UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

NEVUS

TRABAJO FIN DE MÁSTER

Máster en Big Data Analytics

Autor: Pilar Sáez Hernández

Tutor: Jon Ander Gómez
Eduardo Nagore
José Miguel Carot

Curso 2016-2017

Resum

????

Paraules clau: ????, ?????????, ????, ??????????????????

Resumen

????

Palabras clave: Aprendizaje Automático, Redes neuronales, vectores de soporte

Abstract

????

Key words: Machine Learning, Artificial Neural Networks, Support Vector Machine

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

El cáncer de piel es uno de los más comunes a nivel mundial, el cual viene experimentando un importante aumento en países desarrollados desde los años cincuenta en paises desarrollados. Este crecimiento está motivado especialmente por la exposición solar. Pero además existen otros motivos que intervienen en el desarrollo de la enfermedad, como son la información genética, fenotípica y otras características del paciente, sin olvidar algunos factores del entorno.

Dentro del término Cáncer de piel se engloban diferentes tipos de tumor, cada uno de los cuales tiene síntomas, tratamientos y gravedad diferentes.

- Carcinoma Basocelular: Es el tipo de cáncer de piel más frecuente y el menos peligroso, dado que es excepcional que desarrolle metástasis.
- Carcinoma escamoso o espinocelular: Es el segundo tipo de cáncer de piel más común.
- Queratosis Actínica: Lesiones precancerosas
- Melanoma: El tipo de cáncer de piel más peligroso

Por otro lado las técnicas de clasificación automatizada y la búsqueda de patrones en pacientes con patologías de este tipo puede ayudar a la detección precoz y en la aplicación de tratramientos adecuados para la enfermedad.

1.1 Motivación

En este trabajo nos centraremos en pacientes enfermos de melanoma, que aunque es el menos común de los tipos de cáncer citados, es el más peligroso por su riesgo de metástasis, en cuyo caso es determinante el rápido diagnóstico.
El melanoma representa menos del 5 % de los casos de cáncer de piel, pero es la causa de la mayoría de muertes.
Se dispone de información de 844 pacientes diagnosticados de melanoma. La información de la que disponemos se comprende de características generales como la edad, el sexo, datos de fenotipo como el tipo de piel,el color de pelo o color de ojos. También se dispone de algunos datos propios del melanoma como la localización y la profundidad, además de información genética relacionada con la pigmentación, la nevogénica y sensibilidad a la exposición solar.

1.2 Objetivos

Es conocido que la mutación en el gen BRAF está presente en un 66 % de los casos de melanoma mientras que la frecuencia en otros tipos de cáncer no es tan elevada. Este gen elabora la proteína que participa en el envío de señales en las células y en su crecimiento. El objetivo principal que nos ocupa es determinar que variables participan en desencadenar la mutación específica (cambio) en el gen BRAF.

1.3 Estructura de la memória

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 2
??? ???? ??????

????? ????????????? ????????????? ????????????? ????????????? ?????????????

2.1 ?? ???? ???? ? ?? ??

????? ????????????? ????????????? ????????????? ????????????? ?????????????

CAPÍTULO 3
Análisis

En el presente capítulo se analiza el problema planteado, empezando por el desglose de la información de la que se dispone y continuando por su estructuración con un enfoque encaminado a la resolución de los objetivos concretados.

3.1 Análisis del dataset

Se dispone de un dataset con información de 1509 pacientes diagnosticados de cáncer de piel. Los cuales han sido sometidos a intervención por melanoma al menos una vez. De cada paciente se dispone de un total de 136 variables, de las que 104 pertenecen a información genética, 30 incluyen datos identificatívos, de fenotipo, de melanoma u otros datos de interés propios del paciente y las 2 restantes nos indican si se ha desarrollado mutación BRAF y NRAS.
Como en cualquier colección de datos reales existen datos faltantes, ya sea por incorporación posterior de nuevas variables o por desconocimiento del paciente.

