Ciclo Autonómico de Análisis de Datos para el Diseño de Descriptores para Algoritmos de Aprendizaje Automático

Ricardo Vargas
Posgrado en Computación-CEMISID
Universidad de Los Andes
Mérida, Venezuela
ricardo.servitechs@gmail.com

Jose Aguilar
CEMISID Universidad de Los Andes
Universidad de Los Andes
Mérida, Venezuela
aguilar@ula.ve

Eduard Puerto
Grupo de Investigación GIDIS
Universidad Francisco de Paula Santander
Cúcuta, Colombia
eduardpuerto@ufps.edu.co

Mérida, 2018

Contenido

- Motivación
- Ingeniería de Características
  - Extracción
  - Construcción
  - Selección y/o Reducción
- Analítica de Datos
  - Ciclo Autonómico
  - MIDANO
- Especificación del Ciclo Autonómico
- Caso de Estudio
- Resultados
- Conclusiones

Motivación

- La Ingeniería de Características es muy importante para los algoritmos de Aprendizaje Automático.
- Al aplicar técnicas de analítica de datos se pueden organizar los diferentes tipos de tareas de Análisis de Datos que componen el área de la Ingeniería de Características.
- La metodología MIDANO provee un marco de trabajo que facilita este proceso.

Ingeniería de Características

- Extracción
- Construcción
- Selección y/o Reducción

Extracción de Características

Se realizan transformaciones sobre los datos en base a distintos criterios como:

- Maximizar Varianza
- Reducir Correlaciones

Métodos comúnmente utilizados:

- Análisis de Componente Principal (PCA)
- Análisis de Discriminantes Lineales (LDA)
- Máquinas de Soporte Vectorial (SVM)

Construcción de Características

Se busca generar nuevas características, generalmente en casos donde:

- Hay datos incompletos
- Se desea complementar la información disponible
- Se desea buscar relaciones ocultas entre características

Métodos comúnmente utilizados:

- Programación Genética (GP)
- Operadores Lógicos y Algebraicos
- Motores de Inferencia

Selección de Características

Se busca reducir el conjunto de características.

Esta reducción puede ser mediante:

- Selección de un subconjunto de características
- Agregación/Fusión de características

Los criterios de selección del subconjunto óptimo se agrupan en dos modelos:

- Filtros (Filtering)
- Envoltorios (Wrapping)

Analítica de Datos

- Ciclo Autonómico
- MIDANO

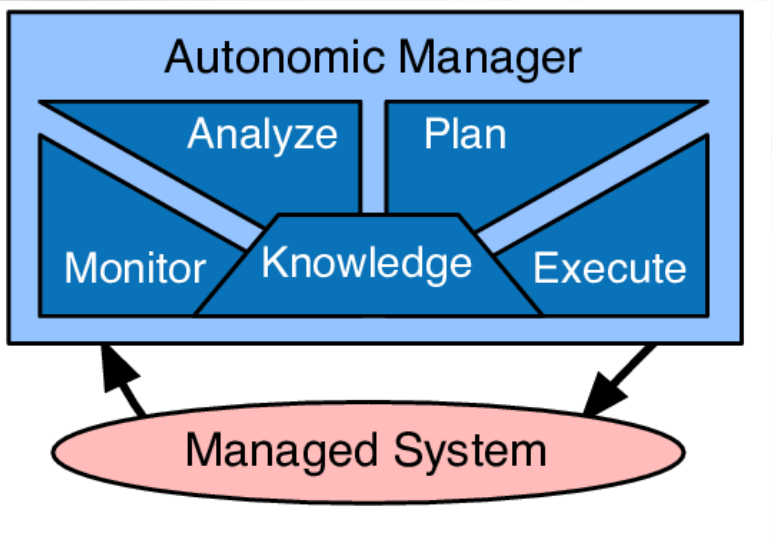
Ciclo Autonómico

Es un ciclo cerrado de tareas de análisis de datos, que supervisa constantemente el proceso bajo estudio.

Organiza los diferentes tipos de tareas de Análisis de Datos.

