Introducción al Diseño de Experimentos para el Reconocimiento de Patrones

Capítulo 3: Redes Neuronales Artificiales

Curso de doctorado impartido por
Dr. Quiliano Isaac Moro
Dra. Aranzazu Simón Hurtado
Marzo 2004

Contenido

Repaso de conceptos

Una red neuronal es un procesador masivamente paralelo distribuido que es propenso por naturaleza a almacenar conocimiento experimental y hacerlo disponible para su uso.

- Aprende por ejemplos, ajustando los pesos de las conexiones entre los elementos que la constituyen.

- Elemento de proceso.
- Capa.
- Regla o algoritmo de aprendizaje
  - Paradigma de aprendizaje: forma en que la RNA interactúa con su entorno.
    - supervisado, no supervisado, por refuerzo, híbrido.
  - Modo de operación: síncrono / asíncrono
  - Sin realimentación / Con realimentación

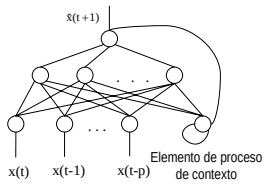
RNA frente a otros modelos de procesamiento

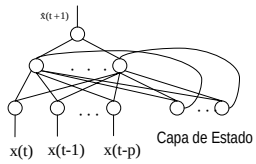
- Eminentemente paralela.
  - Habitualmente no se aprovecha esta característica
- Enfoque intrínsecamente modular.

  Neurona → capas → redes → sistema
- Procesamiento no simbólico de la información.
- Representación distribuida de la solución.
  - Tolerancia a fallos.
  - Robustez ante entradas ruidosas o incompletas.
- Es un modelo de "caja negra".
  - Por lo general no justifica las respuestas

Tipos de RNA más utilizados

- Perceptrón.
  - Algoritmo de aprendizaje: Regla Delta.
  - Separabilidad lineal.
  - Teorema de convergencia del MLP.
- MLP.
  - Algoritmos de aprendizaje: BP, Gradiente conjugado, ...
    - Retropropagación en el tiempo.
  - MLP como generador de universal de funciones.
  - Sobreentrenamiento.
  - Variantes con realimentación:
    - Jordan,
    - Elman.
- SOM.
- LVQ.
- RBF.



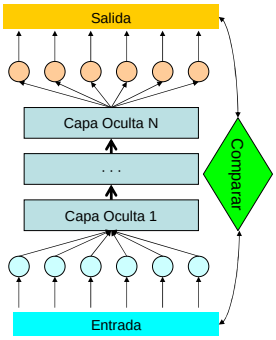


Técnicas de Clasificación con RNA

- Enfoque básico.
  - No supervisado.
    - SOM.
  - Por refuerzo.
  - Supervisado.
    - Establecimiento de las clases: codificación
      - LVQ.
      - MLP.
  - Híbrido.
    - RBF.

Técnicas de Clasificación con RNA

- Enfoque autoasociativo.
  - Se entrena la red con casos SÓLO de la clase a detectar:
    entrada = salida
  - En prueba,
    - Si el vector de entrada pertenece a la clase con la que ha sido entrenado, generará una salida parecida a él (idealmente sería una salida igual)
    - (vector de entrada – vector salida generada) próximo a cero si la entrada pertenece a la clase para la que fue entrenado.
    - (vector de entrada – vector salida generada) diferente de cero para otras clases distintas a la de entrenamiento.



7

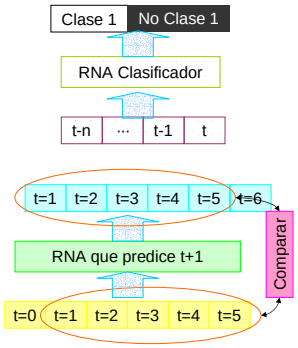
Técnicas de Clas. con RNA: Series Temporales

- El problema de las series temporales se puede considerar como:
  - Clasificación.
  - Pronóstico.
  - Descripción.
  - Transformación.
  - Todos los problemas son intercambiables.

