Clasificación de semillas de malezas utilizando Redes Neuronales Artificiales

Sebastián Gavarini[1], Martín Trípode[1], Marcelo Tosini[1], Alejandro Ceccatto[2]

[1]Instituto Computación aplicada (INCA) – Fac. de Cs. Exactas
Universidad Nac. del Centro de la Prov. de Bs. As. - Tandil – Argentina
TE: (2293) 432466 – email: mtosini@exa.unicen.edu.ar

[2]Instituto de Fisica Rosario (IFIR) - CONICET/UNR
Blvd. 27 de Febrero 210 Bis, 2000 Rosario, Argentina
TE: (341) 485-3222 Ext. 19 FAX: (341) 482-1772

1 Introducción

El presente trabajo es continuación del estudio llevado a cabo en [***Cec2000***], trabajo en el que se desarrolló una herramienta para procesar imágenes de semillas de malezas capturadas por una cámara de video, a las que se les extraían características morfológicas, de color y de textura. Parte del trabajo fue precisamente la selección de las características más relevantes. Posteriormente se analizaron métodos de clasificación por Naïve Bayes [***Alf2002***] y Redes Neuronales Artificiales FeedForward con el usual método de entrenamiento *backpropagation* [***Rie1993***].

El análisis y clasificación de semillas contribuyen al valor agregado final en la producción de cultivos, actividades que son llevadas a cabo en diferentes etapas del proceso global. Las semillas de malezas son identificadas por estaciones de prueba y corporaciones productoras de semillas para medir la pureza de la cosecha, y por estaciones de investigación para detectar cambios en bancos de semillas en el suelo. La identificación automática de semillas de especies salvajes es diferente de la identificación de semillas de variedades de una única especie. Para ser aprobada como una variedad, las plantas cultivadas tienen que ser homogéneas con respecto a ciertos caracteres de plantas. Las especies salvajes, por el contrario, suelen tener mayores variaciones intra-especies. Además, la variación entre especies de malezas será en general mayor, pero las semillas de algunas especies cercanamente relacionadas pueden ser muy similares. Del punto de vista del color, muchas semillas de malezas son de marrón claro a oscuro o negro. Todas estas características hacen a la clasificación automática de semillas de malezas a priori un problema de clasificación difícil.

2 Objetivos del trabajo

En el presente trabajo se considerarán diversos clasificadores aplicados al problema de identificación de semillas de malezas, tanto estadísticos como también redes neuronales. Para la selección de las redes, se analizarán exhaustivamente y compararán varias alternativas y gran parte del trabajo será justamente el estudio de su desempeño variando sus múltiples parámetros. Los parámetros de las redes neuronales son inherentemente empíricos, (prueba y error). Aunque hay una fuerte base teórica que sustenta los algoritmos neuronales [***Per2000***], no existen métodos que garanticen la correcta selección de parámetros en la práctica.

Se analizará la precisión de clasificación obtenida. Los resultados estarán respaldados por *pruebas t* [***Wal1999***] para valorar estadísticamente las hipótesis formuladas. Asimismo se estudiarán los errores incurridos por los clasificadores, y también se analizará un enfoque mixto para automatizar lo más posible la clasificación, sin pérdida de precisión.

Los clasificadores seleccionados (de un conjunto mucho mayor) fueron los estadísticos clásicos; tales como Naïve Bayes, Clasificador Cuadrático, y el de k vecinos más próximos (KNN)

[***Mic1994***], y las Redes Neuronales Radiales [***Bla2003***] Aproximadas de 400 neuronas, las SOM y LVQ [***Koh2001***].

3 Análisis de los clasificadores

Los datos de entrenamiento fueron generados mediante 30 particiones independientes de una base de datos de semillas original de 3163 muestras, y 12 características por cada una. Las particiones separan a la base original en dos subconjuntos, uno de entrenamiento con el 80% de las muestras, y otro de prueba con el 20% restante. A éstas se les aplico normalización de varianza a 1, centrado de media en 0, y un algoritmo *greedy* de escalado con *KNN* 1. El escalado es una ponderación de las características de acuerdo a su poder discriminante. La distancia Euclídea, que utilizan algunos clasificadores, se ve afectada si las magnitudes de las características son dispares, o no están relacionadas con el poder discriminante particular de cada característica. La optimización de parámetros de cada clasificador que lo requirió se realizó mediante un proceso de tres etapas, ver [***Mic1994***], donde el conjunto de entrenamiento se subdivide en dos, un 80% para entrenamiento en la fase de optimización, y un 20% para validación, una vez hallado el óptimo, el clasificador final se entrena con el conjunto completo de entrenamiento con los parámetros elegidos, y se prueba con el conjunto de prueba, todavía no visto nunca por las reglas de clasificación.

