ARTÍCULO

La Técnica Mimética en Ausencia de Datos Originales: Aprendizaje y Revisión de Modelos

Ricardo Blanco-Vega, José Hernández-Orallo, María José Ramírez-Quintana

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, C. de Vera s/n,
46022 Valencia, España.
{rblanco, jorallo, mramirez}@dsic.upv.es

Resumen

La técnica mimética (o *Combined Multiple Models*, CMM) básicamente consiste en usar un modelo, generalmente preciso pero incomprensible, como oráculo para generar un conjunto de datos aleatorios que luego se utiliza junto con el conjunto de datos de entrenamiento inicial para entrenar a un segundo modelo comprensible, conocido como el modelo mimético. Esta técnica se ha empleado para dotar de comprensibilidad a modelos de caja negra sin sacrificar considerablemente su precisión. En este trabajo estudiamos la aplicación del mimetismo en un escenario en el que los datos originales de entrenamiento no están disponibles. En este marco primero determinamos el tamaño óptimo del conjunto de datos aleatorios siguiendo el principio de longitud mínima de mensaje (MML). Este resultado puede ser empleado en la adquisición de conocimiento para la creación de sistemas expertos. En segundo lugar aplicamos la técnica mimética a la revisión de modelos y mostramos cómo en determinadas situaciones de cambio el modelo mimético puede usarse como modelo de transición entre el modelo original y el nuevo modelo.

Palabras clave: Técnica mimética, *Combined Multiple Models*, revisión de teorías, optimización de modelos, sistemas expertos, adquisición de conocimiento, adaptación de modelos.

1. La Técnica Mimética

En muchas aplicaciones reales no están disponibles los datos originales de entrenamiento del modelo que se emplea, ya sea porque el modelo fue construido hace mucho o porque el modelo fue construido por un experto humano sin aplicar ninguna técnica de aprendizaje computacional. También puede ser que los datos usados para construir el modelo simplemente ya no existan por pérdida o daño de los mismos. Esta es una situación real en muchas áreas tales como ingeniería, diagnosis, procesos de fabricación, negocios, etc.

Habitualmente, muchas técnicas y procedimientos del aprendizaje automático se evalúan en escenarios mucho más restrictivos, donde, por ejemplo, sólo se pueden aplicar algunas técnicas del aprendizaje automático, o que requieren el disponer de suficientes datos de entrenamiento.

Por contra, el método mimético puede aplicarse sin virtualmente ninguno de esos supuestos: no hay necesidad de los datos originales y no tenemos que saber su distribución original. Además, cualquier técnica de aprendizaje computacional puede ser utilizada. A esto se suman las otras ventajas ya mostradas por las anteriores aplicaciones del método: su capacidad de recuperar con el modelo mimético la comprensibilidad de la que carecen ciertos modelos altamente precisos y la posibilidad de disponer de un conjunto de datos aleatorio tan grande como sea necesario. Finalmente destacamos su fácil

implementación en diferentes ambientes y librerías de minería de datos.

Nuestro objetivo en este trabajo es doble. En primer lugar estimar el tamaño óptimo del conjunto de datos. El segundo objetivo es aplicar la técnica mimética para la revisión de modelos.

Dentro del primer objetivo damos contestación a dos interrogantes de gran trascendencia en la construcción de los modelos miméticos, como son: ¿existe un tamaño del conjunto de datos de entrenamiento del modelo mimético que maximice precisión y comprensibilidad?; si existe, ¿cómo se puede estimar ese tamaño óptimo?.

Respecto al segundo objetivo, la necesidad de la revisión del modelo es accionada por un aumento significativo en el error del mismo. Pero ¿quién detecta todos estos nuevos errores?. Generalmente los errores tienen que ser supervisados por un regulador o un supervisor que, obviamente, no puede etiquetar o corregir todos los datos. De no ser así, no habría necesidad de usar un modelo y directamente utilizaríamos al supervisor. Por lo tanto, en situaciones reales la cantidad de datos disponibles para la revisión del modelo (datos nuevos) es generalmente pequeña. Demostraremos cómo en estos casos el modelo mimético puede usarse como modelo de transición entre el modelo original y el nuevo modelo que se podría generar cuando el número de datos nuevos fuera lo suficientemente grande como para garantizar la calidad de este nuevo modelo.

Este documento está organizado como sigue. En la sección 2 presentamos la técnica mimética e introducimos las nuevas adaptaciones de esta técnica en el caso de no contar con los datos originales. En la sección 3 presentamos una aproximación de cómo proceder en la aproximación del tamaño óptimo en modelos miméticos considerando el principio MML y curvas de aprendizaje. En la sección 4 presentamos la revisión de teorías a través del aprendizaje mimético, proponiendo dos escenarios donde es factible emplear el método mimético. Finalmente, la sección 5 presenta las conclusiones y el trabajo futuro.

2. La Técnica Mimética

En 1998 Domingos [Domingos98] propuso y evaluó la combinación de múltiples modelos (*Combined Multiple Models*, CMM), un algoritmo de meta-aprendizaje que intenta conservar la mayoría del incremento en la precisión obtenida en la combinación de múltiples modelos, pero generando un solo modelo comprensible. CMM se basa en reaplicar un clasificador base para obtener un modelo simple (μ) que muestre un comportamiento similar, en cuanto a precisión, al presentado por la combinación de varios modelos (Ω). Esto se realiza dando al clasificador base un nuevo conjunto de datos de entrenamiento, integrado por una gran cantidad de ejemplos generados y clasificados por el modelo combinado (R), más los ejemplos originales (R+T). Este proceso se observa en la Figura 1. El principal objetivo de esta técnica es recuperar la comprensibilidad del modelo sin sacrificar demasiado la precisión.

