Estrategia efectiva para el aprendizaje activo multi-etiqueta

Oscar Reyes, Sebastián Ventura

Departamento de Informática y Análisis Numérico, Universidad de Córdoba, Campus de Rabanales, 14071 Córdoba, España.
ogreyesp@gmail.com,sventura@uco.es

Resumen El etiquetado de datos es un proceso costoso. Este costo aumenta considerablemente cuando los ejemplos deben ser etiquetados varias veces, lo cual ocurre en datos multi-etiqueta. Las técnicas de aprendizaje activo permiten construir modelos precisos mediante la selección iterativa de ejemplos no etiquetados, lo cual reduce los costos de etiquetado y de entrenamiento de los modelos. En este trabajo se presenta una nueva estrategia de aprendizaje activo multi-etiqueta. Se han definido y combinado dos medidas para la selección de los ejemplos no etiquetados en cada iteración. La estrategia propuesta fue comparada contra varios métodos del estado del arte y evaluada en 18 conjuntos de datos multi-etiqueta. Los resultados experimentales fueron validados mediante test estadísticos no paramétricos y se confirmó la efectividad de la estrategia propuesta para la resolución del problema de aprendizaje activo multi-etiqueta.

Keywords: Aprendizaje multi-etiqueta, aprendizaje activo, clasificación multi-etiqueta, ranking multi-etiqueta

1. Introducción

En la última década, los problemas que involucran datos que están asociados a un conjunto de etiquetas al mismo tiempo (problemas multi-etiqueta) han sido ampliamente estudiados por la comunidad de aprendizaje automático [7]. El aprendizaje multi-etiqueta se refiere a la construcción de un modelo capaz de predecir un conjunto de etiquetas para un ejemplo nunca antes visto. En el aprendizaje multi-etiqueta han sido estudiadas principalmente dos tareas [7]: la clasificación multi-etiqueta (MLC, Multi-label Classification) y el ranking de etiquetas (LR, Label Ranking). El objetivo de la tarea MLC es aprender un modelo capaz de, dado un ejemplo de prueba, retornar conjuntos de etiquetas relevantes y no relevantes. Por otra parte, la tarea de LR persigue, dado un ejemplo de prueba, retornar un ranking de etiquetas acorde a sus relevancias.
La mayoría de los algoritmos multi-etiqueta propuestos en la literatura están diseñados para trabajar en escenarios de aprendizaje supervisado, es decir escenarios donde todos los ejemplos de entrenamiento están etiquetados. Sin embargo, en algunos escenarios el etiquetado de datos es un proceso costoso que

requiere la intervención de expertos humanos. Este costo se incrementa considerablemente en los datos multi-etiqueta, pues cada ejemplo debe ser etiquetado varias veces. En consecuencia, hoy en día es común encontrar escenarios reales donde se dispone de un pequeño conjunto de datos etiquetados y un enorme conjunto de datos no etiquetados.

El aprendizaje semisuperisado y el aprendizaje activo (AL, Active Learning) son las dos áreas principales de estudio que tienen como objetivo el aprendizaje de modelos a partir de datos etiquetados y no etiquetados [10]. El AL persigue aprender clasificadores precisos mediante la selección iterativa de ejemplos no etiquetados. Por lo tanto, los costos del etiquetado de datos y del entrenamiento de un modelo preciso son reducidos considerablemente. En general, el ciclo de AL incluye los siguientes pasos: 1) una estrategia de consulta selecciona de forma iterativa los ejemplos no etiquetados más informativos para el modelo actual, 2) los ejemplos seleccionados son clasificados por un etiquetador (por ejemplo un experto humano), 3) los ejemplos son insertados en el conjunto de datos etiquetados, y 4) el modelo se reconstruye a partir del conjunto de datos etiquetados [10].