Tabla 3.1 *Resultados finales de las 30 repeticiones de los clasificadores aplicados a datos escalados.*

	Tiempo (seg)		**Aciertos Entrenamiento (%)**				**Aciertos Prueba (%)**			
Clasificador	**Entr.**	**Prue.**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Cuadrático	0	0.71	99.81	100	100	100	**97.61**	99.18	99.46	99.62
Aproximada 400	1614.4	1.17	98.81	99.85	99.91	99.95	**96.84**	99.25	99.68	99.80
KNN 1	0	25.59	100	100	100	100	**96.28**	97.88	98.58	98.93
Naïve Bayes	0.3	27.52	97.67	99.49	99.80	99.89	**96.27**	98.88	99.34	99.54
SOM	55	1	93.50	96.17	97.25	97.81	**95.58**	96.97	97.68	98.19
LVQ	8	1	97.87	98.14	98.43	98.72	**95.33**	97.32	98.13	98.60

La columna rotulada *Aciertos de Entrenamiento*, bajo el título *'1'*, es el porcentaje de aciertos del clasificador entrenado con el conjunto de 'entrenamiento', y probado con el mismo conjunto. Las restantes 3 columnas, rotuladas, '2', '3', y '4' dan por válida la clasificación si el valor real de la muestra está entre los 2, 3, o 4 valores predichos por el clasificador, respectivamente. Los *Aciertos de Prueba* son los porcentajes de acierto del clasificador entrenado con el mismo conjunto de 'entrenamiento', pero probado con el restante conjunto de 'prueba'. Igualmente la *Tabla 3.1* se ordenó por la columna de aciertos de prueba, con la primera opción de clasificación, por ser la más directamente relevante para un clasificador automático. Notar que un clasificador que muestre las cuatro opciones más probables a un operador es de gran utilidad, porque actualmente se requiere de técnicos altamente calificados que una a una seleccionan y rotulan las semillas de malezas halladas. Con la variante de 4 opciones se puede contar con técnicos menos calificados que elijan una opción dadas cuatro fotos y la nueva foto a clasificar.

4 Análisis de errores

Se utilizaron 5 grupos de las *30 repeticiones*, y solamente cuatro clasificadores de los 6 mostrados en la *Tabla 3.1*, (Cuadrático, Aproximada, Naïve Bayes y SOM). Esta selección de clasificadores y conjuntos de muestras es necesaria dada la naturaleza exponencial de las pruebas cruzadas entre ellos.

El promedio de errores comunes a todos los clasificadores es de 1.75 muestras mal identificadas, en cambio el promedio de los errores particulares de cada clasificador es de 23.38, o sea que solo el 7.48% de los errores son comunes a todos. Esto último permite utilizar una

estrategia de clasificadores combinados, pudiendo obtener lo mejor de cada uno de ellos. Si los errores comunes hubieran sido muchos, la combinación de clasificadores no mejoraría la situación, y además podría sospecharse de la extracción de características o captura de imágenes, pudiendo dirigir hacia allí la atención para mejorar la capacidad de clasificación. Solamente para poner en números optimistas lo dicho de la combinación de errores, el máximo que podría alcanzarse, si se considera como error solamente el promedio de errores comunes por conjunto de prueba, 1.75, sería del 99.72% de aciertos. Claro está que es un número optimista, y además tomado sobre cuatro de los treinta conjuntos de prueba, pero es una idea interesante para continuar investigando. Hay dos estrategias que parecen posibles, a simple vista, para la combinación de clasificadores, la primera es utilizar una votación entre ellos, y la segunda tener una medida de la precisión de la predicción de cada uno, (un valor de confianza asociado). Escapa al alcance del presente trabajo un análisis de combinaciones de clasificadores.

5 Resultados de *pruebas t*

La intención es justificar numéricamente si existe o no evidencia para afirmar, por ejemplo, que el clasificador Cuadrático es mejor que la red SOM [***Wal1999***].

Tabla 5.1. Resultados de las pruebas t.

Pares de clasificadores	Valor de distribución	Decisión sobre H_0
Cuadrático - Aproximada 400	1.1954E-05	rechazar
Aproximada 400 - KNN 1	0.004201	rechazar
KNN 1 - Naïve Bayes	0.48944021	no rechazar
Naïve Bayes - SOM	0.0007596	rechazar
SOM - LVQ	0.17255467	no rechazar

En la *Tabla 5.1* se tomaron a los clasificadores de a pares adyacentes para compararlos. Puede verse que hay solo dos casos donde no pueda afirmarse que un clasificador es superior al otro, KNN 1 contra Naïve Bayes, y SOM contra LVQ. Si se miran los porcentajes de aciertos en la *Tabla 3.1*, se verá que son muy próximos. Estadísticamente queda demostrado que no existe evidencia para rechazar la hipótesis nula de igualdad entre la capacidad de clasificación de dichos clasificadores.

