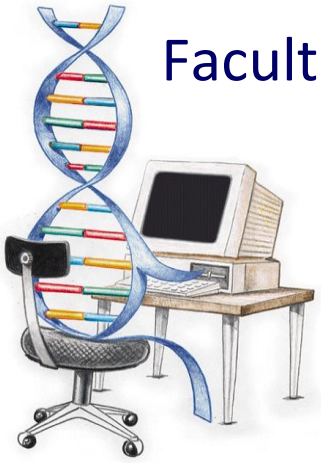
ALGORITMOS EVOLUTIVOS
Curso 2024

Tema 12: Evaluación Experimental de Algoritmos Evolutivos

Centro de Cálculo, Instituto de Computación
Facultad de Ingeniería, Universidad de la República, Uruguay



cecal

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Evaluación experimental de AEs

Contenido

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

1. Objetivos de la evaluación experimental
2. ¿Qué se debe reportar?
3. Instancias de evaluación
4. Ajuste de parámetros
5. En la práctica

Evaluación experimental de AEs

Objetivos del análisis experimental

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Los algoritmos evolutivos no son deterministas
- Si se realizan varias ejecuciones independientes se obtendrán resultados diferentes
- Es necesario realizar una evaluación estadística de los resultados obtenidos a partir de un conjunto de ejecuciones diferentes
- Es importante que la cantidad de ejecuciones sea significativa para poder extraer conclusiones. Se recomienda realizar 50, aunque para problemas complejos pueden aceptarse 30.
- Las ejecuciones deben ser independientes: se deben utilizar diferentes valores de la semilla para el generador de números aleatorios utilizado en cada una de las ejecuciones.

Evaluación experimental de AEs

¿ Qué se debe reportar ?

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

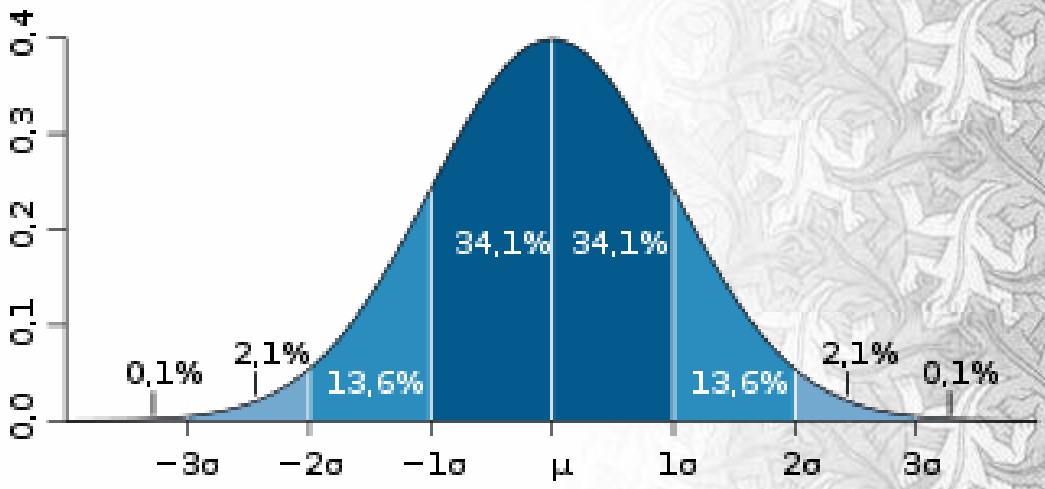
- Deben presentarse de modo sistemático resultados numéricos que permitan evaluar la calidad de resultados obtenida por el AE y la eficiencia computacional del algoritmo
- Existen múltiples valores que pueden ser reportados para evaluar la calidad de las soluciones obtenidas
- Los valores más comúnmente reportados son:
  - el mejor fitness o costo de la solución obtenido en todas las ejecuciones
  - el valor promedio (media o mediana) de los mejores fitness o de los mejores costos de las soluciones obtenidos en cada ejecución
  - la desviación estándar de los mejores fitness o de los mejores costos de las soluciones obtenidos en cada ejecución

