Clasificación de imágenes basada en subconjuntos de subgrafos frecuentes aproximados

Por

Niusvel Acosta Mendoza

Tesis sometida como requisito parcial para obtener el grado de

MAESTRO EN CIENCIAS EN LA ESPECIALIDAD DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica
Tonantzintla, Puebla

Julio 2013

Supervisada por:

Dr. Jesús Ariel Carrasco Ochoa
Investigador titular del INAOE

Dr. José Francisco Martínez Trinidad
Investigador titular del INAOE

Dr. Andrés Gago Alonso
Investigador agregado del CENATAV

Dr. José E. Medina Pagola
Investigador titular del CENATAV

Resumen

En los últimos años se ha podido observar un incremento en el uso de la minería de subgrafos frecuentes aproximados (MSFA). Este tipo de minería, poco a poco, se ha convertido en una importante línea de investigación con un amplio espectro de aplicaciones en varios dominios de la ciencia. Ejemplos de tales aplicaciones son el procesamiento y análisis de bases de datos de imágenes, de componentes químicas, redes de citas, redes biológicas, etc. Para resolver el problema de la MSFA se han propuesto varios algoritmos con diferentes funciones de similitud entre grafos. De este grupo de algoritmos para la MSFA, pocos se han aplicado en la clasificación de imágenes, los cuales han reportado resultados superiores a los alcanzados utilizando minería de subgrafos frecuentes basado en isomorfismo (algoritmos exactos), en esta tarea.

En esta tesis se realiza un estudio de la MSFA, específicamente de los métodos que tratan las aproximaciones entre etiquetas manteniendo la topología de los grafos. Esto se debe a que este tipo de algoritmos son los que se han aplicado en clasificación de imágenes, tarea en la que se basa esta tesis. En dicho estudio, se hace un resumen de los resultados obtenidos al utilizar los subgrafos frecuentes aproximados como atributos para la clasificación de las imágenes, donde se han reportado buenos resultados. Sin embargo, la dimensionalidad de los vectores de atributos crece mientras más pequeño sea el umbral de soporte utilizado en la MSFA, por lo que se reduce la aplicabilidad de estas técnicas

en bases de datos con mayor número de instancias y/o grafos de mayor tamaño.

Esta tesis se enfoca en reducir la dimensionalidad de los atributos utilizando un subconjunto de subgrafos frecuentes aproximados (patrones) para la clasificación de imágenes. Con esta estrategia, de acuerdo con los resultados obtenidos, se logra mejorar la eficiencia y eficacia del proceso de clasificación.

Abstract

In recent years, there has been a significant increase in the use of frequent approximate subgraph mining. This kind of mining has become an important research line and is currently applied to several domains. Examples of such applications are the processing and analysis of image databases, chemical components, citation networks, biological networks, etc. Several algorithms have been proposed to compute the frequent approximate subgraphs for different similarity functions among graphs. Only a few of these algorithms have been applied to image classification, reporting better results than using frequent subgraphs based on isomorphism.

In this thesis, a study of frequent approximate subgraph mining is carried out. This mining technique has tackled the image classification problem successfully; specifically methods addressing label's approximations while maintaining graphs topology. In this study, significant results obtained by using frequent approximate subgraphs as features for image classification are presented.

This thesis focuses on reducing attributes dimensionality by using a subset of frequent approximate subgraphs (patterns) suitable for image classification. Following this idea, according to the experimental results, it is possible to improve the efficiency and effectiveness of the classification process.

A mi esposa

A mi hija

A mi mamá

A mis abuelos

A mi papá

A Baldiovino

A mis tíos

A mi hermana

Agradecimientos

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACyT), al Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), y también al Centro de Aplicaciones de Tecnología de Avanzada (CENATAV) por el apoyo proporcionado para y durante la realización de este trabajo de tesis. En particular, al INAOE por permitirme desarrollar en sus instalaciones este trabajo de investigación, y al CENATAV por haber confiado en mí.

Agradezco a mis guías durante esta investigación: Dr. Jesús Ariel Carrasco Ochoa (Ariel), Dr. José Francisco Martínez Trinidad (Pancho), Dr. Andrés Gago Alonso (Gago), y Dr. José Eladio Medina Pagola (Medina), cuya asesoría fue indispensable para el desarrollo de esta tesis. A Gago y Medina les agradezco de corazón la paciencia y dedicación para conmigo, y por confiar en mí cuando otros no creían. A los Drs. Ariel y Pancho les agradezco por haberme apoyado en todo durante mis estancias en México. Y a todos ellos por extenderme sus manos y apoyarme incondicionalmente en mis dificultades (de cualquier índole).

Le expreso mi gratitud al Dr. Jose Ruiz Shulcloper por haber sido el principal promotor de la formación de nuevos doctores en el CENATAV, su exigencia y optimismo me sirvieron como fuente de inspiración. Agradezco también a los demás compañeros del CENATAV y del INAOE ya que muchos pusieron su granito de arena para la terminación

de este trabajo.

Agradezco a mis padres, a Baldiovino, y abuelos por haberme inculcado el espíritu de estudio, por haberme criado en un ambiente de personas útiles y preparadas, y por ser mis guías en la vida. Agradezco a mi esposa y a mi niñita por el esfuerzo que han tenido que hacer para mantenerme junto a ellas aún no estando a su lado. A cada uno de los miembros de mi familia por su apoyo y los sacrificios que han hecho para que yo pudiera cumplir una de las metas de mi vida profesional.

Incluyo además, mi gratitud a los tantos amigos que compartieron conmigo durante el desarrollo de esta investigación. A mis hermanos: Yosvel, Carly, Yaciel, Ronny, Jorge, y Livan, que aunque no estuvieron a mi lado físicamente siempre me estuvieron apoyando en la distancia. A Voro por enseñarme que las mejores Coca Colas se venden en los OXXOs y aguantarme como compañero de apartamento tanto tiempo. A Miguel Ángel (Mike), Cruz, y Silver por aceptarme en su círculo de amigos de verdad, por considerarme un mexicano más y por enseñarme a verlos como mis hermanos (como diríamos: *–Dios nos hace y nosotros nos juntamos*). A Laura, por ayudarme a completar la tesis proporcionándome su base de datos y su ayuda en la preparación de la misma para mis experimentos.

Agradezco, en especial, a aquellos que me apoyaron en las buenas y en las malas, y que sentirán la ausencia de sus nombres al leer esta sección del documento, pero los llevo en mi corazón.

Quiero agradecer a mis sinodales: Dra. Alicia Morales Reyes, Dr. Miguel Octavio Árias Estrada, y Dr. Manuel Monte y Gómez por su tiempo, observaciones y sugerencias realizadas durante el proceso de revisión de este trabajo.

Finalmente, agradezco a todos los que NO CONFIARON en mi, ya que esos pensamientos me dieron fuerzas en mi trabajo, hicieron que mis logros fueran aún mayores,

me mostraron ante el mundo, y aumentaron mis satisfacciones cuando lograba mis metas paso a paso.

Índice general

1. Introducción **1**
1.1. Problemática actual . . . 3
1.2. Motivación . . . 4
1.3. Objetivo general . . . 5
1.4. Objetivos específicos . . . 5
1.5. Organización de esta tesis . . . 6

2. Marco teórico **7**
2.1. Conceptos básicos usados de la teoría de grafos . . . 7
2.2. Conceptos básicos de aprendizaje automático . . . 13
2.3. Síntesis y conclusiones . . . 16

3. Trabajo relacionado **18**
3.1. Algoritmo VEAM para la MSFA . . . 20
3.2. Síntesis y conclusiones . . . 24

4. Método de clasificación propuesto **26**
4.1. Módulo de representación . . . 27
4.2. Módulo de extracción de patrones . . . 29
4.3. Módulo de reducción de patrones . . . 30
4.4. Módulo de clasificación . . . 34
4.5. Síntesis y conclusiones . . . 34

5. Resultados experimentales **35**
5.1. Bases de datos sintéticas . . . 35
5.1.1. Resultados experimentales . . . 37
5.2. Base de datos de esqueletos estructurales . . . 56
5.2.1. Resultados experimentales . . . 59
5.3. Síntesis y conclusiones . . . 62

6. Conclusiones, aportaciones y trabajo futuro **64**
6.1. Conclusiones . . . 64

6.2. Aportaciones del trabajo de investigación 65
6.3. Trabajo futuro 66

Anexos **67**
Notaciones 67
Acrónimos 69

Referencias **70**

Índice de figuras

2.1. Grafo etiquetado. 8
2.2. Ejemplo de subgrafo y supergrafo. 9
2.3. Ejemplo de semejanza entre dos grafos. 10
2.4. Ejemplo de semejanza entre dos grafos. 11
2.5. Ejemplo de patrones emergentes y contrastantes en un universo de objetos U compuesto por tres clases ($C1$, $C2$ y $C3$). 12

4.1. Método de clasificación de imágenes basada en grafos propuesto en esta tesis. 27
4.2. Ejemplo de un árbol de cuadrantes dada una imagen. 28
4.3. Ejemplo de un grafo construido a partir del quad-tree de la imagen de la figura 4.2. 29
4.4. Flujo del módulo de reducción de patrones. 33

5.1. Ejemplo de imágenes de la colección obtenido de (Acosta-Mendoza *et al.*, 2012a) usando el generador aleatorio de imágenes de Coenen. 36
5.2. Ejemplo de imágenes de la base de datos SIS. 57
5.3. Ejemplo de imágenes y sus esqueletos de la base de datos SIS. 58

Índice de tablas

5.1. Características de las bases de datos sintéticas. 37
5.2. Número de patrones utilizados como atributos en el proceso de clasificación. 40
5.3. Resultados de la clasificación, en términos del porciento de aciertos (*accuracy*), alcanzados utilizando diferentes clasificadores en varias colecciones de imágenes con y sin el uso de los patrones emergentes calculados con γ =0.3, 0.4 y 0.5. 41
5.4. Resultados de la clasificación (F-measure) alcanzados utilizando diferentes clasificadores en varias colecciones de imágenes con y sin el uso de los patrones emergentes calculados con γ =0.3, 0.4 y 0.5. 42
5.5. Pruebas de significancia estadística para los diferentes clasificadores en varias colecciones de imágenes con y sin el uso de los patrones emergentes. Cada celda de la tabla indica cuál opción fue estadísticamente mejor (entre los comparados en la columna 1), "–" indica que no hubo diferencia estadísticamente significativa. 44
5.6. Número de patrones utilizados como atributos en el proceso de clasificación. 47
5.7. Resultados de la clasificación (accuracy) alcanzados utilizando diferentes clasificadores en varias colecciones de imágenes con y sin el uso del algoritmos de selección de atributos. 48
5.8. Resultados de la clasificación (F-measure) alcanzados utilizando diferentes clasificadores en varias colecciones de imágenes con y sin el uso del algoritmos de selección de atributos. 49
5.9. Pruebas de significancia estadística para los diferentes clasificadores en varias colecciones de imágenes con y sin el uso de los algoritmos de selección de atributos. Cada celda de la tabla indica cuál opción fue estadísticamente mejor (entre los comparados en la columna 1), "-" indica que no hubo diferencia estadísticamente significativa. 50
5.10. Número de patrones utilizados como atributos en el proceso de clasificación. 53
5.11. Resultados de la clasificación alcanzados utilizando diferentes clasificadores en varias colecciones de imágenes realizando la reducción de patrones con E(0.4) y GRAE para la selección de atributos. 55

5.12. Pruebas de significancia estadística para los diferentes clasificadores en varias colecciones de imágenes utilizando los patrones emergentes E(0.4) y utilizando el algoritmo de selección de atributos GRAE. Cada celda de la tabla indica cuál opción fue estadísticamente mejor (para la prueba mostrada en la columna 1), "-" indica que no hubo diferencia estadísticamente significativa. 56
5.13. Número de patrones utilizados como atributos en el proceso de clasificación sobre la base de datos SIS. 60
5.14. Resultados de la clasificación alcanzados utilizando el método propuesto en la colección de imágenes SIS con y sin el uso de los patrones emergentes como atributos para varios clasificadores con diferentes valores de γ. . . . 61

Índice de algoritmos

3.1. Algoritmo VEAM. 23
3.2. Procedimiento appLSet de VEAM. 24

Capítulo 1

Introducción

Los grafos son estructuras poderosas y generales que pueden ser usadas para representar diversos tipos de objetos; en múltiples dominios como biología molecular, visión por computadora, astronomía, cartografía, mercadotécnia, detección de fraudes y comportamientos inusuales, entre otros (Riesen & Bunke, 2008; Acosta-Mendoza *et al.*, 2012c). Varias han sido las técnicas basadas en grafos desarrolladas por diferentes autores para satisfacer la necesidad de convertir grandes volúmenes de datos en información útil (Alves *et al.*, 2010; Eichinger & Böhm, 2010; Jiménez *et al.*, 2010; Jiang *et al.*, 2012). Un ejemplo de estas técnicas es la minería de subgrafos frecuentes aproximados (MSFA) (Holder *et al.*, 1992; Ketkar *et al.*, 2006; Jia *et al.*, 2009; Acosta-Mendoza *et al.*, 2012b; Gago-Alonso *et al.*, 2013). Dicha técnica se ha convertido en un tema de gran importancia en las tareas de minería donde los subgrafos son detectados teniendo en cuenta distorsiones en los datos. La información que brindan estos subgrafos permite una mejor representación de los datos debido a que en las aplicaciones prácticas no es común la existencia de dos objetos exactamente iguales. Por este motivo, evaluar la similitud entre grafos permitiendo diferencias estructurales o correspondencia inexacta, se ha convertido en una necesidad práctica importante (Conte *et al.*, 2004; Koyutürk *et al.*, 2004; Hossain

& Angryk, 2007). La correspondencia inexacta consiste en encontrar un mapeo aproximado entre vértices o aristas de dos grafos para determinar su similitud permitiendo diferencias entre su estructura o etiquetas.

Teniendo en cuenta este hecho, varios algoritmos han sido desarrollados para la MSFA en diferentes dominios de la ciencia (Song & Chen, 2006; Chen *et al.*, 2007; Jia *et al.*, 2011; Acosta-Mendoza *et al.*, 2012a,c). De estos algoritmos, solo unos pocos se han aplicado en la clasificación de imágenes (Acosta-Mendoza *et al.*, 2012a,c). En estos trabajos los autores han obteniendo buenos resultados; sin embargo, dichos algoritmos calculan un gran número de subgrafos frecuentes aproximados (patrones) los cuales son utilizados como atributos para la clasificación de imágenes. Esto puede afectar el desempeño de los clasificadores, debido a la alta dimesionalidad de la representación de las imágenes. Además, algunos de estos patrones no proveen información útil para la clasificación. Para solucionar este problema es necesario lograr una reducción de la dimensionalidad de la representación de las imágenes sin afectar la efectividad de los clasificadores. En esta tesis se estudian dos maneras de reducir dicha dimensionalidad: (1) siguiendo una estrategia para reducir el conjunto de subgrafos frecuentes aproximados (SFA) bajo algún criterio que permita mantener solo los patrones que brinden información útil para la clasificación (Dong & Bailey, 2011; Jin & Wang, 2011; Poezevara *et al.*, 2011; Dhifli *et al.*, 2013; Kong *et al.*, 2013), o bien (2) utilizando los enfoques convencionales de selección de atributos (Pudil *et al.*, 1994; Liu *et al.*, 2002; Xue-wen, 2003; Pineda-Bautista *et al.*, 2011; Norshafarina *et al.*, 2013; Rodríguez-Bermúdez *et al.*, 2013).

Los SFA emergentes (Novak *et al.*, 2009; Garcia-Borroto, 2010) son un ejemplo de los patrones que pueden ser utilizados en la primera estrategia mencionada para la reducción de la dimensionalidad de la representación de las imágenes. Un patrón es emergente si aparece mayormente en una clase, mientras que rara vez aparece en el resto de las clases

(Kong *et al.*, 2013). Este tipo de patrones constituyen un subconjunto del conjunto de patrones calculados por los algoritmos para la MSFA, por lo que en el marco de esta tesis se estudiará el uso de los mismos para seleccionar un subconjunto de patrones útiles para la clasificación de imágenes. Por otro lado, dado que los métodos de selección de atributos reducen la dimensionalidad de los vectores de atributos utilizados en la clasificación (Pudil *et al.*, 1994; Xue-wen, 2003), en esta tesis también se estudiará la aplicación de estas técnicas para seleccionar un subconjunto de atributos que permita una mejor y más compacta representación de las imágenes.