6 Clasificador mixto

En esta sección se introduce un enfoque mixto, en el que se intenta automatizar la clasificación, sin descuidar la precisión. Como se vio en la *Tabla 3.1*, se alcanza un porcentaje de aciertos de prueba del 97.61% con un clasificador Cuadrático. Transformando el mismo clasificador, en una herramienta de asistencia a un operador, mostrando las cuatro mejores posibles categorías, alcanza un 99.62% de precisión, pero el operador debe seleccionar una a una las opciones. Para poder automatizar se introduce una nueva respuesta del clasificador, *'no sé'*. Aunque el mejor clasificador automático es el Cuadrático, el mejor a cuatro categorías, por leve margen, es la red Radial Aproximada de 400 neuronas, con un 99.80% de aciertos. En las redes Radiales puede definirse qué se considera como buena clasificación. Para ello se tomaron dos definiciones para la predicción. La primera es dependiente del valor absoluto de la categoría ganadora, (por eso se la llamará escala absoluta), sin importar el valor de las restantes, de esta manera se crea una escala de 12 valores, en las que figura cuanto vale la salida *u* de la neurona ganadora. La segunda definición toma en cuenta el valor relativo entre la ganadora y la segunda mejor categoría, (de ahí el nombre escala relativa), se restan ambos y el resultado *u* da 11 valoraciones.

Tabla 6.1. Valores de umbral absoluto.

Tabla 6.2. Valores de umbral relativo.

Umbral	Aciertos	Cantidad
u>1	99.4879533	23.2016767
u>0.9	99.4951598	40.8635933
u>0.8	99.4610682	57.2985633
u>0.7	99.3833007	71.93785
u>0.6	98.9796252	82.9647073
u>0.5	98.526901	91.0795053
u>0.4	97.7637172	96.7298477
u>0.3	97.0972333	99.4154713
u>0.2	96.8394349	99.9578627
u>0.1	96.8014474	99.9999901
u>0	96.8014474	99.9999901
u>-Inf	96.8014474	99.9999901

Umbral	Aciertos	Cantidad
u>1	99.87245	7.67772567
u>0.9	99.8970721	19.3364957
u>0.8	99.9226983	31.906278
u>0.7	99.8705907	45.1974913
u>0.6	99.84839	58.1253453
u>0.5	99.8049754	69.347053
u>0.4	99.7775325	78.0990273
u>0.3	99.5258619	85.413403
u>0.2	99.0902489	91.6482637
u>0.1	98.2417746	96.4560563
u>0	96.841296	100.000028

La última fila de cada tabla debería ser 100% en la columna *Cantidad*, pero por problemas de redondeo da valores cercanos pero diferentes. Puede verse en estas tablas el compromiso entre capacidad de clasificación y cantidad de muestras clasificadas automáticamente.

A modo de experimento, se utilizaron las redes Radiales Aproximadas de 400 neuronas entrenadas en las *30 repeticiones*, y se planteó un valor *umbral* de certeza que garantice un 99.5%, aproximadamente, de aciertos automáticos. El valor exacto de aciertos, en promedio, hallado por un algoritmo de búsqueda binaria para cada red particular, fue 99.54961% al 84.18114333% de los datos. El restante 15.8% de los datos se clasifican con la variante de cuatro mejores predicciones por operador. Como es de esperarse, el porcentaje de aciertos para el clasificador manual baja comparado a la *Tabla 3.1*, porque las muestras más fáciles ya fueron clasificadas por el automático, quedando así las más confusas. La tabla siguiente muestra justamente eso:

Tabla 6.3. Comparación de datos filtrados y completos.

	Aciertos			
Datos	**1**	**2**	**3**	**4**
Filtrados	78.95593	95.3904933	97.9556667	98.7566933
Completos	96.8404367	99.2469733	99.6787633	99.79989

La columna de datos en la *Tabla 6.3* presenta dos opciones, los *Completos*, que es una copia, con mas cifras decimales, de la segunda fila de la *Tabla 3.1* de Aciertos de Prueba de la red Radial Aproximada, y los *Filtrados*, que son las muestras que fueron clasificadas como *'no se'* por la Aproximada automática y realimentadas en la Aproximada manual de 4 categorías. Como puede observar el lector, el porcentaje de aciertos de *Filtrados* baja dramáticamente para la primer predicción con respecto a *Completos*, pero en la cuarta, es solo un 1.04% menor. Recordar que en este caso no importa la primera predicción, porque la elección quedará a cargo de un operador, lo que importa es que la categoría correcta esté entre las cuatro mostradas, por lo tanto el porcentaje de aciertos de datos *Filtrados* es 98.76%.

