

Revisión sobre métricas de complejidad en el modelado de clústers de un sistema CBR

Núria Macià, Ester Bernadó, Albert Fornells, Elisabet Golobardes,
Josep M. Martorell y Josep M. Garrell

Grup de Recerca en Sistemes Intel·ligents

Enginyeria i Arquitectura La Salle, Universitat Ramon Llull

Quatre Camins 2, 08022 Barcelona (España)

{nmacia,esterb,afornells,elisabet,jmmarto,josepmg}@salle.url.edu

WWW home page: <http://www.salle.url.edu/GRSI>

Resumen

El análisis de la complejidad de los datos mediante métricas de complejidad es útil para predecir el comportamiento de un sistema de aprendizaje. Este trabajo ilustra cómo las métricas permiten definir un espacio a través del cual puede modelarse la bondad de los clústers creados a partir de un mapa autoorganizativo. Unos clústers que se utilizan para organizar la memoria de casos de un sistema de razonamiento basado en casos con el objetivo de mejorar el tiempo empleado en explorar dicha memoria.

Palabras clave: Complejidad de los datos, Métricas de complejidad, Razonamiento basado en casos, Clusterización.

1. Introducción

Con frecuencia, los sistemas de minería de datos deben manejar grandes volúmenes de datos, lo que implica un coste computacional elevado. Esto es especialmente crítico en sistemas de clasificación como el CBR (*Case-Based Reasoning* [1]), puesto que para predecir la clase de un ejemplo debe compararse el mismo con todos los ejemplos disponibles y recuperar los más semejantes. Una posible solución que reduce el elevado coste computacional consiste en organizar la memoria de casos (MC) en

clústers [10]. Se trata de agrupar en patrones los casos que muestran propiedades similares, permitiendo al CBR realizar una recuperación más selectiva. A partir de una MC estructurada, el CBR selecciona el clúster más parecido a la entrada y de éste recupera algunos casos. Así pues, deben fijarse dos criterios: cuántos clústers han de seleccionarse y cuántos casos recuperarse del clúster. Sin embargo, el rendimiento del sistema puede quedar afectado si las agrupaciones no se han definido correctamente. Por lo tanto, es posible que se produzca una mejora en el tiempo de computación a la vez que la precisión de la clasificación disminuye.

Este artículo revisa la aplicación de las métricas de complejidad en el modelado de clústers de un sistema CBR. Sabiendo que la complejidad de los datos influye en la construcción de las agrupaciones, a partir de la información proporcionada por las métricas de complejidad sobre la distribución de los datos se intenta entender el comportamiento de las estrategias de recuperación. Esto ayuda a identificar para qué tipo de problemas es útil organizar la MC en agrupaciones [5] y a determinar qué estrategia alcanza mayor rendimiento según los requisitos establecidos por el usuario [6].

El artículo se estructura de la siguiente manera. La sección 2 describe brevemente el CBR y la organización de la memoria de casos me-

diante mapas autoorganizativos. En la sección 3 se introduce las fuentes de complejidad de un problema y cómo estimar su complejidad geométrica. La sección 4 recoge la experimentación realizada así como los resultados obtenidos. La sección 5 presenta una discusión sobre los resultados. Finalmente, la sección 6 expone las conclusiones y las líneas futuras.

2. CBR y organización de la memoria de casos

2.1. Razonamiento basado en casos

El CBR [1] es una técnica de aprendizaje automático, concretamente de aprendizaje analógico, que trata de resolver un nuevo caso recuperando, de una base de conocimiento, otros casos similares previamente resueltos. De los casos recuperados, se adecúa su solución para ajustarla a la resolución del nuevo caso. Este proceso constituye un ciclo en el que se diferencian cuatro fases: (1) la **fase de recuperación**, en la que se extraen de la base de casos los más parecidos al caso de entrada según una función de similitud, (2) la **fase de adaptación**, donde se ajusta la solución del caso recuperado, (3) la **fase de revisión**, que consiste en evaluar la bondad de la solución obtenida y (4) la **fase de almacenaje**, en la que si el nuevo caso se considera información relevante, se introduce en la memoria de casos.