1.1. Problemática actual

La clasificación de imágenes ha sido un tema que ha mantenido ocupado a muchos investigadores, partiendo desde los enfoques estadísticos de reconocimiento de patrones (Duin & Pekalska, 2005; Pekalska *et al.*, 2006) hasta los estructurales basados en cadenas de caracteres (Spillmann *et al.*, 2006) y en estructuras más complejas como los grafos (Riesen *et al.*, 2007; Bahadir & Selim, 2010a; Morales-González & García-Reyes, 2010, 2011; Bunke & Riesen, 2012). De igual manera se han ido incrementado los esfuerzos en los métodos de clasificación de imágenes basados en minería de grafos, donde mejorar la calidad de la clasificación continúa siendo un reto dentro del reconocimiento de patrones (Jiang & Coennen, 2008; Jiang *et al.*, 2010). Se han desarrollado varios modelos de clasificación, bajo diversos criterios, que utilizan MSFA encaminados a mejorar dicha calidad (Acosta-Mendoza *et al.*, 2012a,c). Estos últimos trabajos han reportado buenos resultados en algunas tareas de clasificación de imágenes; sin embargo, presentan algunos problemas, descritos a continuación, que impiden un mejor desempeño y pueden estar influyendo en la eficacia de los modelos propuestos.

Problema 1. Identifican un elevado número de patrones (atributos) en el proceso de la minería. Dicho número de patrones aumenta el consumo de recursos computacionales lo cual hace poco eficiente el proceso de clasificación.

Problema 2. No se hace ningún procesamiento de los patrones (atributos) con el fin de realizar una selección de los más adecuados para la clasificación. Esto puede estar influyendo negativamente en la toma de decisiones de los algoritmos de clasificación, lo cual reduce la eficacia de los mismos.

1.2. Motivación

Como se mencionó anteriormente, el uso de la MSFA en la clasificación de imágenes es una necesidad en aplicaciones prácticas. Las distorsiones que se permiten con este tipo de minería mejoran los resultados de la clasificación en algunas de las aplicaciones en imágenes. Sin embargo, los modelos de clasificación reportados tienen los problemas 1 y 2 mencionados en la sección 1.1. Por lo tanto, el desarrollo de un nuevo modelo de clasificación que incluya métodos para la reducción de la dimensionalidad, sin comprometer la eficacia alcanzada, sigue siendo una línea de investigación abierta. Esto se debe a que no siempre es práctico considerar todos los SFA debido a que pueden ser demasiados, además de que no todos son útiles para obtener una buena clasificación.

En esta tesis, se centrarán los esfuerzos en la combinación de la MSFA y estrategias de reducción de la dimensionalidad de los patrones o atributos a utilizar en la clasificación de imágenes. Esto se desea realizar de forma tal que se reduzca dicha dimensionalidad sin afectar la efectividad reportada en la literatura, y de ser posible, lograr superar dicha efectividad.

1.3. Objetivo general

Con base en lo antes mencionado, el objetivo general de esta tesis es:

"Diseñar e implementar un método para la clasificación de imágenes utilizando un subconjunto de patrones (atributos) que permita aumentar la eficacia y eficiencia de la clasificación respecto a los métodos similares reportados en la literatura".

1.4. Objetivos específicos

Los objetivos específicos de esta tesis son los siguientes:

1. Identificar alguno o algunos algoritmos de selección de patrones o selección de atributos que sean adecuados para la clasificación de imágenes utilizando la minería de subgrafos frecuentes aproximados.
2. Seleccionar un algoritmo de minería de subgrafos frecuentes aproximados adecuado para el problema de clasificación de imágenes.
3. Proponer un método para la clasificación de imágenes utilizando un subconjunto de patrones o atributos.
4. Realizar un estudio comparativo entre el método desarrollado utilizando patrones emergentes y dicho método utilizando algoritmos de selección de atributos tipo filtrado.
5. Evaluar la eficiencia y eficacia del método desarrollado.

1.5. Organización de esta tesis

La manera en que está organizado el contenido de este documento es la siguiente:

En el capítulo 2 se presentan los conceptos básicos requeridos para definir el problema de la MSFA y para entender el resto del documento. También se incluye la definición de patrón emergente utilizada en esta tesis.

En el capítulo 3 se describen los trabajos más relevantes relacionados con esta línea de investigación. Se hace énfasis en los algoritmos utilizados para la clasificación de imágenes ya que son el centro de atención de esta tesis.

En el capítulo 4 se presenta el método de clasificación de imágenes propuesto en este trabajo de investigación.

El capítulo 5 muestra los resultados experimentales obtenidos al evaluar el desempeño del método de clasificación de imágenes propuesto y una comparación experimental contra los resultados reportados en el estado del arte sobre las mismas bases de datos. Adicionalmente, se muestran los resultados obtenidos por el método de clasificación propuesto en una base de datos de objetos reales.

Finalmente, se exponen las conclusiones y algunas direcciones a seguir como trabajo futuro.

Capítulo 2

Marco teórico

En este capítulo se dan los conceptos básicos necesarios para definir el problema de la minería de subgrafos frecuentes aproximados (MSFA), se dan los conceptos de clasificación y una breve presentación de algunos criterios de selección necesarios para entender el resto del documento.

2.1. Conceptos básicos usados de la teoría de grafos

En esta tesis se pretende clasificar colecciones de imágenes utilizando como atributos los patrones calculados mediante un algoritmo para la MSFA. Por esta razón, dicho algoritmo debe ser capaz de procesar colecciones de grafos etiquetados, simples y no dirigidos. En adelante, cuando se hable de grafos se suponen este tipo de grafos, en otro caso se especificará.

Definición 2.1 (Grafo etiquetado)**.** Un *grafo etiquetado* es una 5-tupla, $G = (V, E, L_V, L_E, I, J)$, donde:

- V es un conjunto cuyos elementos son conocidos como *vértices*

- $E \subseteq \{\{u, v\} \mid u, v \in V, u \neq v\}$ es un conjunto cuyos elementos son conocidos como *aristas* (la arista $\{u, v\}$ conecta el vértice u con el vértice v)
- L_V es el conjunto de etiquetas para los vértices
- L_E es el conjunto de etiquetas para las aristas
- $I : V \to L_V$ es una *función etiquetadora* encargada de asignar etiquetas a los vértices
- $J : E \to L_E$ es una *función etiquetadora* encargada de asignar etiquetas a las aristas

En la figura 2.1 se muestra un ejemplo de grafo etiquetado con $L_V = \{A, B, C\}$ y $L_E = \{0, 1\}$.

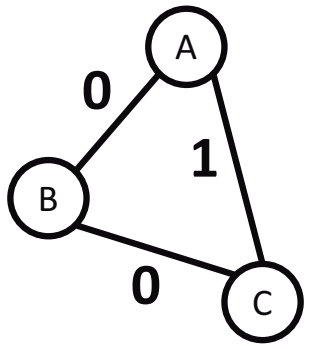

Figura 2.1: Grafo etiquetado.

Definición 2.2 (Subgrafo y supergrafo)**.** Sean $G_1 = (V_1, E_1, L_{V_1}, L_{E_1}, I_1, J_1)$ y $G_2 = (V_2, E_2, L_{V_2}, L_{E_2}, I_2, J_2)$ dos grafos, se dice que G_1 es un *subgrafo* de G_2 si $V_1 \subseteq V_2$, $E_1 \subseteq E_2$, $\forall u \in V_1, I_1(u) = I_2(u)$, y $\forall e \in E_1, J_1(e) = J_2(e)$. En este caso, se utiliza la notación $G_1 \subseteq G_2$ y además se dice que G_2 es un *supergrafo* de G_1.

En la figura 2.2 se muestra un ejemplo de subgrafo y supergrafo dados dos grafos etiquetados G_1 y G_2, donde el grafo G_1 es un subgrafo de G_2 y por tanto el grafo G_2 es supergrafo de G_1.

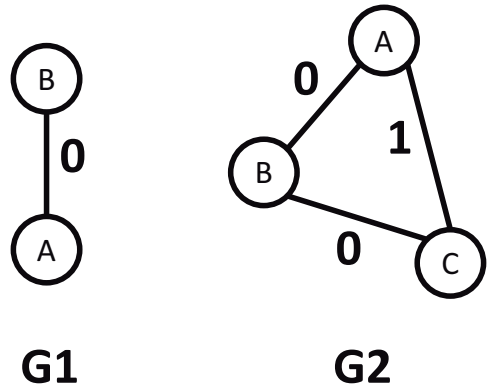

Figura 2.2: Ejemplo de subgrafo y supergrafo.

Definición 2.3 (Isomorfismo)**.** Dados dos grafos G_1 y G_2, se dice que f es un *isomorfismo* entre esos grafos si $f : V_1 \to V_2$ es una función biyectiva, donde:

- $\forall u \in V_1 : f(u) \in V_2 \wedge I_1(u) = I_2(f(u))$
- $\forall \{u, v\} \in E_1 : \{f(u), f(v)\} \in E_2 \wedge J_1(\{u, v\}) = J_2(\{f(u), f(v)\})$

Si existe un isomorfismo entre G_1 y G_2, se dice que G_1 y G_2 son *isomorfos*. Si G_1 es isomorfo a G_3 y $G_3 \subseteq G_2$, entonces se dice que existe un *sub-isomorfismo* entre G_1 y G_2, y además se dice que G_1 es sub-isomorfo a G_2.

Definición 2.4 (Soporte)**.** Siendo $D = \{G_1, \ldots, G_{|D|}\}$ una colección de grafos y G un grafo etiquetado en L, el valor del *soporte* de G en D se define como el conjunto de grafos $G_i \in D$, tal que exista un sub-isomorfismo entre G y G_i. Este valor de soporte se obtiene mediante la ecuación (2.1):

$$supp(G, D) = \frac{|\{G_i \in D\text{: } G \text{ es sub-isomorfo a } G_i\}|}{|D|} \tag{2.1}$$

Definición 2.5 (Semejanza)**.** Siendo Ω el conjunto de todos los posibles grafos etiquetados en el dominio de todas las posibles etiquetas L, la *semejanza* entre dos grafos $G_1, G_2 \in \Omega$ se define como una función $sim : \Omega \times \Omega \to [0, 1]$. Se dice que los grafos son diferentes si $sim(G_1, G_2) = 0$, mientras mayor sea el valor de $sim(G_1, G_2)$ más similares son los grafos, y si $sim(G_1, G_2) = 1$ entonces existe un sub-isomorfismo entre los grafos.

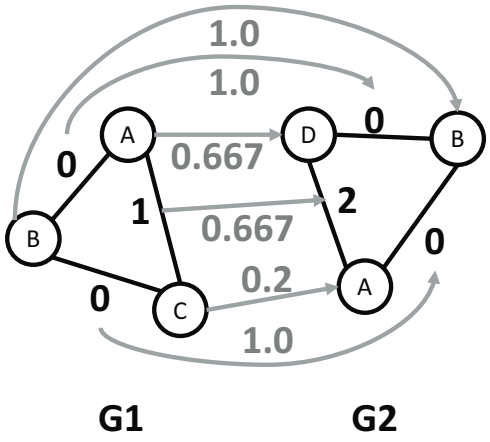

Figura 2.3: Ejemplo de semejanza entre dos grafos.

En la figura 2.3 se muestra un ejemplo de la evaluación de la semejanza entre dos grafos. Supongamos que la función de semejanza se basa en el producto de las semejanzas entre los vértices y entre las aristas de cada grafo manteniendo la topología de los mismos. Supongamos además, que las etiquetas "A" y "1" pueden sustituir a las etiquetas "D" y "2", respectivamente, con una semejanza de 0.667, la etiqueta "C" puede sustituir a la etiqueta "A" con una semejanza de 0.2, y las etiquetas "B", "0" y "2" no sustituyen a ninguna otra, excepto a ellas mismas con una semejanza de 1.0. Entonces, el grafo G_1 es similar al grafo G_2 con una semejanza de 0.09.

Entre dos grafos etiquetados puede existir más de una correspondencia entre sus vértices y aristas. En la figura 2.4 se muestra otra correspondencia entre los grafos G_1

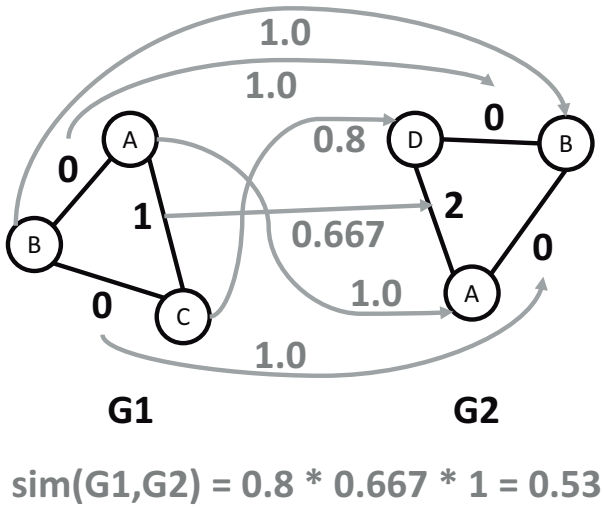

Figura 2.4: Ejemplo de semejanza entre dos grafos.

y G_2 del ejemplo anterior, suponiendo que la etiqueta "C" puede sustituir a la etiqueta "D" con una semejanza de 0.8. Entonces, utilizando la misma función de semejanza de dicho ejemplo (producto de la semejanza entre los vértices y entre las aristas) se obtiene que el grafo G_1 tiene una semejanza de 0.53 con el grafo G_2.

Como se puede observar en las figuras 2.3 y 2.4, pueden existir varias correspondencias entre dos grafos. Por lo que $sim_{max}(G_1, G_2) = max\{sim(G_1, G_2)\}$ se define como el mayor valor de semejanza que se puede obtener entre las diferentes correspondencias entre G_1 y G_2. Definida sim_{max} utilizando la definición de semejanza entre dos grafos, se puede definir un soporte que permita utilizar correspondencia inexacta entre grafos.

Definición 2.6 (Soporte aproximado)**.** Sea $D = \{G_1, \ldots, G_{|D|}\}$ una colección de grafos y G un grafo, el valor del soporte aproximado (denotado por appSupp) de G en D, en términos de la semejanza, se obtiene mediante la ecuación (2.2):

$$appSupp(G, D) = \frac{\sum_{G_i \in D} sim_{max}(G, G_i)}{|D|} \tag{2.2}$$

Definición 2.7 (Subgrafo frecuente aproximado)**.** Un grafo G es un *subgrafo frecuente*

aproximado (SFA) en D si $appSupp(G, D) \geq \delta$ utilizando (2.2).

El valor del umbral de soporte δ está entre $[0, 1]$ dado que la similiaridad está definida entre $[0, 1]$.

Definición 2.8 (Minería de subgrafos frecuentes aproximados)**.** La *minería de subgrafos frecuentes aproximados* consiste en encontrar todos los SFA en una colección de grafos D, utilizando una función de semejanza *sim* y un umbral de soporte δ.

Definición 2.9 (Patrón emergente y patrón contrastante)**.** Sea $D = \{G_1, \ldots, G_{|D|}\}$, sea C un conjunto de clases $C = \{c_1, \ldots, c_{|C|}\}$, donde $c_i \subset D$, $c_i \neq c_j$, $U_{c_i} = D$ y sea G un grafo de D, se dice que G es un *patrón emergente* (en inglés, *emerging pattern*) (Dong & Li, 1999; Li *et al.*, 2000; Garcia-Borroto, 2010; Kong *et al.*, 2013) para c_i si $appSupp(G, c_i) \geq \gamma$ y $appSupp(G, D - \{c_i\}) < \gamma$; $\gamma \in (0, 1)$. Sea G un patrón emergente en D para la clase $c_i \in C$, si $appSupp(G, D - \{c_i\}) = 0$, entonces G es un *patrón contrastante* (en inglés, *contrast pattern*) (Borgelt & Berthold, 2002; Zhao *et al.*, 2011).

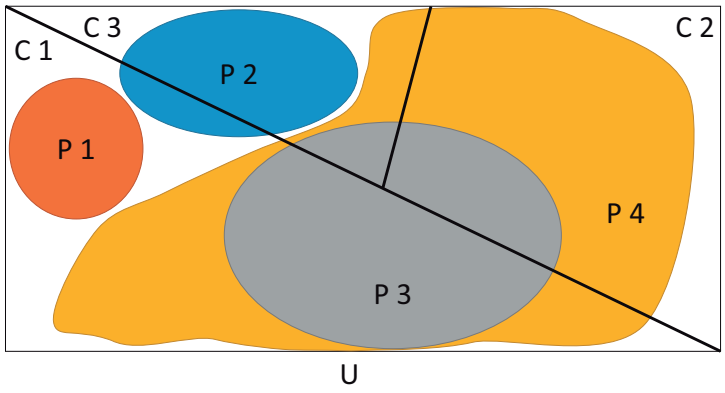

Figura 2.5: Ejemplo de patrones emergentes y contrastantes en un universo de objetos U compuesto por tres clases ($C1$, $C2$ y $C3$).

