)	80,20%	Regres. Logística
	77,40%	Modelo MARS e PPR
	77,00%	RN
	75,30%	k-NN para k =9
	78,90%	OLVQ
	75,60%	CART
Carperter e Markuzon (1998)	77,00%	K-NN
	77,00%	regressão logística

	66,00%	ARTMAP-IC
Khan (1998)	78,00%	(MFN)
	76,90%	DWN
	78,40%	CWN
Eklund e Hoang (1998)	71,02%	C4.5
	71,55%	regra C4.5
	73,16%	ITI
	73,51%	LMDT
	72,19%	CN2
Liu (1998)	75.50%	CAR
King, Elder IV et al. (1998)	76,00%	CART
	30,00%	Scenario
	73,00%	See5
	79,00%	S-Plus
	74,00%	WizWhy
	69,00%	DataMind
	67,00%	DMSK
	77,00%	NeuroShell2
	81,00%	PcOlPARS
	80,00%	PRW
	77,00%	MQ Expert
	78,00%	PolyNet
	81,00%	Gnosis
	78,00%	KnowledgerMiner
Wong (2000)	75,00%	CAEP
Kayaer et al. (2005)	80,47%	GRNN
Barakat et al. (2005)	82,00%	rule-extraction from SVMs
Polat et al. (2007)	89,47%	PCA-ANFIS

do total de óbitos; ou seja, naquele ano 255.585 pessoas morreram em consequência de doenças do aparelho circulatório[5].

O diabetes mellitus constitui um dos principais fatores de risco para as doenças do aparelho circulatório. Entre as suas complicações mais frequentes encontram-se o infarto agudo do miocárdio, o acidente vascular cerebral (AVC), a insuficiência renal crônica, a insuficiência cardíaca, as amputações de pés e pernas, a cegueira definitiva, abortos e mortes perinatais.

A identificação precoce dos casos e o estabelecimento do vínculo entre os portadores da doença e as unidades de saúde são elementos imprescindíveis para o sucesso do controle desses agravos. O acompanhamento e o controle do diabetes

mellitus no âmbito da atenção básica pode evitar o surgimento e a progressão de complicações, reduzindo o número de internações hospitalares, bem como a mortalidade devido a esses agravos.

No Brasil, como exemplo, as doenças cardiovasculares são responsáveis por 1.150.000 das internações/ano, com um custo aproximado de 475 milhões de reais, sendo que nestes números não estão inclusos os gastos com procedimentos de alta complexidade.

Além dos custos financeiros, o diabetes também acarreta um prejuízo social, já que é responsável pelo aumento da mortalidade precoce e por muitas incapacitações. Um estudo realizado em 1987 em nove capitais brasileiras [1] mostrou que metade dos portadores de diabetes desconhece sua condição. Outros estudos [2][4] mostraram que os portadores de diabetes já apresentavam complicações microvasculares ao diagnóstico.

1.2 Trabalhos realizados com bases de dados de diabéticos

Existem muitos trabalhos utilizando bases de dados de diabéticos, estes são utilizados como ferramenta de auxílio ao tratamento, observação da doença em hospitais e centros de controle e prevenção da doença. Dentre estes trabalhos que utilizaram bases de dados de diabéticos destacam-se: Michel e Begin, que em 1994, utilizaram uma dessas bases para consulta da doença [6]; em 1995, Flack et al. para melhorar a comunicação entre diabéticos e médicos; em 1996, Kolpelman e Sanderson para fornecer contínua melhoria na qualidade dos cuidados e prevenção do diabetes. Em 1998, Pogach e Hawley, alunos de administração desenvolveram registros em diabetes de uma base de dados de uma farmácia ambulatorial e dos números do seguro social de administradores, encontraram 139.646 administradores com diabetes.

A base de dados de diabetes Belga foi criada pela secretaria do país. Esta

retorna relatórios de todos os casos incidentes de diabetes do tipo 1 e de seus familiares de primeiro grau com idade inferior a 40 anos. Desde então, estudos epidemiológicos e genéticos tem sido facilitados [7]. Um hospital britânico relacionou seus 7000 pacientes a uma base de dados central da secretaria nacional de serviços de saúde para identificar mortalidade e identificou que o diabetes agravou em 36% dos atestados de óbitos [8].

Particularmente, o diabetes é uma doença oportuna para trabalhar com a extração de informação de bases de dados por várias razões. Primeiro, porque há um grande número de dados disponível para tratamento da informação. Segundo, porque a diabetes é uma doença muito comum na população e seu custo para os governos é muito alto, e por isso tem atraído vários gestores e contribuintes na busca por poupar dinheiro. Terceiro, porque a diabetes pode trazer muitas complicações como cegueira, isuficiência renal e etc, como já foi dito.

Neste trabalho propõem-se um método para a classificação dos pacientes em diabéticos e não-diabéticos. Desta maneira, espera-se contribuir para a classificação do diabetes e possíveis estudos de outras opções de classificação da doença com a ajuda de métodos computacionais, como os utilizados por vários pesquisadores na tabela 1.1 e por outros institutos como descrito neste texto.

1.3 Organização do texto

Este trabalho está dividido na seguinte forma:

No capítulo 2, são descritos os fundamentos teóricos utilizados neste trabalho. Primeiro a base de dados utilizada, segundo o método de extração de características conhecido como Codificação eficiente através da análise de componentes independentes e finalmente descreve-se o classificador utilizado que é o One-Class SVM (*máquinas de vetor de suporte para uma classe*). Finalmente, o capítulo 3 descreve a metodologia e os resultados obtidos. Por fim, o capítulo 7 traz as discussões dos resultados apresentados.

Capítulo 2

Fundamentos teóricos

2.1 Descrição da Base de Dados

2.1.1 Considerações Gerais sobre a Base de Dados

Os índios Pima são nativos americanos e habitam no centro-sul do Arizona ao longo dos rios Gila e Salt. Estes têm a mais alta prevalência em diabetes tipo 2 no mundo, muito mais do que o observado em outras populações dos Estados Unidos. Apesar de não terem um maior risco do que outras tribos, os índios Pima têm sido objeto de estudo intensivo para o diabetes, em parte porque formam um grupo homogêneo.