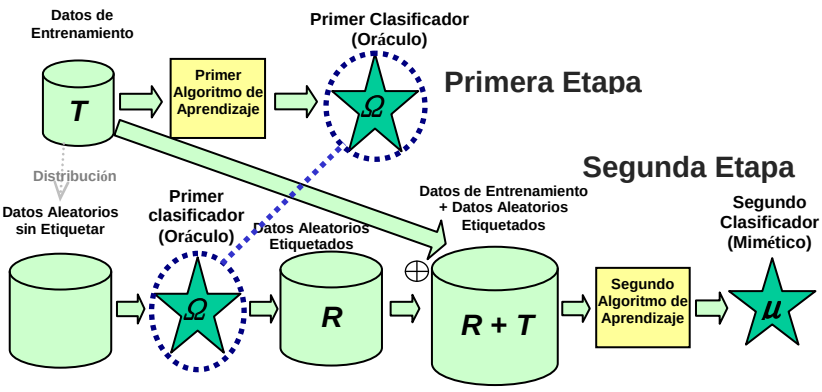

Figura 1. Técnica mimética o Combined Multiple Models

El CMM no solo se puede aplicar a una combinación de modelos sino a cualquier modelo considerado como oráculo, es decir no sólo se pueden emplear algoritmos como *bagging* [Breiman96][Quinlan96], o *boosting* [Freund et al. 96][Quinlan96], sino también modelos que tienen muy buena precisión pero con problemas de comprensibilidad como es el caso de las redes neuronales [McClelland et al. 1986] o las máquinas de vectores soporte [Boser et al. 1992].

Como el método CMM se basa en imitar a un oráculo a través de un modelo comprensible se ha rebautizado con el nombre de método mimético.

El método mimético [Estruch et al. 03] consiste en dos etapas. La primera etapa corresponde al aprendizaje del oráculo. Para ello, se utilizan los datos de entrenamiento originales ***T*** como entrada al primer algoritmo de aprendizaje, el cual creará este primer clasificador, que es preciso pero incomprensible. Ejemplos de algoritmos que se pueden utilizar en esta primer etapa son: las redes neuronales, boosting, bagging, multiclasificadores (cualquier combinación de clasificadores) y máquinas de vectores soporte.

En la segunda etapa se realiza el aprendizaje mimético. El objetivo de esta etapa es obtener un modelo que imite el comportamiento del modelo de caja negra, buscando las propiedades de conservar el nivel de precisión obtenido con el modelo de caja negra y generar un conjunto de reglas comprensibles que permitan el entendimiento de dicho modelo. Tomando como base los datos de entrenamiento originales y empleando algún método simple de muestreo se genera un conjunto de datos aleatorio sin etiquetar. Luego ese conjunto de datos ***R*** se etiqueta utilizando el oráculo. Los conjuntos de datos etiquetados por el oráculo y el conjunto de datos de entrenamiento original se unen para formar el total de los datos de entrada ***T*** + ***R*** para el segundo algoritmo de aprendizaje y así crear el modelo mimético. El segundo algoritmo de aprendizaje puede ser cualquier clasificador que genere reglas, como por ejemplo J48 (la implementación en WEKA [Witten et al. 05] del conocido algoritmo de inducción de árboles de decisión C4.5 [Quinlan93]) y C5 (el sucesor del algoritmo C4.5). En particular, Domingos empleó *bagging* como oráculo y C4.5rules como el modelo comprensible final.

Sin embargo, el caso de aplicar la técnica mimética cuando no se dispone de los datos originales no ha sido estudiada. En este caso la técnica mimética se reduce al proceso mostrado por la Figura 2. Primero se genera un conjunto de datos utilizando la distribución uniforme (aunque no se conocen los datos originales se supone conocido el rango de valores de los atributos). Luego se utiliza el modelo original como oráculo para etiquetarlos formando lo que conocemos como el conjunto de datos inventados. Después, Después, este conjunto se utiliza para aprender un árbol de decisión, como C4.5, que genera en última instancia un modelo en forma de reglas.

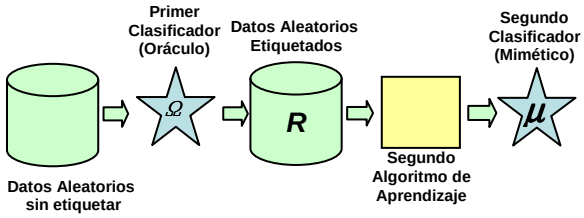

Figura 2. Técnica mimética sin datos originales

3. Tamaño Óptimo en Modelos Miméticos

En [Blanco et al. 04][Domingos98] se mostró que a mayor número de datos inventados se obtiene una mejor precisión y a menor número de datos inventados un menor número de reglas (mayor comprensibilidad). Estos resultados sugieren que es necesario encontrar un compromiso entre la precisión y la comprensibilidad del modelo mimético. El factor que mejor sirve para encontrar ese compromiso es el tamaño del conjunto de datos de entrenamiento.

Para decidir el tamaño óptimo del conjunto artificial, la idea es considerar el mimetismo como lo que realmente es, un segundo problema de aprendizaje, donde tanto un sobreajuste como un subajuste son perniciosos. No obstante, el uso de un conjunto de validación no es útil aquí porque siempre favorecería el sobreajuste ya que podemos generar conjuntos arbitrariamente grandes. Una solución más flexible la podemos encontrar en el principio de la longitud mínima de mensaje (*Minimum Message Length*, MML) [Baxter96][Wallace et al. 68].