Así pues, se observa que las cuatro fases mantienen constante relación con la memoria de casos lo que indica que su organización es un factor importante que condiciona el tiempo de las operaciones. El manejo de un gran volumen de datos obliga a configurar estrategias de recuperación para mejorar el rendimiento de los sistemas en términos de tiempo de cómputo.

2.2. Mapa autoorganizativo

El SOM es una técnica de clusterización basada en redes neuronales. La finalidad es generar un mapa topológico que agrupe los casos similares en patrones. Este comportamiento se aprovecha en el CBR para organizar la memoria de casos obteniendo un sistema SOMCBR

[7]. La figura ilustra el ejemplo de una memoria de casos estructurada en un mapa de dos dimensiones con $M \times M$ patrones. El SOM está formado por dos capas: (1) una capa de entrada compuesta por N neuronas, donde cada neurona representa un atributo del caso de entrada y (2) una capa de salida compuesta por $M \times M$ neuronas, donde cada neurona contiene un conjunto de casos similares representado por un vector director. Cada neurona de la capa de entrada está conectada con todas las neuronas de la capa de salida. Para cada nuevo caso C que se introduce, la capa de salida computa el grado de similitud entre el caso C y su vector director aplicando una función de similitud. Nuestra función es el complementario de la distancia euclidiana.

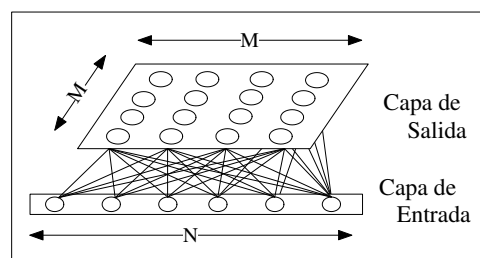


Figura 1: La memoria de casos está organizada por el SOM que distribuye los datos en $M \times M$ clústers, creando grupos de propiedades similares. Esta organización reduce el tiempo de cómputo de la fase de recuperación.

La recuperación consiste en: (1) buscar el patrón más parecido al caso de entrada y (2) compararlo con los casos del patrón seleccionado. En consecuencia, el SOMCBR reduce el tiempo de cómputo puesto que solamente usa los datos de un único clúster. No obstante, los patrones construidos pueden no estar bien definidos debido a la complejidad de los datos y entonces comprometer la precisión de la clasificación.

3. Complejidad de los datos

La complejidad del problema atañe a la distribución de los datos e influye en la precisión

de los sistemas clasificadores. Habitualmente, se estima el error (o precisión) del clasificador como medida para evaluar la calidad del mismo. Sin embargo, esta medida no sólo depende del clasificador sino que también depende de la complejidad inherente al problema que a menudo es la causa del error.

3.1. Causas de la complejidad de los datos

La dificultad de la clasificación está sujeta a tres causas [9][3]:

Ambigüedad de las clases. Algunos problemas de clasificación contienen clases que no pueden distinguirse dado que hay instancias que, perteneciendo a clases opuestas, toman los mismos valores en todos los atributos. Este fenómeno es propio de la ambigüedad intrínseca del problema o de la falta de atributos discriminantes. En el primer caso es necesario conocer el contexto para poder desambiguar y en el segundo se debería redefinir o ampliar el conjunto de atributos.

Complejidad de la frontera. Otros problemas pueden presentar una frontera de separación entre clases compleja que requiere un conjunto de descripción amplio o un algoritmo complejo para representarla. La complejidad de la frontera puede caracterizarse con la complejidad de Kolmogorov o la longitud mínima del programa que se necesita para reproducirla. Esto es independiente de la ambigüedad de las clases y del tamaño del conjunto de entrenamiento, ya que disponiendo de suficientes puntos sin ambigüedad en la clase, la descripción de la frontera puede ser todavía compleja.