Evaluación experimental de AEs

Análisis estadístico

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Si la distribución de los mejores fitness o de los mejores costos de las soluciones obtenidos en cada ejecución sigue una distribución N(μ, σ) se cumple:
  - en el intervalo [μ - 2σ, μ + 2σ] se encuentra, aproximadamente, el 95,44% de la distribución.
- Se debe realizar un test de normalidad para comprobar que la distribución es normal
- Tests utilizados comúnmente:
  - Kolmogorov-Smirnov
  - Lilliefors
  - Anderson-Darling



Evaluación experimental de AEs

Métricas relevantes: calidad de soluciones

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Valores relevantes para evaluar la calidad de las soluciones obtenidas:
  - los valores promedio (media o mediana) de fitness (o costo) de las soluciones obtenidos en cada ejecución
  - el promedio del número de generaciones necesarias para alcanzar el mejor resultado obtenido (eventualmente, debe presentarse la desviación estándar del número de generaciones)
  - cantidad de veces que se obtuvo el fitness óptimo o el costo de la solución óptima
  - análisis gráfico de valores de fitness (mejor fitness y/o fitness promedio) para uno (o varios) casos representativos

Evaluación experimental de AEs

Métricas relevantes: eficiencia computacional

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Para evaluar la eficiencia computacional, debe reportarse el tiempo de ejecución promedio de las ejecuciones realizadas sobre la plataforma computacional usada (generalmente, en segundos) y su desviación estándar
- Para evaluar la eficiencia computacional, es necesario presentar los detalles sobre la plataforma computacional utilizada:
  - Deben ofrecerse datos completos sobre hardware y software.
  - Por ejemplo: modelo de computadora, velocidad de CPU, memoria, sistema operativo, características de la red, lenguaje de desarrollo, bibliotecas utilizadas.

Evaluación experimental de AEs

Métricas relevantes: eficiencia computacional

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Otros valores relevantes:
  - Tiempo requerido para alcanzar diferentes niveles de mejora sobre una solución de referencia (heurística para el problema, mejor resultado conocido hasta el momento, etc.)
  - Evolución de valores de fitness promedio/mejor con el tiempo
- En el caso de AE paralelos debe estudiarse la calidad de resultados y eficiencia computacional al utilizar diferente número de recursos de cómputo

Evaluación experimental de AEs

Análisis comparativo de algoritmos

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- En el caso de comparaciones entre dos o más algoritmos, operadores, configuraciones paramétricas o modelos:
  - se deben reportar todos los datos necesarios para cada algoritmo
  - la comparación debe ser realizada en situaciones idénticas, en lo que concierne a los casos de prueba y al entorno de ejecución
- El análisis comparativo de algoritmos debe incluir la presentación del resultado de un test de hipótesis que permita determinar que los resultados obtenidos por un algoritmo son mejores (o peores) que los del otro

Análisis comparativo: tests estadísticos

- Test paramétricos:
  - Los conjuntos de muestras deben tener distribuciones normales
  - Se debe hacer previamente un test de normalidad de las muestras
  - Los test paramétricos se usan para saber si existen diferencias significativas entre las medias de conjuntos de muestras que tienen una distribución normal
  - Para dos conjuntos de muestras: test de Student
  - Para más de dos conjuntos de muestras: Análisis de varianza (ANOVA)