En la figura 2.5 se presentan gráficamente ejemplos de patrones emergentes y contrastantes en un conjunto de objetos divididos en tres clases ($C1$, $C2$ y $C3$). Como se puede

observar el patrón "$P1$" es un patrón que solo se encuentra en la clase $C1$ y no aparece en ninguna otra clase, por lo que se identifica como un patrón contrastante para la clase $C1$. Supongamos que γ =0.4, entonces los patrones "P2 y "P3" se pueden identificar como patrones emergentes para las clases $C3$ y $C1$ respectivamente, ya que más de 40 % de sus ocurrencias le pertenecen a una clase y el resto de sus ocurrencias que pertenecen a otras clases no sobrepasa ese 40 %. No se cumple esta definición para el patrón "P4" ya que es un patrón que aparece en varias clases con más del 40 % de ocurrencias.

2.2. Conceptos básicos de aprendizaje automático

Dentro del aprendizaje automático, se han desarrollado diversas técnicas de clasificación que nos permiten agrupar objetos de acuerdo a diferentes criterios o métodos. Estas técnicas pueden ser divididas en dos categorías: *clasificación supervisada* y *clasificación no supervisada*. Dichas técnicas tienen como objetivo la asignación de un objeto o un fenómeno físico a una de las diversas categorías o clases especificadas.

La clasificación no supervisada no cuenta con conocimiento a priori, sino que se encuentran objetos o muestras que tiene un conjunto de características, de las que no se sabe a qué clase o categoría pertenece. La finalidad es el descubrimiento de grupos de objetos cuyas características afines nos permitan separar las diferentes clases. Lo ideal es agrupar los objetos de tal manera que los objetos que están en un mismo grupo sean lo más parecidos entre sí, mientras que no compartan semejanzas entre los objetos que se encuentren en otros grupos.

La clasificación supervisada cuenta con un conocimiento a priori, es decir, para la tarea de clasificar un objeto dentro de una categoría o clase se cuenta con modelos ya clasificados de antemano. Dentro de este tipo de clasificación se pueden diferenciar dos etapas: (1) proceso de entrenamiento dependiente del diseño del clasificador. Como

resultado de esta etapa se obtiene un modelo o conjunto de reglas generales que permiten clasificar nuevos objetos; y (2) proceso que se da a la tarea de clasificar los objetos o muestras de las que se desconoce la clase a las que pertenecen utilizando la información de los objetos que se les asemejan.

Definición 2.10 (Clasificación supervisada)**.** La *clasificación supervisada* consiste en encontrar una función f tal que: $f : Obj \rightarrow C$, donde Obj es un objeto de entrada a clasificar y C es el conjunto de etiquetas (categorías) que describen las clases del conjunto de datos dado.

La complejidad de la tarea de clasificación supervisada depende del tamaño del conjunto de objetos y la cantidad de atributos que estos contengan. En muchas aplicaciones estos números son grandes e impiden un buen desempeño de los clasificadores. Por este motivo las investigaciones en aprendizaje automático, minería de datos, reconocimiento de patrones y estadística han desarrollado una serie de métodos para reducción de dimensionalidad basado en utilidad y precisión de la clasificación, esta tarea se le conoce como *selección de atributos*. Esta intenta discriminar los atributos más relevantes de un conjunto de datos dado. Inclusive, la mayoría de esquemas de aprendizaje utilizan selección de atributos, ya sea como técnica de pre-procesamiento o como una etapa dentro del proceso de clasificación.

Definición 2.11 (Selección de atributos)**.** Sea S un conjunto de datos, donde A es un conjunto de atributos tal que $|A| = n$ y $X \subseteq A$, se define una función $R(X)$ que evalúa la relevancia del subconjunto de atributos X y el problema de *selección de atributos* consiste en encontrar un subconjunto de atributos Z tal que $R(Z) = max_{X \subseteq A} R(X)$.

Donde en el peor de los casos se hacen 2^n comparaciones para explorar todo el espacio de búsqueda, por lo que una búsqueda exhaustiva es inviable. Por esta razón se han

propuesto alternativas (i.e. algoritmos de filtrado, de envoltura y embebidos) para atacar este problema. Los enfoques de filtrado son considerados como uno de las primeras propuestas surgidas en la literatura para selección de atributos. Estos métodos intentan descartar los atributos que sean irrelevantes basándose únicamente en la evaluación de las propiedades intrínsecas de los atributos y sus relaciones con las clases del conjunto de datos. La principal ventaja de estos métodos es su bajo costo computacional debido a los criterios (i.e. ganancia de información, chi-cuadrado y cociente de la ganancia de información) que utiliza para la evaluación. Unos de los criterios clásicos que se utiliza para la evaluación es el de *ganancia de información* (*information gain*). Este algoritmo evalúa los atributos midiendo la ganancia de información de cada uno con respecto a la clase. Cuando el conjunto de datos tiene atributos continuos se lleva a cabo un pre-procesamiento donde se discretizan los valores.

Definición 2.12 (Ganancia de información)**.** La *ganancia de información* consiste en calcular la información mútua de un conjunto de atributos X relativa al conjunto de clases C definida como:

$$IG(X, C) = H(X) - H(X|C) \tag{2.3}$$

$$H(X) = -\sum_{x \in X} p(x) log p(x) \tag{2.4}$$

$$H(X|C) = -\sum_{x \in X, c \in C} p(x, c) log \frac{p(x, c)}{p(c)} \tag{2.5}$$

donde $H(X)$ y $H(X|C)$ son la entropía de X y la entropía condicional de X dado C, respectivamente.

El criterio que evalúa cada atributo midiendo su razón de beneficio respecto a la clase es conocido como *cociente de la ganancia de información* (*gain ratio*). Este criterio se

calcula mediante la ecuación (2.6):

$$IGR(X, C) = \frac{IG(X, C)}{H(C)} \tag{2.6}$$

El criterio conocido como *Chi-cuadrado* (*Chi-square*) calcula el valor estadístico chi-cuadrado de cada atributo respecto a la clase y de esta manera se obtiene el nivel de correlación entre la clase y cada atributo. Es considerado como una prueba no paramétrica que mide la discrepancia entre una distribución observada y otra teórica, indicando en qué medida las diferencias existentes entre ambas, de haberlas, se deben al azar en el contraste de hipótesis. Este criterio se calcula mediante la ecuación (2.7):

$$CHI = \sum_i \frac{(vo_i - ve_i)^2}{ve_i} \tag{2.7}$$

Donde vo_i es el valor obtenido y ve_i es el valor esperado.

Cuando el valor de CHI tiende a cero quiere decir que los valores obtenidos se parecen mucho a los valores esperados.

2.3. Síntesis y conclusiones

En este capítulo se han definido formalmente los conceptos que son necesarios para un mejor entendimiento del resto del documento. Primero se presentaron los conceptos básicos de la teoría de grafos que caracterizan el problema de la minería de grafos. Además, se presentaron algunas definiciones importantes para la MSFA como soporte aproximado y subgrafo frecuente aproximado. Luego, se presentaron los conceptos de clasificación y una breve explicación de algunos criterios de selección que se utilizan en esta tesis. Todos estos conceptos son utilizados como base para la definición de los

algoritmos del estado del arte que se presentan en el capítulo 3, así como la definición de patrones emergentes y los criterios de selección de atributos forma parte de la propuesta de esta tesis.

Capítulo 3

Trabajo relacionado

En la literatura se han reportado varios algoritmos para la MSFA en colecciones de grafos, los cuales usan diferentes funciones de similaridad en la correspondencia entre grafos. Existen varios enfoques para la MSFA, por ejemplo:

- Algoritmos basados en distancia de edición (Holder *et al.*, 1992; Song & Chen, 2006), donde todos los posibles caminos de edición de un grafo son explorados durante el proceso de generación de los candidatos. En el algoritmo *SUBDUE* (Holder *et al.*, 1992) se buscan sub-estructuras frecuentes en un solo grafo mediante la identificación de los caminos de menor costo explorados, mientras que el algoritmo *RNGV* (Song & Chen, 2006) no busca el camino de menor costo, solamente busca uno que satisfaga la inexactitud específicada.

- Algoritmos basados en β-arista sub-isomorfismo (Zhang *et al.*, 2007; Zhang & Yang, 2008), el cual solamente permite distorsiones entre aristas y etiquetas de aristas.

- Algoritmos basados en sub-homeomorfismo con vértices/aristas disjuntas (Xiao *et al.*, 2007, 2008), los cuales calculan estructuras aproximadas con topología invariante.

- Algoritmos basados en sub-isomorfismo entre grafos inciertos (Zou *et al.*, 2009, 2010; Papapetrou *et al.*, 2011), donde el soporte esperado para cada candidato es calculado sobre una colección de subgrafos construida utilizando las probabilidades de no ocurrencias en la colección original.

- Algoritmos basados en probabilidades de sustitución (Chen *et al.*, 2007; Jia *et al.*, 2009, 2011; Acosta-Mendoza *et al.*, 2012a,b), donde no siempre una etiqueta de vértice o una etiqueta de arista puede reemplazar o ser reemplazada por otra. El algoritmo *gApprox* (Chen *et al.*, 2007) está desarrollado para procesar un solo grafo, mientras que los algoritmos *VEAM* (Acosta-Mendoza *et al.*, 2012a,b) y *APGM* (Jia *et al.*, 2009, 2011) utilizan matrices de sustitución para realizar la MSFA en colecciones de grafos, preservando la topología de los grafos. En APGM, solamente se tratan las variaciones entre etiquetas de vértices mientras que en VEAM se permiten variaciones entre etiquetas de vértices y aristas.

Los trabajos mencionados anteriormente han sido aplicados en diferentes dominios tales como: análisis de estructuras bioquímicas (Xiao *et al.*, 2007; Zhang *et al.*, 2007; Xiao *et al.*, 2008; Zhang & Yang, 2008; Jia *et al.*, 2009, 2011; Zou *et al.*, 2009, 2010); análisis de redes genéticas regulatorias (Song & Chen, 2006); análisis de redes sociales y de vínculos (Holder *et al.*, 1992), entre otros.

La minería de subgrafos frecuentes ha sido utilizada satisfactoriamente para la clasificación de imágenes (Jiang & Coennen, 2008; Bahadir & Selim, 2010a,b; Jiang *et al.*, 2010; Elsayed *et al.*, 2010a,b; Acosta-Mendoza *et al.*, 2012a,c); sin embargo, solo unos pocos trabajos están enfocados en el cálculo de patrones (subgrafos) en grafos utilizando correspondencia inexacta (Acosta-Mendoza *et al.*, 2012a,c). Estos últimos trabajos han reportado mejores resultados que los trabajos basados en correspondencia exacta, siendo VEAM el algoritmo que mejores resultados reporta. Por este motivo, en este trabajo se

utilizará VEAM como algoritmo para la MSFA.

3.1. Algoritmo VEAM para la MSFA

Como el algoritmo VEAM (***V****ertex and* ***E****dge* ***A****pproximate graph* ***M****iner*) será utilizado para calcular los SFA que se utilizarán como atributos en el modelo de clasificación propuesto en esta tesis, en esta sección se darán más detalles de dicho algoritmo para una mejor comprensión del mismo.

El algoritmo VEAM introduce las aproximaciones entre las etiquetas de las aristas y los vértices utilizando matrices de sustitución (Acosta-Mendoza *et al.*, 2012a). Como se mencionó anteriormente, dichas aproximaciones se tratan manteniendo la topología de los grafos. La definición 3.1 nos permite entender las matrices de sustitución.

Definición 3.1 (Matriz de sustitución y matriz de sustitución estable)**.** Una *matriz de sustitución* $M = (m_{i,j})$ es una matriz $|L| \times |L|$ indizada por el conjunto de etiqueta L, donde una celda $m_{i,j}$ en M $(0 \leq m_{i,j} \leq 1, \sum_j m_{i,j} = 1)$ corresponde a la probabilidad de que la etiqueta i sea reemplazada por la etiqueta j. Si M es diagonal dominante (i.e. $m_{i,i} \geq m_{i,j}, \forall i \neq j$) entonces se dice que M es una *matriz de sustitución estable*.

Una matriz de sustitución para las etiquetas de las aristas y otra para las etiquetas de los vértices son utilizadas para calcular las aproximaciones tratadas en VEAM. La definición 3.2 es la utilizada por VEAM para calcular la similitud entre los subgrafos.

Definición 3.2 (Similitud utilizando sub-isomorfismo aproximado)**.** Sean $G_1 = (V_1, E_1, I_1, J_1)$ y $G_2 = (V_2, E_2, I_2, J_2)$ dos grafos etiquetados en L, siendo MV y ME las matrices de sustitución indizadas por L_V y L_E respectivamente, y τ el umbral de mínimo isomorfismo. La similaridad basada en sub-isomorfismo aproximado entre G_1 y G_2 se define como:

- $S_h(G_1, G_2) = \prod_{u \in V_1} \frac{MV_{I_1(u), I_2(h(u))}}{MV_{I_1(u), I_1(u)}} * \prod_{e=\{u,v\} \in E_1} \frac{ME_{J_1(e), J_2(\{h(u), h(v)\})}}{ME_{J_1(e), J_1(e)}} \geq \tau$.

En caso de que exista un sub-isomorfismo aproximado entre G_1 y G_2, se dice que G_1 es sub-isomorfo aproximado a G_2 y se denota como: $G_1 \subseteq_A G_2$.

El producto normalizado de las probabilidades de sustitución de h, denotado por $S_h(G_1, G_2)$, se conoce como el *grado de sub-isomorfismo aproximado* (approximate sub-isomorphism score). En un grafo pueden encontrarse más de una ocurrencia de un sub-grafo que cumpla con τ según la definición 3.2. Por lo que para realizar el conteo del soporte se utiliza la ocurrencia con mayor grado de sub-isomorfismo aproximado, denotada por $S_{max}(G_1, G_2)$, aunque se utilicen todas (las que cumplan con τ) para realizar el crecimiento del patrón en el proceso de la minería.

Definición 3.3 (Ocurrencia y conjunto de ocurrencias)**.** Dados los grafos $G_1 = (V_1, E_1, I_1, J_1)$, $G_2 = (V_2, E_2, I_2, J_2)$ y $T = (V_T, E_T, I_T, J_T)$, donde $T \subseteq G_2$. Se dice que T es una *ocurrencia* de G_1 en G_2 si $G_1 \subseteq_A T, |V_1| = |V_T|$ y $|E_1| = |E_T|$. El *conjunto de ocurrencias* de G_1 en G_2 se denota por $O(G_1, G_2)$.

Definición 3.4 (Extensión hacia delante y hacia atrás)**.** Dados dos grafos $G_1 = (V_1, E_1, I_1, J_1)$ y $G_2 = (V_2, E_2, I_2, J_2)$, donde $G_1 \subseteq G_2$, se dice que $e = \{u, v\} \in E_2$ es una *extensión* de G_1 si: $V_2 = V_1 \cup \{v\}$ y $E_1 = E_2 \setminus \{e\}$, denotado por $G_2 = G_1 \diamond e$. Se dice que e es una *extensión hacia atrás* (backward extension) si $v \in V_1$, en otro caso se dice que es una *extensión hacia delante* (forward extension) si extiende el conjunto de vértices de G_1.

Definición 3.5 (Conjunto de extensiones)**.** Siendo T un embebido de G_1 en $G_2 = (V_2, E_2, I_2, J_2)$. Entonces el *conjunto de extensiones* de T se denota como $ExtSet(T) = \{e \in E_2 | e \text{ es una extensión de } T\}$.

El algoritmo VEAM se muestra mediante tres procedimientos principales (ver algoritmos 3.1 y 3.2). Dicho algoritmo explora el conjunto de grafos de una colección de grafos dada partiendo de los vértices frecuentes aproximados (ver procedimiento VEAM).

El procedimiento VEAMSearch se encarga de realizar una búsqueda recursiva, creciendo en profundidad un patrón dado. Tomando en cuenta solo el conjunto de ocurrencias (ver definición 3.3) y el conjunto de extensiones en los grafos de la colección (ver definición 3.5) de dicho patrón a crecer, se construyen los candidatos utilizando los conjuntos de posibles etiquetas, obtenidas por el procedimiento *appLSet*, con las que se obtienen grafos aproximados. Luego, del conjunto de candidatos aproximados se almacenan y se continúan creciendo los que cumplan con el soporte aproximado dado un δ (ver definición 2.6) y que no hayan sido calculados en búsquedas anteriores. De esta manera se obtienen todos los SFA de la colección de grafos dada, siendo éstos la respuesta del algoritmo VEAM para la MSFA.