Conjunto de ejemplos reducido y dimensión del espacio de atributos. La cantidad de instancias del conjunto de entrenamiento y su representatividad condicionan la capacidad de generalización del clasificador. Con un conjunto de entrenamiento pequeño se puede errar en la estimación de la complejidad del problema y catalogarlo de simple; es probable que los datos no sean representativos. Un número de ejemplos de entrenamiento insuficiente con una dimensionalidad elevada, es decir un número grande de atributos, pueden

provocar que, aunque se obtenga un buen resultado de clasificación en entrenamiento, los aciertos en la fase de test con nuevas instancias sean bajos e inestables.

La combinación de estos aspectos fija un umbral mínimo de error. En [9] se propusieron un conjunto de métricas de complejidad que determinan la complejidad de la frontera. La ambigüedad de las clases y la representatividad de los ejemplos son difíciles de estimar. Por este motivo, los estudios sobre la complejidad se centran principalmente en caracterizar la geometría de la frontera entre clases.

3.2. Estimación de la complejidad geométrica

Las métricas de complejidad son una herramienta que permite evaluar la distribución geométrica de los datos con la finalidad de estimar su complejidad. Basu y Ho [2] presentan un conjunto de métricas que aproximan la complejidad de los datos desde diferentes aspectos: analizando los atributos de manera individual, midiendo la separabilidad de las clases o a partir de la topología de los datos. No obstante, cada métrica ofrece información de una característica concreta del problema que no siempre corresponde a la complejidad real, lo que significa que para una definición precisa se requiere la combinación de estas métricas. Incluso es posible que falten otro tipo de métricas para completar la estimación de la complejidad.

Actualmente, su aplicación se centra en: (1) predecir el error cometido por un clasificador para un conjunto de datos concreto [3] y (2) caracterizar la dificultad de un problema y relacionarla con el rendimiento del clasificador. Esto permitiría construir un espacio de complejidad sobre el que se puede representar el dominio de competencia de los distintos esquemas de aprendizaje.

4. Estimación de complejidad en SOMCBR y su aplicación

4.1. Experimentación

Este trabajo revisa la aplicabilidad del análisis de la complejidad de los datos en la clusterización de la memoria de casos. En una primera fase, se quiere estudiar cómo aplicar el análisis de la complejidad para predecir si es posible realizar una clusterización de la memoria que reduzca el coste computacional sin una pérdida significativa de precisión. Para ello, se llevó a cabo un estudio sobre 28 problemas de clasificación [5], de distinto dominio y con diferentes características, extraídos del Repositorio UCI [4]. Debido a que las métricas de complejidad están definidas para problemas de dos clases, el estudio se limitó a este tipo de problemas. De este modo, los problemas formados por m clases se transformaron en m problemas biclase, partiendo cada clase con respecto al resto. Como trabajo futuro, se quiere ampliar el estudio a problemas con más de dos clases.

Para cada conjunto de datos se ha ejecutado por un lado el CBR y por el otro el SOMCBR con diferentes configuraciones, en las que varían dos parámetros: (1) el número de clústers definidos y (2) el número de casos recuperados del clúster seleccionado, siendo un determinado porcentaje de casos o bien la totalidad de ellos. La precisión de cada clasificador se ha estimado utilizando la técnica *stratified 10-fold cross-validation* y las diferencias entre CBR y SOMCBR se han evaluado mediante el test *t Student*. El SOMCBR se ha ejecutado diez veces para cada configuración y el resultado promedio ha sido usado en la comparación con el CBR.

Además de la precisión del CBR y SOMCBR, se ha evaluado el porcentaje de reducción en el número de comparaciones en el SOMCBR. Los resultados indican que hay problemas en los que la reducción de operaciones es significativa.

