Relatório sobre variações de arquitetura de uma PMC para classificação do Dermathology Data Set

Matheus M Korb

[1]Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
São Carlos – SP – Brazil

{v8korb}@gmail.com

1. Informações Gerais sobre o Data Set

Para a classificação dos dados foi escolhida a base de dados Dermathology [REF], esta base de dados classifica um conjunto de sintomas de um paciente em uma determinada doença dermatológica. Para tal são utilizados 34 atributos que resultam em 6 classes distintas. As tabelas 1 e 2 demonstram os 34 atributos, divididos em 12 atributos clínicos e 22 atributos histopatológicos e na tabela 3 pode-se encontrar as 6 classificações e o número de instâncias de cada classe.

Tabela 1. Atributos clínicos utilizados pelo Dermathology Data Set

1	erythema
2	scaling
3	definite borders
4	itching
5	koebner phenomenon
6	polygonal papules
7	follicular papules
8	oral mucosal involvement
9	knee and elbow involvement
10	scalp involvement
11	family history, (0 or 1)
34	Age (linear)

O atributo *Age* foi excluído do processo de classificação por estar com valores faltantes em 8 pacientes do database.

O atributo *family history* tem o valor de 1 se qualquer destas doenças tiver sido observada na família, e 0 caso contrário. A característica idade representa simplesmente a idade do paciente. Para todas as outras características (clínicas e histopatológicas) foi dado um grau na escala de 0 a 3. Onde 0 indica que o recurso não estava presente, 3 indica a maior quantidade possível, e 1, 2 indicam os valores intermediários relativos.

Para o trenameito e teste de validação, a base de dados foi separada aleatóriamente em 2 partes, uma de treinamento contendo 75%(274 instâncias) e outra para o teste dos resultados contendo 25%(91 instâncias) da base de dados. A quantidade de instâncias de cada classe utilizadas para o treinamento e teste podem ser visualizadas na tabela 4.

Tabela 2. Atributos histopatológicos utilizados pelo Dermathology Data Set

12	melanin incontinence
13	eosinophils in the infiltrate
14	PNL infiltrate
15	fibrosis of the papillary dermis
16	exocytosis
17	acanthosis
18	hyperkeratosis
19	parakeratosis
20	clubbing of the rete ridges
21	elongation of the rete ridges
22	thinning of the suprapapillary epidermis
23	spongiform pustule
24	munro microabcess
25	focal hypergranulosis
26	disappearance of the granular layer
27	vacuolisation and damage of basal layer
28	spongiosis
29	saw-tooth appearance of retes
30	follicular horn plug
31	perifollicular parakeratosis
32	inflammatory monoluclear inflitrate
33	band-like infiltrate

2. Ambiente de implementação

Para a implementação da rede neural foi escolhida a linguagem R. Foi escolhida esta linguagem por ela dar uma grande facilidade com a manipulação de dados, por ser interpretada e poder rodar em terminal, facilitando a implementação e verificação por etapas, além de ser uma linguagem fortemente utilizada para trabalhos de redes neurais e aprendizado de máquinas em geral. Para a implementação foi utilizada a versão 3.2.4 revised do R. Em conjunto foram utilizados os pacotes *neuralnet* e *clusterSim*, o primeiro para configuração da arquitetura e treinamento da rede e o segundo para normalização dos dados.

As redes neurais implementadas foram baseadas no algoritmo MLP contido no pacote *neuralnet*.

3. Configurações para treinamento

Inicialmente, antes de colocar os dados para serem treinados, foi preciso normalizá-los. Os valores dos atributos da base de dados estavam no intervalo entre 0 e 4, portanto decidiu-se colocá-los no intervalo entre -1 e 1, com o intuito de facilitar a classificação. Esta normalização foi feita de forma simples, através do pacote *clusterSim* onde utilizamos a normalização *n5*, a qual se baseia na seguinte equação:

$$((x - mean)/max(abs(x - mean))) \quad (1)$$

Tabela 3. Classificações da doença dermatológica

Número	Nome	Quantidade de instâncias
1	psoriasis	112
2	seboreic dermatitis	61
3	lichen planus	72
4	pityriasis rosea	49
5	cronic dermatitis	52
6	pityriasis rubra pilaris	20

Tabela 4. Quantitade de instâncias para treinamento e teste

Classe	Treinamento	Teste
psoriasis	79	32
seboreic dermatitis	54	11
lichen planus	55	17
pityriasis rosea	32	13
cronic dermatitis	39	13
pityriasis rubra pilaris	30	5

4. As Redes Neurais

As redes neurais foram definidas contendo 33 valores de entrada, 6 neurônios na camada de saída (um para representar cada classe) e 1 neurônios na camada escondida. A variação nas arquiteturas deu-se na camada escondida, onde foram utilizadas 3 varições na quantidade de neurônios.