Estas tareas tienen diferentes roles:

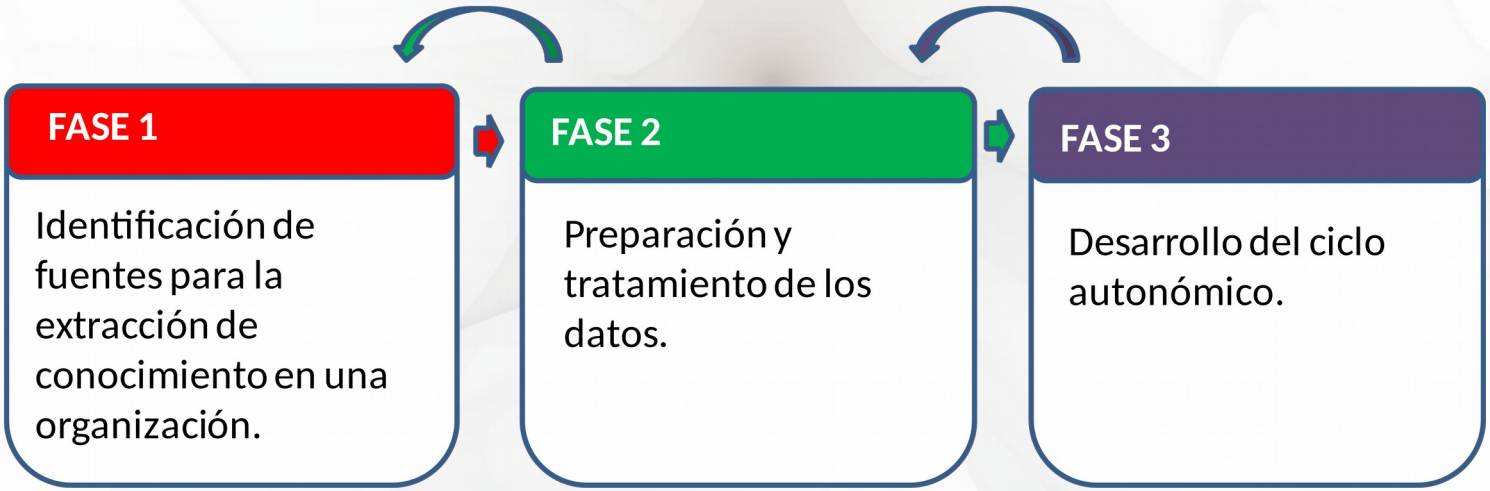
- Observar el proceso
- Analizarlo
- Tomar decisiones.



MIDANO

Metodología para el Desarrollo de Aplicaciones de Minería de datos

Está compuesta por tres fases:



Especificación del Ciclo Autonómico

Tarea	Nombre	Fuentes generales de datos requeridas	Indicadores generados	Efectos esperados sobre el objetivo estratégico
Monitoreo	Captura de datos	Tablas VMC, ETL (Extracción, Tratamiento y Carga) y CCA (colección, curetaje y agregación)	Datos del Experimento	Se obtienen los datos recogidos en etapas anteriores sobre los cuales se quiere realizar la clasificación
Análisis y Toma de decisiones	Construcción de características	Datos obtenidos del paso anterior	Datos tratados y depurados	Se emplean los primeros métodos y técnicas de preparación y tratamiento de datos
	Extracción de características	Datos depurados	Medias, medianas, modas, mínimos, máximos, entre otros valores numéricos necesarios	Conjunto de técnicas para extraer valores numéricos, métricas, etc. que mejor representen los datos
	Selección y reducción de características	Representación numérica de los datos	Conjunto final de características	Se terminan de depurar las características extraídas, reduciendo descriptores redundantes, o descartando algunas características excedentes