Efectivamente, las técnicas de clusterización favorecen el tiempo de recuperación de casos, pero en cuanto los clústers no se construyen debidamente, la precisión de la clasificación decae.

El objetivo de aplicar las métricas de complejidad es doble. En primer lugar, se pretende predecir a partir de la complejidad de los datos en qué casos es útil la clusterización de la MC, y por consiguiente saber cuando aplicar el SOMCBR. En segundo lugar, se intenta relacionar cuál es la mejor estrategia de recuperación para satisfacer los requisitos establecidos por el usuario en base a la complejidad del problema.

Las métricas más relevantes son:

Eficiencia del atributo (F3). En un problema caracterizado por un gran número de dimensiones, la información relevante se distribuye a través de los diferentes atributos. Esta métrica determina la eficiencia de cada atributo de manera individual y estima en qué medida el atributo participa en la separabilidad de las clases. Se parte de una heurística de continuidad local que asume que, para cada atributo, las instancias de la misma clase oscilan entre el mínimo y el máximo de esa clase. Entonces, cuando dos instancias de clase opuesta toman el mismo valor, se produce un solapamiento y por lo tanto se considera ambigua la región de esa dimensión. Esta ambigüedad se resuelve eliminando los puntos que se sitúan en la zona solapada y la eficiencia se calcula como el cociente entre los puntos restantes y el total de puntos. La métrica corresponde al valor máximo de las eficiencias calculadas para cada dimensión.

Distancia de la frontera (N1). Proporciona el porcentaje de nodos que conectan clases opuestas en un árbol de expansión mínimo (*Minimum Spanning Tree*, MST). El árbol se construye a partir del grafo que generan las relaciones de cada instancia con el resto mediante el cálculo de la distancia euclidiana. Esta métrica es un indicador de la separabilidad de las clases y de la tendencia a los clústers. Cuanto mayor es el valor de la métrica, más se acentúa la presencia de puntos cercanos de clases opuestas. En cambio, cuanto menor es, más agrupadas están las instancias de la misma clase y menos dificultad aparente denota el problema. Esta métrica no es adecuada para identificar la separabilidad lineal del problema, puesto que en un conjunto en

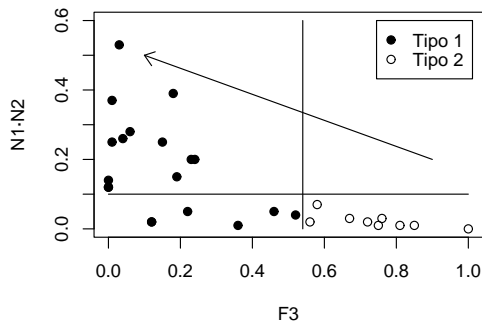


Figura 3: Espacio de complejidad. Cada conjunto de datos se representa en función del valor de $F3$ y $N1 \cdot N2$. Este mapa permite predecir la aplicabilidad del SOMCBR.

bajos.

Los problemas de Tipo 2 corresponden a problemas en que el porcentaje de reducción de operaciones es bajo y coincide, para todos los casos, que muestran una complejidad baja ($F3$ con un valor elevado que indica que hay atributos discriminantes y un valor bajo de $N1 \cdot N2$ que eleva separabilidad entre clases). En cambio, los problemas de Tipo 1 presentan altos porcentajes de reducción en los que no se refleja una tendencia estable. A veces el SOMCBR es equivalente al CBR, es decir que no hay diferencias significativas en cuanto a porcentaje de aciertos, y en otras es peor. Estas dos situaciones que se producen en el Tipo 1 no se pueden diferenciar mediante las métricas de complejidad. Por lo tanto, sólo cabe concluir que si el problema es fácil, desde el punto de vista de complejidad, no es útil aplicar clusterización en la MC ya que no habrá una reducción de operaciones significativa. Para los problemas de complejidad elevada la aplicabilidad del SOMCBR dependerá del problema en concreto.

