CARACTERIZACIÓN DE COALICIONES DE VARIABLES EN DATOS COMPLEJOS MEDIANTE MEDIDAS DE INCERTIDUMBRE

Código: ING362
Tipo de Investigación: Aplicada
Carrera que se vincula: Ing. Electrónica, Lic. En Ciencias de la Computación
Período: 2011 - 2014
Director: Bulacio, Pilar Estela Email: pbulacio@gmail.com
Integrantes: Angelote, Laura M; Pons, Alfonso F; Tapia Paredes, Elizabeth; Ornella, Leonardo A; Coronel, Jose Luis; Murillo, Javier Ivan; Iglesias, Natalia C

Objetivos

1. Realizar un análisis preliminar sobre estructuras de datos complejos sometidos a problemas de clasificación. Corroborar la validez de la definición del conjunto de clases predefinidas sobre los datos, y realizar la redefinición del problema en caso de ser necesario (análisis de tipos y subtipos).
2. Identificar las características de las coaliciones de variables asociadas a distintos tipos de comportamiento de grupos: redundancia, disonancia, contradicción, sinergia. Evaluar cuales de estas características posibilitan un comportamiento efectivo y consistente, sujeto a una posterior clasificación.
3. Realizar un análisis comparativo de medidas de incertidumbre en vista de su uso en la caracterización de coaliciones de variables. Estudiar su interpretación semántica y las posibilidades descriptivas que puedan brindar asociadas al comportamiento de coaliciones de variables: sinergia, disonancia, contradicción, etc.
4. Diseño de un flujo de trabajo que permita ponderar las características de comportamiento de coaliciones de variables descriptas en el objetivo 2, a través de medidas de incertidumbre, seleccionar coaliciones, y validar estadísticamente los resultados. En este contexto, se pretende introducir una selección sistemática no exhaustiva de coaliciones de variables potencialmente relevantes en la representación del conjunto de alternativas (clases), para luego quedarse con un único subconjunto que optimice la clasificación considerada.
5. Diseñar e implementar una aplicación de software que implemente el flujo de trabajo planteado en el objetivo anterior y realizar simulaciones con datos complejos. Dentro de este punto se contemplará la validación estadística de la metodología implementada dentro de un protocolo de prueba de la aplicación. Con este objetivo se usará el potencial estadístico de la herramienta GSEA (Gene Set Enrichment Analysis) y su repositorio de datos. En una segunda etapa se considerará además el tratamiento de datos genómicos de girasol provistos por INTA Castelar.

Resumen Técnico

El presente Proyecto aborda aspectos teóricos y prácticos del diseño de metodologías para la caracterización sistemática (y no exhaustiva) de conjuntos de variables en datos complejos. En particular, la actividad del proyecto se centra en el diseño de un proceso basado en técnicas de Minería de Datos, que permitan la identificación del aporte de coaliciones de variables para abordar el problema de clasificación.
Los datos complejos, es decir, datos que son afectados por uno o varios de los siguientes factores: elevada dimensionalidad, multi-dimensionalmente correlacionados, ruidosos, e insuficientes en número de muestras; son el punto de partida de procesos de Clasificación en diversos dominios, por Ej., bioinformática (datos genómicos), agricultura (mapas de suelos en agricultura de precisión), materiales (datos espectroscópicos),…
El objetivo último considerado en este proyecto es alcanzar la clasificación de datos complejos sobre un conjunto de clases predefinidas. La complejidad de estos tipos de datos requiere antes de abordar el problema de clasificación, una etapa de preprocesamiento que tiene por objetivo alcanzar un nuevo conjunto de datos “limpios”. Este preprocesamiento es realizado generalmente en dos pasos. El primero se ocupa de quitar muestras y variables espurias, limitación del rango de variables y normalización. El segundo, que es donde nos centraremos, realiza un análisis de la estructura de datos, considerando la posibilidad de redefinición de clases tratadas para luego dar lugar a la selección de variables. En referencia a la posible redefinición de clases, debemos notar que las clases predefinidas pueden no estar fuertemente vinculadas con las muestras. Esto se puede deber a falta completitud de

los datos o a las relaciones complejas que los datos presentan, siendo conveniente el análisis de tipos y subtipos inferidos desde el análisis de los datos.
Por lo arriba expuesto, se puede ver que el análisis de la estructura de datos es un punto crucial en el problema de clasificación de datos complejos. Un análisis suficientemente descriptivo que pondere a través de medidas de incertidumbre el comportamiento de las variables por coaliciones en vez de individualmente, posibilita modelar características de comportamiento que se dan a nivel de grupo, por Ej., sinergia, contradicción, disonancia. Pero debemos notar además, que un análisis a nivel de grupo debe considerar la complejidad que conlleva y debe descartar soluciones exhaustivas.

Disciplinas: Ing. comunicaciones electrónica y control
Especialidad: Computación
Palabras Clave: Medidas Borrosas - Incertidumbre - Selección de - Variables - Coaliciones