Análisis comparativo: tests estadísticos

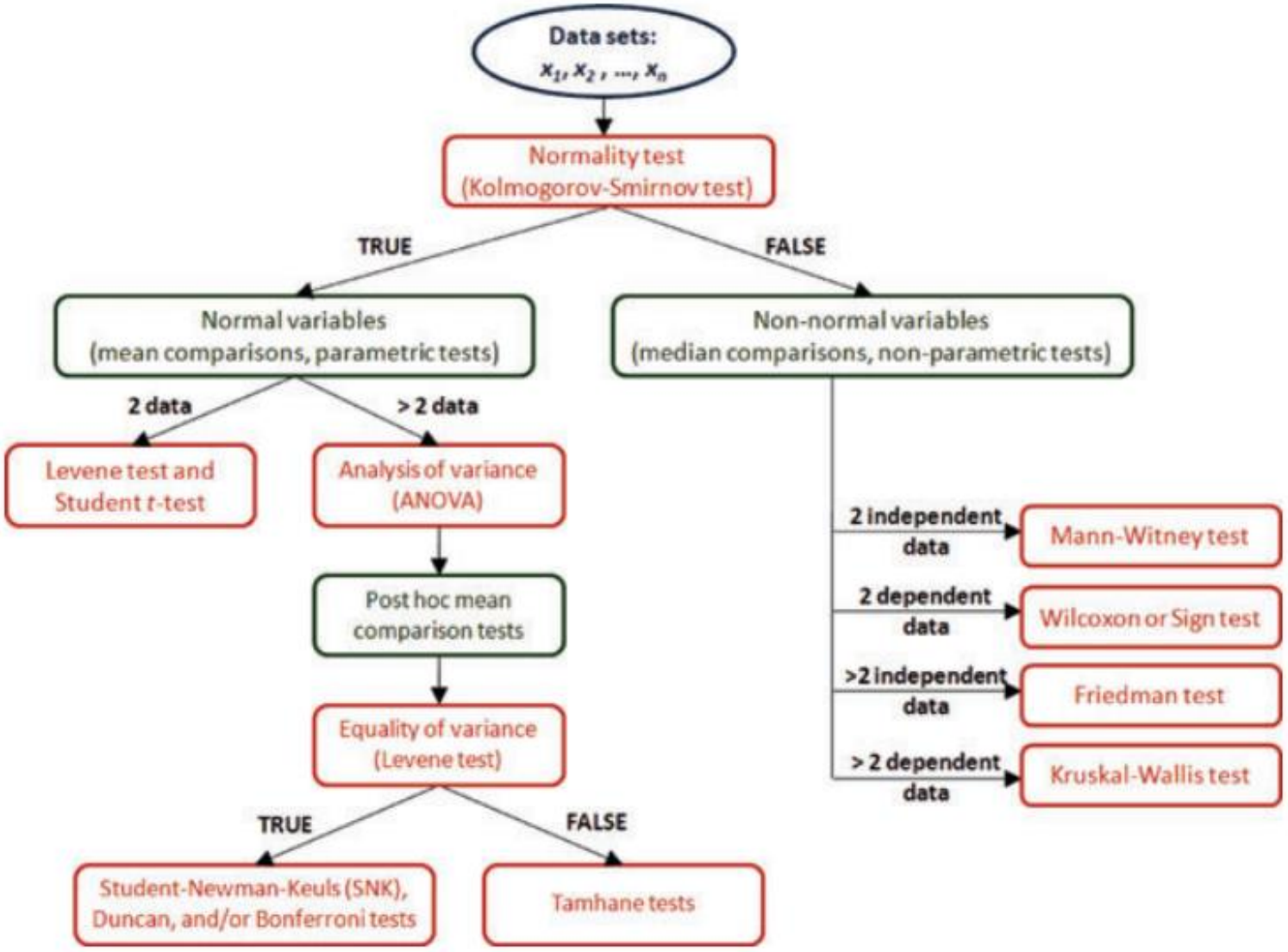
- Test no paramétricos:
  - Usados cuando los conjuntos de muestras no siguen una distribución normal o no se conoce su distribución
  - Se utilizan para saber si existen diferencias significativas entre las medianas de conjuntos de muestras sin imponer restricciones sobre la distribución de las muestras
  - Para dos conjuntos de muestras: test de Mann-Witney
  - Para más de dos conjuntos de muestras: test de Kruskal-Wallis

Análisis comparativo: tests estadísticos

- Test de rangos:
  - Usados para determinar el rango promedio en la comparación de varios algoritmos al resolver el mismo conjunto de instancias de evaluación
  - Se realiza una comparación de a pares de los resultados, y se ordenan los algoritmos de acuerdo a sus valores de fitness, determinando un rango
  - Ejemplo: test de rangos de Friedman

Evaluación experimental de AEs

Análisis comparativo: tests estadísticos



Análisis comparativo: tests estadísticos

- No siempre es posible realizar el test de hipótesis:
  - Si se deben comparar los resultados obtenidos contra resultados publicados previamente, seguramente no sea posible contar con cada una de las muestras. Solamente se suele disponer del promedio y la desviación estándar.
- En caso de omitirse el test de hipótesis, se suele utilizar como criterio de significancia estadística:
  - El algoritmo A es mejor que el algoritmo B si los resultados de A y B cumplen que

$$|f_{AVG}(A) - f_{AVG}(B)| > \max(\text{std}(f_A), \text{std}(f_B))$$

(si un algoritmo es determinista, su desviación estándar será nula)