El principio MML puede emplearse como un método de comparación de modelos. Del teorema de Bayes sabemos que la probabilidad de una hipótesis (H) dada la evidencia (D) es proporcional a $p(D|H)\cdot p(H)$, lo cual es justamente $p(H \cap D)$.

Por lo tanto, deseamos el modelo que genera la descripción más corta de los datos. Definimos la longitud de mensaje como

$$\text{MsgLen}(H \cap D) = -\log_2(p(H \cap D)) \quad (1)$$

entonces el modelo más probable tendrá el mensaje más corto. El mensaje se separa en dos partes:

$$-\log_2(p(H \cap D)) = -\log_2(p(H)) + (-\log_2(p(D|H))) \quad (2)$$

La primera parte es la longitud del modelo, y la segunda es la longitud de los datos dado el modelo (los errores que comete el modelo) [Baxter96]. Reescribiendo la ecuación (2) queda como:

$$\text{Coste(Modelo)} = \text{MsgLen(Modelo)} + \text{MsgLen(Datos|Modelo)} \quad (3)$$

El problema parece que vuelve a aparecer, ya que para un conjunto de datos muy grande (y lo podemos crear infinitamente grande), el segundo sumando se haría muy grande y el tamaño del modelo sería despreciable por muy grande que fuera. La cuestión aquí es establecer un compromiso entre el tamaño del modelo y el tamaño de los datos para que los costes estén equilibrados (llamaremos a este parámetro *K* y será decidido por el usuario).

A la hora de implementarlo usaremos una aproximación muy sencilla. Considerando que el Modelo consta de *R* reglas y que el coste unitario de cada regla es *cr*, entonces:

$$\text{MsgLen(Modelo)} \approx R^{*}cr \quad (4)$$

De la misma forma la longitud de los datos que son excepción al modelo son aproximadamente proporcionales al error *E* y al coste unitario del error (*ce*):

$$\text{MsgLen(Datos/Modelo)} \approx E^{*}ce \quad (5)$$

Si sustituimos (4) y (5) en la ecuación (3), se obtiene la expresión:

$$\text{Coste(Modelo)} \approx R^{*}cr + E^{*}ce \quad (6)$$

la cual representa la función objetivo.

Las curvas de aprendizaje son funciones que relacionan dos parámetros del modelo mimético que son de nuestro interés en la optimización: tamaño contra error y tamaño contra el número de reglas. Estimamos estas curvas en el escenario descrito por la Figura 2, es decir, sin considerar los datos de entrenamiento iniciales. De esta forma, los resultados obtenidos son generales e independientes de un problema concreto y pueden aplicarse en cualquier situación.

La curva de aprendizaje del tamaño del conjunto de datos inventado (*S*) contra el número de reglas (*R*) es lineal, como se ha mostrado también en otros trabajos anteriores [Oates et al. 97]. La función que describe esta curva es

$$R = \delta^{*}S + \lambda \quad (7)$$

siendo δ la pendiente de la recta y λ la coordenada en el origen.

Para el cálculo de una función similar a la (7) pero que sirva para relacionar los errores y el tamaño se realizaron experimentos usando 18 conjuntos de datos (ver Tabla 1) del repositorio UCI [Black et al. 98].

Tabla 1. Conjuntos de datos usados en los experimentos para las curvas de aprendizaje. Dato se refiere al conjunto de datos, Atr. Num. son los atributos numéricos, Atr. Nom. son los atributos nominales y n es el tamaño del conjunto de datos.

No.	Dato	Atr. Num.	Atr. Nom.	Clases	n
1	anneal	6	32	6	898
2	audiology	0	69	24	226
3	balance-scale	4	0	3	625
4	breast-cancer	0	9	2	286
5	cmc	2	7	3	1,473
6	colic	7	15	2	368
7	diabetes	8	0	2	768
8	hayes-roth	0	4	3	132
9	hepatitis	6	13	2	155
10	iris	4	0	3	150
11	monks1	0	6	2	556
12	monks2	0	6	2	601
13	monks3	0	6	2	554
14	sick	7	22	2	3,772
15	vote	0	16	2	435
16	vowel	10	3	11	990
17	waveform	40	0	3	5,000
18	zoo	1	16	7	101

Los experimentos consistieron en generar 100 modelos miméticos siguiendo el proceso mostrado en la Figura 2. Se varió el tamaño del conjunto de datos aleatorios de un 5% hasta un 500% del tamaño del conjunto de entrenamiento original. La Figura 3 muestra como ejemplo una curva de aprendizaje "error vs tamaño" para el conjunto de datos balance-scale.

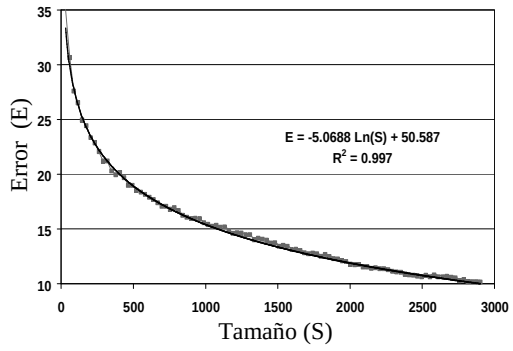

Figura 3. Error vs. Tamaño para el conjunto de datos balance-scale

Con estos experimentos obtuvimos una función que describe la relación entre el error (E) y el tamaño (S) con un promedio del coeficiente de determinación de 0.82; esto significa que el 82% de la variación promedio de los errores son explicados a través del tamaño. La función obtenida es:

$$E = \alpha^* \mathrm{Ln}(S) + \beta \qquad (8)$$

donde α es la pendiente de la recta en una escala semilogarítmica y β la coordenada en el origen.