Entendiendo que la complejidad de los datos afecta a la construcción de los clústers y por lo tanto al comportamiento de las diferentes estrategias, se considera el análisis de la com-

plejidad de la frontera [9] para evaluar su influencia en las estrategias de recuperación.

Refinando el estudio anterior sobre una experimentación ampliada a 56 conjuntos de datos, se obtiene un nuevo espacio que clasifica los problemas en tres tipos de complejidad (ver Fig. 4). En este espacio el punto (1,0) se corresponde al punto de mínima complejidad (mCP) mientras que el punto (0,1) alcanza la máxima complejidad posible (MCP).

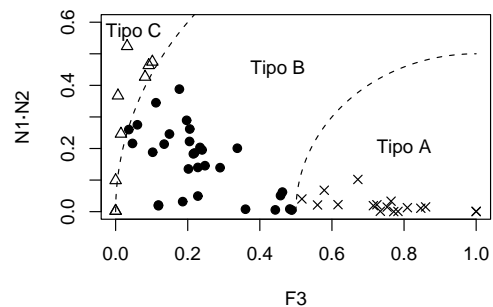


Figura 4: Espacio de complejidad. Divide en tres regiones el grado de complejidad de los datos, siendo C el mayor posible.

La distancia desde el punto mCP divide los conjuntos de datos estudiados en tres grupos: (1) problemas de complejidad baja (Tipo A, la distancia respecto al punto mCP es menor a 0.5), (2) problemas de complejidad elevada (Tipo C, con un valor mayor o igual a 0.9) y (3) problemas de complejidad intermedia (Tipo B). Con esta clasificación se puede describir el rendimiento de cada estrategia con mayor precisión.

Este mapa de complejidad supone una nueva variable que se incorpora en la elección de la configuración de la estrategia de recuperación. El rendimiento es un compromiso entre el tiempo de cómputo y la precisión de clasificación deseada. El resultado del análisis se resume en la tabla 1. Para problemas de complejidad A, es decir fáciles, hay una alta correlación entre el número de operaciones del

SOMCBR respecto al CBR y la precisión de la estrategia de recuperación concreta, que alcanza un coeficiente de correlación de 0.96. Significa que cuantas más operaciones, mayor es el porcentaje de aciertos, lo que se traduce en que la clusterización perjudica la precisión del sistema. En los problemas de complejidad B, complejidad intermedia, se constatan dos efectos: (1) el número de operaciones disminuye en la mayoría de estrategias y (2) el coeficiente de correlación decrece hasta un 0.86. Así pues, se obtiene una mejora en el rendimiento sin perder precisión. Finalmente, para los problemas de tipo C, de mayor complejidad, se pronuncian los efectos enunciados en los problemas de tipo B siendo ya la correlación de 0.76.

Tipo de complejidad	Coefficiente de correlación
A	0.96
B	0.86
C	0.76

Cuadro 1: Coeficientes de correlación entre la precisión y la reducción del número de operaciones según el grado de complejidad.

Analizando cuál es la estrategia más adecuada, se puede decir que para los problemas de tipo A y B, que corresponden a una complejidad baja e intermedia respectivamente, la estrategia All_All es la que ofrece mejor porcentaje de aciertos. Esta configuración equivale a un sistema CBR. En problemas de complejidad elevada, tal como se observa en la figura 5, la mejor estrategia se encuentra en la configuración Eq3_05_All. Esta se basa en recuperar un porcentaje de los casos de cada clúster en función de su bondad para representar el patrón de entrada. La bondad se calcula mediante la diferencia entre el caso de entrada y el vector director del patrón.