8

Técnicas de Clas. con RNA: Series Temporales

- Uso de ventanas.
  - Generalmente para MLP y sistemas no recurrentes.
  - Anchura de la ventana.
- Modelo predictivo.
  - Entrenar para predecir el siguiente valor de la serie formada por las instancias de la clase a detectar.
  - Se recolecta la salida pronosticada al alimentar la red con una secuencia desconocida.
    - Si la salida pronosticada es parecida a la salida real es que la secuencia presentada es de la clase para la que fue entrenada la red.
      - Se puede usar un criterio de distancia entre vectores.



9

Esquema básico de funcionamiento

- Fases
  - Entrenar.
    - Atención al sobreentrenamiento.
      - Weight decay.
  - Validar. Pretende determinar si el modelo es bueno.
    - Evitar el sobreentrenamiento
      - Early stopping.
  - Evaluar
    - Pretendemos obtener una estimación de la precisión de la clasificación.
  - Uso con datos de un problema real.
- Dependiendo del caso:
  - Entrenar + Validar.
  - Entrenar + Evaluar.
  - Entrenar + Validar + Evaluar.
- Dependiendo del número de fases, se debe hacer una división correcta de los datos disponibles.

10

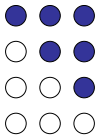
Preprocesamiento de datos

- Reescalar
  - Casi siempre es recomendable para mejorar convergencia.
  - Es obligatorio, por ejemplo para las salidas.
    - El reescalado depende de la función de activación usada.
    - ¿Normalizar a [0,1] las salidas?
      - NO si el criterio de parada de entrenamiento es el error de aprendizaje y la función de activación de salida es sigmoide.
  - Estandarización (p.ej. a media 0 y varianza 1) es recomendable cuando están involucradas medidas de distancias, (p.ej. RBF).
- Normalización.
  - Los vectores de datos se normalizan, p.ej. dividiendo por su módulo.
    - Ej. Para SOM.
- No lineales para estudiar zonas específicas.
  - Ej.: uso del logaritmo cuando
    - Interesa una medida relativa de los valores.
    - Se tiene una idea de que las entradas actúan de forma multiplicativa.

11

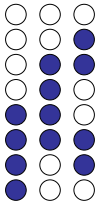
Preprocesamiento de los datos: Codificación

- Codificación.
  - Magnitudes progresivas.
    - Es costoso en cuanto elementos de proceso → conexiones.



Valor alto
Valor medio
Valor bajo
Valor muy bajo

  - Magnitudes cíclicas.
    - Usar códigos continuos y cíclicos.
      - P.ej. para representar ángulos.
    - Es costoso en elementos de proceso → conexiones.

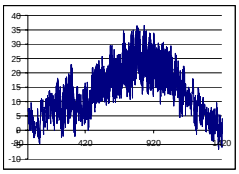


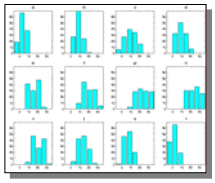
0-44º
45-89º
90-134º
135-179º
180-224º
225-269º
270-314º
315-359º

12

Preprocesamiento de datos: Series Temporales

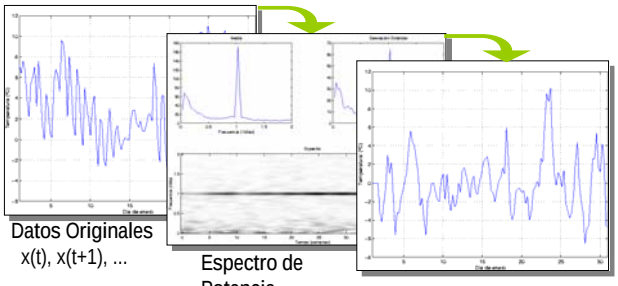
- Estudio de la serie temporal:
  - Muy interesante la representación gráfica.
  - Determinar
    - Estacionaridad.
    - Estacionalidad.
  - Dominio temporal
    - Autocorrelación.
  - Dominio frecuencial
    - Transformada de Fourier (espectro frecuencial)
- Preprocesamiento
  - Eliminación de la tendencia
    - Determinar su existencia.
  - Eliminación de la estacionalidad.
    - Filtrado





13

Estacionalidad: Filtrado



Datos Originales
x(t), x(t+1), ...

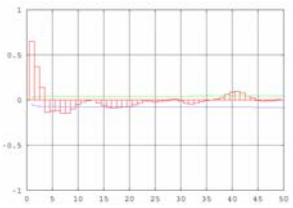
Espectro de Potencia
Periodicidad cada 4 muestras.