El procedimiento *appLSet* mostrado en el algoritmo 3.2 es el encargado de calcular la similaridad entre los candidatos y sus ocurrencias utilizando la definición 3.2. De esta manera se calcula el conjunto de posibles etiquetas a utilizar para la confección de los candidatos aproximados por cada grafo de la colección.

En (Acosta-Mendoza *et al.*, 2012a) se puede obtener mayor información respecto al algoritmo VEAM. También, en (Acosta-Mendoza *et al.*, 2012b) se presenta una mejora de dicho algoritmo mediante dos podas enfocadas a la reducción del espacio de búsqueda de las etiquetas de las aristas y los vértices.

Los SFA calculados por VEAM han mostrado ser útiles para la clasificación de imágenes; sin embargo, dicho algoritmo calcula un gran número de patrones en el proceso de la minería. Muchos de estos patrones calculados pudieran no aportar información útil para la clasificación, puesto que no se verifica que sean representativos para alguna clase

Procedimiento `VEAM`(D, MV, ME, δ, τ, F)

Input: D - Colección de grafos, MV - Matriz de sustitución indizada por L_V, ME - Matriz indizada por L_E, τ - Umbral de mínimo isomorfismo, δ - Umbral de frecuencia mínima.

Output: F - Conjunto de subgrafos frecuentes aproximados.

```
1 F ← C ← {Conjunto de todos los vértices etiquetados en L_V que son frecuentes aproximados en D};
2 forall T ∈ C do
3 |   VEAMSearch(T, D, MV, ME, δ, τ, F);
4 end
```

Procedimiento `VEAMSearch`(T, D, MV, ME, δ, τ, F)

Input: $T = (V_t, E_t, I_t, J_t)$ - Un subgrafo aproximado frecuente, D - Colección de grafos, MV - Matriz de sustitución indizada por L_V, ME - Matriz indizada por L_E, τ - Umbral de mínimo isomorfismo, δ - Umbral de frecuencia mínima.

Output: F - Conjunto de subgrafos frecuentes aproximados.

Local vars: $O(T, G_i)$, - Conjuntos de ocurrencias de T en G_i, $ExtSet(o_j)$, - Conjuntos de extensiones de o_j.

```
1  forall o_j ∈ O(T; G_i), donde G_i ∈ D do
2  |   forall e = {u, v}, e ∈ ExtSet(o_j) do
3  |   |   CL ←appLSet(T, D, MV, ME, G_i, o_j, e, τ);
4  |   |   forall (eLabel, vLabel) ∈ CL do
5  |   |   |   Se construye el candidato X utilizando la tupla (eLabel, vLabel);
6  |   |   |   Se calcula el código CAM de X y se almacena en codeCAM(X);
7  |   |   |   C ← C ∪ {(X, codeCAM(X), score)};
8  |   |   end
9  |   end
10 end
11 forall T_1 ∈ C do
12 |   if appSupp(T_1, D) ≥ δ y codeCAM(T_1) ∉ F then
13 |   |   Se inserta T_1 en F;
14 |   |   VEAMSearch(T, D, MV, ME, δ, τ, F);
15 |   end
16 end
```

Algoritmo 3.1: VEAM.

Procedimiento `appLSet`(T, MV, ME, G, G', e, τ)

Input: T - Un grafo candidato, MV - Matriz de sustitución indizada por L_V, ME - Matriz indizada por L_E, G - Un grafo de la colección, G' - Embebido de T en G, $e = \{u, v\}$ - Una extensión de G', τ - Umbral de mínimo isomorfismo.

Output: CL - Conjunto de tuplas candidatas $(eLabel, vLabel)$.

```
1  forall j ∈ L_E do
2      scoreE ← S_max(T, G') * ME_{j,J(e)} / ME_{j,j};
3      if e es una extensión hacia delante de G' then
4          forall i ∈ L_V do
5              score ← scoreE * MV_{i,I(v)} / MV_{i,i};
6              if score ≥ τ then CL ← CL ∪ {(j, i)};
7          end
8      end
9      else scoreE ≥ τCL ← CL ∪ {(j, ∅)}
10 end
```

Algoritmo 3.2: appLSet.

en específico y por lo tanto no serían de utilidad como atributos para la clasificación. Además, el número de patrones crece a medida que disminuye el umbral de soporte y/o el umbral de similaridad, lo cual afecta negativamente el rendimiento de los clasificadores.

3.2. Síntesis y conclusiones

En este capítulo se han descrito los trabajos relacionados con esta investigación, dándole cumplimiento parcial al objetivo particular 1 de este trabajo. Se ha hecho énfasis en los algoritmos para la MSFA, específicamente en el algoritmo VEAM ya que ha demostrado ser el más exitoso en la clasificación de imágenes basada en SFA. De esta manera se le da cumplimiento al objetivo particular 2 de esta investigación. Sin embargo, el algoritmo VEAM calcula un gran número de SFA especialmente para valores pequeños de los umbrales de soporte y de mínimo isomorfismo. Dicho número de patrones puede

afectar el rendimiento de los clasificadores y por este motivo en este trabajo se estudian algunas alternativas para la reducción de la dimensión del conjunto patrones a tener en cuenta en la clasificación.

Capítulo 4

Método de clasificación propuesto

Como se ha comentado en el capítulo anterior, la cantidad de patrones calculados por VEAM generalmente es muy grande y esto afecta el desempeño de la clasificación. El método propuesto parte del presentado por Acosta-Mendoza *et al.* (2012a), al cual se le agregará un módulo para la reducción de la dimensionalidad del conjunto de atributos a usar en la clasificación. Para esto, las dos líneas de investigación que se estudiarán serán: (1) la selección de un subconjunto de patrones emergentes a partir del conjunto de SFA, y (2) la selección de subconjuntos de atributos sobre la representación vectorial siguiendo algunos criterios clásicos de selección.

El método de clasificación propuesto en esta tesis está compuesto por cuatro módulos: módulo de representación, módulo de extracción de patrones, módulo de reducción de patrones, y módulo de clasificación. En la figura 4.1 se muestra gráficamente el flujo del método de clasificación propuesto.

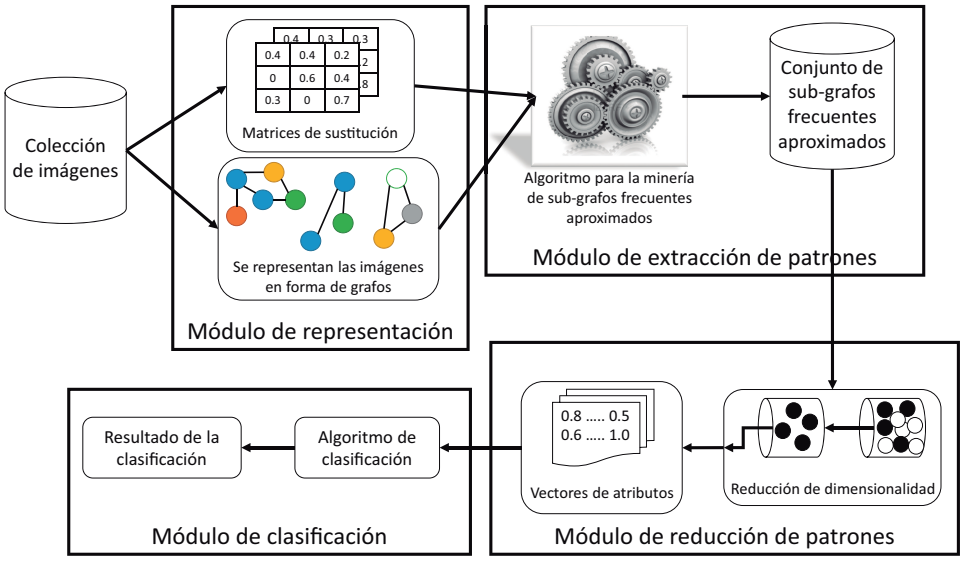

Figura 4.1: Método de clasificación de imágenes basada en grafos propuesto en esta tesis.

4.1. Módulo de representación

Dada una colección de imágenes pre-etiquetadas en este módulo se obtiene una colección de grafos donde cada grafo representa una imagen de la colección. Este proceso de representación es el mismo método propuesto en (Acosta-Mendoza *et al.*, 2012a). En este método, una imagen es dividida en cuatro cuadrantes de dimensiones iguales: noroeste (NW), noreste (NE), suroeste (SW) y sureste (SE); utilizando un método de árboles de cuadrantes (en inglés, *quad-trees*) similar al propuesto en (Finkel & Bentley, 1974). Este método de quad-trees divide recursivamente cada cuadrante mientras que el umbral de máximos niveles de profundidad no se haya alcanzado o hasta que los cuadrantes sean homogéneos respecto a alguna propiedad (en nuestro caso es el color). Nótese que en nuestro caso se utiliza el color como propiedad predominante debido a que las imágenes de la colección se obtienen de forma sintética y no poseen texturas, y los colores están distribuidos de forma homogénea. Cuando ya está dividida la imágen, la propiedad predominante de cada cuadrante es tomada como atributo para el nodo correspondiente del

quad-tree. En la figura 4.2 se muestra un ejemplo de un quad-tree correspondiente a una imagen dada. Como se puede observar en la figura, cada cuadrante se representa como un nodo del árbol, comenzando por el cuadrante que ocupa la imagen completa, el cual representa el nodo raíz del árbol. Al dividir un cuadrante se crea un sub-árbol donde la raíz de dicho sub-árbol es el cuadrante y los nodos hijos son los nuevos sub-cuadrantes. Este proceso se realiza en cada cuadrante como se mencionó anteriormente hasta que se obtiene un quad-tree como el que se muestra en la figura 4.2.

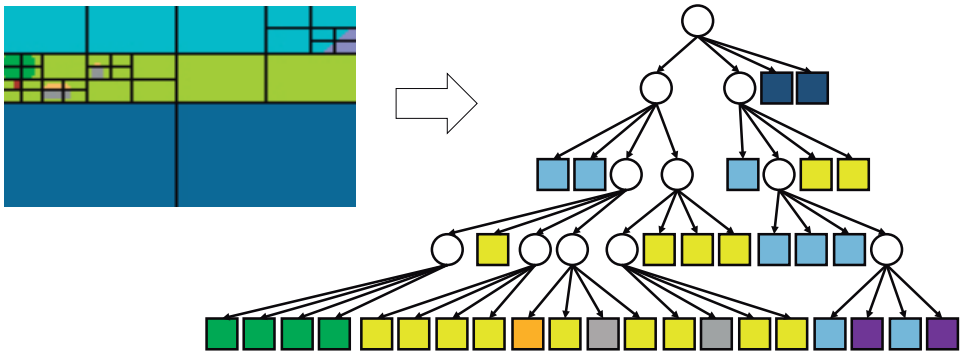

Figura 4.2: Ejemplo de un árbol de cuadrantes dada una imagen.

De dicho árbol, se genera un grafo que representa a la imágen de la siguiente manera:

1. Los vértices del grafo se construyen con las hojas del quad-tree (sub-cuadrantes) y sus atributos. Los atributos de las hojas son el punto medio y la propiedad predominante (en nuestro caso es el color) del cuadrante.

2. Cada vértice (cada hoja del quad-tree) se conecta mediante una arista con sus vecinos en las direcciones norte (N), sur (S), este (E) y oeste (W).

   Por ejemplo, considerando el quad-tree mostrado en la figura 4.2 se genera el grafo mostrado en la figura 4.3. El vértice etiquetado con el número 3 se conecta mediante una arista con cada vértice del conjunto $\{4,\ 19,\ 20,\ 23,\ 24,\ 27,\ 28\}$.

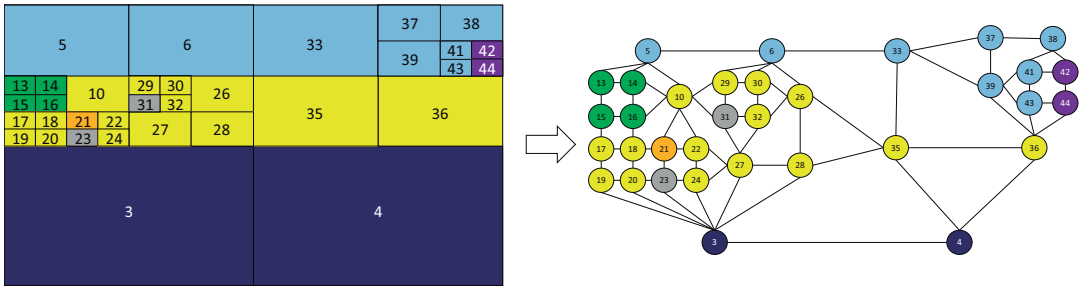

Figura 4.3: Ejemplo de un grafo construido a partir del quad-tree de la imagen de la figura 4.2.

3. Las etiquetas de cada arista e, denotadas por $label(e)$, son un índice obtenido mediante la ecuación (4.1) sobre el ángulo α entre la horizontal y la arista hacia el otro vértice. Esta función depende del número n de etiquetas diferentes utilizadas para categorizar los posibles ángulos. Los posibles ángulos tienen una covertura de 180^o, suponiendo que son aristas no dirigidas.

$$label(e) = \begin{cases} \lfloor \frac{\alpha * n}{\pi} \rfloor & \text{si } 0 \leq \alpha < \pi \\ \\ 0 & \text{si } \alpha = \pi \end{cases} \tag{4.1}$$

Para mayor información acerca del proceso de representación se recomienda consultar la propuesta de Acosta-Mendoza *et al.* (2012a).

4.2. Módulo de extracción de patrones

Una vez que se tienen las imágenes representadas en forma de grafos se calculan los subgrafos frecuentes aproximados (SFA) utilizando el algoritmo VEAM propuesto en (Acosta-Mendoza *et al.*, 2012a). Las matrices de sustitución necesarias para VEAM y que son utilizadas en esta tesis se obtuvieron de (Acosta-Mendoza *et al.*, 2012a).

Estos patrones (SFA) extraídos de la colección de grafos son considerados como una representación alternativa de las imágenes de la colección. Siguiendo la analogía de los vocabularios obtenidos en los enfoques de bolsas de palabras (en inglés, *bag-of-words*) para la clasificación de documentos, en nuestra propuesta se utilizan los SFA para la confección del conjunto de atributos que describen a la colección de imágenes.

Basados en los resultados reportados en (Acosta-Mendoza *et al.*, 2012a), el uso de los patrones teniendo en cuenta variaciones semánticas en los vértices y aristas, i.e. los patrones calculados por VEAM, permite una mejor clasificación que el uso de los patrones calculados por APGM (Jia *et al.*, 2011), y que utilizando los patrones calculados por algoritmos exactos (Yan & Huan, 2002; Gago-Alonso *et al.*, 2009). Por este motivo, para la MSFA en este trabajo se utilizará el algoritmo VEAM.

4.3. Módulo de reducción de patrones

Para evitar el uso de patrones que no contribuyen positivamente a la tarea de clasificación en este trabajo utilizaremos dos estrategias para la reducción del número de patrones (atributos) a tener en cuenta en la clasificación:

Identificación de patrones representativos: En el estado del arte se han propuesto varios trabajos que tratan sobre patrones representativos (Borgelt & Berthold, 2002; Ohara *et al.*, 2009; Garcia-Borroto, 2010; Dong & Bailey, 2011; Fang *et al.*, 2011; Jin & Wang, 2011; Poezevara *et al.*, 2011; Zhao *et al.*, 2011; Dhifli *et al.*, 2012, 2013; Keneshloo & Yasdani, 2013; Kong *et al.*, 2013). Estos patrones son de diferente tipo: patrones que proveen mayor información a los clasificadores (en inglés, *information gain*) (Ohara *et al.*, 2009; Kong *et al.*, 2013); patrones emergentes (Dong & Li, 1999; Li *et al.*, 2000; Garcia-Borroto, 2010; Poezevara *et al.*,

2011; Dhifli *et al.*, 2012, 2013); patrones contrastantes (Borgelt & Berthold, 2002; Dong & Bailey, 2011; Zhao *et al.*, 2011); patrones discriminativos (Fang *et al.*, 2011; Jin & Wang, 2011), entre otros. Muchos de estos criterios han sido utilizados en tareas de clasificación donde los atributos son subgrafos frecuentes (Borgelt & Berthold, 2002; Ohara *et al.*, 2009; Jin & Wang, 2011; Dhifli *et al.*, 2013; Kong *et al.*, 2013); sin embargo, en ninguno de estos casos se ha utilizado la MSFA para la clasificación. En este trabajo se utilizarán los patrones emergentes ya que han mostrado su utilidad al ser usados en varios trabajos arrojando buenos resultados (Dhifli *et al.*, 2012, 2013) y nuestra hipótesis en esta tesis es que siguiendo este mismo enfoque se puede mejorar el proceso de clasificación con un subconjunto de los SFA como atributos.