O grande aumento da prevalência do diabetes neste povo foi hipotetizado como resultado da interação da predisposição genética com uma súbita mudança da dieta tradicional de produtos agrícolas para alimentos industrializados no século passado. Para comparação, existem Pima geneticamente semelhantes no méxico que tem taxa zero em casos de diabetes tipo 2. Os índios do méxico e os índios do Arizona foram separados por questões ambientais, uns migraram para o Arizona e outros para o México.

O Instituto Nacional de doenças digestivas e de diabetes no Canadá disponibilizou uma base de dados de índios Pima diabéticos (Pima Indian Diabetes Database) ao UC-Irvine Machine Learning repository [9] em 1990, desde então esta

têm sido extensivamente utilizada como visto na tabela 1.1.

2.2 Base de Dados dos índios Pima

Os dados são constituídos de 768 pacientes do sexo feminino com idade acima de 21 anos. A base de dados consiste de 8 variáveis clínicas que são a seguir listadas:

1 Número de vezes em que a mulher engravidou

2 Concentração de glicose no sangue (mg/dl)

3 Pressão Arterial (mmHg)

4 Medida do triceps (mm)

5 Quantidade de insulina em 2 hs de jejum (mu U/ml)

6 IMC (Índice de Massa corporal)=(Peso em Kg/altura em m^2)

7 Função de ocorrência de casos da doença na família

8 Idade (anos)

sendo, que existem duas classes, representadas na base de dados pela nona variável com descrição:

- 0- Não-Diabéticos, com um número de pacientes igual a 500;
- 1- Diabéticos, com um número de pacientes igual a 268.

Estas variáveis clínicas podem ser divididas em não-invasivas e invasivas. Assim tem-se como evasivas as características 1, 3, 4, 6, 7 e 8 e as invasivas 2 e 5.

Vale ressaltar que esta é uma base de dados desbalanceada. No domínio de classificação, isso significa que existem muito menos casos de algumas classes do que de outras[23]. Classificadores desenvolvidos com métodos tradicionais são sensíveis

a este tipo de desbalanceamento e tendem a valorizar classes predominantes e a ignorar classes de menor representação (também chamadas de classes raras) [24].

2.3 Informações úteis dos dados

Tabela 2.1: tabela referente a valores de informações úteis da base de dados

N. do atributo	Média	Desvio Padrão	MIN/MAX
1	3,8	3,4	0/17
2	120,9	32,0	0/199
3	69,1	19,4	0/122
4	20,5	16,0	0/99
5	79,8	115,2	0/846
6	32,0	7,9	0/67,1
7	0,5	0,3	0,078/2,42
8	33,2	11,8	21/81

2.4 Codificação eficiente ou esparsa (CE)

2.4.1 Introdução

O método de Codificação Eficiente foi proposto por Horace Barlow em 1961, como sendo uma das supostas estratégias utilizadas pelo cérebro para representar a informação sensorial. É um modelo teórico para a codificação das informações sensoriais pelo sistema nervoso. Para Barlow, um modelo eficiente é aquele que minimiza a quantidade de impulsos nervosos utilizados para transmitir a informação desejada.

Barlow foi inspirado pela teoria da informação, definindo que os caminhos neurais percorridos pela informação sensorial são similares a canais de comunicação. Através de conceitos desta teoria, como capacidade de canal e redundância, Barlow, então sugeriu que a codificação neural é realizada maximizando a capacidade do canal e reduzindo redudâncias na informação transmitida.

O conceito de codificação eficiente pode ser descrito, biologicamente, em duas perspectivas: Para uma única célula com a idéia de esparsidade, e para múltiplas

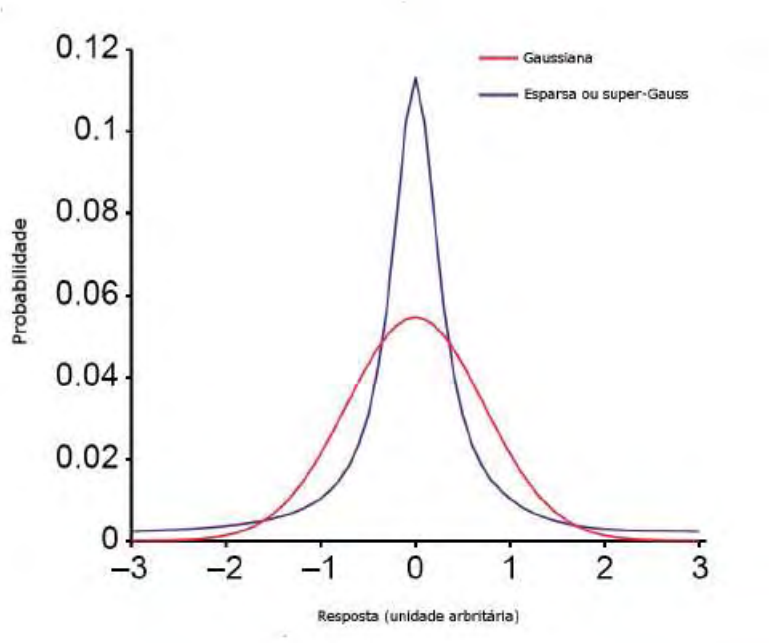

Figura 2.1: Comparação entre a distribuição esparsa e gaussiana. Uma distribuição esparsa é caracterizada por possuir um pico em zero e longas extremidades

células, com a idéia de redução de redundâncias. Na primeira perspectiva, o padrão de resposta de um neurônio é esparsamente distribuído.

Para um melhor entendimento desta idéia de esparsidade considera-se uma distribuição de probabilidade esparsa, que pode ser caracterizada por possuir um pico em zero e longas extremidades. As longas extremidades fazem com que as probabilidades de respostas a célula sejam pequenas. Desta forma, para um dado estímulo, é provável que apenas um pequeno conjunto de células respondam, como visto na figura 2.1.

Baddeley fez um experimento com primatas que demonstrou distribuições esparsas quando células do cortex visual primário eram estimuladas com sequências de imagens [13], o mesmo comportamento foi demonstrado por experimentos feitos por Deweese com células do cortex auditivo primário. Estas últimas células podem produzir um único disparo em resposta a um estímulo sonoro, este comportamento foi caracterizado como codificação binária porque estas células produzem 0 ou 1 em

resposta a um estímulo. No entanto, a probabilidade de disparo é muito pequena ao longo do tempo, o que é consistente como conceito de esparsidade, que parece ser um dos princípios da codificação neural [14].