Datos Filtrados
x'(t)=x(t)-x(t-4)

14

Preprocesamiento de datos: Series Temporales

- Coeficientes y funciones de de autocorrelación
  - ¿Existe relación o influencia entre los valores de una muestra y las que la precedieron en el tiempo?
  - Son una medida de la relación (lineal) entre observaciones separadas K periodos de muestreo.
    - Usada para determinar la anchura de la ventana temporal.



15

Estudio de la Importancia de los datos

- Determinar qué datos de entrada son los más discriminantes.
- En teoría permitiría construir RNA más sencillas
  - Aumentaría su potencia generalizadora.
- Análisis de la importancia de las entradas
  - A priori:
    - Análisis de Componentes Principales.
      - ¡Se puede hacer ACP incluso con RNA! - ACP no lineal.
    - Análisis Factorial Discriminante.
      - » ¿Las variables medidas permiten realizar la clasificación buscada?
      - » ¿Cómo se comporta cada variable en cuanto a su efecto sobre la clasificación?
      - » ¿Cuáles son las variables o grupos de variables que mejor ayudan a la clasificación?
    - Busca
      - » Máxima distancia entre clases.
      - » Mínima distancia intra-clase.
  - A posteriori:
    - Análisis factorial.
    - Análisis de pesos (después del entrenamiento).
    - Poda

16

Diseño del conjunto de entrenamiento

- Abundancia relativa de tipos en la población.
  - Técnicas de igualación:
    - Al menos numeroso.
    - Al más numeroso.
    - Generar casos próximos por agregación de ruido (jitter).
- Mantener la variabilidad.
- Evitar el sobreentrenamiento usando muchas muestras.
  - Siempre acorde con el tamaño de red.
- Prestar atención a los casos de frontera (Bordeline cases)
  - Los no-borderline aportan información poco relevante.
  - Posibles alternativas:
    - Replicar los borderline.
    - Dos fases de entrenamiento:
      1. Con todos.
      2. Después con los difíciles.

17

Diseño del conjunto de entrenamiento

- Casos atípicos
  - Datos fuera de rango
    - ¿Son realmente casos válidos o fallos de lectura?
    - ¿Descartarlos?
  - Datos perdidos
    - Representarlos por entradas especiales.
    - Reemplazarlos por datos "estándar".
  - En cualquier caso valorar si es posible tomarlos de nuevo.

18

Evaluación de la Clasificación

- La medida más adecuada depende de la tarea.
- Error cuadrático medio.
  - No parece adecuado para tareas de clasificación.
  - Todas las salidas tienen el mismo peso.
- Función de coste

$$COSTE = \sum q_i p_i c_i$$

q_i = probabilidad a priori de la clase i

p_i = probabilidad de que la red falle al detectar la clase i.

c_i = coste de los fallos al detectar la clase i.

  - Se puede particularizar para cada tipo de fallo.
    - Una forma de ver esto es con la Matriz de Confusión.

Matriz de Confusión

- Tantas filas como clases en los datos.
- Tantas columnas como clasificaciones pueda realizar el sistema.
  - una columna de casos no reconocidos.
    - Se la suele asociar a un umbral sobre la salida de las otras clases.
- En la celda el número de casos tipo i, que clasificados como j.
  - también probabilidades.
- Lo ideal es una matriz diagonal.
- Más detalle si se usa función coste.
  - El punto clave es decidir los ci.
    - Ligeras variaciones pueden hacer que cambie mucho el coste.
- Esta misma idea se puede usar para optimizar el aprendizaje (usando funciones de coste en vez de RMSE)

	Normal	Benigno	Maligno
Normal	96	3	1
	.960	.030	.010
	.576	.018	.006
	0	.1018	.060
Benigno	2	47	1
	.040	.0940	.020
	.012	.282	.006
	.024	0	.028
Maligno	1	2	27
	.033	.067	.900
	.003	.007	.090
	.333	.333	0

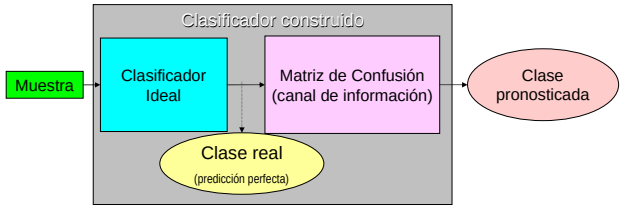
Matriz de Confusión y Teoría de la Información