Evaluación experimental de AEs

Instancias para la evaluación experimental

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- La evaluación experimental debe realizarse sobre instancias específicas del problema que se resuelve.
- Las instancias deben ser realistas y de dimensión razonable para asegurar la complejidad combinatoria del escenario abordado.
- El conjunto de instancias debe tener una cardinalidad apropiada para poder extraer conclusiones sobre la aplicabilidad de los AE para la resolución del problema
  - NO ES CORRECTO evaluar un AE utilizando una única instancia del problema. Al menos tres casos de prueba son requeridos.
- De exisitir, deben resolverse conjuntos estándares de prueba, benchmarks o instancias ya abordadas en trabajos previos

Evaluación experimental de AEs

Ajuste de parámetros

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- El ajuste de parámetros se realiza mediante un análisis experimental sobre instancias específicas
- Las configuraciones paramétricas se evalúan sobre un conjunto de instancias de configuración (diferente al conjunto de problemas de validación), para evitar sesgos en la parametrización
- Deben contemplarse todas las posibles combinaciones de los parámetros estudiados: como consecuencia de la complejidad del análisis, en general se trabaja sobre instancias más pequeñas que las de evaluación
- Los parámetros que se estudian habitualmente son el tamaño de la población, la probabilidad de cruzamiento y la probabilidad de mutación. Otros parámetros pueden ser analizados también

Evaluación experimental de AEs

Ajuste de parámetros

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Otros parámetros y criterios a evaluar:
  - Diferentes criterios de parada tendrán parámetros específicos
    - Esfuerzo prefijado: número de generaciones ejecutadas, tiempo para hallar el mejor
    - Variación en el fitness: número de generaciones, tasa de variación
  - Características y parámetros del modelo de evolución
    - Precisión de la codificación, valor de gap generacional, factores de escalado
  - Parámetros de operadores específicos y avanzados
    - Participantes y sobrevivientes en torneo, proporción de individuos elegidos en una selección por rango, criterio de crowding, valor de $\sigma_{SHARING}$, etc

Evaluación experimental de AEs

Ajuste de parámetros

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- También puede realizarse un estudio de diferentes alternativas para el AE:
  - Estudio comparativo de operadores
  - Tipos de selección, operadores de cruzamiento y de mutación, estrategias de reemplazo
  - Mecanismos de escalado, técnicas para preservar la diversidad, modelos de penalización de soluciones no factibles
- En todos los casos el análisis estadístico es mandatorio

Evaluación experimental de AEs

En la práctica

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Existen dos etapas claramente diferenciadas. La primera etapa donde se evalúa la configuración paramétrica, y la segunda donde se evalúan los resultados de la propuesta
- ¿ Qué actividades se llevan a cabo en la práctica, en la etapa de evaluación de configuración ?
  - Se realiza un análisis de parámetros simple, sobre un conjunto de casos de prueba reducido
  - Siempre se suelen incluir las probabilidades de aplicación de los operadores evolutivos (cruzamiento y de mutación)
  - Es habitual incluir el tamaño de la población
  - Es poco frecuente incluir otros parámetros (salvo que las características de la propuesta lo hagan necesario)

Evaluación experimental de AEs

En la práctica

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Para el análisis paramétrico deben generarse todas las configuraciones posibles y comparar los resultados obtenidos para cada configuración de los parámtros estudiados (tradicionalmente p_C, p_M y tamaño de población).
- El rango de valores candidatos para cada parámetro depende del problema y del algoritmo
  - Habitualmente p_C entre 0.5 y 1.0, p_M entre 0.001 y 0.1, tamaño de población entre 50 y 200
  - Algoritmos específicos pueden tener otros rangos paramétricos
- Ortogonalidad: deben analizarse todas las combinaciones
- Puede incorporarse un estudio acotado de operadores con parámetros fijos (obtenidos en una etapa previa)