Se sustituyen las ecuaciones (7) y (8) en (6), luego se deriva con respecto a *S* y finalmente se calculan los puntos críticos (mínimos y máximos) obteniendo así una estimación del tamaño óptimo con la ecuación

$$S_{opt} = -K^* \alpha / \delta \qquad (9)$$

donde *K* es una constante de proporcionalidad entre los costes unitarios *ce* y *cr*.

Para saber si se trata de un valor máximo o mínimo se requiere el signo de la curvatura que se obtiene a través de la segunda derivada, la cual es

$$-K^* \alpha / S^2$$

El signo de esta expresión coincide con el signo de α dado que *K* y *S* siempre son positivos. Si α es negativo entonces el signo de la doble derivada es positivo, esto significa que hemos encontrado un valor mínimo, es decir, tenemos el punto donde disminuimos el coste de representación del modelo. Si el signo de alfa es positivo esto nos indica que es un punto máximo y no interesa para nuestro problema. El signo positivo de alfa nos estaría indicando que al ir aumentando el número de datos de entrenamiento la precisión del modelo obtenido tendería a bajar y esto típicamente es una anomalía. Por lo tanto queda demostrado que sí es posible estimar con la ecuación (9) el tamaño óptimo del conjunto de datos de entrenamiento en un modelo mimético.

Dentro de los mismos experimentos hemos probado que es posible estimar los parámetros que definen a las curvas de aprendizaje con tan sólo tres modelos miméticos. Estas evaluaciones experimentales se encuentran en [Blanco06].

Uno de los beneficios que se pueden obtener al aplicar el método presentado es usarlo para la adquisición del conocimiento en el diseño de sistemas expertos. La idea es reducir al mínimo el número de entrevistas al experto (oráculo) sin exceder el número de los casos requeridos que se etiquetarán. El proceso general consistiría en pedir que el experto etiquete 20 ejemplos (generando un modelo con 10 ejemplos y otro con 20 ejemplos). Con esto computamos una primera curva, de la cual podemos estimar el valor óptimo. Si el tamaño estimado es más pequeño que el número de ejemplos pedidos al experto (situación extraña) o si la diferencia entre el valor estimado y el número de ejemplos etiquetados es pequeño entonces terminamos. Si no, pedimos al experto los casos restantes hasta el tamaño estimado.

4. Revisión de Teorías

En general, cualquier sistema (teoría o modelo) debe ser eventualmente revisado. El aprendizaje computacional provee de métodos muy poderosos para obtener modelos a partir de los datos, pero su uso en "retocar" o adaptar un método existente depende de la técnica actual.

En general, la revisión de teorías se define como el proceso de modificar una teoría inicial para producir una nueva teoría que sea más precisa en una nueva situación. En los sistemas de revisión de teoría estándares (RTLS [Ginsberg90]; [Taylor et al. 97]; DUCTOR [Cain91]; [Matwin et al. 91]; EITHER [Ourston94]; KBDistAl [Yang et al. 99]) se emplean los casos que son incorrectamente cubiertos por la teoría para calcular el cambio mínimo necesario para corregirla. Esto se hace normalmente agregando y/o suprimiendo reglas y/o antecedentes a las mismas.

El aprendizaje computacional parece no proporcionar muchas herramientas para adaptar el modelo existente cuando el contexto comienza a

cambiar y el antiguo modelo comienza a empeorar su comportamiento. Aprender un nuevo modelo usando los nuevos datos es obviamente una opción, pero el modelo aprendido estará generalmente sobreajustado porque al principio no tendremos suficientes datos nuevos o tendremos solamente datos de las "nuevas áreas" donde el contexto está cambiando.

Ante situaciones como éstas, en general, podemos proceder de las siguientes formas (ver Figura 4):

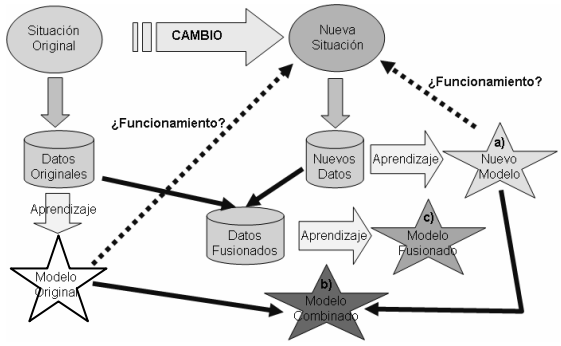

Figura 4. Modelos alternativos en un escenario de cambio.

a) Entrenar un nuevo modelo con los nuevos datos. En este caso, el modelo original es substituido por el nuevo. La debilidad principal de esta aproximación es que se requiere tener a mano una gran cantidad de nuevos datos para que el nuevo modelo sea más preciso o al menos con una precisión similar a la del modelo original.

b) Entrenar un nuevo modelo con los nuevos datos (opción a) y combinarlo con el modelo antiguo. Aunque, en general, un modelo combinado tiene una precisión mayor que la precisión de sus modelos componentes, este resultado se obtiene cuando se combinan una gran cantidad de modelos simples diferentes. Sin embargo es incierto qué sucede con la combinación de solamente dos modelos. Otra desventaja de esta opción es que un modelo combinado sería incomprensible, lo cual puede ser un serio inconveniente para varios dominios de aplicación.

c) Entrenar un nuevo modelo con la fusión de los datos originales y nuevos. Esta opción requiere que los datos originales y nuevos estén disponibles.