- La Matriz de confusión puede interpretarse como la matriz que define un canal que emborrona la salida de un hipotético clasificador perfecto.
- Un clasificador será "bueno" cuanta mayor certeza sobre la clase real al observar la clase pronosticada.
  - Cambiar certeza por entropía (o incertidumbre media):
    $H(\Omega)= -\sum p_i \log(p_i)$, siendo p_i la probabilidad de pronosticar la clase i.
  - Interesa que el hecho de observar la salida del clasificador haga disminuir la incertidumbre sobre la verdadera clase.
    - Información mutua: disminución de la incertidumbre sobre la clase de los datos de entrada después de observar la salida del clasificador.

      entropía a priori - entropía a posteriori

Matriz de Confusión y Teoría de la Información

- Idea interesante pero poco usada
  - requiere un gran número de muestras (de entrenamiento y prueba) para poder evaluar las correspondientes probabilidades.
- Puede dar origen a técnicas de "des-emborronamiento"
  - P.ej.: casos "no reconocidos" pueden ser asignados a la clase más probable



Curvas ROC

- En problemas de clasificación el sistema da una salida:
  - Valor Bajo → tipo A
  - Valor Alto → tipo No A
- Se puede fijar un umbral para hacer la separación (A, No A)
  - Si y < u → Tipo A
  - Si y ≥ u → Tipo B
- Tipos de errores:
  - Error de tipo I: decir que la condición está presente, cuando en verdad es que no → falsos positivos ó falsos aceptados.
  - Error de tipo II: decir que la condición no está presente, cuando en verdad es que sí → falsos negativos ó falsos rechazados.
- La probabilidad de estos errores depende del umbral.

Curvas ROC

- Sensibilidad: habilidad para detectar los verdaderos positivos.

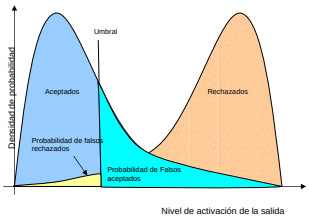
  S=a/(a+b)
  - Todo verdadero positivo ha de poder ser detectado.
    - En este empeño pueden aparecer falsos positivos.
- Especificidad: capacidad de detectar aquello para lo que ha sido creado el clasificador.

  E=d/(c+d)
  - Todo falso negativo debería poderse detectar.
    - En este empeño pueden colarse falsos
- Idealmente S = 1.0 y E = 1.0, pero es muy difícil de obtener.

	Clase Real		
Pronóstico	X	No X	Total
X	a	c	a+c
No X	b	d	b+d
Total	a+b	c+d	

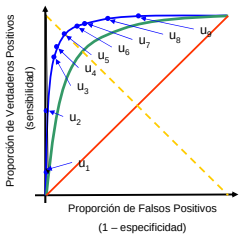
Curvas ROC

- Supongamos que hemos determinado las distribuciones de Clase A y No Clase A en función de un umbral de decisión.
  - Salida < umbral → Clase A
  - Salida ≥ umbral → No A.
- En situación real siempre hay solapamiento de ambas distribuciones.



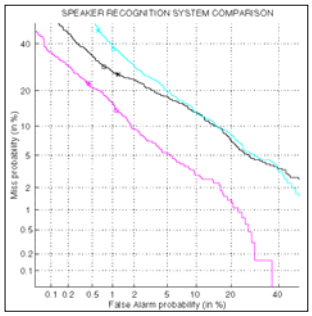
Curvas ROC

- Curva ROC (Característica Operativa del Receptor).
  - (falsas aceptaciones, falsos rechazos) parametrizada por el umbral.
  - Hay un balance entre sensibilidad y especificidad.
    - Un aumento en una se traduce en una reducción de la otra.
- Habrá una curva por cada clasificador.
- La eficiencia se mide por la superficie bajo la curva.
  - Las curvas exteriores son mejores.
- Tasa de equierror:
  - Medida muy popular.
  - Misma proporción de falsos acertados y de verdaderos rechazados.
  - Define un umbral.