El desarrollo de métodos de AL para datos multi-etiqueta ha sido escaso en comparación con el considerable número de métodos de AL que han sido propuestos para datos mono-etiqueta. El principal desafío al cual se enfrentan los métodos de AL en el contexto multi-etiqueta es medir eficientemente el potencial informativo de un ejemplo no etiquetado teniendo en cuenta todas las posibles etiquetas. Entre los métodos más relevantes en el área de aprendizaje activo multi-etiqueta (MLAL, Multi-label Active Learning) se encuentran los propuestos en [9,2,11,5,8]. La mayoría de las propuestas existentes emplean el enfoque *one vs all*, conocido en la literatura como Relevancia Binaria [7], para descomponer un problema multi-etiqueta en varios problemas de clasificación binaria, lo cual supone un costo considerable en conjuntos de datos con un gran número de etiquetas. Por otra parte, comúnmente los métodos de MLAL han sido evaluados en la tarea de MLC. Sin embargo, el rendimiento de estos métodos en la tarea de LR no ha sido considerado.

En este trabajo se presenta una estrategia efectiva para MLAL. Se definen dos medidas de incertidumbre desde las perspectivas de la predicción del clasificador base y la inconsistencia del conjunto de etiquetas predicho. Se formuló un problema de agregación de ranking para calcular la incertidumbre unificada de un ejemplo no etiquetado. Por otra parte, la inconsistencia de un conjunto de etiquetas predicho se calcula mediante la distancia a los conjuntos de etiquetas existentes en los datos etiquetados.

Los experimentos se realizaron en 18 conjuntos de datos. El rendimiento de las estrategias de MLAL fue analizado mediante siete médidas de evaluación multi-etiqueta. El análisis de los resultados se llevó a cabo mediante test no paramétricos como ha sido propuesto en [3,6]. Los resultados obtenidos muestran la efectividad de la estrategia de MLAL propuesta en este trabajo.

El resto del artículo está organizado como sigue. La Sección 2 presenta los funda-

mentos de nuestra propuesta. El estudio experimental se muestra en la Sección 3. Finalmente, las conclusiones son expuestas en la Sección 4.

2. Estrategia de aprendizaje activo multi-etiqueta

Sea $\mathcal{F}$ un espacio de atributos y $\mathcal{L}$ un espacio de etiquetas con cardinalidad q (número de etiquetas). Un ejemplo multi-etiqueta es representado como una tupla $\langle \mathbf{X_i}, \mathbf{Y_i} \rangle$, donde $\mathbf{X_i}$ es el vector de atributos y $\mathbf{Y_i}$ es el vector categoría del ejemplo i. $\mathbf{Y_i}$ es un vector binario que contiene q componentes, donde el componente $Y_{i\ell}$ representa si el ejemplo i pertenece o no a la etiqueta ℓ. Por otra parte, en el problema de AL se dispone de un pequeño conjunto de datos etiquetados L_s y un enorme conjunto de datos no etiquetados U_s.

2.1. Medida de incertidumbre basada en agregación de ranking

Sea Φ un clasificador multi-etiqueta que, para un ejemplo de prueba, retorna probabilidades de pertenencia para cada una de las posibles etiquetas $\ell \in \mathcal{L}$. La probabilidad de que el ejemplo i pertenece a la etiqueta ℓ se denota como $P_\Phi(\ell{=}1|i)$, mientras que $P_\Phi(\ell{=}0|i)$ denota el caso contrario. El margen de diferencia en la predicción de Φ puede ser calculado de la siguiente manera:

$$m_\Phi^{i,\ell} = |P_\Phi(\ell{=}1|i) - P_\Phi(\ell{=}0|i)| \tag{1}$$

Un ejemplo con un largo margen en la etiqueta ℓ significa que existe poca duda en determinar si el ejemplo pertenece o no a dicha etiqueta. Por otra parte, un ejemplo con un margen pequeño en la etiqueta ℓ significa que es más ambiguo para el clasificador predecir si el ejemplo pertenece o no a dicha etiqueta. Dado un ejemplo i, se puede obtener un vector de valores de márgenes $\mathbf{M_\Phi^i} = \left\langle m_\Phi^{i,1}, m_\Phi^{i,2}, \ldots, m_\Phi^{i,q} \right\rangle$. El problema se resume en cómo agregar la información por cada etiqueta para calcular un valor de incertidumbre unificado. Para ello tendremos en cuenta los vectores de márgenes de cada uno de los ejemplos $i \in U_s$.