Tal y como hemos comentado en la introducción, nuestro interés se centra en escenarios en los que los datos originales no están disponibles. Bajo estas condiciones no todas las opciones que hemos mencionado son aplicables. Así, la opción c) es inviable en este caso por no disponer de los datos originales. Si además, el número de datos nuevos no es muy grande, la opción a) tampoco es muy adecuada.

Nosotros proponemos como alternativa reutilizar el modelo original de una manera diferente a como se usa en la opción b). La idea sería utilizarlo como oráculo, para etiquetar un conjunto de datos amplio y grande generado aleatoriamente y que representa la situación original. A continuación, combinamos este conjunto de datos con los nuevos datos originados de la nueva situación, posiblemente usando sobremuestreo para balancear la importancia de la situación original y la nueva. Entonces, con el conjunto de datos resultante de la unión, aprendemos un modelo que abarcará tanto la situación original (de la cual tenemos mucho más detalle a través de los datos aleatorios) como la nueva situación (de la cual tenemos menos detalle). Se trata, por lo tanto, de modificar el esquema general de la técnica mimética prescindiendo del conjunto de entrenamiento inicial y añadiendo a los datos aleatorios los nuevos datos. Este proceso se ilustra en la.Figura 5.

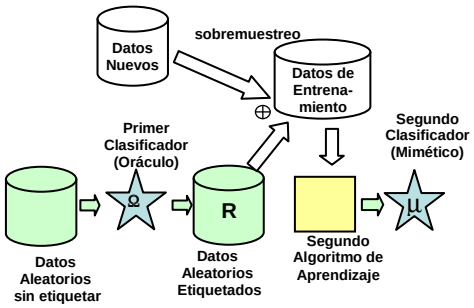

Figura 5.Técnica mimética sin datos originales más datos nuevos, para revisión de teorías.

Por lo que conocemos, la técnica mimética no se ha utilizado para la revisión de teorías. Obviamente, para aplicar la técnica, las dos condiciones generales para la revisión de una teoría deben existir: el modelo antiguo se comporta peor que al inicio cuando se evalúa con los nuevos datos (la revisión es necesaria) y los nuevos datos no son grandes o no son suficientes para obtener un nuevo modelo de calidad (la revisión es apropiada).

A continuación analizamos nuestra propuesta en dos escenarios de revisión: cuando los nuevos datos pertenecen a una porción del dominio no cubierto

por el modelo original (novedad) y cuando la revisión se activa por un cambio generalizado (error). Comparamos los resultados obtenidos por nuestra propuesta con los modelos nuevo (opción a), combinado (opción b) y original.

4.1. Escenario con novedad

En esta sección presentamos un escenario común a la revisión de teorías donde los cambios ocurren sobre todo en las áreas que no eran muy importantes en la situación original. Este escenario es frecuente siempre que un sistema tenga que "cubrir" o "tratar" una nueva área, y generalmente se llama revisión accionada por novedad.

Resumimos este escenario considerando dos situaciones: la situación original y la situación final o futura. La Figura 6 muestra además el momento en el cual la necesidad de cambio es accionada en un cierto punto intermedio entre ambas.

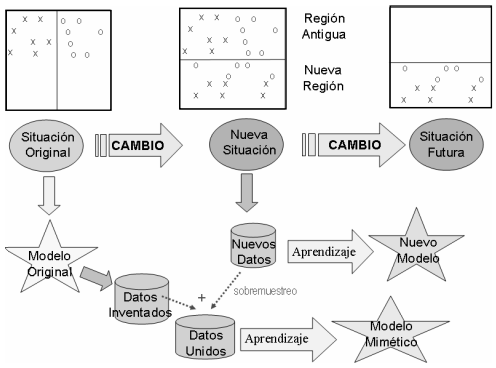

Figura 6. Escenario accionado por novedad.

Como hemos mencionado, la idea es comparar el modelo mimético con el nuevo modelo, el modelo original y su combinación. Para hacer esto simulamos el escenario accionado por novedad como sigue. Dado un conjunto de datos *D*, identificamos artificialmente un subconjunto de ejemplos de *D*, que denotamos como D_C, que pertenecen a una región arbitraria. Esta región se define generando una condición *C* sobre un atributo. Si el atributo seleccionado x es nominal entonces las condiciones son de la forma $x = v$, donde el valor *v* (aleatoriamente elegido) pertenece al dominio de *x*. Similarmente, las condiciones sobre atributos numéricos son de la forma $x <= v$ o $x > v$, donde el valor *v* pertenece al intervalo $[x_{min}, x_{max}]$ siendo x_{min} y x_{max} los valores mínimo y máximo del atributo *x* en *D*. Ambos, el valor y el atributo se determinan aleatoriamente. Ahora, la condición *C* permite definir el conjunto de datos original *O* quitando de *D* los ejemplos que cumplen *C*, es decir, $O = D - Dc$. Por lo tanto, *O* no tiene un área con respecto a *D* (la cual está definida por D_C). Análogamente, para construir el conjunto de datos nuevos, seleccionamos la misma proporción de ejemplos de *O* y de D_C (utilizamos sobremuestreo en caso necesario). La condición *C* se elige de tal forma que el tamaño de D_C esté entre el 20% y el 30% del tamaño de *D*. Esto hace que generalmente la condición *C* se obtenga usando un solo atributo.