5. Discusión

En este trabajo se ha revisado cómo el análisis de la complejidad puede ser útil para estimar la aplicabilidad de la clusterización para sistemas CBR. Por una parte, se ha analizado

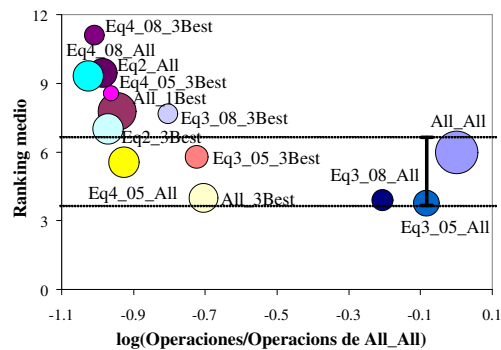


Figura 5: Análisis de las estrategias de recuperación según el tipo de complejidad C.

la posibilidad de predecir cuándo la clusterización puede aportar una mejora importante en el coste computacional sin perder demasiada precisión. Se ha detectado que para problemas de baja complejidad, la clusterización no aporta beneficios. Sin embargo, para problemas de elevada complejidad, la clusterización suele conllevar una reducción importante del número de operaciones aunque no se puede predecir si esta reducción implica pérdida de precisión o no. Con las métricas actuales, no se puede predecir qué tipo de problemas mantendrán la precisión con respecto a un sistema sin clusterizar.

Esta limitación puede venir dada por una caracterización insuficiente de la complejidad del problema. En el trabajo presentado, se usaron las métricas más significativas que son F3, N1 y N2. La primera dedica su atención a estimar en qué grado los atributos individualmente intervienen en la separación de las clases y las dos siguientes describen la complejidad de la frontera entre clases teniendo en cuenta la distancia entre puntos. El estudio de nuevas métricas podría aportar más información y aproximarnos mejor a la caracterización del problema. Asimismo sería necesario aumentar el número de problemas a analizar, que en el estudio presentado cuenta con 28 en una primera fase y 56 en una segunda. Algunos estudios sobre la complejidad de los problemas usan alrededor de 300 *datasets* [8] para gene-

ralizar el comportamiento. Podría suceder que nuestra elección de problemas esté sesgada y limite la generalización de los resultados. De hecho, se ha usado un conjunto reducido de problemas debido al elevado coste computacional que supone ejecutar la clusterización en diferentes configuraciones y con 10 repeticiones cada una de ellas.

La segunda parte de nuestro estudio realiza un paso más allá de la mera predicción de la aplicabilidad de la clusterización. En concreto, trata de analizar qué tipo de estrategia de recuperación del ciclo del CBR es más apropiada según el problema. El estudio pretende obtener como resultado una recomendación del tipo de estrategia a usar dado un problema caracterizado por su complejidad. Para ello, se propuso una metodología que parte de una taxonomía de las estrategias de recuperación donde interviene la definición de dos parámetros: el número de clústers y el número de casos que deben recuperarse. A partir de esta clasificación, cada configuración se ejecuta sobre un conjunto de datos del cual se obtiene, mediante una representación gráfica, un mapa del rendimiento de cada estrategia en función de la reducción del número de operaciones y de la precisión de la clasificación. Para cada conjunto de datos se mide su complejidad mediante tres métricas, F3, N1 y N2, cuya combinación deriva en un mapa sobre el que se identifica a qué tipo de complejidad pertenece el problema. Para cada tipo, se construye un mapa de estrategias en función del tiempo computacional y la precisión de la clasificación.

El estudio presentado se ha realizado únicamente para problemas con dos clases. El motivo es que la definición original de las métricas de complejidad se realizó para tal tipo de problemas [9]. Si el problema es multiclase, se puede analizar la complejidad promedio de los problemas biclase que derivarían de comparar cada clase con el resto. Sin embargo, probablemente sería más preciso analizar la complejidad del problema multiclase en conjunto y no con cada clase por separado. Actualmente, ya existe alguna propuesta de métricas de complejidad para problemas multiclase [2], aunque

no existen trabajos que lo usen. Creemos que sería útil extender este estudio para problemas multiclase.

