Comparación de algoritmos para la clasificación de tumores con base en datos de expresión génica

Laura Brenes & Izayana Sandoval

Contenido

Introducción

Objetivos

Metodología

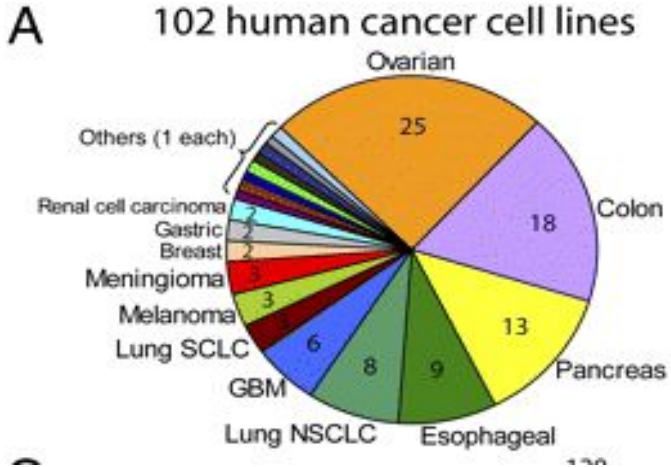
Resultados:

-Análisis exploratorio en weka

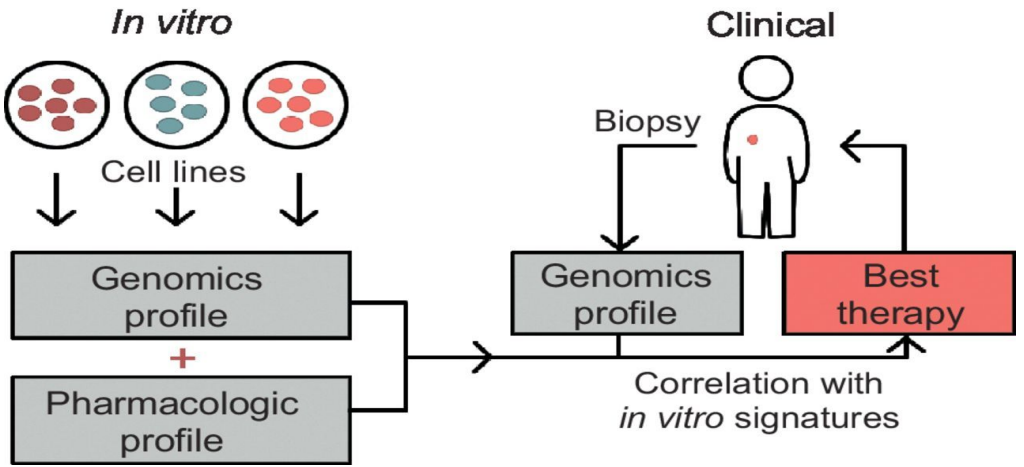
-Workbook: resultados diferentes algoritmos de clasificación scikit-learn

Introducción

- Datos clínicos e histopatológicos para tratamiento: dificultades identificación
- Amplio espectro de tumores que tienen características atípicas
- Confusión en el diagnóstico- tratamiento temprano
- Diagnóstico molecular: preciso, objetivo, clasificación sistemática



Cheung et al. 2011. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer



Goodspeed et al. 2016. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics

Objetivos

Encontrar el mejor algoritmo de clasificación de tumores basado en perfiles de expresión de 16063 genes.

Determinar el mejor algoritmo de clasificación en un análisis exploratorio de los datos con el software weka.

Comparar los diferentes algoritmos de clasificación en scikit-learn

Metodología

Microarreglos: datos de expresión
16,063 genes: 14,030 GenBank y 475 The Institute for Genomic Research (TIGR)