Luego, en esta variante, para reducir el número de SFA calculados por VEAM, se calcularán los patrones emergentes. Una vez calculados los patrones emergentes se procederá de la misma manera que en (Acosta-Mendoza *et al.*, 2012a), es decir, a partir de los SFA que sean patrones emergentes se construirán los vectores de atributos para representar a las imágenes. Para la construcción de estos vectores, solamente se tendrán en cuenta patrones calculados por VEAM que sean patrones emergentes.

Utilización de un algoritmo convencional de selección de atributos: Los algoritmos de selección de atributos tienen como idea principal seleccionar un subconjunto de atributos de entrada mediante la eliminación de atributos que contengan poca o nula información predictiva para la clasificación (He *et al.*, 2006; Solorio-Fernández *et al.*, 2010; Pineda-Bautista *et al.*, 2011; Bermejo *et al.*, 2012; Bolón-Canedo *et al.*, 2013; Rodríguez-Bermúdez *et al.*, 2013; Tan *et al.*, 2013; Ye *et al.*, 2013; Zhao *et al.*, 2013). Los algoritmos de selección de atributos pueden

dividirse en tres clases principales: algoritmos de filtrado (*filters*) (He *et al.*, 2006; Ferreira & Figueiredo, 2012; Ye *et al.*, 2013), algoritmos de envoltura (*wrappers*) (Bermejo *et al.*, 2012) y algoritmos embebidos (*embeddings*) (Duval *et al.*, 2009; Rodríguez-Bermúdez *et al.*, 2013). Debido a la cantidad de atributos (patrones) extraídos, en esta tesis se centrarán los esfuerzos en los algoritmos de filtrado ya que, como no utilizan información de los clasificadores para escoger el subconjunto de atributos, por lo general son los más rápidos. El subconjunto de atributos se escoge según la cantidad de información útil que brindan o el poder predictivo que contengan, utilizando una función objetivo independiente del clasificador.

En esta variante para reducir el número de patrones (atributos) también se procederá de la misma manera que en (Acosta-Mendoza *et al.*, 2012a), es decir, a partir de todos los SFA calculados por VEAM, se construirán los vectores de atributos para representar a las imágenes. Luego, se aplicará un algoritmo de selección de atributos para obtener vectores de menor dimensionalidad, los cuales serán utilizados para la clasificación.

Los vectores de atributos que representarán las imágenes en la clasificación se construyen a partir de los patrones seleccionados (ya sean los emergentes o todos los SFAs calculados por VEAM, en dependencia de la estrategia que se escoja). Para la construcción de estos vectores se tiene en cuenta la semejanza de cada patrón seleccionado con cada imágen de la colección (representada en forma de grafos). Dados estos patrones, una imagen es representada en forma de un vector de atributos $V = (v_1, \ldots, v_x)$, donde x representa la cantidad de patrones utilizados. Se construye una matriz donde el número de filas ($1 \leq i \leq |D|$) corresponde al número de imágenes en la colección, y el número de columnas ($1 \leq j \leq x$) corresponde al número de patrones. Cada valor de un atributo es asignado utilizando una configuración de semejanzas, es decir, cada celda $v_{i,j}$ de la

matriz contiene el mayor valor de semejanza de las ocurrencias del atributo (patrón) j en la imagen i de la colección (utilizando $sim_{max}(G, G_i)$ suponiendo que G es el patrón j y G_i la imagen i) y $v_{i,j} = 0$ en caso de que el atributo j no ocurra en la imagen i.

En este módulo de reducción de patrones se puede utilizar cualquiera de las dos estrategias descritas anteriormente con el objetivo de reducir la dimensionalidad del conjunto de atributos a utilizar en la clasificación de imágenes. Como se mencionó anteriormente, el número de patrones calculados mediante VEAM es mayor para un umbral de soporte pequeño, así como para un umbral de mínimo isomorfismo pequeño. Cuando estos patrones son utilizados como atributos para representar imágenes, existen varios atributos (patrones) que no contribuyen a una buena clasificación, por lo que eliminarlos ayudaría al proceso de clasificación. De esta manera se representarían las imágenes con los atributos que brinden información útil para la clasificación, eliminando patrones innecesarios y redundantes.

En la figura 4.4 se muestra el flujo del módulo de reducción de patrones.

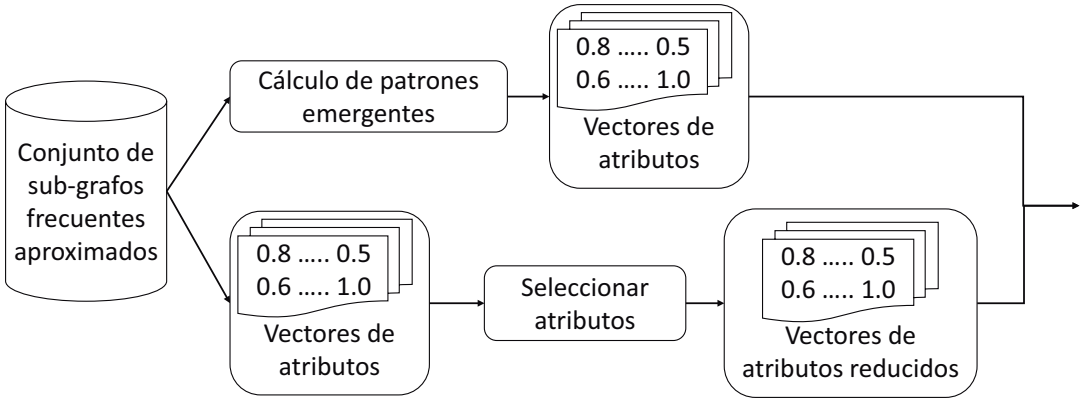

Figura 4.4: Flujo del módulo de reducción de patrones.

4.4. Módulo de clasificación

Una vez obtenidos los vectores de atributos que representan la colección de imágenes se aplica uno o varios algoritmos de clasificación.

4.5. Síntesis y conclusiones

Con el método de clasificación de imágenes descrito en este capítulo se logra cumplir el objetivo específico 3 de esta investigación. El cálculo de los patrones emergentes y el uso de los algoritmos de selección basados en métodos de filtrado permitirán reducir considerablemente las dimensiones de los conjuntos de atributos utilizados para la clasificación manteniendo o mejorando la eficacia.

Capítulo 5

Resultados experimentales

En este capítulo se presentan los resultados experimentales obtenidos al aplicar el método de clasificación de imágenes propuesto en esta tesis sobre varias colecciones de grafos obtenidas de una base de datos sintética y una base de datos de esqueletos estructurales de imágenes. Se presenta una comparación entre los resultados obtenidos por nuestro método y los reportados en la literatura sobre las bases de datos sintéticas.

5.1. Bases de datos sintéticas

El enfoque de los algoritmos para la MSFA fue probado en tareas de clasificación de imágenes utilizando varias bases de datos bien conocidas. Con el objetivo de comparar los resultados obtenidos con nuestra propuesta respecto a los reportados por Acosta-Mendoza *et al.* (2012a), la colección de imágenes que se utilizará en esta tesis es la misma utilizada en (Acosta-Mendoza *et al.*, 2012a). Esta colección de imágenes está compuesta por 700 imágenes obtenidas mediante el Generador aleatorio de imágenes de Coenen[1], la cual fue dividida en seis sub-colecciones con diferentes cantidades de imágenes (desde

[1] www.csc.liv.ac.uk/∼frans/KDD/Software/ImageGenerator/imageGenerator.html

200 hasta 700 con un incremento de 100). Estas imágenes están divididas en dos clases "landscape" y "seascape", en correspondencia con su contenido (ver figura 5.1). Para la representación de estas imágenes en forma de grafos se utilizó el proceso descrito en la sección 4.1 con un nivel máximo de profundidad o límite de subdivisiones igual a 4. Se utilizó 4 como nivel de profundidad porque fue con los que se alcanzaron los mejores resultados comparados con el uso de 3 y 5, y 4 fue el que se propuso y utilizó en (Acosta-Mendoza *et al.*, 2012a).

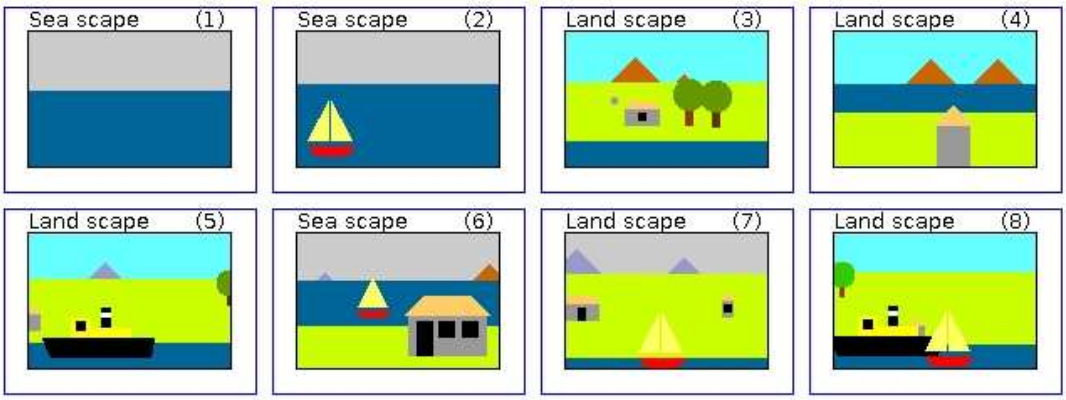

Figura 5.1: Ejemplo de imágenes de la colección obtenido de (Acosta-Mendoza *et al.*, 2012a) usando el generador aleatorio de imágenes de Coenen.

Para la confección de las imágenes de estas colecciones no se utilizan texturas, los colores son distribuidos de forma homogénea en cada objeto que se dibuja y es muy común que se encuentren los mismos colores en imágenes de las dos clases. Por esta razón, se considera que esta base de datos es viable para el problema de clasificación utilizando subgrafos frecuentes debido a que la información espacial entre los objetos se debe tener en cuenta en en proceso de clasificación. Ejemplos de esto se pueden observar en la figura 5.1, donde existen imágenes de vista terrestre que contienen elementos marítimos (i.e. agua, barcos, veleros, etc.), y de este mismo modo existen imágenes de vistas marítimas que contienen elementos terrestres. Este hecho nos lleva a pensar que solo el uso de

histogramas de colores no sería viable para atacar el problema sin tener información espacial de los objetos de las imágenes.

En la tabla 5.1 se muestran las características de las bases de datos sintéticas. En la primera columna de esta tabla se muestran los identificadores de las colecciones de grafos (imágenes), en la segunda y tercera columna se muestran las cantidades de etiquetas para los vértices y las aristas, respectivamente. En la cuarta y quinta columna se muestran los tamaños promedios de las colecciones en términos de la cantidad de aristas y de vértices, respectivamente. Finalmente, en las dos últimas columnas se muestran las cantidades de imágenes por clase que conforman las colecciones.

Tabla 5.1: Características de las bases de datos sintéticas.

Colección	$\|L_V\|$	$\|L_E\|$	P_E	P_V	$\|C_1\|$	$\|C_2\|$
Coenen–200			45	24	76	124
Coenen–300			44	24	114	186
Coenen–400	18	24	45	25	159	241
Coenen–500			46	25	200	300
Coenen–600			47	25	241	359
Coenen–700			47	26	283	417

Los atributos de las imágenes de la colección se extrajeron mediante el proceso descrito en la sección 4.2 y se aplicó el proceso de reducción descrito en la sección 4.3.

5.1.1. Resultados experimentales

En esta sección se presenta una comparación experimental entre el método propuesto en esta tesis y el método de clasificación propuesto en (Acosta-Mendoza *et al.*, 2012a) que utiliza todos los SFA obtenidos por VEAM, ya que es el que mejores resultados, en términos del porcentaje de aciertos (*accuracy*), reporta de acuerdo a la comparación hecha en (Acosta-Mendoza *et al.*, 2012a). En dicha comparación intervienen los resultados de la clasificación alcanzados utilizando gSpan (Yan & Huan, 2002) como representante

de los algoritmos exactos, y los algoritmos para la MSFA APGM (Jia *et al.*, 2011) y VEAM (Acosta-Mendoza *et al.*, 2012a).

En nuestros experimentos se utilizaron varios clasificadores para evaluar la propuesta. Dichos clasificadores se escogieron por ser de diferente naturaleza: máquinas de soporte vectorial (libSVM), red bayesiana (BayesNet), árboles de decisión (J48graft), basados en reglas (Decision table), boosting (AdaBoostM1), y regresión (ClassificationViaRegression). Todos estos clasificadores, con la excepción de SVM, fueron tomados de Weka v3.6.6 (Hall *et al.*, 2009) utilizando los parámetros por defecto. Para SVM se utilizó la misma libSVM[2] usada en (Acosta-Mendoza *et al.*, 2012a) pero en este caso solamente se usa el kernel lineal dado que los mejores resultados fueron reportados con dicho kernel. Para evaluar los resultados de la clasificación se utilizó validación cruzada con un tamaño de ventana igual a 10 (en inglés, *10 fold cross-validation*) sobre dichos vectores de atributos.

Como se mencionó anteriormente se utiliza VEAM para calcular los SFA de la colección de imágenes representadas en forma de grafos. Para dicho algoritmo se requiere la especificación de dos parámetros: el umbral de soporte ($0 < \delta \leq 1$), y el umbral de isomorfismo ($0 < \tau \leq 1$). Se utilizaron siete valores de δ, desde 0.20 hasta 0.50 con un incremento de 0.05, y se utilizó τ =0.4 de acuerdo con Acosta-Mendoza *et al.* (2012a).

Debido a que la diferencia fundamental entre el método propuesto y el método contra el que comparamos está en el módulo de reducción, en las secciones 5.1.1 y 5.1.1 se muestran experimentos con cada variante propuesta para este módulo: utilizando patrones emergentes y selectores de atributos, respectivamente.

[2]www.csie.ntu.edu.tw/∼cjlin/libsvm.

Utilizando patrones emergentes

En esta sección se presenta una comparación entre el método de clasificación propuesto en (Acosta-Mendoza *et al.*, 2012a) y el método propuesto en esta tesis utilizando patrones emergentes para la reducción de la dimensión del conjunto de patrones usados como atributos en la clasificación. Esta comparación se hace utilizando al algoritmo VEAM para la extracción de los patrones de las colecciones de grafos detalladas en la sección 5.1 utilizando $\tau = 0{,}4$ y varios valores de δ. Se utilizaron diferentes clasificadores para mostrar que los resultados no dependen del uso de un clasificador en particular.

Del conjunto de SFA obtenidos mediante VEAM se calculan los patrones emergentes que representarán cada clase de la colección usando como umbral para seleccionarlos a γ =0.3, 0.4 y 0.5. En nuestros experimentos se utilizaron varios valores de γ (i.e. 0.1, 0.2, 0.3, 0.4, 0.5 y 0.6) pero los mejores resultados fueron alcanzados con 0.3, 0.4 y 0.5. Finalmente, utilizando estos patrones emergentes se construyen los vectores de atributos que serán utilizados por los clasificadores.

En la tabla 5.2 y en el resto de la tesis, los patrones emergentes con γ =0.3, 0.4 y 0.5 son representados como "E(0.3)", "E(0.4)" y "E(0.5)", respectivamente. La primera y la segunda columna de esta tabla muestran el nombre de la colección de imágenes utilizada y el valor del umbral de soporte respectivamente. Las otras cuatro columnas consecutivas muestran el número de patrones usados como atributos para la clasificación. La primera de esas cuatro columnas indica el número de patrones calculados por VEAM, y las otras tres columnas indican el número de patrones emergentes calculados utilizando γ =0.3, 0.4 y 0.5 a partir de los patrones calculados por VEAM.

El primer resultado de nuestra propuesta es la reducción de la cantidad de patrones usados como atributos (patrones emergentes) para representar las imágenes de la colección. Como se puede observar en la tabla 5.2, la dimensionalidad de los vectores

Tabla 5.2: Número de patrones utilizados como atributos en el proceso de clasificación.