En resumen, este trabajo ha revisado una posible aplicabilidad del estudio de la complejidad del problema en el análisis de la clusterización para sistemas CBR. En general, el estudio de la complejidad del problema puede ayudar a comprender el comportamiento de determinadas estrategias de aprendizaje, mejorar el rendimiento de las mismas y predecir la aplicabilidad de éstas *a priori*. Aún así, son necesarios más esfuerzos para estimar mejor la complejidad de los problemas y relacionarla debidamente con el rendimiento de los esquemas de aprendizaje.

6. Conclusiones

La necesidad de mejorar el tiempo de cómputo sin mermar la precisión de la clasificación ha llevado a plantear una metodología que ayude a decidir la estrategia de recuperación de casos apoyándose en la información proporcionada por las métricas de complejidad. Así pues, la complejidad de los datos es una nueva variable que ayuda a predecir en qué casos es útil aplicar el SOMCBR y a un nivel superior cuál debe ser la configuración de la estrategia, en lo que se refiere a número de clústers a seleccionar y número de casos a recuperar. Cabe destacar que el beneficio del estudio de la complejidad de los datos se encuentra en que las decisiones se toman únicamente con los datos y sin la necesidad de aplicar el sistema clasificador. Por lo tanto, la aportación de las métricas de complejidad es la posibilidad de predecir *a priori* el rendimiento del SOMCBR para un problema determinado.

Las líneas de futuro apuntan hacia el estudio de nuevas métricas que puedan completar la definición de la complejidad de los datos, y consolidar así la metodología sobre memorias de casos organizadas por otras técnicas.

Agradecimientos

Este trabajo ha sido soportado en parte por los proyectos TIN2006-15140-C03-03 y TIN2005-

08386-C05-04 gracias al Ministerio de Ciencia y Tecnología, y al Fondo Europeo de Desarrollo Regional (FEDER). También se debe al *Departament d'Universitats, Recerca i Societat de la Informació* (DURSI) por el apoyo proporcionado mediante las becas 2005SGR-302 y 2007FIC-00976. Asimismo, los autores agradecen a *Enginyeria i Arquitectura La Salle*, de la Universitat Ramon Llull, el soporte a nuestro Grupo de Investigación en Sistemas Inteligents

Referencias

- [1] A. Aamodt and E. Plaza. Case-based reasoning: Foundations issues, methodological variations, and system approaches. *AI Communications*, 7:39–59, 1994.
- [2] M. Basu and T.K. Ho. *Data Complexity in Pattern Recognition*. Advanced Information and Knowledge Processing. Springer, 2006.
- [3] E. Bernadó and T.K. Ho. Domain of competence of XCS classifier system in complexity measurement space. *IEEE Transaction Evolutionary Computation*, 9(1):82–104, 2005.
- [4] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.
- [5] A. Fornells, E. Golobardes, J.M. Martorell, J.M. Garrell, E. Bernadó, and N. Macià. Measuring the applicability of self-organization maps in a case-based reasoning system. In *3rd Iberian Conference on Pattern Recognition and Image Analysis*, volume 4478 of *LNCS*, pages 532–539. Springer-Verlag, 2007.
- [6] A. Fornells, E. Golobardes, J.M. Martorell, J.M. Garrell, E. Bernadó, and N. Macià. A methodology for analyzing the case retrieval from a clustered case memory. In *7th International Conference on Case-Based Reasoning*, LNAI. Springer-Verlag, 2007. In press.
- [7] A. Fornells, E. Golobardes, D. Vernet, and G. Corral. Unsupervised case memory organization: Analysing computational time and soft computing capabilities. In *8th European Conference on Case-Based Reasoning*, volume 4106 of *LNAI*, pages 241–255. Springer-Verlag, 2006.
- [8] T.K. Ho. A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis and Applications*, 5:102–112, 2002.
- [9] T.K. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(3):289–300, 2002.
- [10] T. Kohonen. *Self-Organization and Associative Memory*, volume 8 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1984. 3rd ed. 1989.