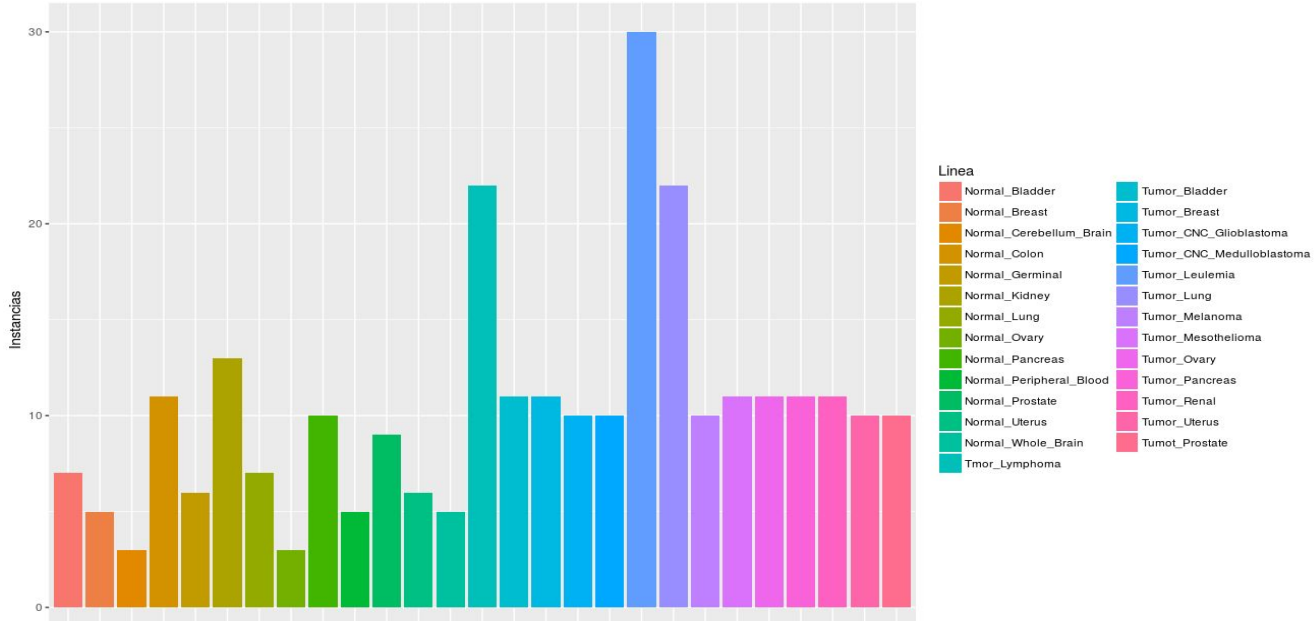
Clasificación del tumor
Clase II (Tipo de tumor-Tejido Normal) -- Clase III (Normal-Tumor)

Preprocesamiento de los datos
Análisis exploratorio Weka-algoritmos de clasificación
Algoritmos de clasificación scikit-learn
-Clasificación correcta de líneas tumores según patrones de expresión de genes
-Clasificación de líneas tumorales de líneas normales utilizando patrones de expresión génica.

Resultados

280 instancias - 16062 atributos (datos expresión de cada gen)

Clases: 14 diferentes tipos de tumor- 13 líneas normales



Análisis exploratorio weka

1021 atributos - 280 instancias
Clase II (Tipos de tumor: 27 clases)