Na segunda perspectiva, a representação da informação sensorial é eficiente se as respostas neurais forem estatisticamente independentes. Desta forma, não existe "informação redundante"entre as células da população. De acordo com experimentos de Hubel e Wiesel células neurais emitem fortes respostas quando estimuladas com estruturas não redundantes, como barras e bordas, assim é sugerido que estímulos visuais são representados de forma eficiente no cortex visual primário [**?**].

2.4.2 Extração de características por CE

O uso de estatística para a extração de características tem sido influenciado pelo modelo de processamento do estímulo da informação neural [11]. Estudos em Neurociência sugerem que os neurônios processam o estímulo da informação de acordo com o conceito de Codificação Eficiente. Neste conceito as respostas dos neurônios são estatisticamente independentes entre si. Isto significa que não há informação redundante, como visto anteriormente.

O objetivo computacional da codificação eficiente é extrair dos padrões um código compacto que consiga reduzir a redundância nos padrões com o mínimo de perda de informação. Os dados são transformados por um conjunto de filtros lineares $\mathbf{W}^{-1}$, de saída $\mathbf{x}$, no modelo matemático:

$$\mathbf{x} = \mathbf{W}^{-1}.\mathbf{s} \tag{2.1}$$

em que $\mathbf{s}$ é uma estimativa das componentes independentes. Um método para derivar este código eficiente no modelo da equação acima é chamado de Análise de Componentes Independentes.

2.5 Análise de componentes independentes

2.5.1 Introdução

A análise de componentes independentes, do inglês independent component analysis (ICA), é uma técnica baseada no modelo de independência, que é um requisito do código eficiente [15], esta técnica foi desenvolvida para atender ao requisito de trabalhar com componentes independentes e não-gaussianas.

Foi difundido e popularizado pela resolução do problema de separação de fontes cegas (BSS, blind source separation), onde o problema está na estimação da saída de uma fonte conhecida, sendo que esta fonte recebe vários sinais misturados e desconhecidos.

Esta técnica tem sido aplicada em áreas como áudio, radar, comunicação móvel, engenharia biomédica e outras. Como a técnica é baseada no modelo de independência, as fontes devem ser mutuamente estatisticamente independentes, definiremos independência estatística para entender o modelo de ICA.

2.5.2 O que é independência estatística

Duas variáveis aleatórias s_1 e s_2 são estatisticamente independentes se a partir de s_1 não é possível estimar ou inferir algum valor ou informação de s_2. Exemplos simples e rotineiramente utilizados de variáveis aleatórias independentes são sinais de eletrocardiograma de um feto e ruídos em sistemas de comunicação[**?**].

Matematicamente, a independência estatística de s_1 e s_2 ocorre satisfazendo-se a seguinte condição :

$$P_{s_1,s_2}(s_1, s_2) = P_{s_1}(s_1)P_{s_2}(s_2) \tag{2.2}$$

Assim, a probabilidade conjunta de duas variáveis estatísticamente independentes pode ser calculada apenas como o produto das marginais.

2.5.3 Modelo do método de ICA

O modelo de ICA é similar ao modelo de codificação eficiente de acordo com a equação (3.1) . De uma forma geral, cosideremos que $\mathbf{x} = [x_1; x_2; \ldots; x_n]^T$ e $\mathbf{s} = [s_1; s_2; \ldots; s_n]^T$ são vetores aleatórios, sendo que cada elemento x_i é uma mistura dos elementos de $\mathbf{s}$.

As componentes de $\mathbf{s}$ não podem ser observadas diretamente, pois $\mathbf{s}$ além de independentes, são latentes .

Desta forma para simplificar, utiliza-se um modelo matemático da forma:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2.3}$$

em que A é uma matrix de mistura e corresponde ao W^{-1} da equação (3.1)

O objetivo da técnica, basicamente, é recuperar $\mathbf{s}$, por meio de $\mathbf{x}$, sem nenhuma informação sobre as características de $\mathbf{A}$, (3.3) é um modelo estatístico chamado de análise das componentes independentes, que descreve os dados observados pelo processo de mistura das componentes independentes $\mathbf{s}$, sendo que estas não podem ser observadas diretamente. Assim é preciso estimar tanto $\mathbf{s}$ quanto a matriz de mistura $\mathbf{A}$, porque somente $\mathbf{x}$ é observável.

2.5.4 Particularidades em ICA

A técnica de ICA tem suas particularidades, dentre as quais pode-se citar 3 delas. Como dito acima, ICA foi criada para atender a variáveis aleatórias não-gaussianas, pois para variáveis aleatórias gaussianas as componentes são sempre descorrelacionadas e, consequentemente, independentes.

A técnica não pode atender a dados gaussianos pois a distribuição conjunta de misturas dessas variáveis também é gaussiana, assim haveria informação perdida na aplicação da técnica pois esta distribuição é rotacionalmente simétrica e a informação da rotação da mistura é perdida.

A segunda particularidade é que a informação de variância das componentes independentes é perdida no processo de estimação; e a terceira é que não se pode estabelecer uma ordem para as componentes independentes.

2.5.5 CI por maximização da não-gaussianidade

Um dos problemas chaves de ICA é estimar as componentes independentes $\mathbf{s}$ a partir de $\mathbf{x}$ e a não-gaussianidade é um elemento chave para esta estimação, pois a matriz $\mathbf{A}$ segundo a equação (3.3) não é identificável quando mais de uma das componentes independentes tem distribuições gaussianas.

O teorema do limite central implica que a distribuição de soma das componentes independentes é sempre mais próxima de uma gaussiana do que qualquer uma das distribuições das componentes isoladas. Dessa forma para estimar uma componente independente s_i, considera-se a seguinte soma ou combinação linear dos vetores $\mathbf{x}$

$$y = \mathbf{b}^T\mathbf{x} \tag{2.4}$$

em que $\mathbf{b}$ é um vetor a ser determinado. Se $\mathbf{b}$ for uma das linhas da inversa de $\mathbf{A}$, y será igual a uma das componentes independentes s_i.