El modelo Combinado es obtenido por la combinación del modelo nuevo y el modelo original utilizando las probabilidades de los modelos por votación, es decir un ejemplar se clasifica con la clase que tenga mayor probabilidad de ser en ambos modelos (sumando sus probabilidades).

Debido a que se desconoce la distribución de los datos originales, la generación del conjunto de datos inventados se realizó usando la distribución uniforme para la cual solamente se necesita conocer los límites (valores máximo y mínimo) de los atributos.

Para la evaluación experimental se utilizaron 10 conjuntos de datos (ver Tabla 2) del repositorio UCI [Black et al. 98],

Tabla 2. Información de los conjuntos de datos usados en los experimentos para el escenario accionado por novedad. Num. significa numéricos y Nom. significa nominales.

#	Conjunto de Datos	Atributos		Clases	Tamaño	Datos Faltantes
		Num.	Nom.			
1	anneal	6	32	6	898	No
2	audiology	0	69	24	226	Sí
3	balance-scale	4	0	3	625	No
4	breast-cancer	0	9	2	286	Sí
5	colic	7	15	2	368	Sí
6	diabetes	8	0	2	768	No
7	hepatitis	6	13	2	155	Sí
8	iris	4	0	3	150	No
9	vowel	10	3	11	990	No
10	zoo	1	16	7	101	No

El modelo antiguo evaluado con los datos nuevos se denota por $M_{Old}(New)$ o M_{Old} . El modelo nuevo con M_{New}. Comparemos la precisión que estos modelos obtuvieron. Hay casos donde M_{Old} es preferible a M_{New} y otros casos donde ocurre lo contario, como se puede ver en la Tabla 3 (columnas 2 y 3). Atendiendo a las medias, sin embargo, podemos concluir que M_{Old} es mejor que M_{New}. Esto sucede probablemente, porque el tamaño del conjunto de datos usado para el entrenamiento de M_{New} es

demasiado pequeño (que es precisamente el escenario que deseamos examinar aquí).

Para el modelo mimético el parámetro α es la proporción de ejemplos que tenemos que utilizar de la vieja situación y de la nueva situación. El modelo al azar para los nuevos datos es justamente $M_{Random}(New) = 1/nC$, donde nC es el número de clases en el problema[1]. Así pues, si sabemos la precisión obtenida por el modelo para los datos viejos, es decir $M_{Old}(Old)$ y el "mal" comportamiento para los nuevos datos, es decir $M_{Old}(New)$, entonces podemos estimar la proporción α de ejemplos que vienen de la vieja situación y de la nueva situación como sigue:

$$\alpha = \frac{M_{Old}(New) - M_{Random}(New)}{M_{Old}(Old) - M_{Random}(New)}$$

Tabla 3. Precisión promedio de los diferentes modelos.

#	$M_{Old}(New)$	M_{New}	M_{Comb}	$M_{Comb\alpha}$	$M_{Mim\alpha}$
1	89.23	90.78	93.22	89.44	96.32
2	61.31	53.39	61.75	61.31	64.36
3	75.86	72.43	76.67	76.18	78.01
4	68.75	67.70	68.59	69.13	67.01
5	84.60	79.64	82.70	84.49	84.05
6	71.46	68.97	70.65	71.72	69.40
7	76.00	76.55	78.17	78.28	76.84
8	93.38	81.67	91.76	93.41	93.37
9	57.39	46.27	58.81	58.15	58.76
10	66.20	59.84	70.56	66.26	83.26
Media	***74.42***	***69.72***	***75.29***	***74.84***	***77.14***

Tenemos resultados más interesantes cuando comparamos los modelos que toman en cuenta el antiguo modelo y los nuevos datos, tal como los modelos miméticos y los modelos combinados. En la Tabla 3 mostramos la precisión media obtenida por el modelo mimético usando el α estimado ($M_{Mim\alpha}$), el modelo combinado con iguales pesos (M_{Comb}) (se obtiene con $0.5 \cdot M_{Old} + 0.5 \cdot M_{New}$) y el modelo combinado con pesos derivados de α ($M_{Comb\alpha}$) (usamos $\alpha \cdot M_{Old} + (\alpha\text{-}1) \cdot M_{New}$ para obtener sus predicciones).

Como podemos ver en la Tabla 3, hay casos donde M_{Old} gana, pero generalmente M_{Comb}, y especialmente el modelo mimético $M_{mim\alpha}$, son mejores. El modelo mimético en términos de precisión es levemente mejor que la combinación. No obstante, es importante destacar que mediante la combinación no se produce un modelo comprensible, mientras que con la técnica mimética sí, por lo que existen muchas situaciones donde no será aplicable la técnica combinada.

Los resultados de las pruebas t de Student realizadas a un nivel de confianza del 5% en los 10 conjuntos de datos (ver Tabla 2) del repositorio UCI, muestran que en general, el modelo mimético se comporta mejor que los demás modelos. En [Blanco06] se encuentra una evaluación experimental más extensa.

4.2. Escenario con error

En la sección anterior hemos demostrado que el modelo mimético puede mantener parcialmente el buen funcionamiento en las regiones que no han cambiado, y puede integrar las reglas que cubren la nueva región.