Colección	**Soporte (δ)**	**Todos los patrones**	**Patrones emergentes**		
			E(0.3)	E(0.4)	E(0.5)
Coenen–200	20 %	340	256	133	74
	25 %	143	131	94	48
	30 %	72	61	49	33
	35 %	35	24	26	23
	40 %	18	11	12	15
	45 %	13	6	7	10
	50 %	11	4	5	8
Coenen–300	20 %	374	299	139	76
	25 %	154	141	95	51
	30 %	69	57	52	37
	35 %	31	21	25	25
	40 %	17	8	11	14
	45 %	13	4	7	10
	50 %	11	2	5	8
Coenen–400	20 %	433	292	150	78
	25 %	203	173	102	57
	30 %	79	60	55	45
	35 %	31	17	25	23
	40 %	18	6	12	14
	45 %	13	2	7	10
	50 %	12	1	6	9
Coenen–500	20 %	453	357	140	70
	25 %	238	211	101	56
	30 %	93	71	57	43
	35 %	35	18	26	22
	40 %	19	7	12	14
	45 %	13	2	6	9
	50 %	12	1	5	8
Coenen–600	20 %	498	321	149	69
	25 %	257	222	107	56
	30 %	99	73	62	42
	35 %	35	18	27	21
	40 %	20	8	13	14
	45 %	13	2	6	8
	50 %	12	1	5	7
Coenen–700	20 %	864	326	143	65
	25 %	321	230	109	54
	30 %	116	84	65	41
	35 %	44	23	33	23
	40 %	23	10	14	15
	45 %	13	2	4	8
	50 %	12	1	3	7

de atributos es reducida drásticamente si se representan las imágenes con los patrones emergentes. Basados en esa reducción, se puede afirmar que independientemente del clasificador utilizado para la clasificación de imágenes, dicha representación permite una mejora en la eficiencia del clasificador. Nótese que esta mejora en nuestra propuesta es aún mayor a para un umbral de soporte grande (i.e. $\delta = 50\,\%$), sin embargo, se debe

utilizar $\gamma > \delta$ para obtener patrones emergentes de utilidad.

Tabla 5.3: Resultados de la clasificación, en términos del porciento de aciertos (*accuracy*), alcanzados utilizando diferentes clasificadores en varias colecciones de imágenes con y sin el uso de los patrones emergentes calculados con γ =0.3, 0.4 y 0.5.

		SVM (kernel lineal)				**Bayes-Net**			
Colección	δ	Todos	E(0.3)	E(0.4)	E(0.5)	Todos	E(0.3)	E(0.4)	E(0.5)
Coenen–700	20 %	95.86 %	96.43 %	**96.57** %	95.00 %	90.29 %	90.43 %	**93.86** %	92.29 %
Coenen–600	20 %	95.83 %	96.00 %	**96.50** %	95.33 %	91.17 %	89.00 %	**92.83** %	90.00 %
Coenen–500	25 %	**97.20** %	96.80 %	96.80 %	96.80 %	90.90 %	89.60 %	**93.60** %	**93.60** %
Coenen–400	20 %	96.75 %	98.00 %	**98.25** %	96.25 %	93.25 %	90.75 %	**94.50** %	92.50 %
Coenen–300	20 %	97.33 %	**98.00** %	97.00 %	95.67 %	88.33 %	88.00 %	**94.00** %	92.00 %
Coenen–200	20 %	**97.50** %	97.00 %	95.60 %	92.00 %	88.00 %	88.50 %	**94.00** %	93.00 %
Promedio		96.75 %	**97.04** %	96.79 %	95.18 %	90.27 %	89.38 %	**93.80** %	92.23 %

		AdaBoostM1				**Regression**			
Colección	δ	Todos	E(0.3)	E(0.4)	E(0.5)	Todos	E(0.3)	E(0.4)	E(0.5)
Coenen–700	20 %	94.14 %	93.86 %	94.00 %	**94.14** %	96.57 %	96.14 %	**96.57** %	93.71 %
Coenen–600	20 %	92.67 %	**94.00** %	93.00 %	93.33 %	95.60 %	96.50 %	**96.83** %	95.83 %
Coenen–500	25 %	94.80 %	**95.20** %	94.80 %	93.80 %	95.60 %	95.80 %	96.00 %	**96.60** %
Coenen–400	20 %	94.50 %	**95.50** %	95.00 %	94.25 %	96.75 %	**97.00** %	95.75 %	95.25 %
Coenen–300	20 %	95.00 %	94.67 %	**96.33** %	93.00 %	94.00 %	94.33 %	**95.00** %	**95.00** %
Coenen–200	20 %	**94.00** %	92.50 %	93.50 %	92.50 %	94.50 %	**95.00** %	93.50 %	91.00 %
Promedio		94.19 %	94.29 %	**94.44** %	93.50 %	95.49 %	**95.80** %	95.61 %	94.57 %

		Decision-Table				**J48graft**			
Colección	δ	Todos	E(0.3)	E(0.4)	E(0.5)	Todos	E(0.3)	E(0.4)	E(0.5)
Coenen–700	20 %	92.71 %	92.71 %	93.29 %	**94.14** %	**96.14** %	96.00 %	94.14 %	95.29 %
Coenen–600	20 %	**95.17** %	94.67 %	93.83 %	94.83 %	95.67 %	**96.00** %	94.67 %	94.83 %
Coenen–500	25 %	**95.60** %	91.60 %	94.40 %	94.80 %	95.80 %	**96.60** %	**96.60** %	96.40 %
Coenen–400	20 %	93.25 %	96.25 %	**97.00** %	95.25 %	94.50 %	**97.00** %	96.00 %	94.75 %
Coenen–300	20 %	94.67 %	**94.67** %	**94.67** %	93.00 %	94.33 %	94.00 %	**94.33** %	**94.33** %
Coenen–200	20 %	92.50 %	93.00 %	93.00 %	**94.50** %	91.50 %	93.00 %	**94.50** %	93.00 %
Promedio		93.98 %	93.82 %	94.37 %	**94.42** %	94.66 %	**95.43** %	95.04 %	94.77 %

En las tablas 5.3 y 5.4, y en el resto de esta tesis, "Todos" representa el método que utiliza todos los SFAs como atributos para la clasificación. En estas tablas se resumen los resultados de los experimtentos de esta subsección donde se evalúa el porciento de aciertos y la métrica F-measure, respectivamente. La primera y la segunda columna de estas tablas muestran el identificador de la colección y el valor del umbral de mínimo soporte, respectivamente. Las otras cuatro columnas consecutivas muestran el resultado de la clasificación (*accuracy* o *F-measure*) para el clasificador especificado en la parte superior de estas columnas, usando todos y solamente los patrones emergentes calcula-

Tabla 5.4: Resultados de la clasificación (F-measure) alcanzados utilizando diferentes clasificadores en varias colecciones de imágenes con y sin el uso de los patrones emergentes calculados con γ =0.3, 0.4 y 0.5.

		SVM (kernel lineal)				**Bayes-Net**			
Colección	δ	Todos	E(0.3)	E(0.4)	E(0.5)	Todos	E(0.3)	E(0.4)	E(0.5)
Coenen–700	20 %	94.70 %	95.17 %	**95.92** %	94.02 %	89.18 %	89.36 %	**92.30** %	89.98 %
Coenen–600	20 %	94.82 %	95.05 %	**95.62** %	94.03 %	89.29 %	87.10 %	**90.89** %	86.74 %
Coenen–500	25 %	96.24 %	96.45 %	**96.70** %	90.48 %	88.88 %	86.51 %	**92.53** %	90.19 %
Coenen–400	20 %	97.80 %	**98.16** %	97.76 %	95.56 %	91.77 %	89.16 %	**92.93** %	89.29 %
Coenen–300	20 %	95.07 %	**96.39** %	95.55 %	92.39 %	86.12 %	85.69 %	**91.80** %	88.46 %
Coenen–200	20 %	95.75 %	**96.55** %	95.21 %	88.29 %	85.14 %	85.94 %	**91.73** %	89.58 %
Promedio		95.73 %	**96.30** %	96.13 %	92.46 %	88.40 %	87.29 %	**92.03** %	89.04 %

		AdaBoostM1				**Regression**			
Colección	δ	Todos	E(0.3)	E(0.4)	E(0.5)	Todos	E(0.3)	E(0.4)	E(0.5)
Coenen–700	20 %	92.90 %	92.65 %	92.80 %	**92.96** %	95.74 %	95.21 %	**95.76** %	92.23 %
Coenen–600	20 %	90.96 %	**92.73** %	91.61 %	91.99 %	94.38 %	95.60 %	**95.98** %	94.79 %
Coenen–500	25 %	93.63 %	92.05 %	**93.63** %	90.17 %	94.49 %	94.79 %	**94.98** %	91.25 %
Coenen–400	20 %	93.09 %	**94.34** %	93.68 %	92.76 %	95.98 %	**96.15** %	94.59 %	94.01 %
Coenen–300	20 %	93.49 %	92.94 %	**95.15** %	90.73 %	91.62 %	92.36 %	**93.44** %	93.35 %
Coenen–200	20 %	**91.78** %	89.74 %	91.01 %	89.66 %	92.61 %	**93.52** %	90.91 %	88.11 %
Promedio		92.64 %	92.41 %	**92.98** %	91.55 %	94.14 %	**94.61** %	94.28 %	92.29 %

		Decision-Table				**J48graft**			
Colección	δ	Todos	E(0.3)	E(0.4)	E(0.5)	Todos	E(0.3)	E(0.4)	E(0.5)
Coenen–700	20 %	91.29 %	91.25 %	92.03 %	**92.91** %	**95.35** %	95.13 %	95.26 %	94.21 %
Coenen–600	20 %	**94.17** %	93.34 %	92.26 %	93.65 %	94.63 %	**94.97** %	93.32 %	93.46 %
Coenen–500	25 %	**94.58** %	93.39 %	93.17 %	89.22 %	94.75 %	94.97 %	**95.71** %	92.69 %
Coenen–400	20 %	91.87 %	95.30 %	**96.23** %	93.77 %	93.12 %	**96.32** %	95.08 %	93.34 %
Coenen–300	20 %	92.76 %	**93.15** %	92.92 %	90.62 %	92.60 %	92.27 %	**92.70** %	92.42 %
Coenen–200	20 %	89.92 %	91.17 %	91.17 %	**92.03** %	87.79 %	89.99 %	**92.75** %	90.38 %
Promedio		92.43 %	92.93 %	**92.96** %	92.02 %	93.04 %	93.94 %	**94.14** %	92.75 %

dos con γ =0.3, 0.4 y 0.5, respectivamente. De esta misma manera se distribuyen las últimas cuatro columnas pero para un clasificador diferente. Nótese que estas tablas están divididas en tres sub-tablas, cada una muestra los experimentos realizados con dos clasificadores diferentes utilizando los patrones emergentes.

Como se puede observar en las tablas 5.3 y 5.4, los resultados alcanzados utilizando SVM con nuestra propuesta son mejores que los resultados reportados en (Acosta-Mendoza *et al.*, 2012a). En esta colección de imágenes, con el método propuesto en (Acosta-Mendoza *et al.*, 2012a) se obtiene un promedio del porciento de aciertos de 96.75 % y se obtiene un promedio de F-measure de 95.73 % mientras que con nuestro

método utilizando patrones emergentes se obtienen promedios de 97.04 % y 96.79 % en el porciento de aciertos y promedios de 96.30 % y 96.16 % en F-measure, siendo esta mejora mucho mejor en los resultados del cálculo de F-measure. Es importante señalar que con nuestra propuesta se utiliza menos del 50 % de los patrones (atributos) usados en (Acosta-Mendoza *et al.*, 2012a). Por otro lado, los resultados obtenidos con los demás clasificadores utilizando los patrones emergentes fueron también mejores (en un rango de 0.12 % a 3.53 % en el porciento de aciertos y de 0.34 % a 3.63 % en el F-measure) que los obtenidos utilizando todos los SFAs como atributos, y además, nuestra propuesta utiliza un conjunto mucho menor de atributos. Con estos experimentos se puede observar que los resultados obtenidos usando γ =0.3 y γ =0.5, en general, son peores que los obtenidos usando γ =0.4. En esta comparación realizada tomando en cuenta los resultados del porciento de aciertos y de F-measure, se puede observar que E(0.4) logra superar en un 25.69 % los resultados de E(0.3), E(0.5) y al usar todos los patrones como atributos, seguido por E(0.3) que supera al resto en un 13.19 %, mientras que E(0.5) y el uso de todos los patrones solo logran superar al resto en un 6.94 %. Por esta razón, se propone el uso de γ =0.4, con el cual se obtienen mejores resultados. Estos resultados muestran que utilizar los patrones emergentes como atributos para la clasificación de imágenes permite obtener mejores resultados de clasificación y una considerable reducción de la dimensionalidad del problema.

Adicionalmente, en la tabla 5.5 se presenta una comparación utilizando pruebas estadísticas realizadas entre los resultados obtenidos por los clasificadores con y sin el uso de los patrones emergentes. Esta comparación se compone por la comparación por pares entre nuestra propuesta utilizando patrones emergentes (para los diferentes valores de γ) y el método que utiliza todos los SFAs. Para dicha comparación, se utilizaron tres pruebas de significancia estadística presentadas en (García & Herrera, 2008): Holm (Holm,

Tabla 5.5: Pruebas de significancia estadística para los diferentes clasificadores en varias colecciones de imágenes con y sin el uso de los patrones emergentes. Cada celda de la tabla indica cuál opción fue estadísticamente mejor (entre los comparados en la columna 1), "–" indica que no hubo diferencia estadísticamente significativa.

(a) Resultados obtenidos con α =0.05

Resultados obtenidos utilizando las pruebas Holm y Hommel						
Clasificador	SVM	BayesNet	AdaBoostM1	J48graft	Regression	Decision-Table
Todos vs. E(0.3)	–	–	–	–	–	–
Todos vs. E(0.4)	–	E(0.4)	–	–	–	–
Todos vs. E(0.5)	–	–	–	–	–	–
E(0.3) vs. E(0.4)	–	E(0.4)	–	–	–	–
E(0.3) vs. E(0.5)	E(0.3)	–	–	–	–	–
E(0.4) vs. E(0.5)	E(0.4)	–	–	–	–	–

Resultados obtenidos utilizando la prueba Bonferroni-Dunn						
Clasificador	SVM	BayesNet	AdaBoostM1	J48graft	Regression	Decision-Table
Todos vs. E(0.3)	–	–	–	–	–	–
Todos vs. E(0.4)	–	E(0.4)	–	–	–	–
Todos vs. E(0.5)	–	–	–	–	–	–
E(0.3) vs. E(0.4)	–	E(0.4)	–	–	–	–
E(0.3) vs. E(0.5)	E(0.3)	–	–	–	–	–
E(0.4) vs. E(0.5)	E(0.4)	–	–	–	–	–

(b) Resultados obtenidos con α =0.10

Resultados obtenidos utilizando la prueba Holm						
Clasificador	SVM	BayesNet	AdaBoostM1	J48graft	Regression	Decision-Table
Todos vs. E(0.3)	–	–	–	–	–	
Todos vs. E(0.4)	–	E(0.4)	–	–	–	–
Todos vs. E(0.5)	–	–	–	–	–	
E(0.3) vs. E(0.4)	–	E(0.4)	–	–	–	–
E(0.3) vs. E(0.5)	E(0.3)	E(0.5)	–	–	–	
E(0.4) vs. E(0.5)	E(0.4)	–	–	–	–	–

Resultados obtenidos utilizando la prueba Hommel						
Clasificador	SVM	BayesNet	AdaBoostM1	J48graft	Regression	Decision-Table
Todos vs. E(0.3)	–	–	–	–	–	
Todos vs. E(0.4)	–	E(0.4)	–	–	–	–
Todos vs. E(0.5)	–	–	–	–	–	
E(0.3) vs. E(0.4)	–	E(0.4)	–	–	–	–
E(0.3) vs. E(0.5)	E(0.3)	–	–	–	–	–
E(0.4) vs. E(0.5)	E(0.4)	–	–	–	–	–

Resultados obtenidos utilizando la prueba Bonferroni-Dunn						
Clasificador	SVM	BayesNet	AdaBoostM1	J48graft	Regression	Decision-Table
Todos vs. E(0.3)	–	–	–	–	–	
Todos vs. E(0.4)	–	E(0.4)	–	–	–	–
Todos vs. E(0.5)	–	–	–	–	–	
E(0.3) vs. E(0.4)	–	E(0.4)	–	–	–	–
E(0.3) vs. E(0.5)	E(0.3)	–	–	–	–	–
E(0.4) vs. E(0.5)	E(0.4)	–	–	–	–	–

1979), Hommel (Hommel, 1988), y Bonferroni-Dunn (Simes, 1986). Los valores del α utilizado fueron 0.05 y 0.10, y los resultados de estas pruebas se resumen en la tabla 5.5 respectivamente. En la primera columna de la tabla 5.5 se especifican los métodos que se comparan. Las otras columnas indican cuál método es significativamente mejor que otro; el símbolo "–" indica que no se tienen diferencias estadísticamentes significativas entre los resultados de ambos métodos.

