Universidad
Carlos III de Madrid

Técnicas mono y multiobjetivo para la evolución de transformaciones lineales en problemas de clasificación

Iván Monteagudo García

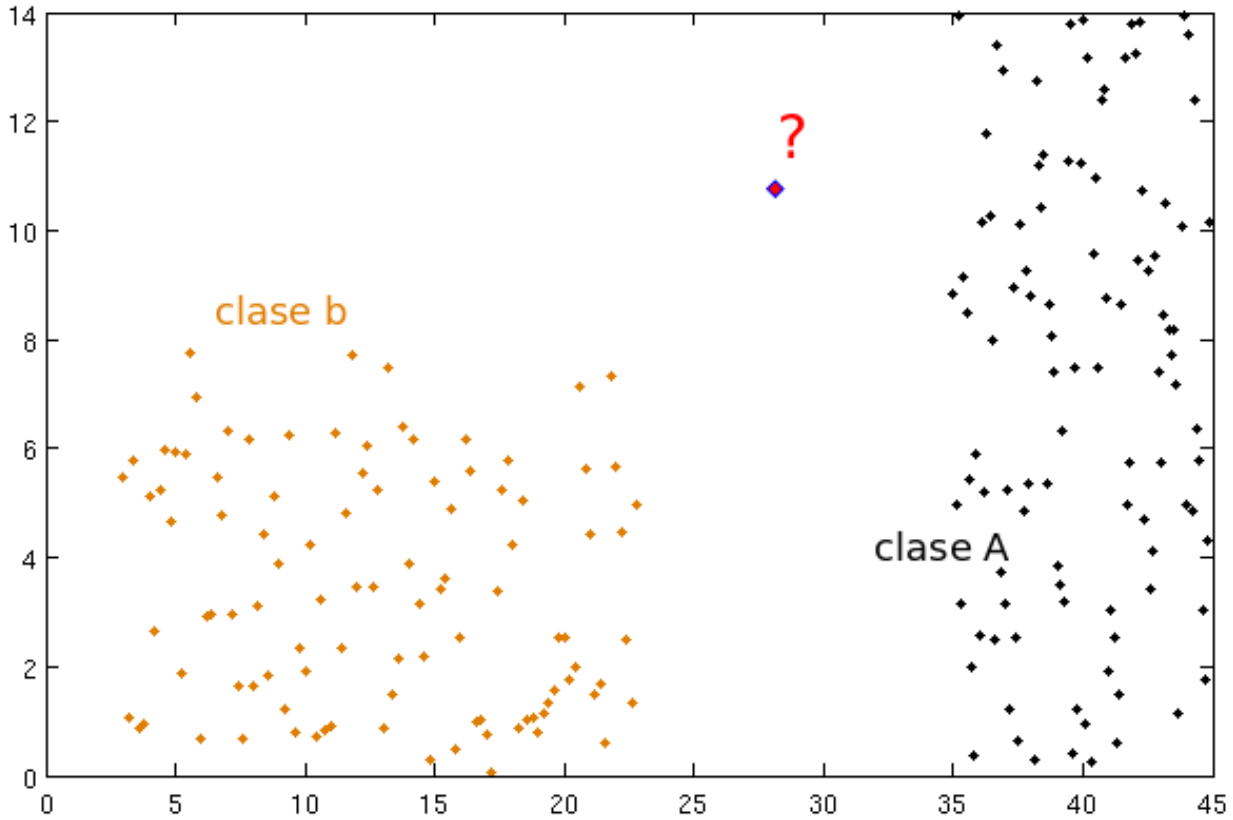
Ingeniería en Informática

27 de julio de 2010

Índice

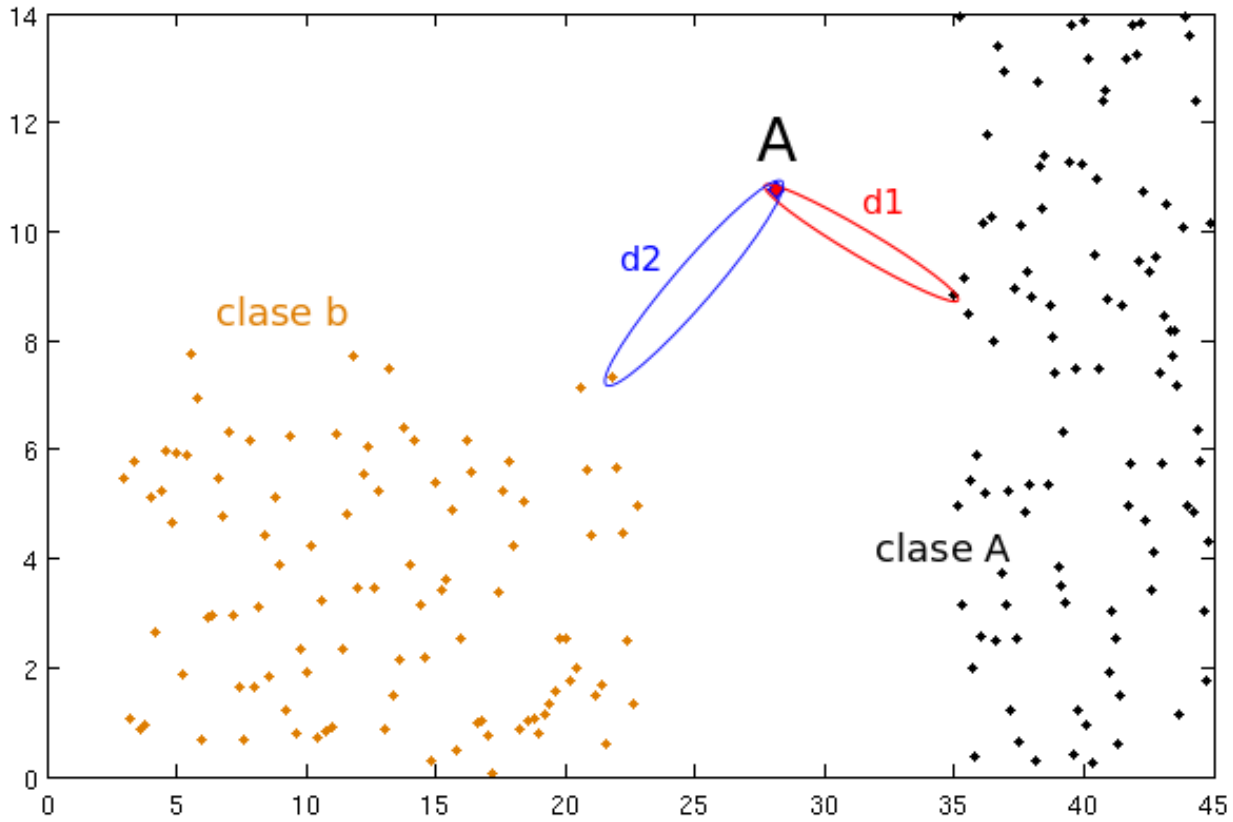
Fundamentos

- Algoritmos de clasificación : Determinan la pertenencia de una *instancia* a una *clase*.



Fundamentos

- Algoritmos de clasificación : Determinan la pertenencia de una *instancia* a una *clase*.
- K-Vecinos lo hace a partir de distancias.



Definición del problema (I)

Problemas:

- Escala
- Importancia

Definición del problema (I)

Problemas:

- Escala
- Importancia

Soluciones – Modificar:

- El cálculo de la distancia.
- La posición de los datos.

Soluciones

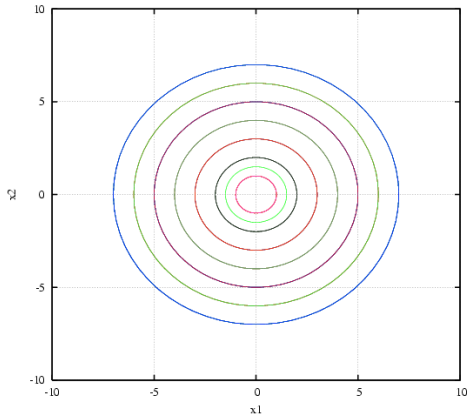
Cálculo de distancias:

- Diversos métodos para cálculo de distancias.
- Euclídea, Hamming, Mahalanobis,
- **Euclídea Generalizada**