Do modelo de ICA, equação (3.3), tem-se que

$$y = \mathbf{b}^T\mathbf{x} \tag{2.5}$$

$$y = \sum_i b_i x_i \tag{2.6}$$

$$y = \mathbf{b}^T\mathbf{A}\mathbf{s} \tag{2.7}$$

Denota-se o produto $\mathbf{b}^T\mathbf{A}$ como $\mathbf{q}$. Assim tem-se que:

$$y = \mathbf{b}^T\mathbf{x} = \mathbf{q}^T\mathbf{s} = \sum_i q_i s_i \tag{2.8}$$

Assim, da equação (3.7), observa-se que y também é uma soma das componentes independentes s_i; Desta forma, pelo teorema do limite central é possível concluir

que a distribuição de y é mais gaussiana do que a distribuição de qualquer outra componente independente s_i. Assim, um dos elementos q_i é diferente de zero (0).

Como na prática os valores q_i são desconhecidos, pode-se através das equações (3.4) e (3.7) dizer que

$$\mathbf{b}^T\mathbf{x} = \mathbf{q}^T\mathbf{s} \tag{2.9}$$

Assim, pode-se variar $\mathbf{b}$ e observar a distribuição de $\mathbf{b}^T\mathbf{x}$. Desta forma, pode-se tomar $\mathbf{b}$, como um vetor que maximiza a não-gaussianidade de $\mathbf{b}^T\mathbf{x}$, sendo que este vetor corresponde a $\mathbf{q} = \mathbf{A}^T\mathbf{s}$, sendo que este vetor possui apenas uma de suas componentes diferente de zero. Daí, pode-se concluir que y da equação (3.4) é igual a uma das componentes independentes. Logo, a maximização da não-gaussianidade de $\mathbf{b}^T\mathbf{x}$ permite encontrar uma das componentes.

2.5.6 Medindo a não-gaussianidade

A Curtose é uma medida clássica de gaussianidade, é também conhecida como cumulante de quarta-ordem. A curtose de uma variável x é definida como

$$curt(x) = E\{x^4\} - 3(E\{x^2\})^2 \tag{2.10}$$

Observa-se que a curtose é uma versão normalizada do quarto momento $E\{x^4\}$, se x é gaussiana, o quarto momento é igual a $3(E\{x^2\})^2$. Assim , a curtose é zero (0) para variáveis gaussianas.

Desta forma, pode-se medir o grau de não-gaussianidade de uma variável x a partir da distância do valor absoluto de sua curtose para zero (0), quanto mais distante de zero (0) , mais não-gaussiana é a variável.

2.6 Máquinas de vetor de suporte

2.6.1 Introdução

Os fundamentos das Máquinas de Vetor de Suporte, do inglês support vector machine (SVM), foram introduzidos por V. Vapnick em 1995 e consiste em um método de classificação em duas classes [17]. A idéia básica destas máquinas é construir um hiperplano com uma superfície de decisão em que a margem de separação entre as duas classes é maximizada.

O termo Máquinas de vetor de suporte surgiu porque que os pontos do conjunto de treinamento que estão mais próximos da superfície de decisão são chamados de vetores de suporte. SVM realiza essa tarefa baseado no princípio de Minimização do risco estrutural que é baseado no fato de a taxa de erro da máquina de aprendizado no conjunto de teste é limitada pelo somatório taxa de erro de treinamento e por um termo que depende da dimensão de Vapnik-Chervonenkis (VC)[18]

A classificação de dados estatísticos pode ser atribuída a um problema de uma, duas ou várias classes. Na classificação binária, os dados de duas classes estão disponíveis, supondo que as amostras da base de dados contenham classes igualmente balanceadas. Uma base de dados desbalançeada poderia levar a resultados insatisfatórios [19]. Um problema comum com esta abordagem é a decisão da criação do limite entre as duas classes, na pior das situações, poderia ter uma grande taxa de erro se as classes não forem bem separadas.

As Máquinas de Vetor de Suporte para uma classe constroem um classificador somente para o conjunto de exemplos positivos, chamados de amostras de treinamento positivas [20]. A classificação é feita basicamente pela geração de uma hiper-esfera para a decisão, limitando apenas uma classe de outras que possam existir. A estratégia é mapear os dados em função do espaço de características e em seguida tentar usar a hiper-esfera para descrever os dados e para a inserção de dados. Por isso, a metodologia necessita aprender apenas sobre uma classe e assim as bases

de dados desbalançeadas podem ser utilizadas nesta abordagem, sem problemas com o desempenho do classificador. Esta é uma vantagem deste em relação ao SVM para duas classes.

2.6.2 Máquinas de vetor de suporte para uma classe

Supondo que uma base de dados tenha uma distribuição probabilidade P no espaço de características. Encontrar um subconjunto S deste espaço de características, tal que a probabilidade que um ponto de P esteja fora de S é determinada por uma condição a priori especificada

$$\upsilon \in (0, 1) \tag{2.11}$$

A solução para este problema é obtida pela estimação da função f, que é positiva em S e negativa no complemento $\overline{S}$. Em outras palavras, Scholkopf desenvolveu um algoritmo que retorna uma função f [19]. Esta função toma valores $+1$ em uma pequena região, a hiper-esfera, capturando o maior número de dados e toma valores -1 em outro local.

$$f(x) = \begin{cases} +1 \ \ se \ \ x \in S \\ -1 \ \ se \ \ x \in \overline{S} \end{cases} \tag{2.12}$$

O algoritmo pode ser resumido como um mapeamento dos dados em um espaço de características H usando uma função kernel apropriada, e então tentar separar os dados mapeados da origem com uma margem máxima.