Especificación del Ciclo Autonómico

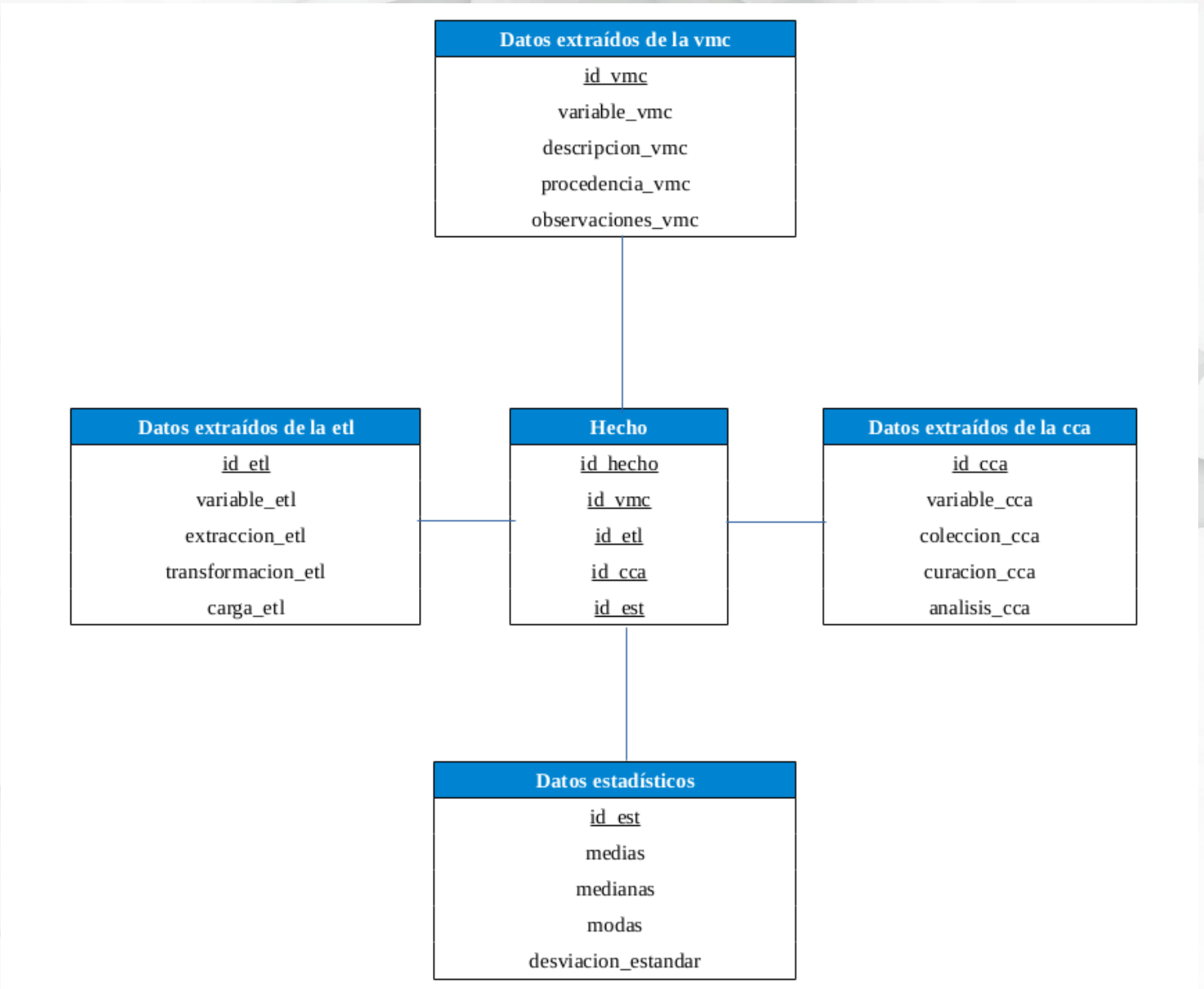
Nombre de la tarea:	**Extracción de características**
Descripción	Obtener representaciones numéricas de los datos procesados para usarlas luego en el proceso de clasificación
Fuente de datos	Datos depurados
Tipo de tarea de analítica de datos	Clasificación, Agrupamiento, Asociación, Regresión
Técnicas de analítica de datos	Cálculos estadísticos, KNN, random forest, regresión, grafos, etc.
Tipo de modelo de conocimiento	Descriptivo, Predictivo
Tareas relacionadas de analítica de datos	Construcción de características Selección y reducción
Tipo de tarea del ciclo (rol)	Análisis / Toma de decisiones