Dado los vectores de márgenes de cada ejemplo no etiquetado, q ranking de ejemplos son calculados $\tau_1, \tau_2, \ldots, \tau_q$; un ranking para cada etiqueta. Dada una etiqueta ℓ, el ranking de ejemplos no etiquetados es calculado como sigue:

$$\tau_\ell = (i_{\pi 1}, i_{\pi 2}, \ldots, i_{\pi |U_s|}) \mid m_\Phi^{i_{\pi 1},\ell} < m_\Phi^{i_{\pi 2},\ell} \ldots < m_\Phi^{i_{\pi |U_s|},\ell} \tag{2}$$

El ranking τ_ℓ es un ordenamiento de los ejemplos no etiquetados acorde a sus valores de márgenes en la etiqueta ℓ. Se desea determinar un ranking de ejemplos τ' que combine los ranking $\tau_1, \tau_2, \ldots, \tau_q$, de tal manera que los ejemplos posicionados en las primeras posiciones del ranking final τ' representen los ejemplos más inciertos para el clasificador.
El problema de agregación de ranking ha sido ampliamente estudiado en la literatura [4]. En este trabajo se propone usar el método de agregación de ranking más simple y antiguo, hasta donde sabemos, el método de Borda [1]. Este último

es un método posicional, se le asigna una puntuación a un elemento según las posiciones en la que aparece dicho elemento en los ranking. Basado en el método de Borda, la puntuación de un ejemplo i se calcula de la siguiente manera:

$$s(i)=\frac{\sum_{\ell\in\mathcal{L}}(|U_s|\text{-}\tau_\ell(i))}{q(|U_s|\text{-}1)} \tag{3}$$

donde $\tau_\ell(i)$ es la posición del ejemplo i en el ranking τ_ℓ, q es el número de etiquetas y $|U_s|$ denota el número de ejemplos no etiquetados. A mayor valor de $s(i)$, mayor incertidumbre del ejemplo i tomando en cuenta la información de todas las etiquetas.

2.2. Medida basada en la inconsistencia del vector categoría

Las técnicas de AL toman como premisa que los conjuntos L_s y U_s son generados a partir de la misma distribución, por lo tanto es de esperar que los conjuntos de etiquetas predichos por el clasificador compartan propiedades comunes con los conjuntos de etiquetas presentes en L_s. El Cuadro 1 muestra una matriz de contingencia dado dos vectores categoría $\mathbf{Y_i}$ y $\mathbf{Y_j}$ de los ejemplos i y j, respectivamente. El número de componentes en los cuales $Y_{i\ell}$=$Y_{j\ell}$=1 se denota como a. Los otros casos posibles que pueden ocurrir entre los elementos de los vectores categoría se denotan como b, c y d.

$\mathbf{Y_i} \diagdown \mathbf{Y_j}$	1	0
1	a	b
0	c	d

Cuadro 1: Tabla de contingencia dado dos vectores categoría.

Dado dos vectores categoría $\mathbf{Y_i}$ and $\mathbf{Y_j}$, la distancia de Hamming normalizada es calculada como sigue:

$$d_H(\mathbf{Y_i},\mathbf{Y_j})=\frac{b+c}{q} \tag{4}$$

La distancia de Hamming representa la cantidad de casos en los cuales dos ejemplos difieren en su clasificación por etiquetas. Sin embargo, se desea además medir la diferencia que existe en la estructura de los vectores categoría. Los conjuntos de etiquetas más frecuentes en datos multi-etiqueta forman estructuras (combinaciones de ceros y unos), y dichas estructuras pueden ser comúnmente encontradas en los vectores categoría de los ejemplos etiquetados.