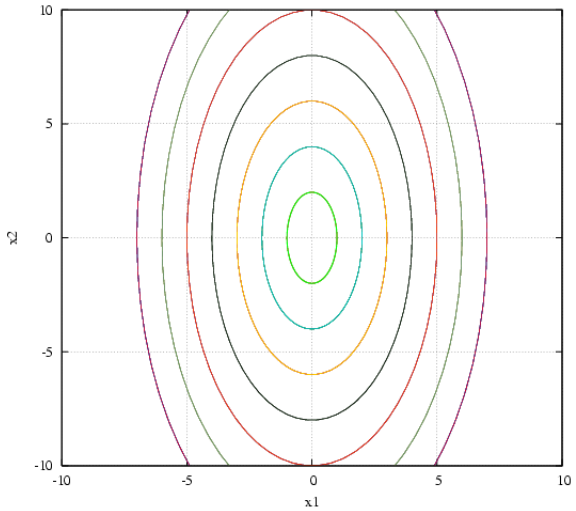
$$\sqrt{\sum_{i=0}^{n}(x_i - y_i)M_d M_d^T (x_i - y_i)^T}$$

Soluciones - Distancias

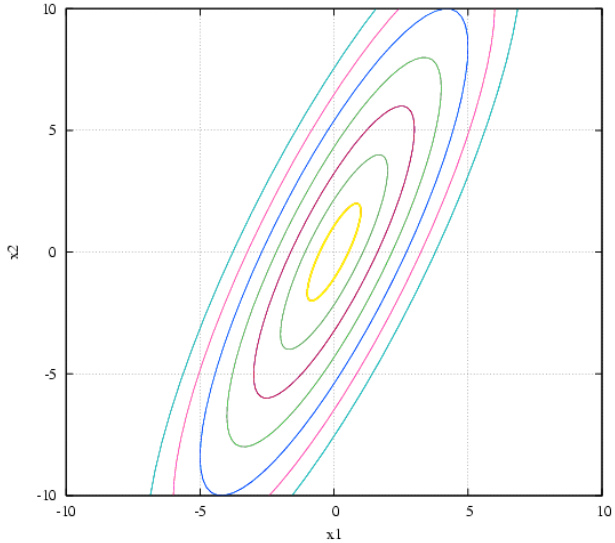
Representación de las distancias



Distancia Euclídea



Matriz Diagonal



Matriz Completa

Soluciones

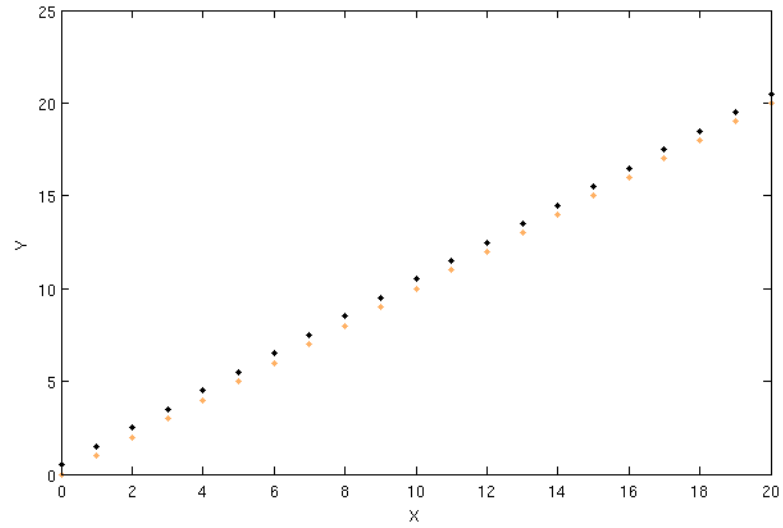
Posición de los datos:

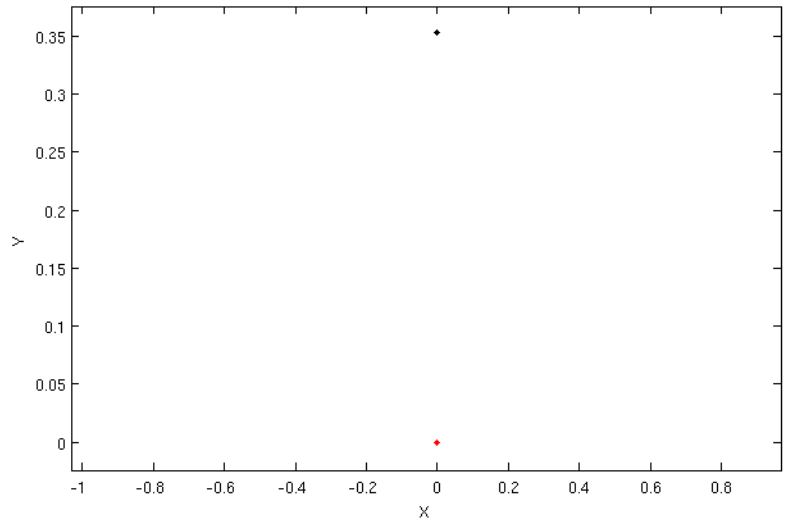
- "Reposicionar" : Aplicación $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$
- Basta una transformación lineal, representable con una matriz $M \in \mathbb{R}^{n \times n}$
- Nueva posición de los datos $\vec{posicion'} = M \times \vec{posicion}$
- Buscamos una **matriz de transformación.**

Soluciones - Transformaciones

Ejemplo de transformación

$$\begin{pmatrix} 0 & -sen(\pi/4) \\ 0 & cos(\pi/4) \end{pmatrix}$$





Definición del problema (II)

El problema a resolver es:

Construir un sistema **general** para encontrar matrices de transformación que, aplicadas a un dominio, mejoren la capacidad de K-Vecinos para clasificar sus instancias.

Método

- Búsqueda en el dominio de las matrices reales $n \times n$.
- El espacio de búsqueda es infinito.
- **Heurística**
- Método que encuentre buenas soluciones únicamente a partir de:
  - Formato de la solución : Matriz de números.
  - *Calidad* de la solución : prueba de clasificación.

Técnicas Evolutivas

Método : Técnicas Evolutivas

- Técnicas basadas en la proximidad referencial y abstracción de conceptos biológicos.
- CMA-ES : Añade conceptos de estadística multidimensional para acelerar y diversificar la búsqueda.
- NSGA-II : Multiobjetivo. Explora los mejores resultados límite de los objetivos.

Obtención de Resultados

Resultados:

- Basados en estadísticas.
- Los resultados tanto de CMA-ES como de K-Vecinos dependen del conjunto de entrenamiento.
- **Validación Cruzada**

Herramientas

- Se quieren probar diversos métodos y ajustar según el resultado.
- Implementación en MATLAB.
- Estructura modular –*Toolboxes*.

Implementación

- Basada en 3 módulos diferenciados:
  - Bucle principal de validación cruzada.
  - Ejecución del evolutivo.
  - Cálculo de *fitness* mediante prueba de clasificación.

Implementación

- Basada en 3 módulos diferenciados:
  - Bucle principal de validación cruzada.
  - Ejecución del evolutivo.
  - Cálculo de *fitness* mediante prueba de clasificación.
- Entrada: Instancias del dominio.
- Salida: Datos estadísticos y *logs* completos con matrices obtenidas, etc.

Pruebas Iniciales

Objetivos – Determinar :

- La utilidad de las transformaciones.
- Diferencias entre varios métodos básicos.

Pruebas Iniciales

Objetivos – Determinar :

- La utilidad de las transformaciones.
- Diferencias entre varios métodos básicos.

Dominios:

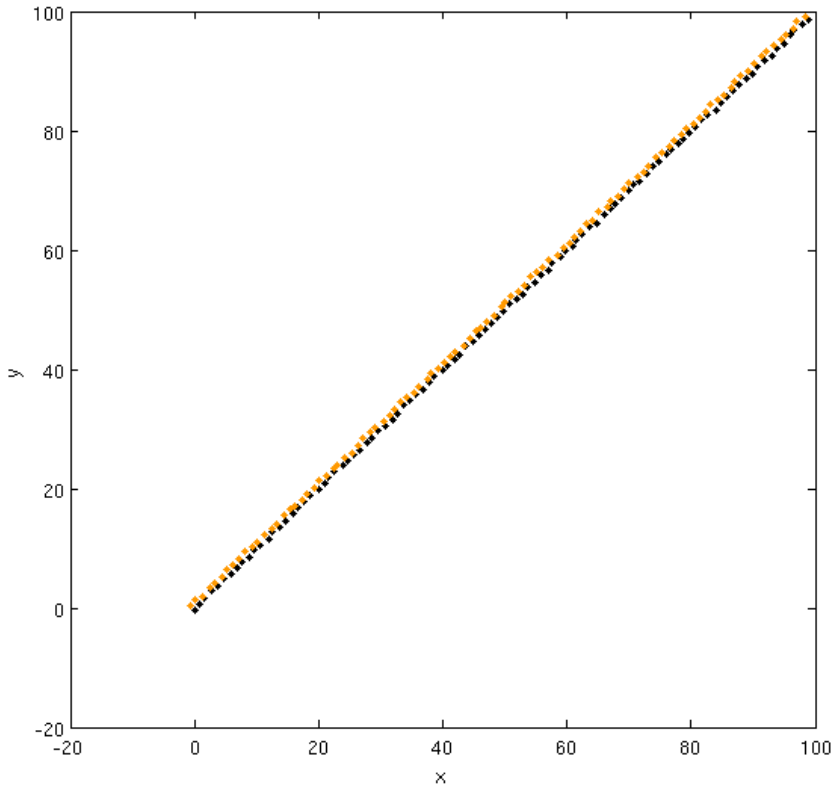
- 3 dominios reales : Iris, Ripley, Wine.
- 2 dominios sintéticos.

Dominios de clasificación

- Aleatorio : 4 atributos
  - 2 tienen importancia en la clasificación $[0, 1]$.
  - 2 no la tienen $[0, 100]$. Confunden las distancias.

Dominios de clasificación

- Aleatorio : 4 atributos
  - 2 tienen importancia en la clasificación $[0, 1]$.
  - 2 no la tienen $[0, 100]$. Confunden las distancias.
- Rectas45



Pruebas Iniciales : Resultados

Dominio	Euclidea	LS-diag	CMA-diag	LS	CMA-comp
Aleatorio	50.96	94.10	**95.70**	55.50	62.23
Iris	**95.86**	94.66	94.26	95.13	95.26
Rectas45	08.30	09.00	08.85	**98.70**	**98.70**
Ripley	88.72	**88.88**	88.76	88.54	88.77
Wine	76.48	94.15	**94.16**	74.35	80.86

- La transformación mejora los resultados.
- Algunos dominios funcionan mejor con matrices diagonales.

Pruebas inciales : Conclusiones

Las matrices diagonales:

- Sólo producen transformaciones de escalado y reflejo.
- El espacio de búsqueda es notablemente menor.

Pruebas inciales : Conclusiones

Las matrices diagonales:

- Sólo producen transformaciones de escalado y reflejo.
- El espacio de búsqueda es notablemente menor.

Método *híbrido* : Comenzar con una matriz completa y pasar a una diagonal.

	Aleatorio	Rectas45	Wine
Ratio Clasificacion	85.93	98.60	93.40

Fitness ponderada

Comprobar el efecto de aumentar la dispersión (ceros) de la matriz.

$$f = ErrorRate \times Peso + (1 - Peso)(1 - \frac{nCeros}{Dimension})$$

Minimizar f :

- minimizar el ratio de error de clasificación.
- maximizar el número de ceros normalizado.
- $Peso$ = importancia de la clasificación.

Además se incluye un umbral a partir del cual los elementos son 0.

Fitness ponderada : Resultados

Resultados para Aleatorio:

Peso	01	0.5	0.7	09	1
Umbral					
0	57.76	62.03	**63.13**	62.36	61.80
0.2	52.46	57.43	60.53	61.50	**62.76**
1	53.30	98.10	98.06	**99.23**	98.16
2	50.00	97.23	97.43	98.43	**99.03**
3	50.00	50.00	50.00	50.00	50.00
5	50.00	50.10	50.16	50.00	50.00

Fitness ponderada : Conclusiones

Conclusión

- Aumentar el número de ceros mejora los resultados.
- Dos parámetros a ajustar.
- 30 ejecuciones frente a 2 : diagonal y completa.

Evolución del umbral

Primer parámetro: Umbral.

- Optimización de un número real : CMA-ES
- Nuevo cromosoma : umbral + matriz de transformación.
  $INDIVIDUO_n = (umbral,\ m_{1,1},\ m_{1,2},\ \ldots,\ m_{k,k})$

Evolución del umbral

Primer parámetro: Umbral.