Como se puede observar en la tabla 5.5, el uso de los patrones emergentes con γ =0.4 (E(0.4)) es la mejor opción en el 5.56 % de los resultados, "E(0.4)" es significativamente mejor que "Todos", mientras que en el 92.59 % de los resultados no existe diferencias estadísticamente significativas. Es importante señalar que con nuestra propuesta se reduce notablemente el conjunto de patrones a usar como atributos en la clasificación. Además, en el 16.67 % de los resultados, E(0.4) es significativamente mejor que E(0.3) y en ese mismo porciento E(0.4) es mejor que E(0.5), mientras que E(0.3) y E(0.5) no son significativamente mejores que E(0.4) en ningún caso.

En general, los resultados muestran que nuestra propuesta utilizando los patrones emergentes como atributos para representar las imágenes de la colección ofrece mejores resultados que utilizando todos los patrones calculados por VEAM.

Utilizando selección de atributos

En esta sección se presenta una comparación entre el método de clasificación propuesto en (Acosta-Mendoza *et al.*, 2012a) y el método propuesto en esta tesis utilizando varios algoritmos de selección de atributos de tipo filtrado (ganancia de información, chi-cuadrado y cociente de evaluación de la ganancia de información) para la reducción del conjunto de atributos a usar en la clasificación. En esta comparación se utiliza el algoritmo VEAM para la extracción de los patrones de las colecciones de grafos obtenidas

de las colecciones de imágenes detalladas en la sección 5.1. En nuestros experimentos se utilizaron diferentes clasificadores para mostrar que los resultados no dependen del uso de un clasificador en particular.

A partir de los SFAs calculados por VEAM se construyen los vectores de atributos que representarán a las imágenes. Posteriormente se obtiene un subconjunto de atributos utilizando varios criterios de selección: ganancia de información (***I****nformation* ***G****ain* IG), chi-cuadrado (CHI-Q) y el cociente de evaluación de la ganancia de información de los atributos (***G****ain* ***R****atio* ***A****ttribute* ***E****valuation* GRAE). En nuestros experimentos se realizó una búsqueda exhaustiva de la cantidad de atributos adecuada para una buena clasificación tomando como semilla la cantidad de patrones emergentes calculados con γ =0.4 mostrados en la tabla 5.2. Partiendo de esta semilla se utilizó un rango de cantidades de atributos entre 50 y 200 con un incremento de 15 para cada algoritmo de selección, en los cuales se identificaron los mejores resultados de clasificación (*accuracy* y *F-measure*) para ser usados en las comparaciones realizadas en estos experimentos. Finalmente, utilizando este subconjunto de atributos se construyen los vectores de atributos que serán usados por los clasificadores.

En la tabla 5.6, la primera y la segunda columna muestran el nombre de la colección de imágenes utilizada y el valor del umbral de mínimo soporte, respectivamente. Las otras cuatro columnas consecutivas muestran el número de atributos usados para la clasificación. La primera de esas cuatro columnas indica el número de patrones calculados por VEAM, y las otras tres columnas indican el número de atributos seleccionados, a partir de los patrones calculados por VEAM, utilizando IG, CHI-Q y GRAE como algoritmos de selección.

En las tablas 5.7 y 5.8 se resumen los resultados, en términos del porciento de aciertos (*accuracy*) y F-measure, de los experimentos de esta subsección. La primera y la

Tabla 5.6: Número de patrones utilizados como atributos en el proceso de clasificación.

Colección	**Soporte (δ)**	**Todos los atributos**	**Atributos seleccionados**			**Clasificador**
			IG	CHI-Q	GRAE	
Coenen–200	20 %	340	200	133	200	SVM
			110	80	95	Bayes-Net
			155	125	155	AdaBostM1
			170	125	95	Regression
			155	155	170	Decision-Table
			133	110	110	J48graft
Coenen–300	20 %	374	140	110	200	SVM
			125	110	125	Bayes-Net
			140	50	140	AdaBostM1
			125	139	139	Regression
			170	50	155	Decision-Table
			125	140	155	J48graft
Coenen–400	20 %	433	200	185	185	SVM
			125	80	65	Bayes-Net
			140	150	150	AdaBostM1
			200	185	185	Regression
			200	170	50	Decision-Table
			125	110	140	J48graft
Coenen–500	25 %	238	140	155	155	SVM
			65	50	50	Bayes-Net
			155	101	110	AdaBostM1
			170	170	200	Regression
			50	50	170	Decision-Table
			95	101	95	J48graft
Coenen–600	20 %	498	185	95	200	SVM
			155	65	50	Bayes-Net
			95	80	125	AdaBostM1
			65	155	200	Regression
			50	125	125	Decision-Table
			65	200	125	J48graft
Coenen–700	20 %	864	200	200	200	SVM
			65	80	65	Bayes-Net
			65	65	155	AdaBostM1
			140	140	143	Regression
			65	110	50	Decision-Table
			200	200	200	J48graft

segunda columna de estas tablas muestran el nombre de la colección y el valor del umbral de soporte, respectivamente. Las otras cuatro columnas consecutivas muestran el resultado de la clasificación (*accuracy* o *F-measure*) para el clasificador especificado en la parte superior de estas columnas, usando todos los atributos y solamente los atributos seleccionados mediante los algoritmos de selección de atributos IG, CHI-Q y GRE, respectivamente. De esta misma manera se distribuyen las últimas cuatro columnas pero para un clasificador diferente.

Tabla 5.7: Resultados de la clasificación (accuracy) alcanzados utilizando diferentes clasificadores en varias colecciones de imágenes con y sin el uso del algoritmos de selección de atributos.

		SVM (kernel lineal)				**Bayes-Net**			
Colección	δ	Todos	IG	CHI-Q	GRAE	Todos	IG	CHI-Q	GRAE
Coenen–700	20 %	95.86 %	96.29 %	**96.43** %	**96.43** %	90.29 %	**94.57** %	**94.57** %	**94.57** %
Coenen–600	20 %	95.83 %	**96.50** %	96.17 %	**96.50** %	91.17 %	**94.83** %	94.67 %	94.50 %
Coenen–500	25 %	97.20 %	97.40 %	**97.60** %	**97.60** %	90.60 %	**94.80** %	**94.80** %	**94.80** %
Coenen–400	20 %	96.75 %	96.50 %	96.75 %	**97.25** %	93.25 %	95.25 %	**95.50** %	**95.50** %
Coenen–300	20 %	**97.33** %	97.00 %	97.00 %	97.00 %	88.33 %	**95.00** %	**95.00** %	**95.00** %
Coenen–200	20 %	97.50 %	97.00 %	95.50 %	**97.50** %	88.00 %	**94.50** %	**94.50** %	**94.50** %
Promedio		96.75 %	96.78 %	96.58 %	**97.05** %	90.27 %	94.83 %	**94.84** %	94.81 %

		AdaBoostM1				**Regression**			
Colección	δ	Todos	IG	CHI-Q	GRAE	Todos	IG	CHI-Q	GRAE
Coenen–700	20 %	94.14 %	94.29 %	**94.43** %	94.14 %	**96.57** %	96.14 %	96.00 %	96.00 %
Coenen–600	20 %	92.67 %	**94.33** %	94.17 %	93.67 %	95.50 %	94.33 %	96.00 %	**96.33** %
Coenen–500	25 %	94.80 %	**94.80** %	**94.80** %	**94.80** %	95.60 %	96.20 %	96.20 %	**96.40** %
Coenen–400	20 %	94.50 %	94.75 %	94.75 %	**95.25** %	96.75 %	**96.75** %	96.50 %	96.25 %
Coenen–300	20 %	95.00 %	95.00 %	**95.33** %	95.00 %	94.00 %	94.67 %	**95.00** %	**95.00** %
Coenen–200	20 %	94.00 %	94.50 %	**95.00** %	94.50 %	**94.50** %	92.50 %	92.50 %	92.00 %
Promedio		94.19 %	94.61 %	**94.75** %	94.56 %	**95.49** %	95.10 %	95.37 %	95.35 %

		Decision-Table				**J48graft**			
Colección	δ	Todos	IG	CHI-Q	GRAE	Todos	IG	CHI-Q	GRAE
Coenen–700	20 %	92.71 %	**94.57** %	**94.57** %	**94.57** %	96.14 %	96.29 %	**96.43** %	**96.43** %
Coenen–600	20 %	95.17 %	94.50 %	**95.67** %	95.50 %	95.67 %	94.50 %	**96.17** %	96.00 %
Coenen–500	25 %	95.60 %	94.40 %	94.40 %	**96.00** %	95.80 %	**96.60** %	96.40 %	**96.60** %
Coenen–400	20 %	93.25 %	**95.57** %	**95.75** %	94.75 %	94.50 %	**96.00** %	95.75 %	95.50 %
Coenen–300	20 %	94.67 %	**94.67** %	**94.67** %	94.33 %	94.33 %	**96.00** %	94.67 %	95.33 %
Coenen–200	20 %	92.50 %	93.50 %	94.50 %	**95.00** %	91.50 %	**95.00** %	**95.00** %	**95.00** %
Promedio		93.98 %	94.54 %	94.93 %	**95.03** %	94.66 %	95.73 %	95.74 %	**95.81** %

Los resultados alcanzados utilizando selectores de atributos (tipo filtrado) como preprocesamiento permite mejorar el desempeño de la clasificación que se obtiene si se usan todos los patrones (atributos) calculados por VEAM. En las tablas 5.7 y 5.8 se puede observar que en la mayoría de los casos se obtienen mejores resultados de clasificación al aplicar dichos algoritmos de selección de atributos. Es importante notar que en los casos que no se mejora la clasificación (i.e. mediante el clasificador *Classification via regression* y en el caso de F-measure también ocurre mediante SVM), se logran resultados de clasificación no tan alejados de los obtenidos utilizando todos los atributos, pero utilizando un conjunto mucho menor de atributos para la clasificación. Estos resultados

Tabla 5.8: Resultados de la clasificación (F-measure) alcanzados utilizando diferentes clasificadores en varias colecciones de imágenes con y sin el uso del algoritmos de selección de atributos.

Colección	δ	SVM (kernel lineal)				Bayes-Net			
		Todos	IG	CHI-Q	GRAE	Todos	IG	CHI-Q	GRAE
Coenen–700	20 %	94.70 %	**95.40** %	**95.40** %	**95.40** %	89.18 %	**93.73** %	**93.73** %	**93.73** %
Coenen–600	20 %	94.82 %	93.94 %	94.34 %	**95.63** %	89.29 %	**93.92** %	93.43 %	93.54 %
Coenen–500	25 %	96.24 %	96.20 %	**96.72** %	**96.72** %	88.88 %	**93.89** %	**93.89** %	**93.89** %
Coenen–400	20 %	**97.80** %	95.53 %	94.55 %	94.55 %	91.77 %	**94.43** %	94.03 %	94.03 %
Coenen–300	20 %	95.07 %	95.02 %	91.15 %	**95.48** %	86.12 %	**94.04** %	93.18 %	**94.04** %
Coenen–200	20 %	**95.75** %	95.56 %	91.44 %	95.56 %	85.14 %	**92.93** %	92.20 %	92.61 %
Promedio		**95.73** %	95.28 %	93.93 %	95.56 %	88.40 %	**93.82** %	93.41 %	93.64 %

Colección	δ	AdaBoostM1				Regression			
		Todos	IG	CHI-Q	GRAE	Todos	IG	CHI-Q	GRAE
Coenen–700	20 %	92.90 %	**93.73** %	93.31 %	92.92 %	**95.74** %	95.25 %	94.84 %	94.62 %
Coenen–600	20 %	90.96 %	**93.54** %	93.00 %	92.38 %	94.38 %	93.24 %	94.97 %	**95.41** %
Coenen–500	25 %	93.63 %	88.88 %	**93.63** %	**93.63** %	94.49 %	95.24 %	95.24 %	**95.46** %
Coenen–400	20 %	93.09 %	93.42 %	**93.47** %	93.34 %	**95.98** %	95.90 %	95.55 %	95.19 %
Coenen–300	20 %	93.49 %	**93.67** %	93.29 %	93.54 %	91.62 %	92.84 %	**93.42** %	93.31 %
Coenen–200	20 %	91.78 %	90.65 %	**92.79** %	92.20 %	**92.61** %	89.19 %	90.09 %	89.44 %
Promedio		92.64 %	92.32 %	**93.25** %	93.00 %	**94.14** %	93.61 %	94.02 %	93.91 %

Colección	δ	Decision-Table				J48graft			
		Todos	IG	CHI-Q	GRAE	Todos	IG	CHI-Q	GRAE
Coenen–700	20 %	91.29 %	93.47 %	93.43 %	**93.70** %	95.35 %	95.41 %	**95.58** %	**95.58** %
Coenen–600	20 %	94.17 %	93.40 %	**94.70** %	94.38 %	94.63 %	93.93 %	**95.28** %	95.04 %
Coenen–500	25 %	94.58 %	93.27 %	93.27 %	**95.03** %	94.75 %	**95.77** %	94.69 %	95.73 %
Coenen–400	20 %	91.87 %	**94.57** %	**94.57** %	92.95 %	93.12 %	**94.91** %	94.64 %	94.36 %
Coenen–300	20 %	**92.76** %	92.64 %	92.66 %	90.46 %	92.60 %	**94.79** %	93.05 %	93.96 %
Coenen–200	20 %	89.92 %	90.46 %	90.46 %	**92.48** %	87.79 %	**93.09** %	**93.09** %	**93.09** %
Promedio		92.43 %	92.97 %	**93.18** %	93.17 %	93.04 %	**94.65** %	94.39 %	94.63 %

muestran que utilizar un subconjunto de atributos, seleccionados mediante la aplicación de los algoritmos de selección de atributos antes mencionados se pueden obtener mejores resultados de clasificación de imágenes. En esta comparación realizada tomando en cuenta los resultados del porciento de aciertos y de F-measure, se puede observar que el algoritmo de selección que mejores resultados obtiene respecto al resto de algoritmos de selección y al uso de todos los atributos es GRAE logrando superarlos en un 25.00 %, seguido por CHI-Q que supera al resto en un 21.53 %, mientras que IG los supera en un 13.89 % y el uso de todos los patrones solo logra superarlos en un 6.25 %.

Adicionalmente, en la tabla 5.9 se presenta una comparación utilizando pruebas de

Tabla 5.9: Pruebas de significancia estadística para los diferentes clasificadores en varias colecciones de imágenes con y sin el uso de los algoritmos de selección de atributos. Cada celda de la tabla indica cuál opción fue estadísticamente mejor (entre los comparados en la columna 1), "-" indica que no hubo diferencia estadísticamente significativa.

(a) Resultados obtenidos con α =0.05

Resultados obtenidos utilizando las pruebas Holm y Hommel						
Clasificador	SVM	BayesNet	AdaBoostM1	Regression	Decision-Table	J48graft
Todos vs. GRAE	–	GRAE	–	–	–	GRAE
Todos vs. CHI-Q	–	CHI-Q	CHI-Q	–	–	CHI-Q
Todos vs. IG	–	IG	IG	–	–	IG
IG vs. GRAE	–	–	–	–	–	–
CHI-Q vs. GRAE	–	–	–	–	–	–
IG vs. CHI-Q	–	–	–	–	–	–

Resultados obtenidos utilizando la prueba Bonferroni-Dunn						
Clasificador	SVM	BayesNet	AdaBoostM1	Regression	Decision-Table	J48graft
Todos vs. GRAE	–	GRAE	–	–	–	GRAE
Todos vs. CHI-Q	–	CHI-Q	CHI-Q	–	–	CHI-Q
Todos vs. IG	–	IG	–	–	–	IG
IG vs. GRAE	–	–	–	–	–	–
CHI-Q vs. GRAE	–	–	–	–	–	–
IG vs. CHI-Q	–	–	–	–	–	–

(b) Resultados obtenidos con α =0.10

Resultados obtenidos utilizando la prueba Holm						
Clasificador	SVM	BayesNet	AdaBoostM1	Regression	Decision-Table	J48graft
Todos vs. GRAE	–	GRAE	–	–	–	GRAE
Todos vs. CHI-Q	–	CHI-Q	CHI-Q	–	–	CHI-Q
Todos vs. IG	–	IG	IG	–	–	IG
IG vs. GRAE	GRAE	–	–	–	–	–
CHI-Q vs. GRAE	–	–	–	–	–	–
IG vs. CHI-Q	–	–	–	–	–	–

Resultados obtenidos utilizando la prueba Hommel						
Clasificador	SVM	BayesNet	AdaBoostM1	Regression	Decision-Table	J48graft
Todos vs. GRAE	–	GRAE	–	–	–	GRAE
Todos vs. CHI-Q	–	CHI-Q	CHI-Q	–	–	CHI-Q
Todos vs. IG	–	IG	IG	–	–	IG
IG vs. GRAE	–	–	–	–	–	–
CHI-Q vs. GRAE	–	–	–	–	–	–
IG vs. CHI-Q	–	–	–	–	–	–

Resultados obtenidos utilizando la prueba Bonferroni-Dunn						
Clasificador	SVM	BayesNet	AdaBoostM1	Regression	Decision-Table	J48graft
Todos vs. GRAE	–	GRAE	–	–	–	GRAE
Todos vs. CHI-Q	–	CHI-Q	CHI-Q	–	–	CHI-Q
Todos vs. IG	–	IG	IG	–	–	IG
IG vs. GRAE	GRAE	–	–	–	–	–
CHI-Q vs. GRAE	–	–	–	–	–	–
IG vs. CHI-Q	–	–	–	–	–	–

significancia estadística realizadas entre los resultados obtenidos por los clasificadores con y sin el uso de los algoritmos de selección de atributos. Esta comparación se compone por la comparación por pares entre nuestro método utilizando selección de atributos y el método que utiliza todos los atributos. Para dicha comparación, se utilizaron tres pruebas de significancia estadística presentadas en (García & Herrera, 2008): Holm (Holm, 1979), Hommel (Hommel, 1988), y Bonferroni-Dunn (Simes, 1986). Los valores del α utilizados fueron 0.05 y 0.10.