Algoritmo		Instancias correctamente clasificadas %
IbK	test	46,42
	training	100
Naive Bayes	test	46,31
	training	86,42
LMT	test	43,15
	training	97,14
J48	test	33,68
	training	91,78
Random Forest	test	40,35
	training	100
Zero R	test	11,57
	training	10,71
One R	test	13,68
	training	25,35
Naive Bayes-kernel estimador	test	48,42
	training	81,42

```
=== Evaluation on training set ===

Time taken to test model on training data: 3.15 seconds

=== Summary ===

Correctly Classified Instances         242               86.4286 %
Incorrectly Classified Instances        38               13.5714 %
Kappa statistic                          0.8572
Mean absolute error                      0.0099
Root mean squared error                  0.0991
Relative absolute error                 14.0868 %
Root relative squared error             52.7737 %
Total Number of Instances              280
```

```
=== Evaluation on test split ===

Time taken to test model on test split: 1.23 seconds

=== Summary ===

Correctly Classified Instances          44               46.3158 %
Incorrectly Classified Instances        51               53.6842 %
Kappa statistic                          0.4291
Mean absolute error                      0.0398
Root mean squared error                  0.1991
Relative absolute error                 56.3275 %
Root relative squared error            106.097  %
Total Number of Instances               95
```

```
=== Confusion Matrix ===

 a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t  u  v  w  x  y  z  aa   <-- classified as
 1  0  0  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  1  0  0  0  0  |  a = Tumor__Breast
 0  1  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  b = Tumor__Prostate
 0  1  3  0  0  0  0  0  0  0  4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  c = Tumor__Lung
 0  0  0  7  0  0  0  1  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  |  d = Tumor__Lymphoma
 0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  e = Tumor__Melanoma
 0  0  1  0  0  1  1  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  f = Tumor__Bladder
 0  0  0  0  0  2  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  g = Tumor__Uterus
 0  0  0  0  0  0 11  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  h = Tumor__Leukemia
 0  1  0  0  0  1  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  i = Tumor__Renal
 0  1  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  |  j = Tumor__Pancreas
 0  0  0  0  0  0  0  1  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  k = Tumor__Ovary
 0  0  0  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  l = Tumor__Mesothelioma
 0  0  0  0  0  0  0  0  0  0  0  3  1  0  0  0  0  0  0  0  0  0  0  0  0  |  m = Tumor__CNS_Glioblastoma
 0  0  0  0  0  0  0  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0  0  |  n = Tumor__CNS_Medulloblastoma
 0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  o = Normal__Breast
 0  2  1  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  p = Normal__Prostate
 0  0  0  0  0  0  0  1  1  0  1  0  0  0  0  0  0  0  0  1  0  0  0  0  0  |  q = Normal__Lung
 0  0  1  0  0  0  0  0  0  0  1  0  0  0  0  0  0  3  0  0  0  0  0  0  0  |  r = Normal__Colon
 0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  s = Normal__Germinal
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  1  0  0  0  0  |  t = Normal__Bladder
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  |  u = Normal__Uterus
 0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  |  v = Normal__Peripheral_Blood
 0  0  0  0  0  0  0  1  1  0  1  0  0  0  0  0  0  0  0  0  0  2  0  0  0  |  w = Normal__Kidney
 0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  0  0  |  x = Normal__Pancreas
 0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  1  0  0  0  0  0  |  y = Normal__Ovary
 0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  |  z = Normal__Whole_Brain
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  | aa = Normal__Cerebellum_Brain
```

Análisis exploratorio weka

1021 atributos - 280 instancias
Clase III (Tipos de tumor: 2 clases)