La distancia entrópica normalizada entre dos vectores categoría $\mathbf{Y_i}$ y $\mathbf{Y_j}$ es calculada como sigue:

$$d_E(\mathbf{Y_i},\mathbf{Y_j})=\frac{2H(\mathbf{Y_i},\mathbf{Y_j})-H(\mathbf{Y_i})-H(\mathbf{Y_j})}{H(\mathbf{Y_i},\mathbf{Y_j})} \tag{5}$$

$$H(\mathbf{Y_i},\mathbf{Y_j})=H_4(\frac{a}{q},\frac{b}{q},\frac{c}{q},\frac{d}{q})$$

$$H_4(\frac{a}{q},\frac{b}{q},\frac{c}{q},\frac{d}{q})=H_2(\frac{b+c}{q},\frac{a+d}{q})+\frac{b+c}{q}H_2(\frac{b}{b+c},\frac{c}{b+c})+\frac{a+d}{q}H_2(\frac{a}{a+d},\frac{d}{a+d})$$

$$H(\mathbf{Y})=H_2(\frac{w}{q},\frac{s}{q})=-\frac{w}{q}log_2(\frac{w}{q})-\frac{s}{q}log_2(\frac{s}{q})$$

donde w y s son la cantidad de unos (etiquetas positivas) y ceros (etiquetas negativas), respectivamente, en el vector categoría $\mathbf{Y}$. Basado en las funciones de distancia d_H y d_E, la inconsistencia de un conjunto de etiquetas predicho para un ejemplo no etiquetado i se calcula como sigue:

$$v(i)=\frac{1}{|L_s|}\sum_{j\in L_s} f_u(\mathbf{Y_i},\mathbf{Y_j}) \tag{6}$$

$$f_u(\mathbf{Y_i},\mathbf{Y_j})=\begin{cases} d_E(\mathbf{Y_i},\mathbf{Y_j}) & d_H(\mathbf{Y_i},\mathbf{Y_j})<1 \\ 1 & d_H(\mathbf{Y_i},\mathbf{Y_j})=1 \end{cases}$$

2.3. Estrategia de aprendizaje activo

Basado en las dos medidas definidas en este trabajo, un ejemplo no etiquetado es seleccionado de la siguiente manera:

$$\underset{i\in U_s}{\operatorname{argmax}}\ s(i)\cdot v(i) \tag{7}$$

Llamamos a esta estrategia Inconsistencia del Vector Categoría y Ranking de Puntuaciones (CVIRS, Category Vector Inconsistency and Ranking of Scores). Esta estrategia selecciona el ejemplo no etiquetado más incierto para el clasificador actual y que tiene el vector categoría predicho menos similar a los vectores categoría presentes en el conjunto etiquetado L_s. La estrategia propuesta puede ser usada con cualquier clasificador base a partir del cual se puedan obtener estimaciones de probabilidades desde sus salidas. Esta estrategia manipula directamente los datos multi-etiqueta, no está restringida a usar un método de transformación de problemas.

3. Experimentación

3.1. Configuración experimental

En el estudio experimental, la estrategia propuesta -CVIRS- fue comparada con las siguientes estrategias del estado del arte: BinMin [2], ML [9], MML [9], MMC [11], CMN [5], MMU [8] y LCI [8]. Además, en el estudio comparativo se incluyó como línea base una estrategia que selecciona de forma aleatoria los ejemplos no etiquetados (denotada como Random). Se espera que las estrategias comparadas superen a la estrategia Random.

En aras de la equidad, todas las estrategias fueron ejecutadas con el clasificador base BR-SVM, es decir por cada etiqueta se entrena un clasificador binario SVM. Se empleó BR-SVM pues la mayoría de las estrategias de AL utilizadas en el estudio comparativo están restringidas al uso de este clasificador base.
Se empleó el método validación cruzada en 10 particiones y se promediaron los resultados. En cada partición de la validación cruzada se empleó el protocolo experimental descrito en el Algoritmo 1. El 5 % del conjunto de entrenamiento T_r se seleccionó aleatoriamente para construir el conjunto etiquetado L_s. El 95 % restante del conjunto de entrenamiento es considerado el conjunto no etiquetado U_s, para ello se ocultan los conjuntos de etiquetas de los ejemplos. El número máximo de iteraciones de AL fue 750. En cada iteración el clasificador base fue evaluado con el conjunto de prueba T_s. El proceso de etiquetado se realizó de manera simulada, es decir el conjunto de etiquetas oculto de un ejemplo $i \in U_s$ es revelado.

Algoritmo 1: Protocolo experimental.