Aunque un escenario en el que la revisión se activa por novedad puede ser habitual en muchas organizaciones, hay otro escenario que puede ser también frecuente y en el que el cambio sucede globalmente y no en una región particular. En este caso, parece más difícil ver cómo los datos del modelo antiguo y los datos del nuevo modelo pueden coexistir.

Por ejemplo, la Figura 7 muestra el modelo antiguo y un nuevo modelo con un cambio global.

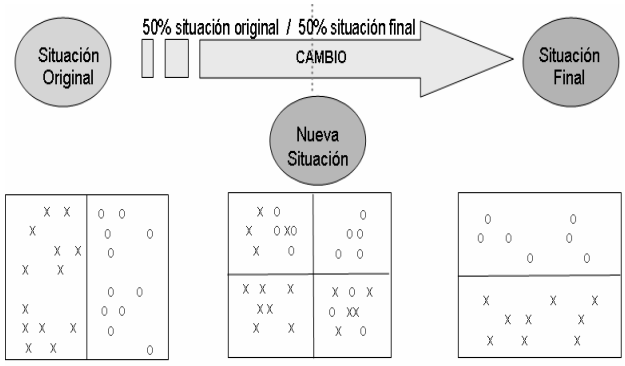

Figura 7. Cambio del modelo de forma global y profunda.

En estos casos, y tal y como reflejan los experimentos realizados, no se puede garantizar que, en general, el modelo mimético presente un comportamiento mejor que el modelo original. No obstante, hay algunos casos donde el modelo

[1] Esto es solamente cierto si en ambos casos las clases están equilibradas. En caso contario, el comportamiento sería peor, aunque si las clases están desequilibradas de forma semejante el comportamiento sería mejor.

mimético podría ser mejor que el antiguo modelo: por ejemplo, la situación reflejada en la Figura 8.

Estos casos dependen sobre todo de la técnica que se use para la técnica mimética y los tipos de patrones que capture. En nuestro caso, usamos árboles de decisión, como hemos dicho.

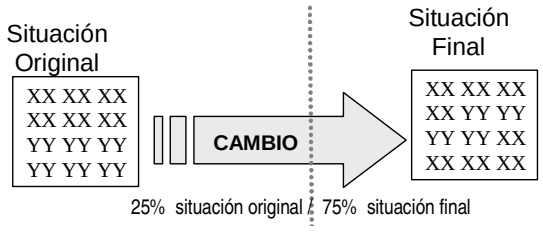

Figura 8. Cambio del modelo de forma global y profunda donde la técnica mimética puede trabajar bien.

5. Conclusiones

Hemos presentado dos aplicaciones de la técnica mimética donde no se requiere el uso de los datos originales.

En la primera parte del trabajo, damos contestación a las dos preguntas de la introducción ¿Existe un tamaño del conjunto de datos de entrenamiento que maximice precisión y comprensibilidad?, la respuesta es que sí existe un tamaño con el cual se puede estimar un coste mínimo en la mayoría de los conjuntos de datos experimentados pero con algunas excepciones. ¿Cómo se puede estimar ese tamaño óptimo? Se presentó un procedimiento para estimar el tamaño óptimo en modelos miméticos usando el principio MML y curvas de aprendizaje. Esto resuelve uno de los problemas del método mimético. Dado un parámetro *K* de compromiso entre los costes, a partir de sólo tres aprendizajes sobre el mismo oráculo con distintos tamaños de conjunto artificial podemos trazar la curva de aprendizaje completa y determinar el tamaño óptimo.

En la segunda parte del trabajo, hemos visto que el modelo mimético se comporta mejor que el modelo antiguo y un nuevo modelo en un escenario donde los cambios suceden en las regiones que no eran relevantes o inexistentes en la situación original. Esto es especialmente así cuando tenemos pocos datos nuevos disponibles para conseguir entrenar un buen modelo. Éste es el caso usual en la revisión de teorías: ya que en la situación opuesta tenemos un montón de nuevos datos, y entonces la revisión de teorías en general es inútil.

Un escenario diferente sucede cuando los cambios ocurren de una manera más global y los nuevos datos contradicen, en cierta manera, la situación original. Aquí, los casos donde es útil la técnica mimética no son tan generales como en el escenario anterior. En cualquier caso, el modelo mimético se puede construir muy fácilmente y se puede comparar con los antiguos y nuevos modelos para considerar si el modelo mimético es eventualmente mejor o no. Como trabajo futuro, quisiéramos investigar varias cuestiones. Primero estudiar si el método de optimización propuesto en la primera parte es válido para otros métodos de aprendizaje. Segundo, realizar un análisis del valor de proporcionalidad de los costes *K*, para saber su incidencia en la optimización. Tanto para la primera como para la segunda parte, quisiéramos analizar más profundamente la influencia de los cambios en la distribución de clases. También quisiéramos demostrar que la técnica también trabaja para la regresión. Finalmente, para la segunda parte, quisiéramos utilizar la técnica cuando la representación del problema cambia (nuevos atributos, nuevos valores, nuevas clases, etc.), un escenario para el cual nuestra técnica mimética de revisión de teorías podría ser especialmente beneficioso.

Resumiendo, este trabajo muestra la posibilidad de estimar el modelo mimético óptimo con tan solo tres modelos y ha demostrado que cualquier técnica de aprendizaje computacional se puede utilizar como método general de revisión de teorías, independientemente de cómo el modelo fue generado y sin los datos originales o su distribución. La aplicación de la técnica mimética es prometedora en el área de sistemas expertos y en revisión. En general la técnica mimética es ideal tanto para la adquisición de conocimiento como para su adaptación.