A continuación se describe la información que nos ofrece cada una de las variables:

- **Sexo** 1:Male,2:Female
- **Edad:** Edad en el momento de la intervención, EdadGrupo: 0-21,21-32,33-42,43-52,53-64,+65
- **Fototipo:** es la capacidad de la piel para asimilar la radiación solar. Su clasificación oscila entre 1 y 5 en nuetro caso
- **Ojos R** Color de ojos. Valores entre 1 y 4
- **Pelo R** Color de pelo, valores entre 1 y 3
- **Quemintcod** Quemaduras graves Valores entre 1 y 4
- **QareaMM** Quemaduras en el área del melanoma. Valores entre 1 y 3
- **Añossolprof** Número de años de exposición al sol por profesión
- **Añospaquete** Paquetes de tabaco fumados por año.
- **Efélides en inf** Pecas 1:No,2:Sí
- **Léntigos** 1:No,2:Sí
- **Léntigos en área de MM** 1:No,2:Sí

- **Segtumor** Segundo tumor (no cutáneo) 1:No,2:Sí
- **CBC** 1:No,2:Sí Carcinoma basocelular
- **CEC** 1:No,2:Sí Carcinoma epidermoide cutáneo
- **Angiomas sen** Angiomas. Tumores benignos de color rojizo. Valores de 1 a 6
- **Q seborreicas** Queratosis seborreica. Valores de 1 a 6
- **Nevmult** Nevus múltiple. Valores entre 1 y 4
- **Nevus atípicos** Número de nevus atípicos
- **MMM** Múltiples melanomas 1:No,2:Sí
- **Foto loc** Relación entre la exposición solar y la localización del melanoma. Valores entre 1 y 3 . 1: Crítica,2: Intermedia, 3: Nula
- **Locali5** Localización del melanoma Valores entre 1 y 5
- **TipoHX** Valores entre 1 y 5 Histiocistosis X??
- **Breslow** Medida Breslow de profundidad de melanoma
- **Ulceración** 1:No,2:Sí
- **Infiltintrat** Linfocitos intratumorales Valores entre 1 y 3 (Limpiar 77)
- **Nevuspre** 1:No,2:Sí Nevus pre??
- **ElastosisHx** 1:No,2:Sí Degeneración de la piel (por exposición solar, envejecimiento,?)
- **CSD** 1:No,2:Sí ??
- **BRAFmut** Mutación en el gen BRAF
- **NRASmut** Mutación en el gen NRAS

CAPÍTULO 4

Aplicación de técnicas de aprendizaje automático sobre los datos

4.1 Support Vector Machine

Aplicamos el algoritmo SVM utilizando el kernel Radial Basis Functions (rbf) y dividimos los datos en training y test.

4.1.1. Mutación BRAF

En la siguiente tabla, 220 muestras tienen la mutación de un total de 568 muestras de training : 38.732 % 0 missclassified samples of 568 Accuracy = 100.0 %

Tabla 4.1: Resultados de aplicar SVM sobre el trainning set

	precision	recall	f1-score	support
No Mutation	1	1	1	348
Mutation	1	1	1	220
avg / total	1	1	1	568

En la siguiente tabla, 72 muestras tienen la mutación de un total de 190 muestras de testing : 37.895 % 66 missclassified samples of 190 Accuracy = 65.3 %

Tabla 4.2: Resultados de aplicar SVM sobre el test set

	precision	recall	f1-score	support
No Mutation	0.68	0.82	0.75	118
Mutation	0.56	0.38	0.45	72
avg / total	0.64	0.65	0.63	190

4.1.2. Mutación NRAS

En la siguiente tabla, 63 muestras tienen la mutación de un total de 568 muestras de training : 11.092 % 4 missclassified samples of 568 Accuracy = 99.3 %

Tabla 4.3: Resultados de aplicar SVM sobre el trainning set

	precision	recall	f1-score	support
No Mutation	1	0.99	1	348
Mutation	0.94	1	0.97	220
avg / total	0.99	0.99	0.99	568