Entrada: $T_r \rightarrow$ conjunto de entrenamiento, $T_s \rightarrow$ conjunto de prueba, $\gamma \rightarrow$ estrategia de AL, $\theta \rightarrow$ etiquetador, $\beta \rightarrow$ número de iteraciones

Inicio
  //Construir el conjunto etiquetado y no etiquetado a partir de T_r
  $L_s \leftarrow$ `Resample` (5 %, T_r);
  $U_s \leftarrow T_r \setminus L_s$;
  Para *iter* $\leftarrow$ 1 **hasta** β
    //Entrenar clasificador Φ con L_s
    $\Phi \leftarrow$ `Train` (L_s,Φ);
    //Evaluar clasificador Φ con T_s
    `Test` (T_s, Φ);
    //Seleccionar ejemplo de U_s
    $i \leftarrow$ `SelectInformativeInstance` (γ, Φ, U_s);
    //Etiquetar ejemplo seleccionado
    `Label` (θ,i);
    //Actualizar los conjuntos etiquetado y no etiquetado
    $L_s \leftarrow L_s \cup \{i\}$;
    $U_s \leftarrow U_s \setminus \{i\}$;
  fin
fin

En este trabajo se utilizaron varias medidas para evaluar el rendimiento de los modelos multi-etiqueta inducidos. En cuanto a la tarea de MLC se emplearon las medidas *Micro-Average* F_1*-Measure* ($M_{iF_1} \uparrow$), *Macro-Average* F_1*-Measure* ($M_{aF_1} \uparrow$), *Hamming Loss* ($H_L \downarrow$) y *Example-based* F_1*-Measure* ($F_{1Ex} \uparrow$). Respecto a la tarea de LR se emplearon las medidas *Ranking Loss* ($R_L \downarrow$), *Average Precision* ($A_P \uparrow$) y *One Error* ($O_E \downarrow$). Los símbolos "↑" y "↓" representan que son medidas de máximo y mínimo, respectivamente. La definición formal e interpretación de todas estas medidas puede ser consultada en [7].
Las estrategias de AL generalmente son evaluadas mediante la comparación visual de curvas de aprendizaje [10]. Sin embargo, cuando se compara un considerable número de estrategias, y además algunas de ellas tienen rendimientos

similares, la comparación visual de curvas se torna una tarea confusa. En este trabajo se evaluó el rendimiento de las estrategias de AL mediante la comparación del área debajo de la curva (ALC, Area under Learning Curve), lo cual permitió llevar a cabo un análisis estadístico de los resultados. El test de Friedman se empleó para determinar si existían diferencias significativas en los resultados. Si el test de Friedman detectó diferencias significativas, entonces procedimos a realizar el test *post-hoc* de Shaffer para realizar comparaciones múltiples de todos contra todos, como fue propuesto en [6].
Las estrategias de AL fueron evaluadas en 18 conjuntos de datos multi-etiqueta. La descripción de estos conjuntos de datos multi-etiqueta, así como estadísticas de los mismos, puede ser consultada en `http://-mulan.sourceforge.net/datasets-mlc.html`.

3.2. Resultados y discusión

En esta sección solo se presenta un resumen de los resultados obtenidos para cada medida de evaluación considerada. Los resultados completos pueden ser consultados en `http://www.uco.es/grupos/kdis/kdiswiki/MLAL`.

Se calcularon los valores de ALC en cada conjunto de datos para llevar a cabo una comparación estadística entre las estrategias de MLAL consideradas. Las Tablas 2 y 3 muestran los valores de ALC para las medidas M_{iF_1} y M_{aF_1}, las tablas restantes pueden ser consultadas en la página Web disponible. Los mejores valores de ALC son resaltados en negrita. La última fila de las tablas muestra el ranking promedio (Rank. Pro.) calculado por el test de Friedman.

