Trabajo Final de Grado
Inteligencia Artificial

Universitat Oberta de Catalunya

Reducción de la dimensionalidad mediante métodos de selección de características en *microarrays* de ADN

Grado de Ingeniería Informática
Maseda Tarin, Miguel

10/01/2019

INDICE

1 INTRODUCCIÓN

2 SELECCIÓN DE CARACTERÍSTICAS

3 MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS

4 CONJUNTO DE DATOS

5 MÉTODOS DE VALIDACIÓN

6 MÉTRICAS

7 *f-score*

8 *mRMR*

9 *Sequential Forward Selection*

10 Método híbrido

11 Conclusiones

1. INTRODUCCIÓN

Los datos por si solos no ofrecen toda la información que en ellos se encuentran

Uso de técnicas de DM y ML para obtener un valor añadido a nuestros datos

Hay que tratar previamente los datos para poder hacer uso de las técnicas

Métodos de tratamiento previo de los datos:

- Imputación de valores ausentes
- Filtrado de ruido
- Reducción de la dimensionalidad
- Reducción de instancias
- Discretización
- Aprendizaje no balanceado

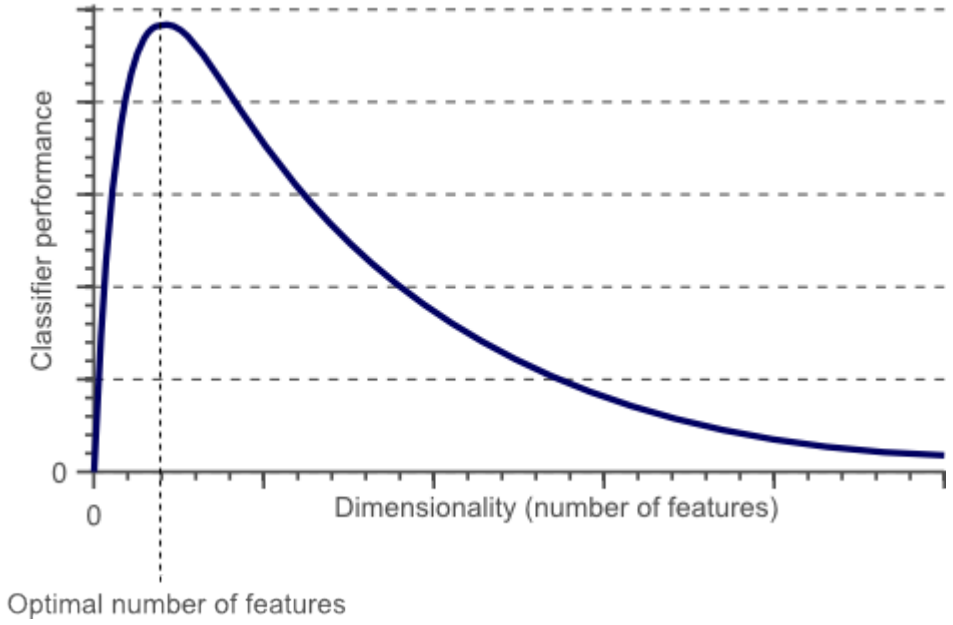
1. INTRODUCCIÓN

Métodos de clasificación supervisada

- Sabemos a qué clase pertenece cada muestra
- ¿a qué clase pertenecerá una nueva muestra?

Maldición de la dimensionalidad

Abordaremos la reducción de la dimensionalidad en los conjuntos de datos



2. SELECCIÓN DE CARACTERÍSTICAS

Dimensionalidad del conjunto de datos:

- Número de muestras: $X=\{x_1,x_2,\ldots,x_m\}$
- Número de características: $x_m=\{x_{m1},x_{m2},\ldots,x_{mn}\}$

Problema de dimensionalidad:

- m > n

Problema de dimensionalidad grave:

- m >> n

Métodos de reducción de la dimensionalidad:

- Selección de características
- Extracción de características

X	1	2	3	...	m
x_1					
x_2					
x_3					
$\vdots$					
x_n					

2. SELECCIÓN DE CARACTERÍSTICAS

Extracción de características:

- Transformación del conjunto original
- Las características originales no son necesarias para la interpretación del modelo
- Interpretabilidad < Exactitud

Selección de características:

- Seleccionan las características relevantes
- Las características originales son necesarias para la interpretación del modelo
- Interpretabilidad = Exactitud

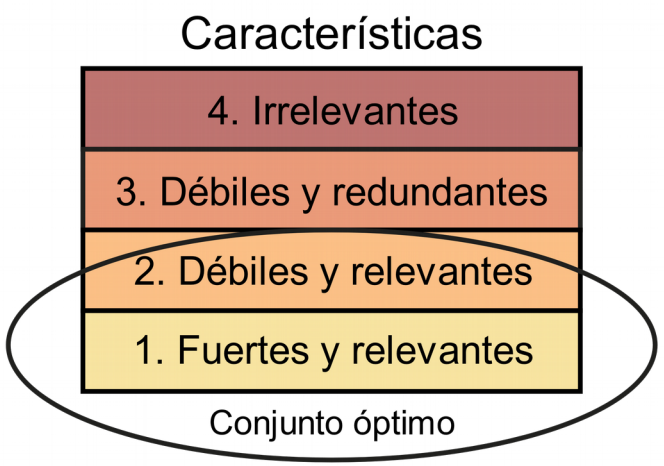
2. SELECCIÓN DE CARACTERÍSTICAS

Tipos de características (descripción de Kohavi):

- Fuertes
- Débiles
- Irrelevantes

Hay que tener en cuenta las características redundantes

1. Relevantes
2. Débiles pero no redundantes
3. Débiles redundantes
4. Irrelevantes



2. SELECCIÓN DE CARACTERÍSTICAS

Beneficios de la selección de características:

- Modelos más precisos y rápidos
- Reducción del espacio de almacenamiento y búsqueda
- Mayor comprensión sobre el conjunto de datos
- Modelos más simples, mejorando su visualización
- Pueden reducir los costes a la hora de recopilar nueva información

3. MÉTODOS DE SELECCIÓN DE CARACTERÍSTICAS

Filtro

- La selección es independiente al algoritmo utilizado
- Coste computacional bajo
- Gran capacidad de generalización
- Univariantes y multivariantes

Wrapper

- Utilizan un algoritmo ML que mide la eficacia de las características
- Gran coste computacional
- Se usan cuando el coste computacional no es un problema

Empotrados

- La selección de características se realiza en el propio algoritmo de ML
- Menor coste computacional que los métodos *wrapper*

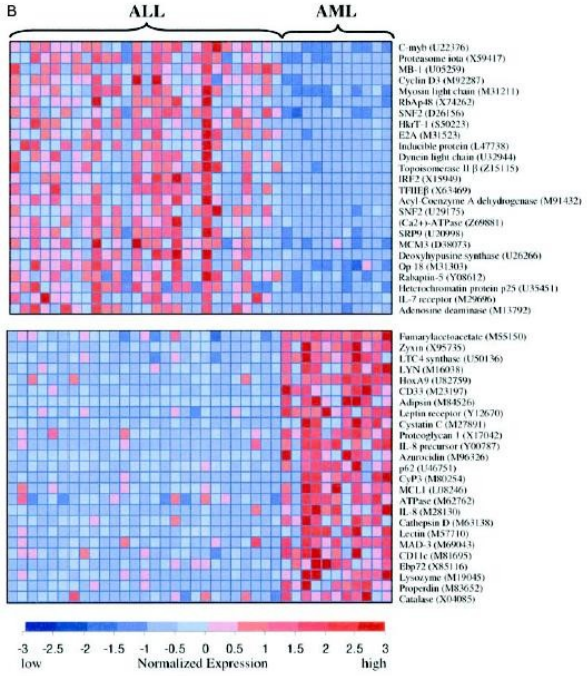
Híbridos

- Combinación de métodos de filtro y *wrapper*
- Buscan obtener las ventajas de estos métodos

4. CONJUNTO DE DATOS

Microarray de ADN: LEUCEMIA

- Contiene un total de 72 muestras
- Cada muestra tiene 7.129 características (genes)
- Se divide en dos clases:
  - Leucemia mieloide aguda, AML (25 muestras)
  - Leucemia linfoblástica aguda, ALL (47 muestras)



5. METODOS DE VALIDACION

Aplicación de los métodos de selección de características

Stratified k-fold (5 hojas):