Agradecimientos

Este trabajo ha sido apoyado parcialmente por SEIT-ANUIES y la licencia beca comisión de la Dirección General de Educación Superior Tecnológica de México. Este trabajo también ha tenido el apoyo de la UE (FEDER) y el MEC, bajo ayuda TIN-2004-7943-C04-02, ICT para EU-India Cross-Cultural Dissemination Project bajo ayuda ALA/95/23/2003/077-054. También fue apoyado por los proyectos de la Generalitat Valenciana: META-MIDAS, MEDIM, GV06/301 y UPV bajo ayuda TAMAT.

Referencias

[Baxter96] Baxter, R. Minimum Message Length Inference: Theory and Applications, PhD thesis, Department of Computer Science, Monash University, Clayton, Australia. (1996).

[Black et al. 98] Black C. L.; Merz C. J. UCI repository of machine learning databases. (1998).

[Blanco06] Blanco-Vega, R. Extracción y contextualización de reglas comprensibles a partir de modelos de cajas negras. Tesis doctoral. Universidad Politécnica de Valencia. (2006).

[Blanco et al. 04] Blanco-Vega R., Hernández-Orallo J., Ramírez-Quintana Ma. J. Analysing the Trade-off between Comprehensibility and Accuracy in Mimetic Models. The 7th International Conference on Discovery Science, DS 2004, LNAI 3245, pp. 338-346. (2004).

[Boser et al. 1992] Boser, B.; Guyon, I.; Vapnik, V. "A Training Algorithm for Optimal Margin Classifiers" in Proceedings of the 5th Annual ACM W4orkshop on Computational Learning Theory, COLT, (1992).

[Breiman96] Breiman L. Bagging predictors. Machine Learning, 24(2):123-140. (1996).

[Cain91] Cain, T. The DUCTOR: A theory revision system for propositional domains. In Proceedings of the Eighth International Workshop on Machine Learning, 485-489. (1991).

[Domingos98] Domingos, P. Knowledge Discovery Via Multiple Models. Intelligent Data Analysis, 2(1-4): 187-202. (1998).

[Esposito et al. 02] Esposito, F.; Ferilli, S,: Fanizzi, N.; Altomare, T.M.; Di Mauro, N. Cooperation of Multiple Strategies for Automated Learning in Complex Environments, Proceedings of the 13th International Symposium on Methodologies for Intelligent Systems, LNAI 2366, pp: 574-582. (2002).

[Estruch et al. 03]Estruch, V.; Ferri, C.; Hernandez-Orallo, J.; Ramirez-Quintana, M.J. Simple Mimetic Classifiers, Proc. of the Third Int. Conf. on Machine Learning and Data Mining in Pattern Recognition, LNCS 2734, pp:156-171. (2003).

[Freund et al. 96] Freund Y. and Schapire R.E. Experiments with a new Boosting algorithm. In Proc. 13th International Conference on Machine Learning, pages 148-146. Morgan Kaufmann. (1996).

[Ginsberg90] Ginsberg, A. Theory reduction, theory revision, and retranslation. In Proceedings of the Eighth National Conference on Artificial Intelligence, 777-782. (1990).

[Matwin et al. 91]Matwin, S., and Plante, B. A deductive-inductive method for theory revision. In Proceedings of the International Workshop on Multistrategy Learning, 160-174. Harper's Ferry. (1991).

[McClelland et al. 1986] McClelland, J.; Rumelhart D. E.; The PDP Research Group. Parallel Distributed Processing, Volume 1 and 2. MIT Press, (1986).

[Oates et al. 97]Oates Tim; Jensen David The Effects of Training Set Size on Decision Tree Complexity. Proceedings of the Fourteenth International Conference on Machine Learning: 254 – 262. Morgan Kaufmann Publishers Inc. (1997).

[Ourston94] Ourston, D.; Raymond J. Theory refinement combining analytical and empirical methods. Volume 66, Issue 2, pp.273-309 Mooney Elsevier Science Publishers Ltd. (1994).

[Quinlan93] Quinlan, J. Ross. C4.5: Programs for machine learning, Morgan Kaufmann Publishers. (1993).

[Quinlan96] Quinlan, J. R. "Bagging, Boosting, and C4.5" in Proceedings of the 30th National Conference on Artificial Intelligence and the 8th Innovative Applications of Artificial Intelligence, pp 725-730, 1996.

[Taylor et al. 97]Taylor Charles and Nakhaeizadeh Gholamreza. Learning in Dynamically Changing Domains: Theory Revision and Context Dependence Issues Source. Lecture Notes In Computer Science; Vol. 1224 pp.353-360. Springer-Verlag. (1997).

[Utgoff94] Utgoff P. An Improved Algorithm for Incremental Induction of Decision Trees. In Proc. 11th ICML, pages 318-325. Morgan Kaufmann. (1994).

[Wallace et al. 68] Wallace, C.S., Boulton, D.M. An information measure for classification. Computer Journal 11 185-194. (1968).

[Witten et al. 05] Witten, Ian H. and Frank, Eibe. Data Mining: Practical machine learning tools with Java implementations. Second edition. Morgan Kaufmann, San Francisco. (2005).

[Yang et al. 99]Yang Jihoon, Parekh Rajesh, Honavar Vasant, Dobbs Drena. Data-Driven Theory Refinement Using KBDistAl Advances in Intelligent Data Analysis: Third International Symposium, IDA-99. (1999).