No nosso contexto, tem-se amostras de treino $x_1, x_2, ..., x_l$ pertencentes a uma classe X, onde X é um pequeno subconjunto de $\mathbb{R}^N$. Tem-se $\phi : X \rightarrow H$ sendo o kernel que transforma as amostras de treinamento para outro espaço. Então, para separar o conjunto de dados da origem tem-se a seguinte função objetivo na forma primária

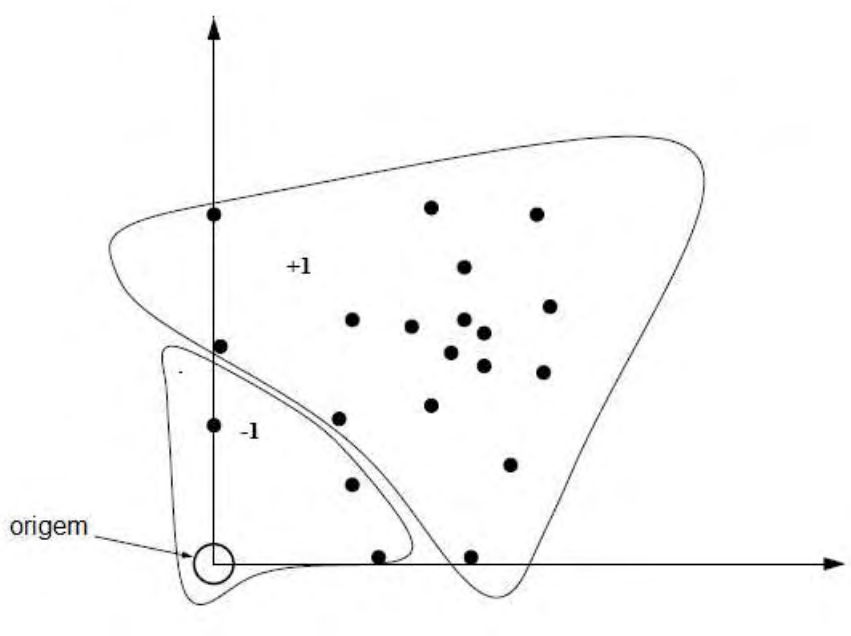

Figura 2.2: Classificador de Máquinas de vetor de suporte para uma classe. A origem é o único membro da segunda classe.

$$\min \mathbf{r}^2 - \rho + \frac{1}{vl} \sum\nolimits_i \zeta_i$$

sujeito a

$$\|\Phi(X_i) - c\|^2 \leq \mathbf{r}^2 + \zeta_i \,, \quad \zeta_i \geq 0 \; para \; i \in [l]$$

Sendo que $v \in [0, 1]$ representa a quantidade total das amostras de treinamento, $\mathbf{r}$ é um vetor ortogonal que separa as amostras de treinamento da origem até um limiar ρ, l representa a parte dos dados de treinamento rejeitados pela hiper-esfera, ζ é usado para rejeitar as amostras de treinamento da hiper-esfera.

Este problema de otimização é resolvido com os multiplicadores de Lagrange:

$$\begin{aligned} L(\mathbf{r},\, \zeta,\, c,\, \alpha,\, \beta) &= \mathbf{r}^2 + \sum_{i=1}^{l} \alpha_i \left[\|\Phi(X_i) - c\|^2 - \mathbf{r}^2 - \zeta_i \right] \\ &+ \frac{1}{vl} \sum_{i=1}^{l} \zeta_i - \sum_{i=1}^{l} \beta_i \zeta_i \end{aligned}$$

$$\frac{\partial L}{\partial \mathbf{r}} = 2\mathbf{r}\left(1 - \sum \alpha_i\right) = 0 \Rightarrow \sum \alpha_i = 1 \tag{2.13}$$

$$\frac{\partial L}{\partial \zeta_i} = \frac{1}{vl} - \alpha_i - \beta_i = 0 \Rightarrow 0 \leq \alpha_i \leq \frac{1}{vl} \tag{2.14}$$

$$\frac{\partial L}{\partial c} = -\sum 2\alpha_i \left(\Phi(X_i) - c\right) = 0$$

$$\Rightarrow c = \sum \alpha_i \Phi(X_i) \tag{2.15}$$

A equação (2.11) e (2.12) coloca para fora da hiper-espera as amostras de treinamento rejeitadas, enquanto que a equação (2.13) informa o c (centro da hiper-esfera) que pode ser expresso como a combinação linear $\Phi(X)$, o que é possível resolver pela forma dual com a função kernel

$$\min \sum\nolimits_{i,j} \alpha_i \alpha_j k\left(X_i,\, X_j\right) - \sum\nolimits_i \alpha_i k\left(X_i,\, X_i\right)$$

sujeito a

$$0 \leq \alpha_i \leq \frac{1}{vl}, \quad \sum\nolimits_i \alpha_i = 1$$

Uma importante família de funções núcleo (kernel) é a função de base radial (RBF, *Radial Basis Function*), muito comumente utilizada em problemas de reconhecimento de padrões e também utilizada neste trabalho, que é definida por

$$K(\mathbf{x}, \mathbf{y}) = e^{-\mathbf{y}\|\mathbf{x}-\mathbf{y}\|^2} \tag{2.16}$$

em que $\gamma > 1$ é um parâmetro que é definido pelo usuário.

No proxímo capítulo iremos descrever a metodologia utilizada com os fundamentos teóricos descritos neste capítulo.

Capítulo 3

Resultados

3.1 Metodologia

A metodologia proposta tem o objetivo de fazer a classificação de pacientes em duas classes :

1 Diabéticos

2 Não-Diabéticos

Para isto, utilizou-se a base de dados de índios Pima que contém 8 variáveis clínicas, que é descrita no capítulo 2. Como exposto na tabela 1.1 vários métodos de classificação utilizaram esta mesma base de dados, desde que esta foi disponibilizada para estudo no UCI-Irvine Machine Learning repository [9]. Todos os métodos utilizados e resultados obtidos desde então, utilizaram a base completa, com suas 8 variáveis disponíveis.

Nosso método diferencia-se na forma de uso das variáveis. Primeiro realizou-se um teste com todas as variáveis da base de dados, como já utilizado extensivamente na literatura; depois realizamos um segundo teste retirando as características invasivas, que são: Concentração de glicose no sangue e quantidade de insulina em 2 hs de jejum, variáveis 2 e 5, respectivamente; como mostra a tabela 3.1. Após isto, realizou-se testes sem as duas características invasivas e retirando uma característica não-invasiva por teste. Desta forma, foram feitos mais 6 testes como mostra a tabela

3.2 para observar a influência da falta de cada uma no resultado.