En la siguiente tabla, 21 muestras tienen la mutación de un total de 190 muestras de testing : 11.053 % 20 missclassified samples of 190 Accuracy = 89.5 %

Tabla 4.4: Resultados de aplicar SVM sobre el test set

	precision	recall	f1-score	support
No Mutation	0.89	1	0.94	118
Mutation	1	0.05	0.09	72
avg / total	0.91	0.89	0.85	190

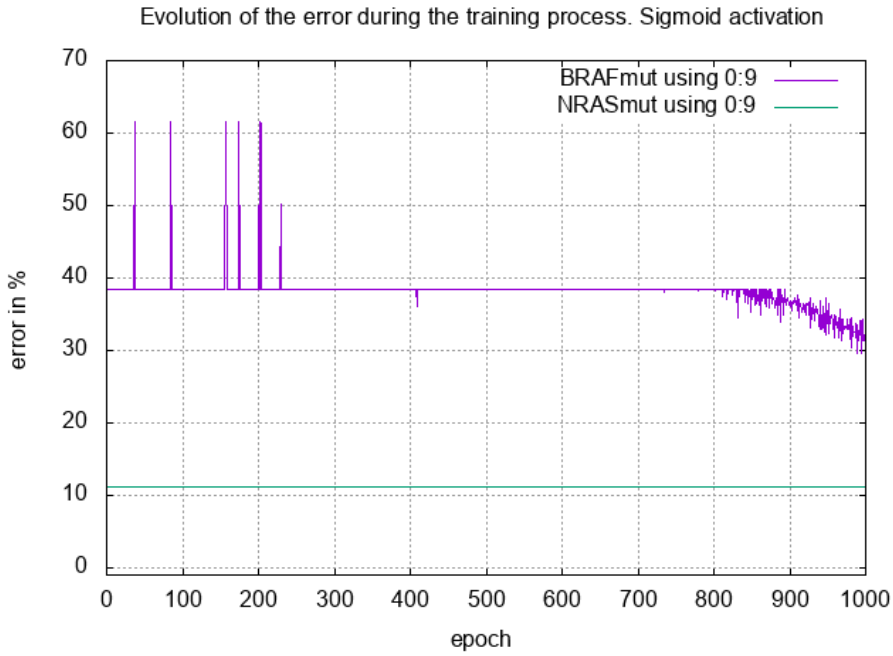
kernel	degree	gamma	C	Accuracy
rbf	1	0.100000	1.000000e-03	54.7 %
rbf	1	0.100000	1.000000e-02	54.7 %
rbf	1	0.100000	1.000000e-01	54.7 %
rbf	1	0.100000	1.000000e+00	62.6 %
rbf	1	0.100000	1.000000e+01	66.3 %
rbf	1	0.100000	1.000000e+02	66.3 %
rbf	1	0.100000	1.000000e+03	66.3 %
rbf	1	1.000000	1.000000e-03	54.7 %
rbf	1	1.000000	1.000000e-02	54.7 %
rbf	1	1.000000	1.000000e-01	54.7 %
rbf	1	1.000000	1.000000e+00	60.0 %
rbf	1	1.000000	1.000000e+01	60.0 %
rbf	1	1.000000	1.000000e+02	60.0 %
rbf	1	1.000000	1.000000e+03	60.0 %
rbf	1	2.000000	1.000000e-03	54.7 %
rbf	1	2.000000	1.000000e-02	54.7 %
rbf	1	2.000000	1.000000e-01	54.7 %
rbf	1	2.000000	1.000000e+00	60.0 %
rbf	1	2.000000	1.000000e+01	60.0 %
rbf	1	2.000000	1.000000e+02	60.0 %
rbf	1	2.000000	1.000000e+03	60.0 %
linear	1	0.100000	1.000000e-03	54.7 %
linear	1	0.100000	1.000000e-02	54.7 %
linear	1	0.100000	1.000000e-01	64.2 %
linear	1	0.100000	1.000000e+00	62.1 %
linear	1	0.100000	1.000000e+01	60.0 %
linear	1	0.100000	1.000000e+02	60.5 %
linear	1	0.100000	1.000000e+03	60.5 %
linear	1	1.000000	1.000000e-03	54.7 %
linear	1	1.000000	1.000000e-02	54.7 %
linear	1	1.000000	1.000000e-01	64.2 %
linear	1	1.000000	1.000000e+00	62.1 %
linear	1	1.000000	1.000000e+01	60.0 %
linear	1	1.000000	1.000000e+02	60.5 %
linear	1	1.000000	1.000000e+03	60.5 %
linear	1	2.000000	1.000000e-03	54.7 %
linear	1	2.000000	1.000000e-02	54.7 %
linear	1	2.000000	1.000000e-01	64.2 %
linear	1	2.000000	1.000000e+00	62.1 %
linear	1	2.000000	1.000000e+01	60.0 %
linear	1	2.000000	1.000000e+02	60.5 %
linear	1	2.000000	1.000000e+03	60.5 %
poly	1	0.100000	1.000000e-03	54.7 %
poly	1	0.100000	1.000000e-02	54.7 %
poly	1	0.100000	1.000000e-01	54.7 %
poly	1	0.100000	1.000000e+00	64.2 %
poly	1	0.100000	1.000000e+01	62.1 %
poly	1	0.100000	1.000000e+02	60.0 %
poly	1	0.100000	1.000000e+03	60.5 %
poly	1	1.000000	1.000000e-03	54.7 %
poly	1	1.000000	1.000000e-02	54.7 %
poly	1	1.000000	1.000000e-01	64.2 %
poly	1	1.000000	1.000000e+00	62.1 %