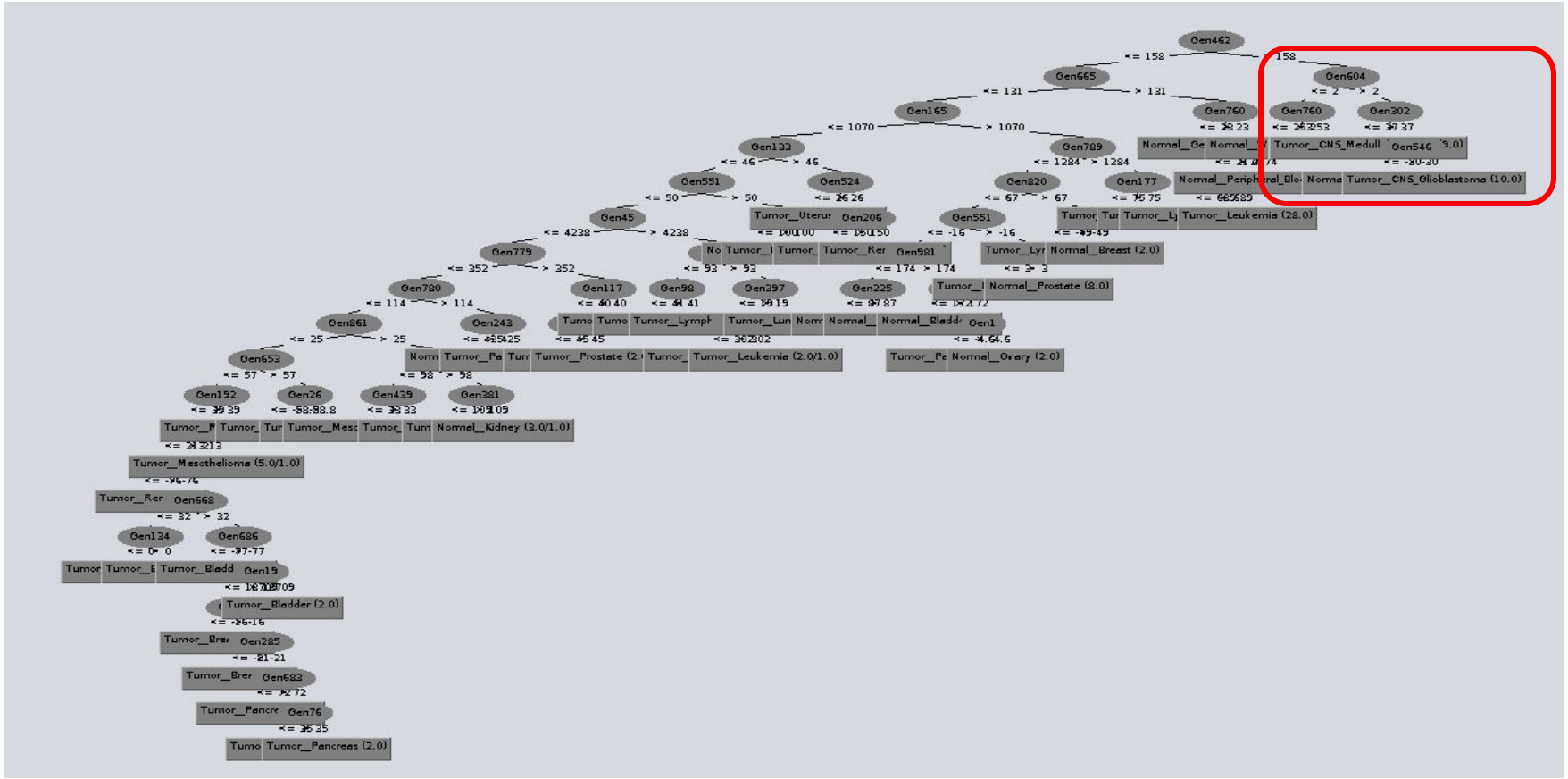
Algoritmo		Instancias correctamente clasificadas %
IbK	test	84,21
	training	100
Naive Bayes	test	77,89
	training	82,85
LMT	test	88,42
	training	99,26
J48	test	64,21
	training	99,28
Random Forest	test	84,21
	training	100
Zero R	test	67,36
	training	67,85
One R	test	61,78
	training	78,57
Naive Bayes-kernel estimador	test	75,78
	training	83,21