Dataset	**Multi-label AL strategy**								
	Random	**BinMin**	**ML**	**MML**	**MMC**	**CMN**	**MMU**	**LCI**	**CVIRS**
Flags	0.541	0.691	0.668	0.671	0.671	0.683	0.688	0.681	**0.692**
Emotions	0.616	0.621	0.640	0.643	0.644	0.658	0.601	0.607	**0.659**
Birds	0.265	0.333	0.384	0.385	0.387	0.412	0.326	0.396	**0.415**
Genbase	0.945	0.949	0.952	0.946	0.923	0.956	0.921	0.940	**0.963**
Cal500	0.330	0.336	0.331	0.330	0.332	0.332	0.329	0.328	**0.346**
Medical	0.648	0.648	0.570	0.556	0.609	0.665	0.665	0.665	**0.667**
Yeast	0.575	0.630	0.618	0.608	0.616	0.640	0.780	**0.784**	0.658
Scene	0.630	0.634	0.618	0.608	0.616	0.640	0.642	0.630	**0.643**
Enron	0.420	0.436	0.372	0.378	0.384	0.457	0.447	0.450	**0.464**
Corel5k	0.101	**0.168**	0.126	0.128	0.120	0.158	0.154	0.157	0.160
Corel16k	0.099	**0.161**	0.145	0.146	0.149	0.152	0.155	0.154	0.158
TMC2007-500	0.598	0.608	0.589	0.584	0.584	0.608	0.597	0.600	**0.620**
Bibtex	0.203	0.299	0.274	0.286	0.289	0.312	0.298	0.314	**0.321**
Arts	0.200	**0.266**	0.260	0.262	0.259	0.265	0.249	0.260	0.264
Business	0.305	0.366	**0.476**	0.391	0.375	0.387	0.411	0.422	0.436
Entertainment	0.259	0.343	0.323	0.304	0.298	0.332	0.334	0.333	**0.350**
Recreation	0.199	0.268	0.265	0.264	0.258	0.268	0.261	0.255	**0.273**
Health	0.301	0.359	0.347	0.332	0.315	0.347	0.357	0.341	**0.371**
Rank. Pro.	7.806	3.583	6.000	6.528	6.639	3.278	5.056	4.722	1.389

Cuadro 2: Valores de ALC para la medida M_{iF_1} (↑). El test de Friedman rechaza la hipótesis nula con un p-valor igual a 6.121E-11 considerando un nivel de significación α=0.05.

Dataset	Multi-label AL strategy								
	Random	**BinMin**	**ML**	**MML**	**MMC**	**CMN**	**MMU**	**LCI**	**CVIRS**
Flags	0.569	0.583	0.572	0.576	0.562	**0.592**	0.575	0.567	0.588
Emotions	0.517	0.520	0.608	0.636	0.636	0.642	0.495	0.498	**0.654**
Birds	0.304	0.255	0.309	0.310	0.311	**0.332**	0.239	0.311	0.330
Genbase	0.751	0.785	**0.806**	0.753	0.699	0.794	0.735	0.788	0.785
Cal500	0.161	0.156	0.154	0.154	0.151	0.162	0.156	0.146	**0.170**
Medical	0.352	0.348	0.310	0.312	0.317	0.376	0.370	0.369	**0.383**
Yeast	0.385	0.413	**0.416**	0.408	0.396	0.393	0.398	0.396	0.400
Scene	0.645	0.640	0.624	0.612	0.628	0.650	0.647	0.634	**0.651**
Enron	0.152	0.173	0.147	0.152	0.154	0.171	0.170	0.166	**0.185**
Corel5k	0.274	0.315	0.303	0.310	0.300	**0.321**	0.300	0.309	0.314
Corel16k	0.033	0.059	0.048	0.054	0.051	0.062	0.060	0.061	**0.065**
TMC2007-500	0.485	0.497	0.479	0.473	0.467	0.500	0.476	0.487	**0.521**
Bibtex	0.111	0.145	0.149	0.154	0.152	0.152	0.150	0.151	**0.156**
Arts	0.132	**0.171**	0.147	0.148	0.147	0.167	0.155	0.159	0.170
Business	0.135	0.158	0.159	0.161	0.158	0.158	0.148	0.149	**0.170**
Entertainment	0.154	0.200	0.191	0.195	0.187	0.197	0.190	0.194	**0.201**
Recreation	0.142	0.209	0.207	0.205	0.204	0.198	0.197	0.190	**0.218**
Health	0.123	0.188	0.171	0.169	0.155	0.174	0.188	0.185	**0.194**
Rank. Pro.	7.528	3.972	5.778	5.306	6.667	2.972	5.750	5.389	1.639

Cuadro 3: Valores de ALC para la medida $M_{aF_1}(\uparrow)$. El test de Friedman rechaza la hipótesis nula con un p-valor igual a 1.029E-10 considerando un nivel de significación α=0.05.