- Rede Neural 3: Nessa rede foram utilizados 39 neurônios na camada escondida, esse valor foi baseado na soma da quantidade de entradas e saídas.
- Rede Neural 2: Nessa rede foram utilizados 26 neurônios na camada escondida, $2/3$ da quantidade de neurônios da rede neural 3.
- Rede Neural 1: Nessa rede foram utilizados 13 neurônios na camada escondida, $1/3$ da quantidade de neurônios da rede neural 3.

As variações na arquitetura buscavam otimizar desempenho, uma vez que menos neurônios implicam em menos operações, e verificar a variação de acurácia entre elas.

O algoritmo utilizado foi o *RPROP+*, uma variação do *backpropagation*, esse algoritmo leva em conta apenas o sinal da derivada parcial sobre todos os padrões, e não o valor de sua magnitude, e atua de forma independente em cada **peso**. Para cada **peso**, se houve mudança no sinal da derivada parcial da função de erro total em comparação com a última iteração, o **peso** é multiplicado por um fator $n-$, com $n < 1$. Se não houver variação no sinal, o **peso** é multiplicado por um fator $n+$, com $n > 1$.

Outros parâmetros de inicialização ajustados manualmente foram:

- learningrate: taxa de aprendizado, ajustado em 0.001
- rep: número de épocas: 5
- linear.output: ajusatado para FALSE, buscando classificação

Os demais parâmetros foram deixados *default*:

- threshold: 0.001

- startweights: random
- err.fct: sse (sum of squares error)
- act.fct: logistic
- likelihood: FALSE
- exclude: NULL
- constant.wights: NULL

A classificação na base de dados é feita através de um valor entre 1 e 6, como são utilizados 6 neurônios de saída, um para representar cada classe, foi utilizado um vetor de tamanho 6 para representar o resultado desejado da saída, onde o vetor era inicializado com 0's e o índice com o número da classe desejada continha o valor 1.

5. Resultados

Devido a um grande número de atributos bem separados e uma pequena quantidade de classificações, a convergência da rede foi muito boa para todas as 3 variações de arquitetura. A rede com maior número de neurônios na camada oculta foi a que obteve melhor resultado.

A partir das matrizes 1, 2 e 3, que são matrizes de confusão, é possível visualizar os resultados de forma a se verificar a quantidade de acertos para cada classe assim como de instancias classificadas de forma errada.

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	32	0	0	0	0	0
[2,]	0	7	0	0	0	0
[3,]	0	0	17	0	0	0
[4,]	0	5	0	12	0	0
[5,]	0	0	0	0	13	0
[6,]	0	0	0	0	0	5

Figura 1. Matriz de Confusão 3

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	32	0	0	0	0	0
[2,]	0	7	0	0	0	0
[3,]	0	0	17	0	0	0
[4,]	0	7	0	10	0	0
[5,]	0	0	0	0	13	0
[6,]	0	0	0	0	0	5

Figura 2. Matriz de Confusão 2

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	32	0	0	0	0	0
[2,]	0	7	0	0	0	0
[3,]	0	0	17	0	0	0
[4,]	0	6	0	11	0	0
[5,]	0	2	0	0	11	0
[6,]	0	0	0	0	0	5

Figura 3. Matriz de Confusão 1

partir destas tabelas é possível também obter-se a acurácia de cada rede:

- Rede 3: $0,945$
- Rede 2: $0,923$
- Rede 1: $0,912$

As redes finais ficaram da seguinte forma:

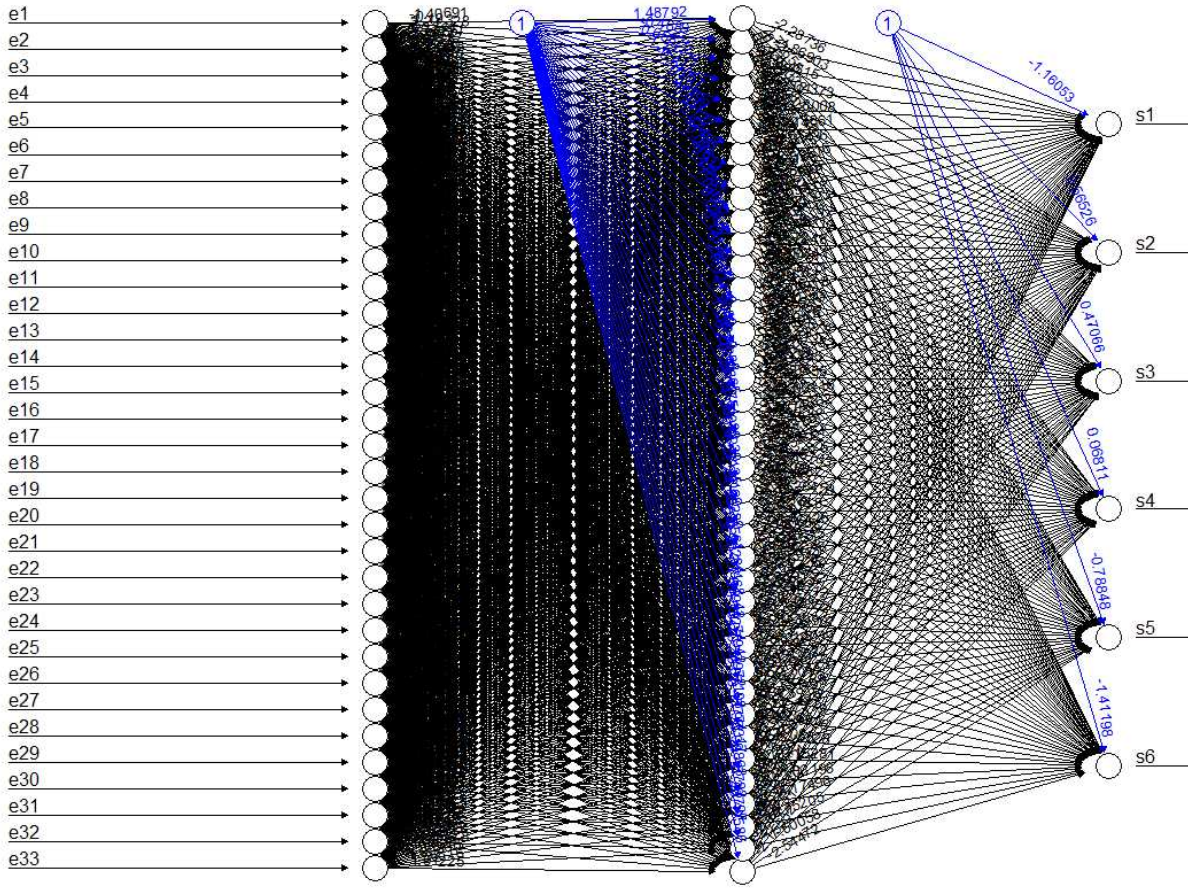

Figura 4. Rede 3

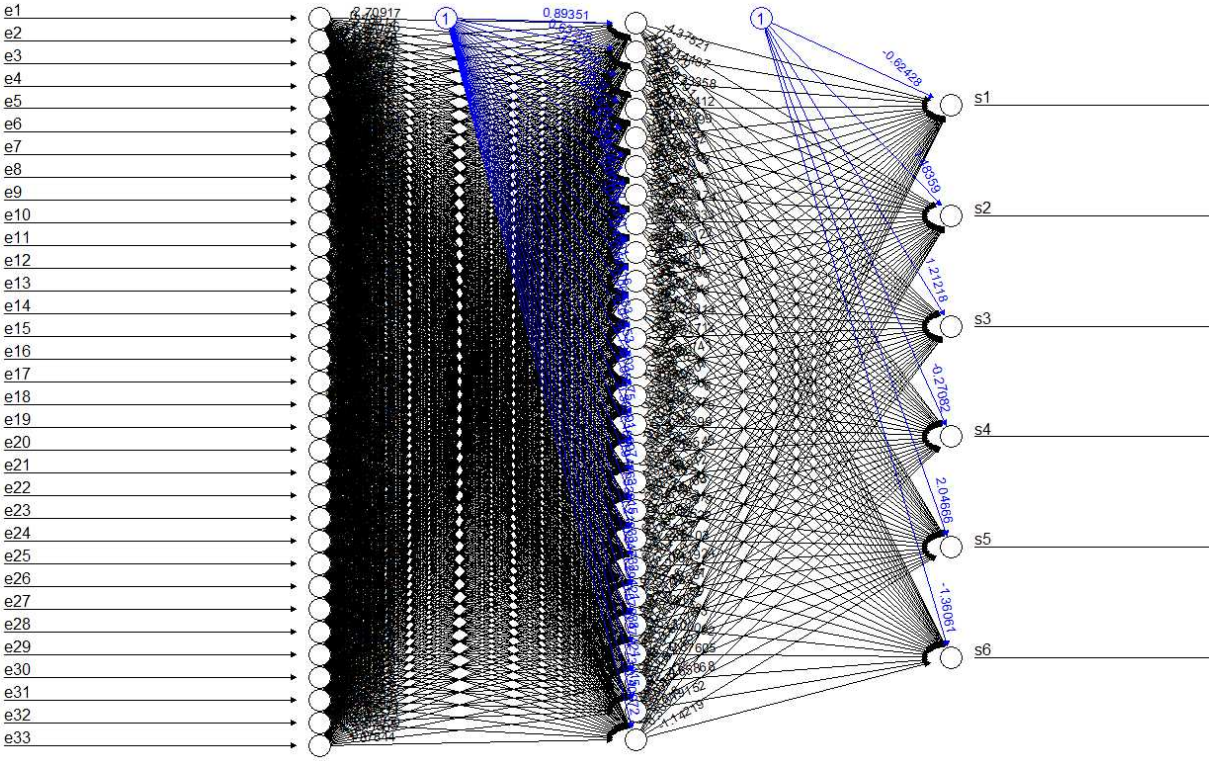

Figura 5. Rede 2

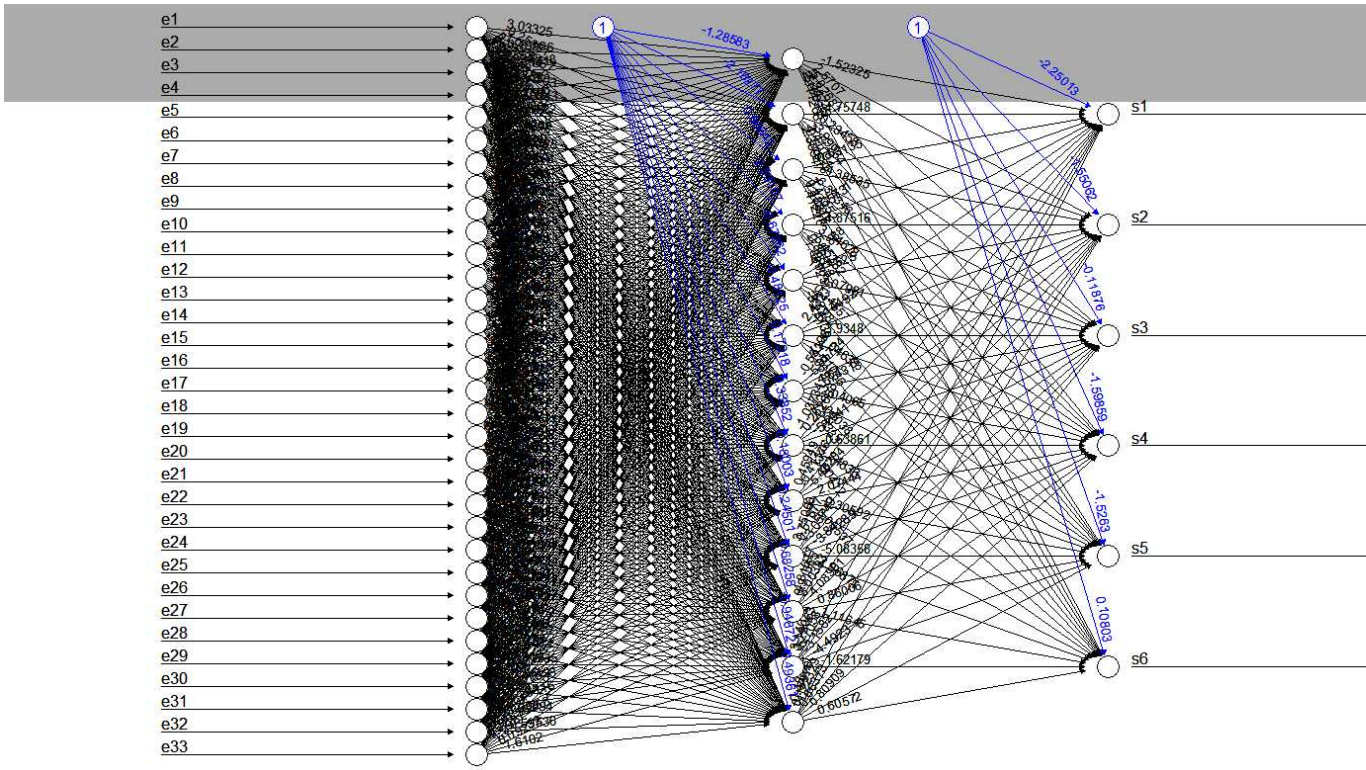

Figura 6. Rede 1

6. Conclusões

Com este trabalho pode-se perceber também a importância de outros métodos aliados com a Rede MLP, métodos que auxiliem na taxa de aprendizagem, como a normalização dos dados.

Pode-se notar que houve pouca diferença na acurácia dos resultados, porém isso se deve a boa separabilidade dos dados. Nem sempre um maior número de neurônios na camada oculta ou mais treinamento trazem melhroes resultados, este foi um caso particular.