kernel	degree	gamma	C	Accuracy
poly	1	1.000000	1.000000e+01	60.0 %
poly	1	1.000000	1.000000e+02	60.5 %
poly	1	1.000000	1.000000e+03	60.5 %
poly	1	2.000000	1.000000e-03	54.7 %
poly	1	2.000000	1.000000e-02	58.9 %
poly	1	2.000000	1.000000e-01	67.9 %
poly	1	2.000000	1.000000e+00	61.1 %
poly	1	2.000000	1.000000e+01	60.5 %
poly	1	2.000000	1.000000e+02	60.5 %
poly	1	2.000000	1.000000e+03	60.5 %
poly	2	0.100000	1.000000e-03	54.7 %
poly	2	0.100000	1.000000e-02	56.3 %
poly	2	0.100000	1.000000e-01	63.2 %
poly	2	0.100000	1.000000e+00	65.3 %
poly	2	0.100000	1.000000e+01	64.7 %
poly	2	0.100000	1.000000e+02	64.7 %
poly	2	0.100000	1.000000e+03	64.7 %
poly	2	1.000000	1.000000e-03	61.6 %
poly	2	1.000000	1.000000e-02	66.3 %
poly	2	1.000000	1.000000e-01	64.7 %
poly	2	1.000000	1.000000e+00	64.7 %
poly	2	1.000000	1.000000e+01	64.7 %
poly	2	1.000000	1.000000e+02	64.7 %
poly	2	1.000000	1.000000e+03	64.7 %
poly	2	2.000000	1.000000e-03	68.9 %
poly	2	2.000000	1.000000e-02	64.2 %
poly	2	2.000000	1.000000e-01	64.2 %
poly	2	2.000000	1.000000e+00	64.2 %
poly	2	2.000000	1.000000e+01	64.2 %
poly	2	2.000000	1.000000e+02	64.2 %
poly	2	2.000000	1.000000e+03	64.2 %
poly	3	0.100000	1.000000e-03	55.8 %
poly	3	0.100000	1.000000e-02	59.5 %
poly	3	0.100000	1.000000e-01	67.9 %
poly	3	0.100000	1.000000e+00	66.3 %
poly	3	0.100000	1.000000e+01	66.3 %
poly	3	0.100000	1.000000e+02	66.3 %
poly	3	0.100000	1.000000e+03	66.3 %
poly	3	1.000000	1.000000e-03	65.8 %
poly	3	1.000000	1.000000e-02	65.8 %
poly	3	1.000000	1.000000e-01	65.8 %
poly	3	1.000000	1.000000e+00	65.8 %
poly	3	1.000000	1.000000e+01	65.8 %
poly	3	1.000000	1.000000e+02	65.8 %
poly	3	1.000000	1.000000e+03	65.8 %
poly	3	2.000000	1.000000e-03	66.3 %
poly	3	2.000000	1.000000e-02	66.3 %
poly	3	2.000000	1.000000e-01	66.3 %
poly	3	2.000000	1.000000e+00	66.3 %
poly	3	2.000000	1.000000e+01	66.3 %
poly	3	2.000000	1.000000e+02	66.3 %
poly	3	2.000000	1.000000e+03	66.3 %
poly	4	0.100000	1.000000e-03	57.4 %