Finalmente se tiene un clasificador automático para el 84.2% de los datos, con una precisión del 99.55%, y para el 15.8% restante de las muestras, un clasificador manual con 98.76% de capacidad de mostrar la categoría correcta entre las cuatro presentadas. Si el operador humano no introduce errores adicionales, (al 15.8% de los datos manuales), la capacidad de clasificación final será del 99.42%, haciendo la suma ponderada respectiva. El lector puede apreciar que esta precisión es casi igual a los mejores clasificadores vistos en la *Tablas 3.1*, pero con el valioso agregado de la automatización del 84.2% de los datos. Puede decirse que el mencionado porcentaje de datos automatizado es inversamente proporcional a la precisión final buscada, dando lugar a ajustes. Como comentario final, también puede utilizarse este método con otros clasificadores, con otros parámetros o tamaños, o con reentrenamiento de los clasificadores con el conjunto *Filtrado* de

datos. Todas estas variantes abren futuros caminos de experimentación y pueden dar lugar a interesantes resultados para llevar a una versión de producción. Nuevamente esas variantes escapan al alcance del presente trabajo.

Conclusiones

Las Redes Neuronales tienen diversos parámetros que deben ser ajustados empíricamente. Hay que aclarar que los clasificadores estadísticos hacen más suposiciones sobre los datos que las Redes Neuronales para tener buena capacidad de clasificación, por ejemplo Naïve Bayes supone independencia de las variables (características) y una distribución normal de los datos, y el clasificador Cuadrático normalidad de los datos y gran cantidad de muestras de entrada. Todas estas condiciones parecen cumplirse para los datos presentados, porque el desempeño de los clasificadores fue excelente, pero pueden no sostenerse en otros conjuntos. Por eso es imprescindible tomar estas conclusiones en función de estos datos y de los parámetros utilizados.

Los seis clasificadores elegidos para las pruebas son muy buenos, separando sólo un 2.28% al mejor de ellos del peor. No puede decirse que un clasificador estadístico es mejor que una Red Neuronal en general, sino que en este caso fue superior.

Se estudiaron los errores incurridos por los clasificadores, y aparentemente no hay una deficiencia en los datos. La independencia de los errores permitiría utilizar una combinación de diversos clasificadores, (ya sean de distintos tipos o del mismo pero con variantes de entrenamiento), para mejorar la capacidad de clasificación. Sobre este punto es posible hacer futuros trabajos de investigación.

Se presentó una alternativa mixta de clasificación que abre las puertas a la experimentación y brinda un parámetro ajustable, un compromiso, entre automatización y precisión. La propuesta presentada es simplemente un ejemplo, y abre las puertas a futuros trabajos.

Referencias

[Alf2002] E. Alfaro Cortés, M. Gámez Martínez, N. García Rubio. "*Una Revisión de los Métodos de Clasificación"*. Área de Estadística. Departamento de Economía y Empresa. Universidad de Castilla - La Mancha.

[Bla2003] E. Blanzieri. "*Theoretical interpretations and application of Radial Basis Function Netowroks"*. 2003

[Cec2000] P. Granitto, H. Navone, P. Verdes, A. Ceccatto. "*Weed Seeds Identification by Machine Vision"*. Instituto de Física Rosario, CONICET y Universidad Nacional de Rosario, Boulevard 27 de Febrero 210 Bis, 2000 Rosario, Argentina.

[Koh2001] T. Kohonen. "*Self Organizing Maps"*. Third Edition, Springer, 2001.

[Mic1994] D. Michie, D.J. Spiegelhalter, C.C. Taylor. "*Machine Learning, Neural and Statistical Classification"*. February 17, 1994

[Per2000] M. I. Acosta Buitrado, C. A. Zuluaga Muñoz. "*Tutorial sobre Redes Neuronales aplicados en la Ingenieria"*. Universidad Tecnologíca de Pereira, Facultad de Ingenieria.

[Rie1993] M. Riedmiller, H. Braun. "*A direct adaptive method for faster backpropagation learning: The RPROP algorithm"*. Proceedings of the IEEE International Conference on Neural Networks, 1993.

[Wal1999] R. Walpole, R. Myers, S. Myers. "*Probabilidad y estadística para ingenieros"*. Sexta edición, 1999.