Tabela 3.1: Características clínicas utilizadas nos testes 1 e 2

N. da variável	Teste 1 (Não-invasivas e Invasivas)	Teste (Não-invasivas)
1	X	X
2	X	
3	X	X
4	X	X
5	X	
6	X	X
7	X	X
8	X	X

Tabela 3.2: Testes feitos sem as características invasivas , retirando-se apenas 1 característica não-invasiva por teste.

N. da Variável	Teste 3	Teste 4	Teste 5	Teste 6	Teste 7	Teste 8
1		X	X	X	X	X
2						
3	X		X	X	X	X
4	X	X		X	X	X
5						
6	X	X	X		X	X
7	X	X	X	X		X
8	X	X	X	X	X	

A metodologia proposta utiliza os seguintes passos: A aquisição dos dados; a extração de características por Codificação Eficiente através da Análise de Componentes Independentes e a classificação através das Máquinas de Vetor de Suporte para uma classe, como mostra o diagrama de blocos da figura 1.1.

3.2 Aquisição de dados

O Software MATLAB foi utilizado para a aquisição dos dados em uma matrix **X**. Obtivemos 768 linhas ou observações, representando cada paciente e 9 colunas, representando as variáveis clínicas. Sendo que a última coluna representa a classe que cada paciente pertence.

Organizamos os pacientes por classe, 0 para diabéticos e 1 para não-diabéticos. Em cada teste utilizamos as variáveis como mostram as tabelas 3.1 3.2. Nos proxímos

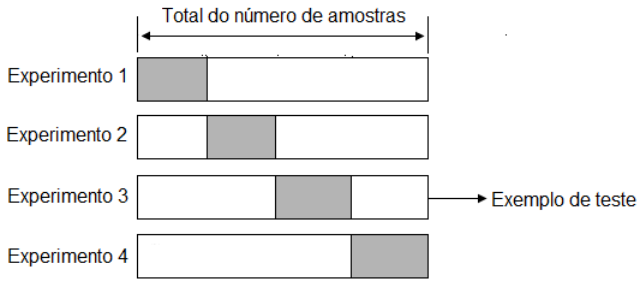

Figura 3.1: Exemplo da divisão das amostras com um K igual a 4 no método de validação cruzada em K divisões

passos separamos os grupos de treino e teste com o método de validação cruzada de K divisões, sendo K igual: 20 e retiramos a última coluna que comtém as informações sobre a classe a que cada paciente pertence.

3.2.1 Validação Cruzada em K divisões

Neste método o conjunto é dividido em K subamostras, que também serão K experimentos. Para cada K, usa-se $K-1$ para o conjunto de treinamento e um dos K subgrupos para o conjunto de teste como mostra a figura 3.1 que exemplifica a divisão das amostras para este método.

A vantagem deste método é que não importa a forma que os dados foram divididos, cada ponto da amostra aparece no conjunto de teste apenas uma vez. A taxa de erro é calculada como a média da taxa de erro dos experimentos, que é dada por :

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i \tag{3.1}$$

3.3 Extração de Características

Nesta fase utilizamos o método de codificação eficiente para a extração de características, como já foi dito o objetivo computacional da codificação eficiente é estimar de padrões um código compacto que consiga reduzir a redundância nos

padrões com o mínimo de perda de informação e os dados são transformados por um conjunto de filtros lineares $\mathbf{W}^{-1}$.

Os modelos de codificação eficiente e de Análise das componentes independentes são ambos baseados em redução de redundância. Dessa forma os filtros estimados pela Análise de componentes independentes podem ser utilizados em um modelo eficiente.

3.3.1 Aprendendo um subespaço através ICA

Dessa forma, tomando $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_n\}$ como o conjunto de amostras separadas para treino, adquirido na fase de aquisição e tratado como explicado na sessão anterior. ICA aprende as funções bases em colunas da matrix $\mathbf{A}$ para as classes de dados de modo que as variáveis que compõem $\mathbf{s}$ são estatisticamente mutualmente independentes.

$$\mathbf{x} = \mathbf{As} \tag{3.2}$$

Neste trabalho utiliza-se o algoritmo FastICA para estimar as componentes independentes.[21].

3.3.2 Projetando os dados sobre o subespaço

Essa etapa consiste no produto interno do conjunto de dados e as funções bases. As amostras originais de treino e as amostras de teste são projetadas sobre o novo subespaço (funções bases) como podemos ver na equação 3.3

$$\hat{\mathbf{x}} = \mathbf{Ax} \tag{3.3}$$

Onde as colunas de A são as funções bases ou características, afirmamos novamente que estas funções bases foram estimadas do conjunto de treino. Agora obtemos as amostras de treino e teste em um novo espaço, sem as redundâncias e com o mínimo de perda da informação.

3.4 Classificação

Nesta fase as amostras de treino e de teste em um novo espaço são novamente rotuladas com as classes a que pertencem, diabéticos e não-diabéticos. Ou seja, a coluna 9 é recolocada com o fim de preparar os dados para serem a entrada do classificador.

Neste trabalho utilizou-se a biblioteca para máquinas de vetor de suporte chamada LIBSVM [22] para realizar as etapas de treinamento e teste. O classificador utilizado foi o de máquina de vetor de suporte para uma classe, que foi explicado no capítulo anterior.

3.5 Validação do método de classificação

Nesta etapa é realizada a avaliação do classificador quanto sua capacidade de diferenciação de outros classificadores para a mesma base de dados utilizada. Nesta avaliação utilizam-se critérios de sensibilidade, especificidade e acurácia. Estes termos possuem variáveis que vamos relacionar para melhor entendimento destes, como veremos a seguir:
As variáveis são:

- Verdadeiro Positivo(VP): Diagnóstico do paciente classificado corretamente como diabético .
- Falso Positivo(FP): Diagnóstico do não-diabético classificado como diabético.
- Verdadeiro Negativo(VN): Diagnóstico do paciente classificado corretamente como não-diabético;
- Falso Negativo(FN): Diagnóstico do diabético classificado como não-diabético;

A sensibilidade indica quão bom é o classificador para identificar os pacientes diabéticos e é definida por

$$sensibilidade = \frac{VP}{(VP + FN)} \tag{3.4}$$

A especificidade indica quão bom é o classificador para identificar os pacientes não-diabéticos e é definida por:

$$especificidade = \frac{VN}{(VP + FP)} \tag{3.5}$$

A acurácia é a taxa de sucesso ou acerto do teste e é dada por:

$$acuracia = \frac{(VP + TN)}{(VP + TN + FP + FN)} \tag{3.6}$$

3.6 Resultados

Com o nosso método obtivemos os seguintes resultados para os 8 testes feitos na base de dados em questão:

Tabela 3.3: Resultados dos testes

Testes	Sensibilidade	Especificidade	Acurácia
teste 1	99,81%	98,34%	98,47%
teste 2	99,05%	98,19%	98,28%
teste 3	98,77%	98,36%	98,39%
teste 4	99,09%	96,55%	96,85%
teste 5	98,67%	97,91%	97,97%
teste 6	98,88%	98,42%	98,45%
teste 7	99,43%	98,43%	98,53%
teste 8	99,43%	98,87%	98,91%

Os paramêtros da função kernel nos testes na maioria das vezes foram:

Tabela 3.4: Paramêtros da função kernel

Paramêtros kernel		
Testes	C	g
Teste 1	0,03125	0,007813
Teste 2	0,03125	0,0078125/0,0001220703125
Teste 3	0,03125	0,03125/0,001953125/0,0078125
Teste 4	0,03125	0,007813
Teste 5	0,03125	0,0078125/0,00048828125
Teste 6	0,03125	0,0078125/0,001953125
Teste 7	0,03125	0,0078125/0,001953125
Teste 8	0,03125	0,0078125/0,00048828125

Capítulo 4

Discussões

Realizou-se o primeiro teste de classificação seguindo a utilização da base de dados completa, com as 8 características invasivas e não-invasivas, como feito por outros pesquisadores , ver na tabela1, comparando com os três últimos trabalhos mais recentes , ver tabela 4.1. O nosso resultado obteve uma acurácia de 99,84%, os outros métodos tiveram uma acurácia menor. Comparando o nosso método com o método mais semelhante, que foi o realizado por Polat em 2007 [25], que também fez a utilização da extração de características e de um classificador obtendo uma acurácia de 89,47%, o nosso método obteve um aumento de acurácia de 9

Tabela 4.1: Comparação das acurácias de trabalhos realizados com a base de dados de índios pima completa

Método	**Acurácia**
Redes neurais de regressão generalizada	80,47%
Regras de extração de máquinas de vetor de suporte	82%
ANFIS e PCA	89,47%
Nosso método	98,47%

A diferença crucial entre os métodos está nas etapas de extração de características e de classificação. Polat utilizou na fase de extração de características o método de análise de componentes principais (PCA) [1] que usa descorrelação, estatística de segunda ordem. Já o método proposto, na mesma fase, utilizou o princípio

[1]PCA é uma técnica que usa estatística de segunda ordem para estudar a correlação entre os dados, uma das principais aplicações desta técnica é a redução da dimensionalidade através da eliminação das variáveis originais de menor variância.

de codificação eficiente que usa estatística de alta ordem, que nos levou a obter componentes estatísticamente independentes, chegando ao objetivo de extrair informação redundante , com o mínimo de perda de informação. Realizou-se testes na fase de extração de características do teste 1 com PCA, mas a acurácia foi insatisfatória em relação ao uso de ICA, ver tabela4.1.

Na etapa de classificação Polat utilizou um classificador tradicional, o sistema de inferência neuro-fuzzy (ANFINS), que separa em duas classes. O classificador utilizado neste trabalho, o de Máquinas de vetor de suporte para uma classe(SVM-One class)faz uma grande diferença, pois como a base de dados de índios pima é desbalançeada, o uso de classificadores tradicionais tende a privilegiar a classe dominante e a ignorar a classe de menor representação, já o SVM-One class aprende só uma classe e não cai nesta sensibilidade dos classificadores tradicionais para duas classes.

O segundo teste foi realizado com o mesmo método do primeiro teste, retirando as características invasivas da base de dados, neste obteve-se uma acurácia de 99,28%. Estas características invasivas são o padrão ouro para o diagnóstico da diabetes na realidade atual do diagnóstico de diabetes; nosso teste, utilizando um sistema de auxílio ao diagnóstico obteve mesmo sem as características invasivas bons resultados, com uma sensibilidade de 99,05% e uma especificidade de 98,19%. Isto quer dizer uma boa capacidade de detectar os verdadeiros-positivos e os verdadeiros-negativos.

Com o objetivo de observar a influência de cada variável no resultado, foram realizados mais 6 testes, sem as características invasivas e retirando uma característica não-invasiva por vez como mostra a tabela 3.2, com estes também foram obtidos bons resultados como mostra a tabela 4.2

Os testes 2, 3, 4, 5, 6, 7 e 8 que foram feitos retirando-se variáveis, continuaram com bons resultados na sensibilidade, especificidade e acurácia o que significa que a boa capacidade de detectar verdadeiros-positivos e verdadeiros negativos continua, para a sensibilidade a média de diferença entre os testes é menor que 1%, para a

Tabela 4.2: Resultados dos testes 3 a 8, sem as variáveis invasivas e retirando-se uma variável não-invasiva por teste

Testes	Sensibilidade	Especificidade	Acurácia
teste 3	98,77%	98,36%	98,39%
teste 4	99,09%	96,55%	96,85%
teste 5	98,67%	97,91%	97,97%
teste 6	98,88%	98,42%	98,45%
teste 7	99,43%	98,43%	98,53%
teste 8	99,43%	98,87%	98,91%

especificidade a média entre os resultados é de menos de 2% e para a acurácia com a diferença média menor do que 2% também.

Finalmente, conclui-se que estes bons resultados deve-se a homogeneidade da população, já que a base é composta somente por descendentes Pima e o método utilizado para o auxílio ao diagnóstico com os benefícios descritos acima. Vale ressaltar a veracidade dos testes, já que foi utilizada a técnica de validação cruzada. Assim , os valores dos resultados obtidos para sensibilidade, especificidade e acurácia são resultado de uma média dos experimentos realizados com um K igual a 20.

Trabalhos futuros

Aplicar o método em uma base de dados heterogênea, mais precisamente em uma base de dados brasileira. Para isso tem-se o objetivo de montar uma base com características clínicas de pacientes maranhenses diabéticos e não-diabéticos com apoio de algum hospital da área.