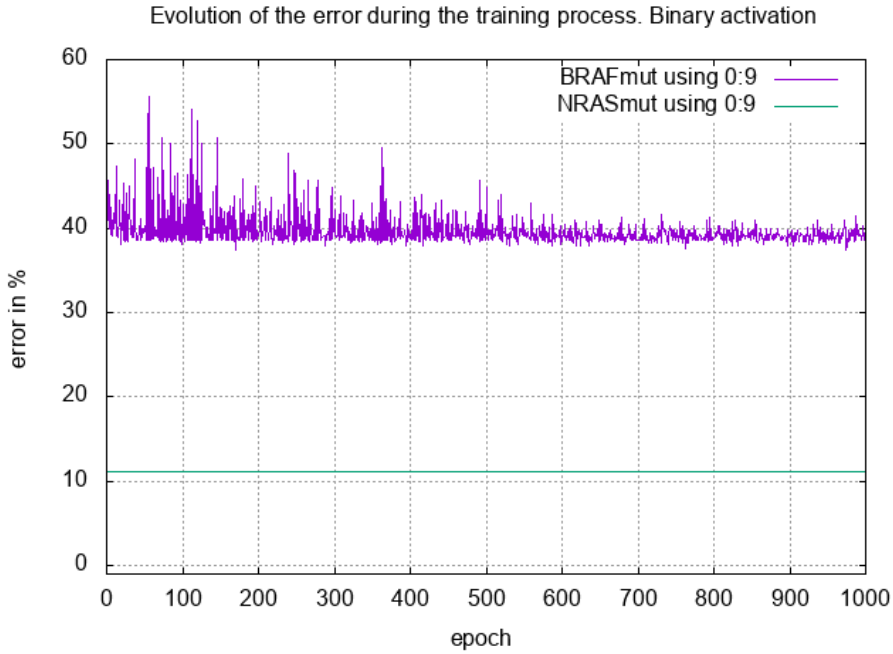
kernel	degree	gamma	C	Accuracy
poly	4	0.100000	1.000000e-02	66.8 %
poly	4	0.100000	1.000000e-01	65.8 %
poly	4	0.100000	1.000000e+00	65.8 %
poly	4	0.100000	1.000000e+01	65.8 %
poly	4	0.100000	1.000000e+02	65.8 %
poly	4	0.100000	1.000000e+03	65.8 %
poly	4	1.000000	1.000000e-03	68.4 %
poly	4	1.000000	1.000000e-02	68.4 %
poly	4	1.000000	1.000000e-01	68.4 %
poly	4	1.000000	1.000000e+00	68.4 %
poly	4	1.000000	1.000000e+01	68.4 %
poly	4	1.000000	1.000000e+02	68.4 %
poly	4	1.000000	1.000000e+03	68.4 %
poly	4	2.000000	1.000000e-03	67.9 %
poly	4	2.000000	1.000000e-02	67.9 %
poly	4	2.000000	1.000000e-01	67.9 %
poly	4	2.000000	1.000000e+00	67.9 %
poly	4	2.000000	1.000000e+01	67.9 %
poly	4	2.000000	1.000000e+02	67.9 %
poly	4	2.000000	1.000000e+03	67.9 %
poly	5	0.100000	1.000000e-03	61.1 %
poly	5	0.100000	1.000000e-02	67.9 %
poly	5	0.100000	1.000000e-01	67.4 %
poly	5	0.100000	1.000000e+00	67.4 %
poly	5	0.100000	1.000000e+01	67.4 %
poly	5	0.100000	1.000000e+02	67.4 %
poly	5	0.100000	1.000000e+03	67.4 %
poly	5	1.000000	1.000000e-03	66.3 %
poly	5	1.000000	1.000000e-02	66.3 %
poly	5	1.000000	1.000000e-01	66.3 %
poly	5	1.000000	1.000000e+00	66.3 %
poly	5	1.000000	1.000000e+01	66.3 %
poly	5	1.000000	1.000000e+02	66.3 %
poly	5	1.000000	1.000000e+03	66.3 %
poly	5	2.000000	1.000000e-03	66.3 %
poly	5	2.000000	1.000000e-02	66.3 %
poly	5	2.000000	1.000000e-01	66.3 %
poly	5	2.000000	1.000000e+00	66.3 %
poly	5	2.000000	1.000000e+01	66.3 %
poly	5	2.000000	1.000000e+02	66.3 %
poly	5	2.000000	1.000000e+03	66.3 %
sigmoid	1	0.100000	1.000000e-03	54.7 %
sigmoid	1	0.100000	1.000000e-02	54.7 %
sigmoid	1	0.100000	1.000000e-01	54.7 %
sigmoid	1	0.100000	1.000000e+00	54.7 %
sigmoid	1	0.100000	1.000000e+01	52.1 %
sigmoid	1	0.100000	1.000000e+02	48.9 %
sigmoid	1	0.100000	1.000000e+03	48.9 %
sigmoid	1	1.000000	1.000000e-03	54.7 %
sigmoid	1	1.000000	1.000000e-02	54.7 %
sigmoid	1	1.000000	1.000000e-01	54.7 %
sigmoid	1	1.000000	1.000000e+00	54.7 %
sigmoid	1	1.000000	1.000000e+01	54.7 %