Especificación del Ciclo Autonómico

Nombre de la tarea:	Construcción de características
Descripción	Preparar los datos obtenidos
Fuente de datos	Datos obtenidos en la etapa anterior
Tipo de tarea de analítica de datos	Clasificación, Agrupamiento, Asociación
Técnicas de analítica de datos	Normalización, estandarización, transformación, reducción, discretización, etc.
Tipo de modelo de conocimiento	Descriptivo, Predictivo
Tareas relacionadas de analítica de datos	Captura de datos Extracción de características
Tipo de tarea del ciclo (rol)	Análisis / Toma de decisiones

Nombre de la tarea:	Selección y reducción de características
Descripción	Asegurar que se obtienen solo las características más relevantes para realizar la clasificación
Fuente de datos	Características extraídas en la etapa anterior
Tipo de tarea de analítica de datos	Clasificación, Agrupamiento, Asociación
Técnicas de analítica de datos	Filtering, Wrapping, análisis bivariantes, etc.
Tipo de modelo de conocimiento	Descriptivo, Predictivo
Tareas relacionadas de analítica de datos	Extracción de características
Tipo de tarea del ciclo (rol)	Análisis / Toma de decisiones

Modelo Multidimensional



Caso de Estudio

El caso de estudio seleccionado para esta implementación consiste en una base de datos de gestos de letras escritos a mano en una tableta digital, por niños que se encuentran aprendiendo a escribir.

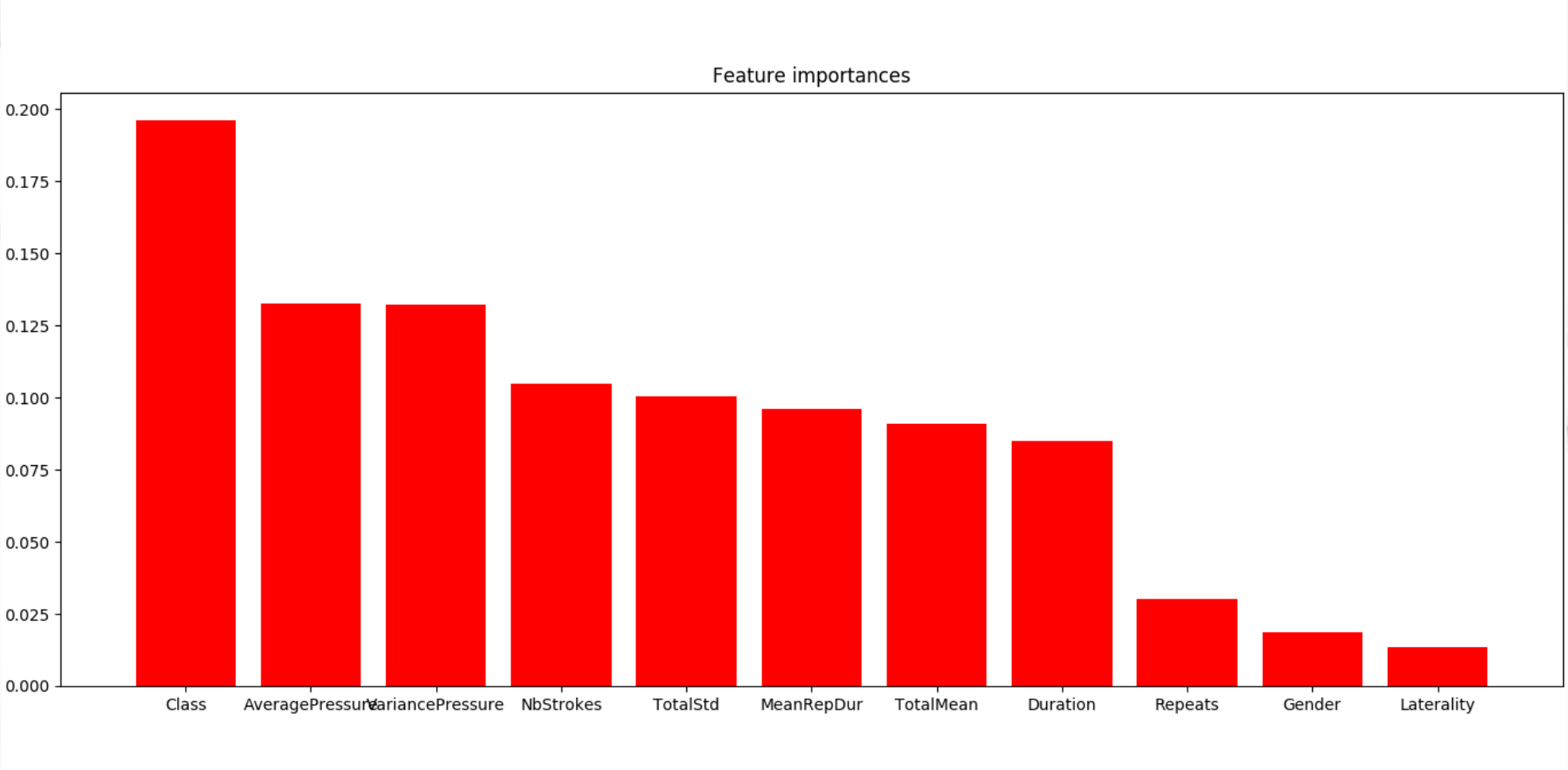
Características iniciales:

- ID: id del niño
- Repeats: número de veces que el niño ha intentado escribir la letra
- Class: la letra escrita
- Gender: genero del niño (masculino/femenino)
- Age: edad del niño
- Laterality: lateralidad del niño (zurdo/derecho)
- Duration: el tiempo en ms que el niño tardó en escribir la letra
- NbStrokes: número de trazos usados para escribir la letra
- AveragePressure: presión promedio aplicada por el niño al realizar el gesto
- VariancePressure: varianza de la presión aplicada por el niño al realizar el gesto

Especificación de los Pasos

- Captura de Datos
- Aplicar Técnicas de Construcción de Características:
  - Media entre Duration y Repeats
  - Desviaciones Estandar
- Aplicar Transformaciones Básicas en los Datos:
  - Transformar Class, Gender y Laterality de texto a valores numéricos
- Aplicar Técnicas de Extracción de Características:
  - Random Forest
  - PCA
- Aplicar Técnicas de Selección de Descriptores:
  - Recursive Feature Elimination (RFE)
  - VarianceThreshold

Importancia de Descriptores con Random Forest



Resultados

Se obtienen diferentes conjuntos de características al aplicar las técnicas descritas anteriormente.

- Las 6 características con los mas altos valores de importancia: RFE_6
- Las 8 características con los mas altos valores de importancia: RFE_8
- Conjuntos usando diferentes valores de VarianceThreshold sobre los conjuntos anteriores:

  RFE_6_V_50, RFE_6_V_80, y RFE_6_V_90: 6 características con varianzas de 50%, 80% y 90%

  - RFE_8_V_50, RFE_8_V_80, y RFE_8_V_90: 8 características con varianzas de 50%, 80% y 90%

Para validar los resultados obtenidos se realiza una clasificación usando los diferentes conjuntos de características obtenidos y calculando la precisión mediante la técnica “k-fold cross”.

Comparación de Resultados

Técnica FS vs FE	RF	PCA	RF + PCA
DNP*	0.8155	0.8109	0.8008
RFE_6	0.8097	0.6471	0.8187
RFE_8	0.8090	0.8017	0.8090
RFE_6_V_50	0.8173	0.6023	0.8109
RFE_6_V_80	0.8133	0.6068	0.8068
RFE_6_V_90	0.8203	0.6221	0.8078
RFE_8_V_50	0.8155	0.7848	0.8080
RFE_8_V_80	0.8157	0.7887	0.8042
RFE_8_V_90	0.8090	0.7936	0.8094

*DNP: Datos No Procesados

Conclusiones

- Importancia de la ingeniería de características
- Organización del proceso de ingeniería de características en un Ciclo Autonómico
- Sensibilidad de los resultados de clasificación según las técnicas usadas
- Realizar trabajos futuros con:
  - Diferentes contextos de aplicación
  - Un mayor número de técnicas en cada una de las fases del Ciclo Autonómico
  - Diferentes métricas de calidad
  - Datos con desbalance entre clases, etiquetados y no etiquetados, con ruidos, etc.
  - Diferentes clasificadores