- Conjunto de entrenamiento
- Conjunto de test

k-fold (10 hojas) al conjunto de entrenamiento:

- Conjunto de entrenamiento
- Conjunto de validación

Algoritmos de aprendizaje:

- *k Neighbours Classifier*
- *Decision Tree Classifier*
- *Support Vector Classifier*

6. METRICAS

Exactitud (*accuracy*)

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Precisión (*precision*)

$$\frac{TP}{TP + FP}$$

Exhaustividad (*recall*)

$$\frac{TP}{TP + FN}$$

TP = true positive
TN = true negative
FP = false positive
FN = false negative

7. *f-score*

Función *f_score* de la biblioteca *scikit-Feature*

Valoración individual de las características

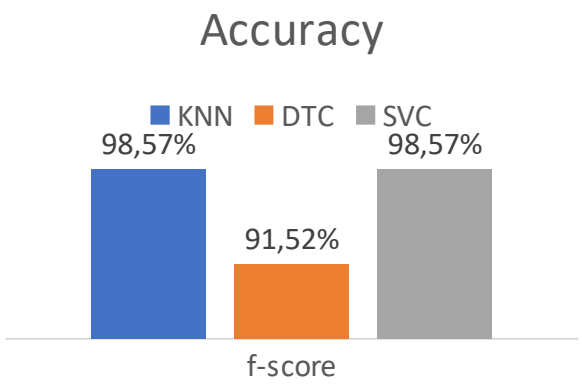
Ordena de mejor a peor

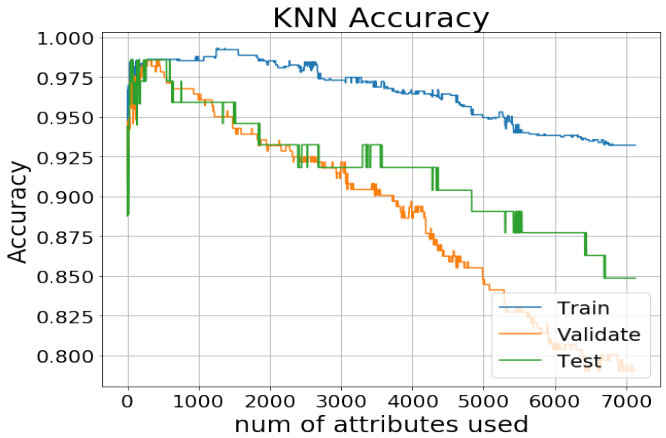
Ventajas:

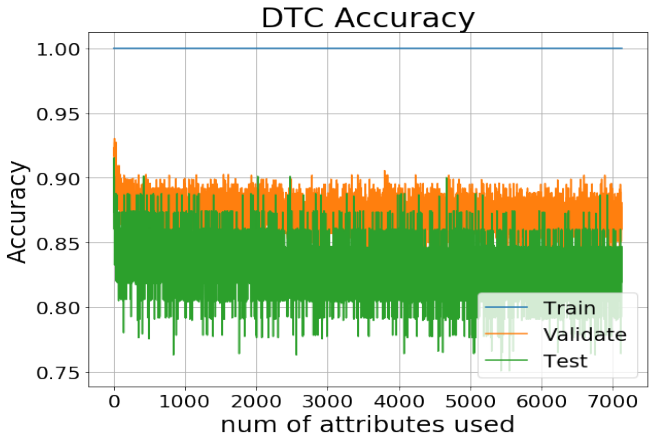
- Tiempo de ejecución casi nulo
- Muy buenos resultados

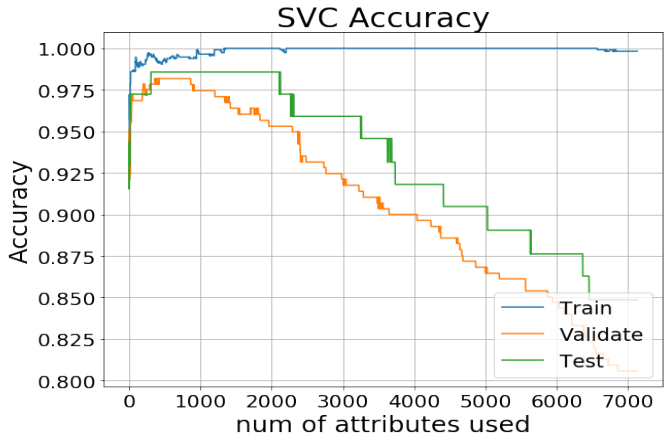
Inconvenientes:

- No elimina redundancia









8. *mRMR*

Función *mRMR* de la biblioteca *scikit-Feature*

Método multivariante

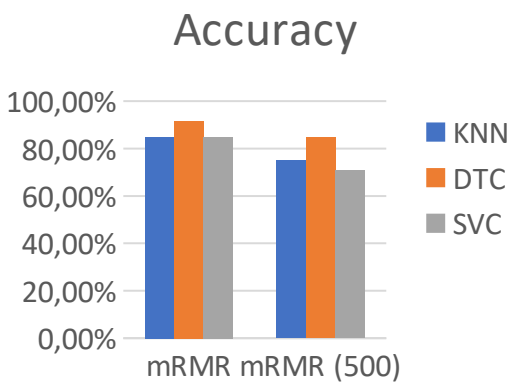
Ordena según relevancia

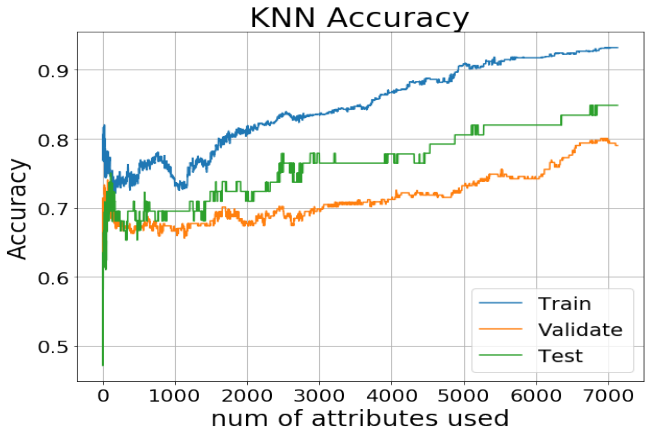
Ventajas:

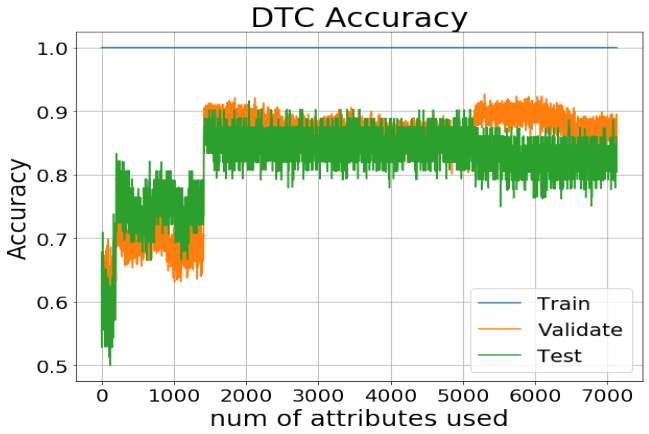
- Descarta características redundantes

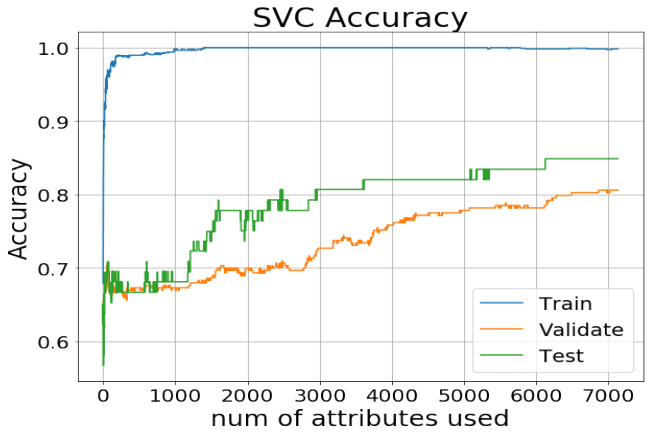
Inconvenientes:

- Tiempos de ejecución elevados
- No tiene tan buenos resultados como *f-score*









9. *Sequential Foward Selection*

Función *SFS* de la biblioteca *mlxtend*

Algoritmo utilizado para la evaluación: SVC

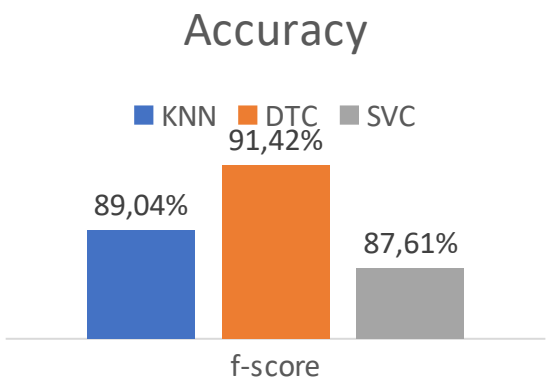
Añade características una a una

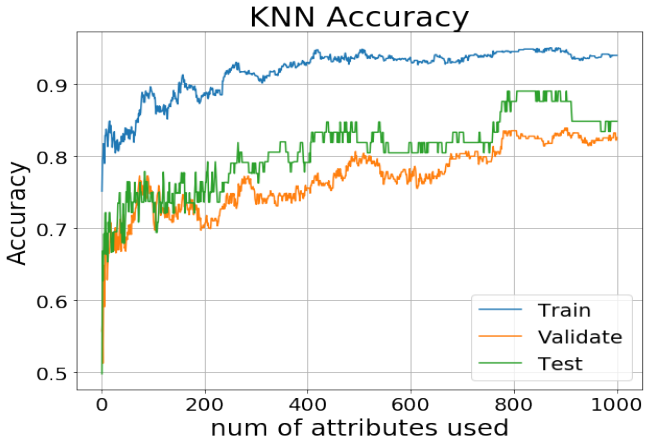
Ventajas:

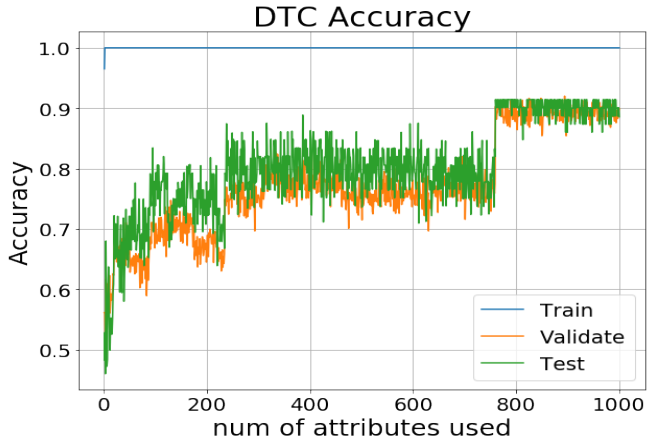
- Niveles de clasificación aceptables

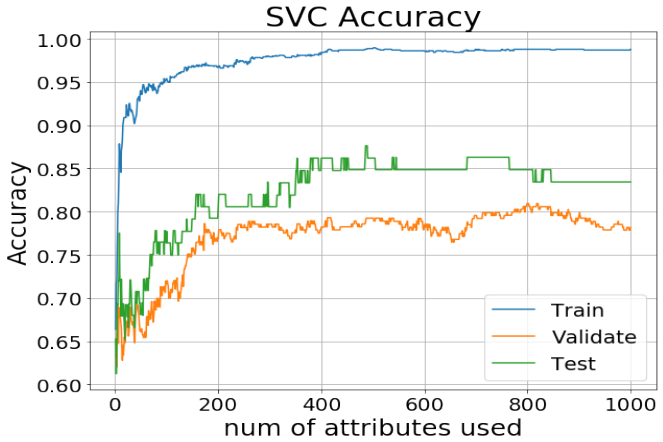
Inconvenientes:

- Tiempos de ejecución muy elevados
- No tiene tan buenos resultados como *f-score*









10. MÉTODO HÍBRIDO

Combinación de los métodos:

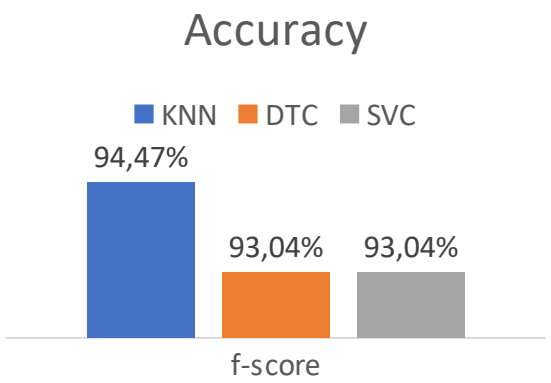
- *f-score*
- *mRMR*
- *SFS_b*
- *SFS_f*