kernel	degree	gamma	C	Accuracy
sigmoid	1	1.000000	1.000000e+02	52.1 %
sigmoid	1	1.000000	1.000000e+03	44.2 %
sigmoid	1	2.000000	1.000000e-03	54.7 %
sigmoid	1	2.000000	1.000000e-02	54.7 %
sigmoid	1	2.000000	1.000000e-01	54.7 %
sigmoid	1	2.000000	1.000000e+00	54.7 %
sigmoid	1	2.000000	1.000000e+01	54.7 %
sigmoid	1	2.000000	1.000000e+02	54.7 %
sigmoid	1	2.000000	1.000000e+03	52.1 %

4.2 Artificial Neural Networks

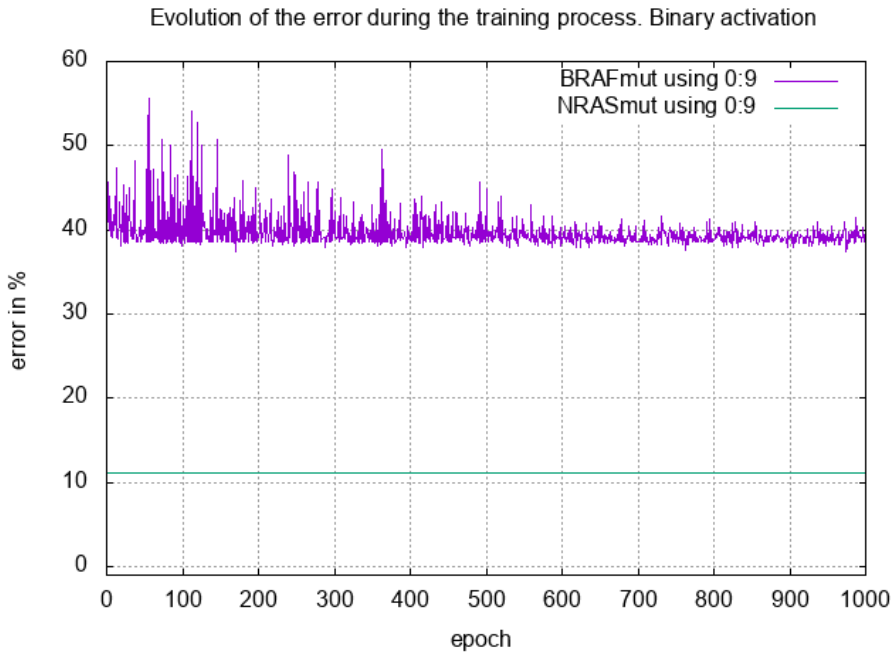


(a) Activation Sigmoid

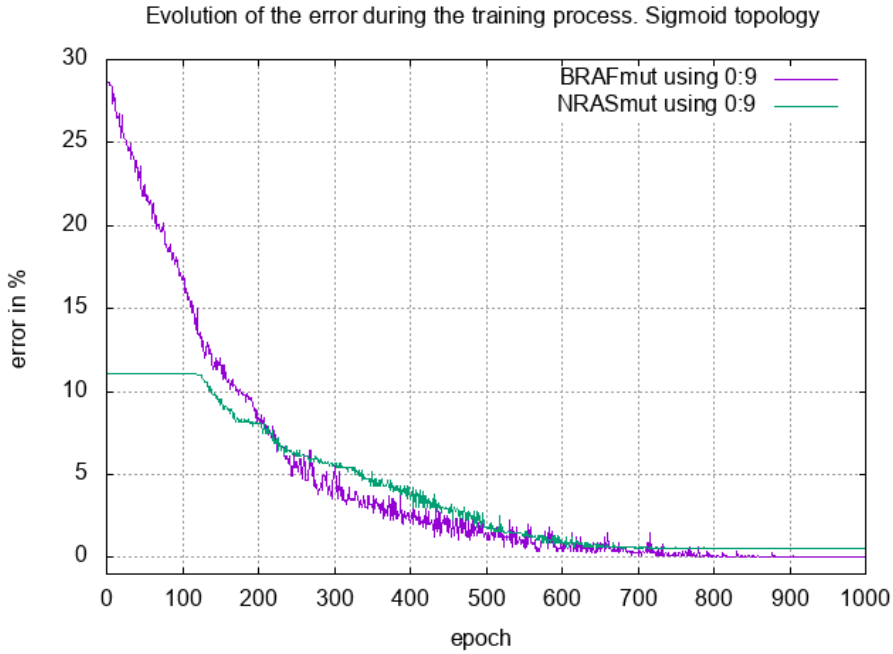


(b) Activation binary

Figura 4.1: Error evolution on training proccess



(a) Activation binary



(b)

Figura 4.2: Error evolution on training proccess

CAPÍTULO 5
Conclusions

????? ????????????? ????????????? ????????????? ????????????? ?????????????

Bibliografía

[1] Jennifer S. Light. When computers were women. *Technology and Culture*, 40:3:455–483, juliol, 1999.

[2] Georges Ifrah. *Historia universal de las cifras*. Espasa Calpe, S.A., Madrid, sisena edició, 2008.

[3] AECC-Asociación española contra el cáncer. Consultado en https://www.aecc.es/SobreElCancer/CancerPorLocalizacion/melanoma

[4] Nature. International weekly journal of science Consultado en https://www.nature.com/nature/journal/v417/n6892/full/nature00766.html.

APÉNDICE A
Configuració del sistema

????? ????????????? ????????????? ????????????? ????????????? ?????????????

A.1 Fase d'inicialització

????? ????????????? ????????????? ????????????? ????????????? ?????????????

A.2 Identificació de dispositius

????? ????????????? ????????????? ????????????? ????????????? ?????????????

APÉNDICE B
??? ????????????? ????

????? ????????????? ????????????? ????????????? ????????????? ?????????????