En general, la estrategia CVIRS obtuvo un buen rendimiento en los 18 conjuntos de datos y las siete medidas de evaluación consideradas. El test de Friedman rechazó la hipótesis nula en los siete casos (un caso por cada medida de evaluación) con un nivel de significación α=0.05. El test de Shaffer se ejecutó con el objetivo de realizar comparaciones múltiples de todos contra todos. En el análisis se tomaron en cuenta los p-valores ajustados (APV, Adjusted p-values), como fue propuesto en [6].

Los resultados del test de Shaffer se muestran en la Tabla 4. En cada celda se muestran las medidas en las que la estrategia ubicada en la fila obtiene mejores resultados que la estrategia ubicada en la columna. Para rechazar las hipótesis nulas se consideró un nivel de significación α=0.05. Los APV son indicados entre paréntesis. En caso que una estrategia no supere a otra estrategia en ninguna de las medidas de evaluación consideradas, aparece en la celda correspondiente el símbolo "-".

Las evidencias muestran que la estrategia propuesta -CVIRS- tuvo un buen rendimiento en las dos tareas, MLC y LR. CVIRS supera significativamente a las otras estrategias consideradas en varias medidas de evaluación. Tomando en cuenta el ranking promedio calculado por el test de Friedman por cada medida de evaluación, se puede concluir que las estrategias que mejores resultados obtuvieron fueron CVIRS, CMN y BinMin. CVIRS supera significativamente en la tarea MLC a la mayoría de las estrategias consideradas. El test de Shaffer no detectó diferencias significativas entre CVIRS, CMN y BinMin en la tarea de MLC. Sin embargo, CVIRS tuvo un rendimiento significativamente mejor que todas las estrategias en la tarea LR. Es importante destacar que ninguna de las otras estrategias consideradas en la comparación superan significativamente a CVIRS. Por otro lado, resulta interesante que, bajo las condiciones en las

vs	Random	BinMin	ML	MML	MMC	CMN	MMU	LCI
BinMin	$M_{iF_1}(0.0)$ $M_{aF_1}(0.0)$ $F_{1Ex}(0.0)$ $H_L(0.0)$ $R_L(0.04)$ $A_P(0.0)$	-	-	$M_{iF_1}(0.03)$ $F_{1Ex}(0.03)$	$M_{iF_1}(0.02)$ $F_{1Ex}(0.01)$	-	-	-
CMN	$M_{iF_1}(0.0)$ $M_{aF_1}(0.0)$ $F_{1Ex}(0.0)$ $H_L(0.0)$ $A_P(0.02)$	-	$M_{aF_1}(0.04)$	$M_{iF_1}(0.01)$ $F_{1Ex}(0.02)$	$M_{iF_1}(0.01)$ $M_{aF_1}(0.0)$ $F_{1Ex}(0.01)$	-	-	-
LCI	$O_E(0.04)$	-	-	-	-	-	-	-
CVIRS	$M_{iF_1}(0.0)$ $M_{aF_1}(0.0)$ $F_{1Ex}(0.0)$ $H_L(0.0)$ $R_L(0.0)$ $A_P(0.0)$ $O_E(0.0)$	$R_L(0.03)$ $O_E(0.04)$	$M_{iF_1}(0.0)$ $M_{aF_1}(0.0)$ $F_{1Ex}(0.0)$ $H_L(0.01)$ $R_L(0.0)$ $A_P(0.0)$ $O_E(0.0)$	$M_{iF_1}(0.0)$ $M_{aF_1}(0.0)$ $F_{1Ex}(0.0)$ $H_L(0.0)$ $R_L(0.0)$ $A_P(0.0)$ $O_E(0.0)$	$M_{iF_1}(0.0)$ $M_{aF_1}(0.0)$ $F_{1Ex}(0.0)$ $H_L(0.0)$ $R_L(0.0)$ $A_P(0.0)$ $O_E(0.0)$	$R_L(0.0)$	$M_{iF_1}(0.0)$ $M_{aF_1}(0.0)$ $F_{1Ex}(0.0)$ $H_L(0.0)$ $R_L(0.01)$ $A_P(0.02)$ $O_E(0.01)$	$M_{iF_1}(0.01)$ $M_{aF_1}(0.0)$ $F_{1Ex}(0.05)$ $H_L(0.01)$ $R_L(0.01)$ $A_P(0.05)$