Evaluación experimental de AEs

En la práctica

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- ¿Qué se hace en la práctica en la etapa de evaluación de resultados de la propuesta?
  - Se lleva a cabo un análisis sobre instancias específicas, diferentes de las utilizadas en la etapa de configuración
  - Se reportan los valores que permiten evaluar la calidad de resultados y la eficiencia computacional del AE
  - Incluyendo análisis estadísticos en la evaluación
- Se deben reportar los resultados de ambas etapas y sus conclusiones
- Se deben comparar los resultados con otras técnicas (deterministas, heurísticas, metaheurísticas) para la resolución del problema
  - Incluyendo análisis estadísticos en la comparación

Evaluación experimental de AEs

En la práctica: problemas multiobjetivo

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Para MOEAs, en la etapa de validación es necesario reportar los valores promedio y desviación estándar de las métricas consideradas para evaluar la convergencia al FP y la diversidad:
  - Convergencia al FP:
    - # puntos no dominados
    - # puntos en el FP real
    - Distancia generacional
    - Tasa de error
  - Diversidad:
    - Spacing
    - Spread
  - Combinadas:
    - Hipervolumen relativo
    - Attainment functions

Evaluación experimental de AEs

En la práctica: problemas multiobjetivo

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- En los problemas en que el frente de Pareto no es conocido:
  - Debe construirse una aproximación del FP real combinando todas las soluciones no dominadas halladas en un número apropiado de ejecuciones (mínimo 30, recomendado 50) de el/los MOEAs y otras técnicas evaluadas
  - Luego se deben reportar los valores promedio y de desviación estándar de cada métrica estudiada para cada ejecución individual de cada MOEA evaluado
  - También suelen reportarse ejemplos gráficos de FP representativos en la resolución del problema

Evaluación experimental de AEs

En la práctica: qué reportar

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

1. Resultados de configuración paramétrica
   - Para cada configuración (combinación de parámetros estudiados): mejor fitness, fitness promedio, desviación estándar para cada instancia de configuración
   - Significancia estadística: p-values del/los tests estadísticos utilizados para comparar los resultados, para cada instancia de configuración
   - Configuración que permitió alcanzar los mejores resultados
   - Tiempos de ejecución

Evaluación experimental de AEs

En la práctica: qué reportar

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

2. Valores de fitness para todas las instancias
   - Mejor fitness, fitness promedio, desviación estándar para todas las instancias estudiadas
   - Mejoras respecto a resultados conocidos o a los calculados con otras técnicas (OT), se reporta como valor porcentual
     - % de mejora = [f(AE) – f(OT)]/ f(OT))
   - Significancia estadística: p-values del/los tests estadísticos utilizados para la comparación de las distribuciones de resultados (para cada instancia de evaluación)
   - GAP (porcentual) respecto a valores teóricos (VT) o cotas del problema, calculadas con métodos exactos
     - % GAP = [f(VT) – f(AE)]/ f(AE))

Evaluación experimental de AEs

En la práctica: qué reportar

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

3. Eficiencia computacional
   - Tiempo de ejecución total, tiempo para alcanzar cierta calidad de resultados o ciertos valores de mejora respecto a otras técnicas, análisis de valores de fitness y tiempo
4. Comparación de algoritmos
   - Número de veces que un algoritmo supera a otro(s), rangos estadísticos de un test de hipótesis de comparación de múltiples algoritmos, p-values de la comparación
5. Gráficos
   - Mejoras respecto a otras técnicas o resultados previos (por instancia o por categorías de instancias), evolución del fitness (mejor/promedio) con el tiempo, análisis comparativo de resultados, etc.

Evaluación experimental de AEs

En la práctica: qué reportar (problemas multiobjetivo)

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Deben reportarse los valores de las métricas para evaluar MOEAs, siguiendo los lineamientos presentados previamente
  - Reportar tablas, test estadísticos y gráficos para cada métrica estudiada
  - Deben estudiarse/reportarse valores mejores, promedios y desviación estándar de los resultados obtenidos en cada ejecución de el/los MOEAs para cada instancia, comparando con el frente de Pareto real (o con la aproximación construida siguiendo la metodología presentada)
  - La comparación y el análisis estadístico debe realizarse por métrica

Evaluación experimental de AEs

En la práctica: qué reportar (problemas multiobjetivo)

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

- Habitualmente, suelen reportarse los mejores resultados obtenidos para cada objetivo
- En ocasiones suelen reportarse también los valores para las mejores soluciones de compromiso (respecto al vector ideal)