- Optimización de un número real : CMA-ES
- Nuevo cromosoma : umbral + matriz de transformación.
  $INDIVIDUO_n = (umbral,\ m_{1,1},\ m_{1,2},\ \ldots,\ m_{k,k})$

Aleatorio	Rectas45	Wine
99.23	98.70	91.29

Multiobjetivo

¿Por qué?

- Encontrar el balance entre clasificación y ceros.
- Posibilidad de añadir nuevos objetivos.
  - Para mejorar la clasificación.
  - Objetivos específicos de la situación: Reducir falsos negativos, etc.

Multiobjetivo : definición de objetivos

Los objetivos son minimizar error de clasificación más...

- Maximizar ceros.
- Minimizar el umbral.
- Maximizar desviación.
- Minimizar media.
- Maximizar ceros fuera de la diagonal.

Multiobjetivo : Resultados

	Aleatorio	Rectas45	Wine
Ceros	**99.60**	**99.55**	91.25
Ceros,Umbral	97.40	99.10	**91.94**
1/Desviación	54.30	98.90	77.33
Media/Desviación	61.16	**99.55**	75.11
Ceros noD	76.86	95.45	91.30

Multiobjetivo : Conclusiones

- Los resultados son tan buenos como los de los demás métodos.
- Pero **no** mejores.
- El coste en recursos es mayor.

Conclusión: Se puede utilizar si se tienen otros objetivos, pero si sólo se quiere el mejor ratio de clasificación hay métodos más eficientes.

Pruebas finales

Comprobaciones:

- ¿Se mejora el resultado de clasificación respecto a no realizar la transformación?
- ¿Es la diferencia significativa?
- ¿Cuál es el mejor método de entre todos los probados?
- ¿Es el método generalizable?

Test de significación

Dominio	Resultado Umbral	Resultado sin transformación	¿Diferencia significativa?
Aleatorio	99.23	49.40	**sí**
Bupa	63.18	62.11	no
Car	97.54	87.35	**sí**
Diabetes	67.18	67.53	no
Ionosphere	89.13	86.98	**sí**
Iris	94.93	96	no
Rectas45	98.7	6.95	**sí**
Ripley	88.56	88.71	no
Wine	91.30	76.14	**sí**
Yeast	51.64	51.97	no

Test de significación

Dominio	Mejor Método Anterior	Resultado Anterior	Resultado Umbral	¿Diferencia significativa?
Aleatorio	CMA-diag	95.70	99.23	**sí**
Bupa	CMA-comp	64.17	63.18	no
Car	CMA-comp	97.51	97.54	no
Diabetes	Euclídea	67.53	67.18	no
Ionosphere	CMA-comp	86.60	89.13	**sí**
Iris	Euclídea	96	94.93	no
Rectas45	CMA-Comp	98.7	98.7	no
Ripley	LS-diag	88.89	88.56	no
Wine	CMA-diag	94.17	91.30	**sí**
Yeast	Euclídea	51.97	51.64	no

Conclusiones

Evolucionando matriz de transformación y umbral:

- Mejoras significativas en 5 de 10 dominios.
- En un dominio se consiguen mejores resultados con una matriz diagonal.
- 2 dominios son artificiales y el resto reales.

Conclusiones

- Mejora significativa en el 50 % de los dominios probados.
- Aumentar la dispersión mejora el resultado en varios dominios.
- Método igual o mejor que todos los demás excepto en un caso.
- Para llegar hasta este método se han analizado varios más.
  - Desde búsqueda local a multiobjetivo.

Líneas futuras

- Mejorar la eficacia de K-Vecinos antes de pasar a dominios mayores.
- Buscar otros parámetros que se puedan modificar (K).
- Definir métodos para transformaciones que no cubra una matriz $n \times n$.
- Viabilidad de realizar predicciones de parámetros en una clase.

Presupuesto

Utilizado la plantilla de la rúbrica de la Universidad Carlos III.
Costes:

Personal 220 días a media jornada. 5 meses-persona.

Material Equipos de desarrollo y pruebas.

Software Matlab + Toolboxes.

Total: 26 127€.

Comentarios y pregunas

Gracias