Cuadro 4: Comparación múltiple de todos contra todos mediante el test de Shaffer.

cuales se llevó a cabo el estudio empírico, hubo estrategias que no superaron significativamente a la estrategia base Random en ninguna de las medidas de evaluación consideradas, por ejemplo las estrategias ML, MML, MMC y MMU. Las estrategias que peores resultados obtuvieron fueron ML, MML y MMC.

En general, CVIRS tuvo un buen rendimiento en conjunto de datos con diferentes características. Sin embargo, se evidencia una mayor efectividad en conjuntos de datos con pocas etiquetas, por ejemplo en Emotions, Birds y Yeast. El rendimiento de CVIRS puede verse afectado en conjuntos de datos con un alto número de etiquetas debido al método usado actualmente para resolver el problema de agregación de ranking en el cálculo de la incertidumbre unificada.

4. Conclusiones

En este trabajo se ha presentado una estrategia de MLAL, llamada CVIRS. La estrategia CVIRS combina dos medidas para la selección de los ejemplos no etiquetados, y puede usar cualquier clasificador base siempre y cuando se puedan obtener estimaciones de probabilidades desde sus salidas. Para comprobar el rendimiento de CVIRS se consideraron 18 conjuntos de datos multi-etiquetas, y se comparó contra siete estrategias de MLAL del estado del arte. Los resultados demuestran que CVIRS funciona bastante bien en conjunto de datos con diversas características, obtiene buenos resultados en las tareas de MLC y LR, y además es competitivo respecto a las estrategias de MLAL del estado del arte. Como trabajo futuro, sería interesante el estudio de otros métodos para la resolución del problema de agregación de ranking formulado en el cálculo de la incertidumbre unificada de un ejemplo no etiquetado.

Agradecimientos

El presente trabajo ha sido financiado por el Ministerio de Economía y Competitividad de España, proyecto TIN-2014-55252-P, y los fondos FEDER.

Referencias

1. Borda, J.C.: Memoire sur les election au scrutin, Histoire de la ácademie Royale des Sciences, Paris, France (1781)
2. Brinker, K.: From Data and Information Analysis to Knowledge Engineering, chap. On Active Learning in Multi-label Classification, pp. 206–213. Springer (2006)
3. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)
4. Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: Rank Aggregation Methods for the Web. In: Proceedings of the 10th World Wide Web Conference. pp. 613–622. ACM (2001)
5. Esuli, A., Sebastiani, F.: Active Learning Strategies for Multi-Label Text Classification. In: Advances in Information Retrieval. pp. 102–113. Springer (2009)
6. García, S., Herrera, F.: An extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all pairwise comparisons. J. Mach. Learn. Res. 9, 2677–2694 (2008)
7. Gibaja, E., Ventura, S.: Multi-label learning: a review of the state of the art and ongoing research. WIREs Data Mining Knowl. Discov. 4, 411–444 (2014)
8. Li, X., Guo, Y.: Active Learning with Multi-Label SVM Classification. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. pp. 1479–1485. AAAI Press (2013)
9. Li, X., Wang, L., Sung, E.: Multi-label SVM active learning for image classification. In: Proceedings of the International Conference on Image processing (ICIP'04). vol. 4, pp. 2207–2210. IEEE (2004)
10. Settles, B.: Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 1 edn. (2012)
11. Yang, B., Sun, J., Wang, T., Chen, Z.: Effective multi-label active learning for text classification. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 917–926. ACM, Paris, France (2009)