Artigo publicado

1 Áurea Celeste Ribeiro ; Daniel Duarte Costa ; Allan Kardec Barros ; D. Guilhon ; S. Comani ; Geraldo Braz Júnior . Diabetes Diagnosis Trough ICA and One Class SVM (Aceito). In: Brain Inspired Cognitive Systems 2008 (BICS 2008), 2008, São Luís. Brain Inspired Cognitive Systems 2008 (BICS 2008), 2008.

Referências Bibliográficas

[1] Malerbi DA, Franco LJ. The Brazilian Cooperative Group on the Study of Diabetes Prevalence. Multicenter Study of the Prevalence of Diabetes Mellitus and Impaired Glucose Tolerance in the Urban Brazilian Population Aged 30-69yr. Diabetes Care 1992; 15:1509-16.

[2] American Diabetes Association. Implications of the United Kingdom Prospective Diabetes Study. Diabetes Care 2002; 25 Suppl 1:28S-32S.

[3] Sociedade brasileira de diabetes. Tudo sobre Diabetes.2009,http://www.diabetes.org.br/diabetes/index.php.

[4] The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993; 329:977-86.

[5] Sistema de cadastramento e acompanhamento de hipertensos e diabéticos. Hiperdia. 2009, http://hiperdia.datasus.gov.br/.

[6] Michel, C. and C. Beguin (1994). Using a database to query for diabetes mellitus. Stud Health Technol Inform 14: 178-182.

[7] Dorchy, H. (1999). Screening, prediction and prevention of type 1 diabetes. Role of the Belgian Diabetes Registry. Rev Med Brux 20(1): 15-20.

[8] Weng, C., D. V. Coppini, et al. (1997). Linking a hospital diabetes database and the National Health Service Central Register: a way to establish accurate mortality and movement data. Diabet Med 14(10): 877-883.

[9] C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases, 1996, http://www.ics.uci.edu/ mlearn/MLRepository.html.

[10] Michie, D., D. J. Spiegelhalter, et al. (1994). Machine learning, neural and statistical classification. New York, Ellis Horwood.

[11] Barros, A.K., Chichocki, A.: Neural Coding by Redundancy Reduction and Correlation. In: Proc. of the VII Brazilian Symposium on Neural Networks (SBRN) (IEEE) (2002)

[12] Simoncelli, E.P., Olshausen, B.A.: Natural Image statistics and Neural Representation. Annu. Rev. Neurosci. 1193-216 (2001)

[13] Baddeley R., Abbott, L.F., Booth, M.C., Sengpiel, F., Freeman, T., Wakeman E.A., Rolls E.T. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. Proc R Soc Lond B Biol Sci 264, 1775-1783 (1998).

[14] DeWeese M., Wehr M., Zador A. Binary spiking in auditory cortex. J Neurosci 23, 7940-7949 (2003).

[15] Comon, P. Independent component analysis, A new concept? Signal Processing 36, 287-314, (1994).

[16] Lewicki M. S. Efficient Coding of natural sounds. Nature Neuroscience, 5(4):356-363, 2002.

[17] Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Kluwer Academic Publishers, 1998.

[18] Zhuang L. et al. Parameter Optimization of Kernel-based One-class Classifier on Imbalance Learning .journal of computers, vol. 1, no. 7, october/november 2006

[19] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, R. Williamson, Estimating the Support of a High-Dimensional Distribution, Neural Computation 13 (7)(2001) 1443 1471.

[20] L. Manevitz, M. Yousef, One-class SVMs for document classification, Journal of Machine Learning Research 2 (2) (2001) 139 154.

[21] Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley and Sons, New York (2001)

[22] Chang, C. C., Lin, C. J., LIBSVM - A Library for Support Vector Machines (2003), available at http://www.csie.ntu.edu.tw/cjlin/libsvm/ accessed in January 2008.

[23] Chawla, N.V. Bowyer, K.W. Hall, L.O. Kegelmeyer, W.P.SMOTE:Synthetic Minority Over-sampling Technique (2002) .

[24] Phua, C. Alahakoon, D. Lee, V. . Minority Report in Fraud Detection: Classification of Skewed Data. ACM SIGKDD Explorations. (2004)

[25] Polat K., Günes S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing, p. 702-710 . June, 2007.

[26] Oates, T. (1994). MSDD as a Tool for Classification. EKSL Memorandum 94-29, Department of Computer Science, University of Massachusetts at Amherst.

[27] Smith, J. W., J. E. Everhart, et al. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Symposium on

Computer Applications and Medical Care (Washington, DC). R. A. Greenes. Los Angeles, CA, IEEE Computer Society Press: 261-265.

[28] Quinlan, J. R. (1993). C4.5: programs for machine learning. San Mateo, Calif., Morgan Kaufmann Publishers.

[29] Wahba, G., C. Gu, et al. (1992). Soft Classification, a.k.a. Risk Estimation, via Penalized Log Likelihood and Smoothing Spline Analysis of Variance. The mathematics of generalization: the proceedings of the SFI/CNLS Workshop on Formal Approaches to Supervised Learning, Santa Fe, Addison-Wesley Pub. Co.: 331- 360

[30] Ripley, B. D. (1996). Pattern recognition and neural networks. Cambridge ; New York, Cambridge University Press.

[31] Bioch, J. C., O. van der Meer, et al. (1996). Classification using Bayesian neural nets. The 1996 IEEE International Conference on Neural Networks, p. 1488-1493, Washington, DC, Institute of Electrical and Electronics Engineers.

[32] Carpenter, G. A. and N. Markuzon (1998). ARTMAP-IC and medical diagnosis: instance counting and inconsistent cases. Neural Networks 11(2): 323-336.

[33] Khan, A. H. (1998). Multiplier-free Feedforward Networks, citeseer.nj.nec.com/6034.html. 1998.

[34] Eklund, P. W. and A. Hoang (1998). Classifier Selection and Training Set Features: LMDT, citeseer.nj.nec.com/309003.html.

[35] Liu, B. (1998). Integrating Classification and Association Rule Mining. KDD-98, Knowledge Discovery and Data Minng, New York: 80-86.

[36] King, M. A., J. F. Elder IV, et al. (1998). Evaluation of Fourteen Desktop Data Mining Tools. IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, citeseer.nj.nec.com/293388.html.