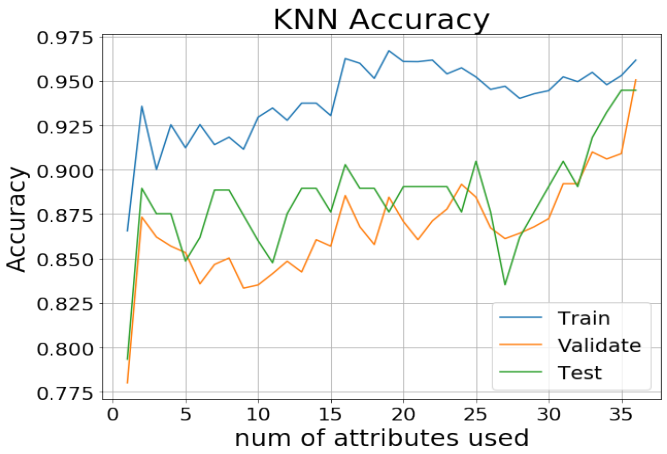
Ventajas:

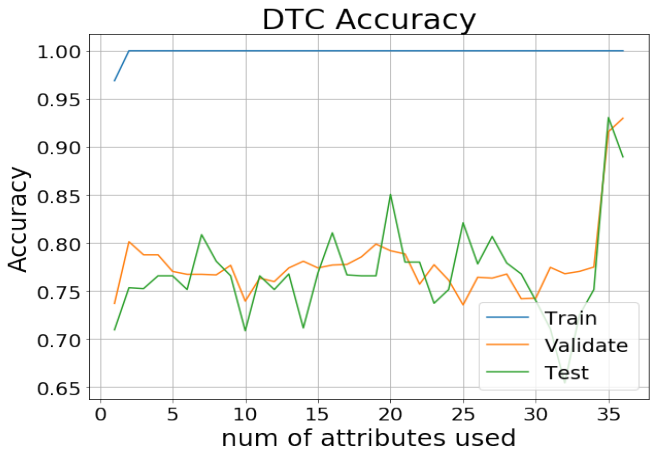
- Buenos niveles de clasificación
- Menor tiempo de ejecución que el método *wrapper*

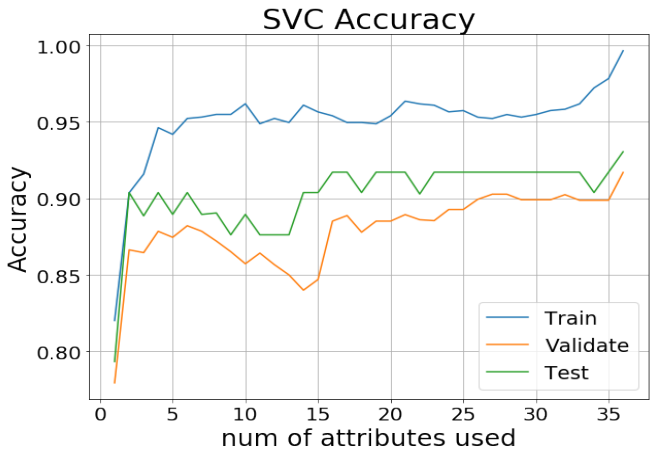
Inconvenientes:

- Tiempos de ejecución no tan rápidos como los de filtro
- No supera los resultados de *f-score*









11. CONCLUSIONES

Los conjuntos de datos *microarray* de ADN proponen un desafío para los algoritmos de aprendizaje

La reducción de la dimensionalidad es un paso previo para obtener los mejores resultados

Los mejores resultados se obtienen con el método *f-score*

Nuestro método híbrido solo supera al método *f-score* con el algoritmo DTC

Método	Atributos	Algoritmo	Accuracy
Filtro: *f-score*	45	KNN	98,57%
Híbrido	35	DTC	93,04%
Filtro: *f-score*	305	SVC	98,57%