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.4 seconds

=== Summary ===

Correctly Classified Instances         232               82.8571 %
Incorrectly Classified Instances        48               17.1429 %
Kappa statistic                          0.5747
Mean absolute error                      0.1715
Root mean squared error                  0.414
Relative absolute error                 39.2824 %
Root relative squared error             88.6547 %
Total Number of Instances              280
```

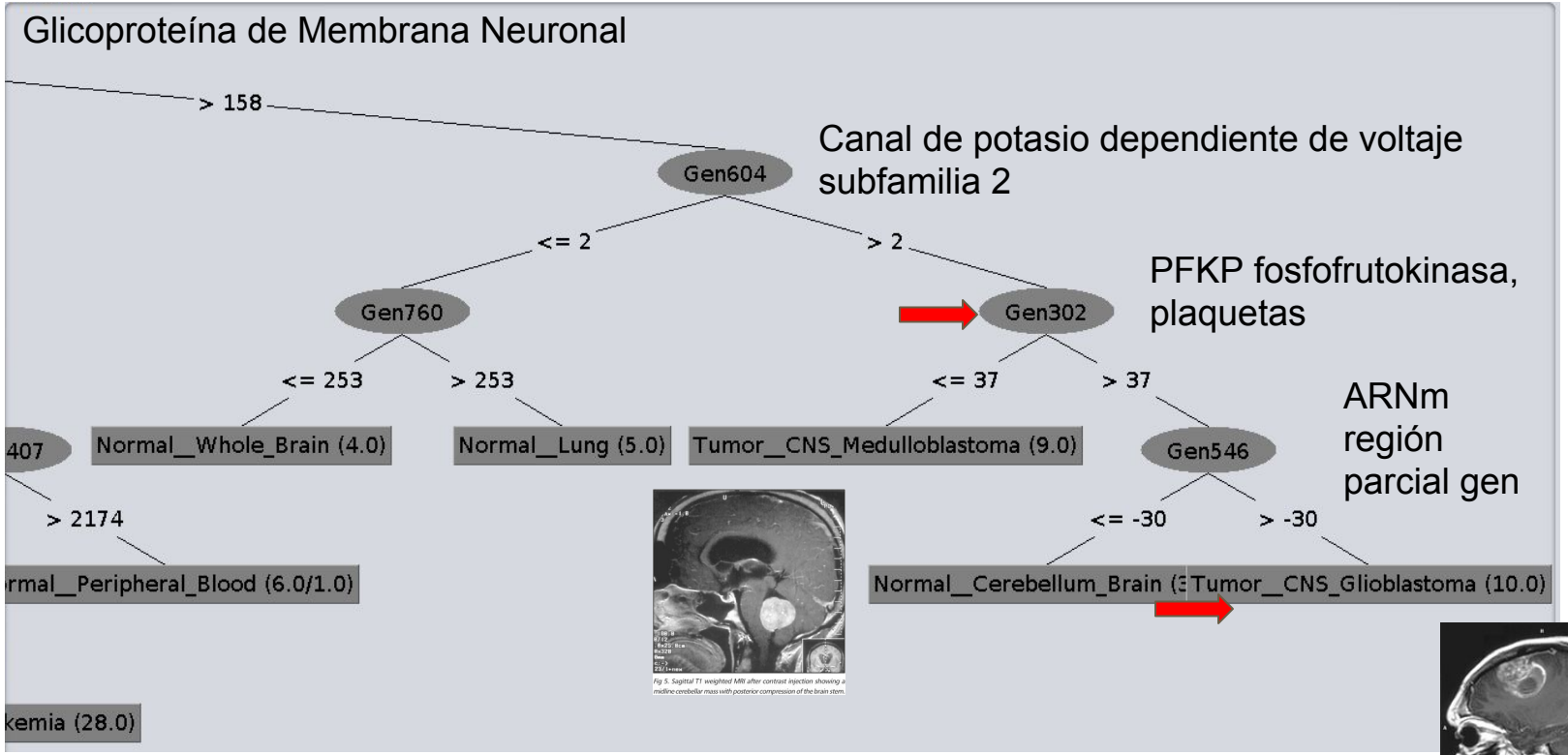
```
=== Evaluation on test split ===

Time taken to test model on test split: 0.09 seconds

=== Summary ===

Correctly Classified Instances          74               77.8947 %
Incorrectly Classified Instances        21               22.1053 %
Kappa statistic                          0.4363
Mean absolute error                      0.224
Root mean squared error                  0.471
Relative absolute error                 51.1768 %
Root relative squared error            100.4409 %
Total Number of Instances               95
```

```
=== Confusion Matrix ===

  a  b   <-- classified as
 60  4 |  a = Tumor
 17 14 |  b = Normal
```

Árbol de Clasificación de tumores (1024 genes)



Árbol de Clasificación de tumores (1024 genes)



2 clases

Comparación entre algoritmos

Algoritmo	Clase	training score	test score
Redes neuronales	III	0.986	0.914
Support vector machine	III	0.990	0.886
Adaboost	III	1.000	0.829
Random forest	III	0.881	0.657
Naive Bayes	III	0.910	0.800

27 clases

Comparación entre algoritmos

Algoritmo	Clase	training score	test score
Redes neuronales	II	0.995	0.471
Support vector machine	II	0.933	0.528
Adaboost	II	0.267	0.229
Random forest	II	0.738	0.200
Naive Bayes	II	0.967	0.514

Conclusiones

Redes neuronales y Naive Bayes fueron los algoritmos que dieron mejores resultados para clasificación de tumores

No todos los 16 mil genes son útiles para clasificación

En comparación con resultados previamente obtenido utilizando Support Vector Machine, con redes neuronales se obtuvo un porcentaje de instancias clasificadas correctamente menor que el reportado (Ramaswamy et al. 2001) para la clase II y mayor para la clase III.