Curvas DET

- Curvas DET (Detection Error Tradeoff)
  - Las curvas ROC son difíciles de comparar
    - distintas curvas pueden tener misma superficie bajo ellas.
  - Se busca una representación cercana a la lineal.
    - Se logra haciendo que los ejes tengan escalas no lineales.
    - Se representan desviaciones normales en la desviación normal estandarizada correspondiente a esa probabilidad.



Clasificadores Multiclase

- Codificación de las salidas:
  - Con una salida se usa umbrales.
  - Con varias salidas.
    - Una salida para cada clase.
    - Cuando hay alguna relación entre las distintas clases (por ejemplo, de orden).
      - Codificación incremental.
- Agregar la clase “desconocida”.
  - Se usa junto con umbrales:

Múltiples Predicciones en Series Temporales

- Predicción (clasificación) a varios horizontes temporales.
  - Clasificación de varias características a la vez.
  - Si los horizontes temporales son “próximos” con una única RNA podría ser suficiente.
  - Sopesar la modularización para evitar la interferencia
    - Una RNA para cada pronóstico / clase.
- Problemas al determinar la bondad de la clasificación
  - Se tiene que evaluar cada clase por diferentes criterios.
    - Considerar este problema incluso en el aprendizaje.

Métodos para determinar la exactitud

- Determinar de manera correcta una estimación de la exactitud de la clasificación.
  - Permite elegir entre varios modelos de clasificadores.
  - NO DETERMINAN LA CONFIGURACIÓN DEL CLASIFICADOR MÁS CORRECTO.
- Obliga a una división bien diferenciada entre datos de entrenamiento y prueba.
  - Los datos de entrenamiento incluirán también (si se precisan)
    - Determinación de umbrales.
    - Validación y selección de otros parámetros.
- Dilema “bias-variance”.
  - Originalmente sólo para problemas de regresión:
    error cuadrático = bias^2 + varianza
  - No se conoce una fórmula equivalente para problemas de clasificación.
    Exactitud = promedio de casos correctamente clasificados

Métodos para determinar la exactitud

- Resustitución
  - Usa los mismos datos para entrenar y para probar.
  - Resultados optimistas.
  - Desaconsejable, salvo casos específicos:
    - Sistemas de clasificadores lineales y muchos ejemplos.
- Holdout
  - Del conjunto de datos disponibles se selecciona (muestrea) aleatoriamente dos conjuntos mutuamente excluyentes:
    - Datos para entrenamiento (habitualmente 2/3 del total).
    - Datos para prueba (habitualmente 1/3 del total).
  - Es considerado como una evaluación pesimista.
    - El número de datos reservados para entrenar es pequeño comparado con el total.
  - Random Subsampling: se repite k veces el procedimiento anterior y se hace la media de las exactitudes obtenidas.
    - Problema: las exactitudes calculadas no son datos independientes.

31

Métodos para determinar la exactitud

- Leave k-Out
  - Separar k datos, entrenar con los demás. Evaluar con los k datos apartados.
    - Repetir lo anterior apartando otros k datos diferentes.
  - Obtener una media de los resultados.
  - Ofrece resultados no sesgados por la elección de los datos de prueba.

32

Métodos para determinar la exactitud

- Validación cruzada
  - Es una modificación del Leave-K-out.
  - Dividir aleatoriamente el conjunto de datos D en K partes disjuntas $\{D_1,... D_k\}$, procurando que tengan tamaños parecidos.
  - El clasificador es entrenado y entrenado K veces, excluyendo en el entrenamiento cada vez uno de los subconjuntos:
    - Entrenado con todos los D_i excepto el k-ésimo
    - Probado con D_k.
  - La precisión será la media de las K obtenidas.
    - Garantiza haber probado con todos los datos.
    - Es ligeramente pesimista.
    - Hay otras posibilidades.
  - Validación cruzada estratificada: conseguir que en cada subconjunto D_i haya igual número de ejemplares de cada clase representada.

33

Métodos para determinar la exactitud

- Bootstrap
  - Selecciona las muestras de entrenamiento de forma aleatoria, permitiendo la repetición de los ejemplos.
    - En un conjunto de entrenamiento puede que no estén representadas todas las clases.
    - En un conjunto de entrenamiento puede haber muestras repetidas varias veces.

34