En la primera columna de la tabla 5.9, "Todos" representa el método que utiliza todos los atributos mientras que nuestro método que utiliza un subconjunto de atributos son representados como "IG", "CHI-Q" y "GRAE" respectivamente. Las otras columnas indican qué método es significativamente mejor que otro; el símbolo "–" indica que no existe una diferencia estadísticamente significativa entre los resultados de ambos métodos.

Como se puede observar en la tabla 5.9, con el uso de los algoritmos de selección de atributos se pueden obtener resultados significativamente mejores que utilizando todos los atributos en un 43.52 % y el uso de todos los atributos no fue mejor significativamente que el uso de los selectores en ningún caso de las pruebas realizadas. Por otro lado, en el resto de las pruebas realizadas (56.48 %) no se obtienen resultados con diferencias estadísticamente significativas entre el uso de los selectores de atributos y el uso de todos los atributos. Por este motivo, concluimos que el uso de nuestro método con los algoritmos de selección de atributos es mejor que utilizar todos los atributos no solo en los resultados de la clasificación sino que además, nuestro método cuenta con una reducción del conjunto de atributos superior al 15.50 %.

Por otro lado, la mejor opción de los algoritmos de selección de atributos es GRAE. Esto se debe a que GRAE es mejor significativamente que IG y el uso de todos los atribu-

tos ("Todos") en 5.56 % y 33.33 % de los resultados, respectivamente, mientras que IG y "Todos" no son mejores significativamente que GRAE en ninguno de los casos. Además, CHI-Q no tiene diferencias estadísticamente significativas respecto a IG y GRAE.

En general, los resultados muestran que nuestra propuesta utilizando un subconjunto de los atributos, obtenidos mediante los algoritmos de selección de atributos, para representar las imágenes de la colección ofrece mejores resultados que utilizando todos los atributos en la clasificación.

Patrones emergentes vs. algoritmos de selección de atributos

En esta sección se presenta una comparación entre el método propuesto en esta tesis utilizando patrones emergentes y dicho método utilizando algoritmos de selección de atributos tipo filtrado. Para esta comparación se utilizaron las mejores opciones identificadas en las secciones 5.1.1 y 5.1.1, donde se utilizaron patrones emergentes y algoritmos de selección de atributos, respectivamente. Estas opciones fueron el uso de patrones emergentes con γ =0.4, denotado por E(0.4), y el uso del algoritmo que calcula el cociente de evaluación de la ganancia de información de los atributos para la selección de atributos, denotado por GRAE.

La primera comparación que se realiza en esta sección es respecto a la reducción de patrones que se logra mediante el uso de las dos estrategias (E(0.4) y GRAE) en el método de clasificación propuesto. Dicha reducción se puede observar en la tabla 5.10, donde se muestran las cantidades de atributos que utilizan cada estrategia en la clasificación. Las primeras dos columnas de esta tabla muestran el identificador de cada colección y el valor del umbral de mínimo soporte, respectivamente. Las otras dos columnas consecutivas muestran las cantidades de atributos que utilizan las diferentes estrategias indicadas en la parte superior de dichas columnas. Dichas cantidades de atributos son utilizados por

los clasificadores especificados en la última columna de esta tabla.

Tabla 5.10: Número de patrones utilizados como atributos en el proceso de clasificación.

Colección	**Soporte (δ)**	E(0.4)	GRAE	**Clasificador**
Coenen–200	20 %	133	200	SVM
			95	Bayes-Net
			155	AdaBostM1
			95	Regression
			170	Decision-Table
			110	J48graft
Coenen–300	20 %	139	200	SVM
			125	Bayes-Net
			140	AdaBostM1
			139	Regression
			155	Decision-Table
			155	J48graft
Coenen–400	20 %	150	185	SVM
			65	Bayes-Net
			150	AdaBostM1
			185	Regression
			50	Decision-Table
			140	J48graft
Coenen–500	25 %	140	155	SVM
			50	Bayes-Net
			110	AdaBostM1
			200	Regression
			170	Decision-Table
			95	J48graft
Coenen–600	20 %	149	200	SVM
			50	Bayes-Net
			125	AdaBostM1
			200	Regression
			125	Decision-Table
			125	J48graft
Coenen–700	20 %	143	200	SVM
			65	Bayes-Net
			155	AdaBostM1
			143	Regression
			50	Decision-Table
			200	J48graft

Como se puede observar en la tabla 5.10, el uso de los patrones emergentes E(0.4) permite reducir el conjunto de atributos en un 8.33 % más de las veces que lo reduce el uso del algoritmo de selección de atributos GRAE. GRAE permite reducir significativamente la dimensionalidad del conjunto de patrones, superando a E(0.4), en el uso de algunos clasificadores como: Bayes-Net, Decision-Table, y en algunos casos de J48graft. Por lo que, desde el punto de vista de la reducción del conjunto de atributos, el uso de E(0.4) es

más recomendable que el uso de GRAE. Por otro lado, para encontrar la reducción más adecuada mediante GRAE fue necesaria una búsqueda exhaustiva teniendo en cuenta los resultados de la clasificación en un rango de cantidades de atributos definido. Este hecho hace más costoso el uso de GRAE que el uso de E(0.4).

En la tabla 5.11 se resumen los resultados de la clasificación de las dos mejores variantes para la reducción de la dimensionalidad del conjunto de atributos. Esta tabla está dividida en dos subtablas, las cuales presentan los resultados de tres clasificadores cada una. En la primera y la segunda columna de las subtablas se muestra el identificador de la colección y el valor del umbral de mínimo soporte, respectivamente. Las otras dos columnas consecutivas muestran los resultados de la clasificación (*accuracy* o *F-measure*) para el clasificador especificado en la parte superior de ambas columnas, usando los patrones emergentes E(0.4) y los atributos seleccionados por GRAE, respectivamente. De esta misma forma se distribuyen las demás columnas pero para clasificadores diferentes.

Como se puede observar en la tabla 5.11, en la mayoría de los casos se obtienen mejores resultados de clasificación al aplicar el algoritmo GRAE para la selección de atributos. Mediante el uso de GRAE se logra superar a E(0.4) en el 58.33 % mientras que E(0.4) supera a GRAE en solo el 29.17 % de los experimentos realizados.

Adicionalmente, en la tabla 5.12 se presenta una comparación utilizando pruebas de significancia estadística realizadas entre los resultados obtenidos por los clasificadores con el uso de los patrones emergentes E(0.4) y con el uso del algoritmo de selección de atributos GRAE. Dicha comparación se consiste en la comparación por pares entre las dos estrategias para la reducción de patrones en nuestro método para la clasificación. En esta comparación se utilizaron tres pruebas de significancia estadística presentadas en (García & Herrera, 2008): Holm (Holm, 1979), Hommel (Hommel, 1988) y Bonferroni-Dunn (Simes, 1986). Los valores del α utilizados fueron 0.05 y 0.10.

Tabla 5.11: Resultados de la clasificación alcanzados utilizando diferentes clasificadores en varias colecciones de imágenes realizando la reducción de patrones con E(0.4) y GRAE para la selección de atributos.

(a) Resultados del porcentaje de aciertos (*accuracy*)

		SVM		BayesNet		AdaBoostM1	
Colección	δ	E(0.4)	GRAE	E(0.4)	GRAE	E(0.4)	GRAE
Coenen–700	20 %	**96.57** %	96.43 %	93.86 %	**94.57** %	94.00 %	**94.14** %
Coenen–600	20 %	96.50 %	96.50 %	92.83 %	**94.50** %	93.00 %	**93.67** %
Coenen–500	25 %	96.80 %	**97.60** %	93.60 %	**94.80** %	94.80 %	94.80 %
Coenen–400	20 %	**98.25** %	97.25 %	94.50 %	**95.50** %	95.00 %	**95.25** %
Coenen–300	20 %	97.00 %	97.00 %	94.00 %	**95.00** %	**96.33** %	95.00 %
Coenen–200	20 %	95.60 %	**97.50** %	94.00 %	**94.50** %	93.50 %	**94.50** %
Promedio		96.79 %	**97.05** %	93.80 %	**94.81** %	94.44 %	**94.56** %

		Regression		Decision-Table		J48graft	
Colección	δ	E(0.4)	GRAE	E(0.4)	GRAE	E(0.4)	GRAE
Coenen–700	20 %	**96.57** %	96.00 %	93.29 %	**94.57** %	94.14 %	**96.43** %
Coenen–600	20 %	**96.83** %	96.33 %	93.83 %	**95.50** %	94.67 %	**96.00** %
Coenen–500	25 %	96.00 %	**96.40** %	94.40 %	**96.00** %	96.60 %	96.60 %
Coenen–400	20 %	95.75 %	**96.25** %	**97.00** %	94.75 %	**96.00** %	95.50 %
Coenen–300	20 %	95.00 %	95.00 %	**94.67** %	94.33 %	94.33 %	**95.33** %
Coenen–200	20 %	**93.50** %	92.00 %	93.00 %	**95.00** %	94.50 %	**95.00** %
Promedio		**95.61** %	95.35 %	94.37 %	**95.03** %	95.04 %	**95.81** %

(b) Resultados de F-measure

		SVM (kernel lineal)		BayesNet		AdaBoostM1	
Colección	δ	E(0.4)	GRAE	E(0.4)	GRAE	E(0.4)	GRAE
Coenen–700	20 %	**95.92** %	95.40 %	92.30 %	**93.73** %	92.80 %	**92.92** %
Coenen–600	20 %	95.62 %	95.63 %	90.89 %	**93.54** %	91.61 %	**92.38** %
Coenen–500	25 %	96.70 %	96.72 %	92.53 %	**93.89** %	93.63 %	93.63 %
Coenen–400	20 %	**97.76** %	94.55 %	92.93 %	**94.03** %	**93.68** %	93.34 %
Coenen–300	20 %	**95.55** %	95.48 %	91.80 %	**94.04** %	**95.15** %	93.54 %
Coenen–200	20 %	95.21 %	**95.56** %	91.73 %	**92.61** %	91.01 %	**92.20** %
Promedio		**96.13** %	95.56 %	92.03 %	**93.64** %	92.98 %	93.00 %

		Regression		Decision-Table		J48graft	
Colección	δ	E(0.4)	GRAE	E(0.4)	GRAE	E(0.4)	GRAE
Coenen–700	20 %	**95.76** %	94.62 %	92.03 %	**93.70** %	95.26 %	**95.58** %
Coenen–600	20 %	**95.98** %	95.41 %	92.26 %	**94.38** %	93.32 %	**95.04** %
Coenen–500	25 %	94.98 %	**95.46** %	93.17 %	**95.03** %	95.71 %	95.73 %
Coenen–400	20 %	94.59 %	**95.19** %	**96.23** %	92.95 %	**95.08** %	94.36 %
Coenen–300	20 %	**93.44** %	93.31 %	**92.92** %	90.46 %	92.70 %	**93.96** %
Coenen–200	20 %	**90.91** %	89.44 %	91.17 %	**92.48** %	92.75 %	**93.09** %
Promedio		**94.28** %	93.91 %	92.96 %	**93.17** %	94.14 %	**94.63** %

Como se puede observar en la tabla 5.12, utilizando el algoritmo de selección de atributos GRAE se pueden obtener resultados significativamente mejores que utilizando los patrones emergentes E(0.4) en un 33.33 % de las pruebas realizadas, mientras que no hay diferencia significativa en el 66.67 % restante.

Tabla 5.12: Pruebas de significancia estadística para los diferentes clasificadores en varias colecciones de imágenes utilizando los patrones emergentes E(0.4) y utilizando el algoritmo de selección de atributos GRAE. Cada celda de la tabla indica cuál opción fue estadísticamente mejor (para la prueba mostrada en la columna 1), "-" indica que no hubo diferencia estadísticamente significativa.

(a) Resultados obtenidos utilizando α =0.05

Clasificador	SVM	BayesNet	AdaBoostM1	Regression	Decision-Table	J48graft
Bonferroni-Dunn	–	GRAE	–	–	–	GRAE
Hommel	–	GRAE	–	–	–	GRAE
Holm	–	GRAE	–	–	–	GRAE

(b) Resultados obtenidos utilizando α =0.10

Clasificador	SVM	BayesNet	AdaBoostM1	Regression	Decision-Table	J48graft
Bonferroni-Dunn	–	GRAE	–	–	–	GRAE
Hommel	–	GRAE	–	–	–	GRAE
Holm	–	GRAE	–	–	–	GRAE

De manera general, el uso de estas dos estrategias para la reducción de atributos son igualmente viables para dar solución al problema de la clasificación de imágenes en las bases de datos sintéticas. Aunque con el uso de GRAE se obtienen resultados mejores, utilizando E(0.4) no se logran resultados alejados de estos, y la reducción es mejor mediante E(0.4) que usando GRAE. El proceso de encontrar la cantidad adecuada de atributos mediante el ranking que brinda GRAE es mucho más costosa (en esfuerzos y tiempo) que utilizando los patrones emergentes E(0.4).

5.2. Base de datos de esqueletos estructurales

La base de datos, identificada como SIS (*structural image skeletons database*), que se utilizará en esta sección está compuesta por 36 siluetas de imágenes reales, obtenidas de diferentes fuentes para confeccionar una base de datos similar a la de (Shen *et al.*, 2013). La base de datos obtenida está dividida en 9 clases: *elephant*, *fork*, *heart*, *horse*, *human*, *L-star*, *star*, *tortoise*, y *whale*. Cada clase con 4 imágenes, las cuales incluyen imágenes

con rotación y deformación producida por las articulaciones (ver figura 5.2).

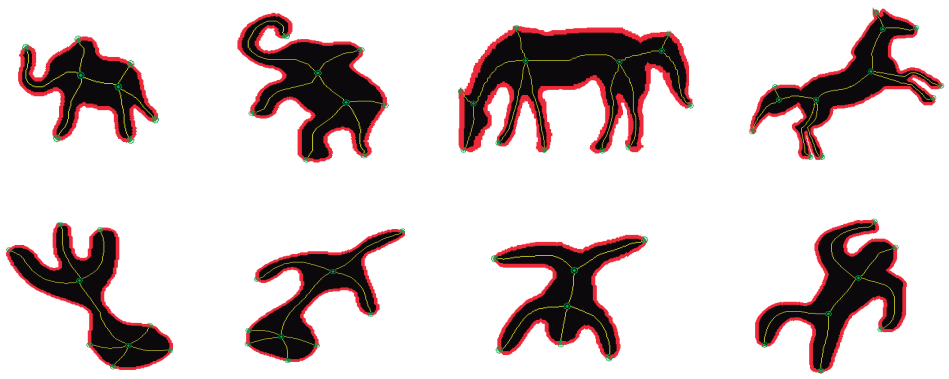

Figura 5.2: Ejemplo de imágenes de la base de datos SIS.

A cada imagen de la base de datos se le calcula el esqueleto de forma semi-automática. Estos esqueletos representarán la colección y para su obtención se utilizó el programa proporcionado por Bai[3] presentado en (Bai & Latecki, 2007). Dicho programa se utiliza para obtener una aproximación del eje medial de las figuras basandose en los radios de los discos máximos inscritos en el interior de las figuras de las imágenes. Luego, de forma manual se podaron las ramas que no representaban la forma estructural de las figuras y se agregaron unas pocas ramas que no fueron construidas por el algoritmo. Como resultado del trabajo manual se obtuvieron esqueletos simples y que representan la información estructural de las figuras de las imágenes (ver figura 5.3).

Una vez confeccionados los esqueletos que representan las imágenes, se etiquetan las aristas con la distancia entre los vértices. Los esqueletos se tratan como grafos y con estos se tiene la colección de grafos. La colección contiene 13 etiquetas de vértices, 211 etiquetas de aristas, el tamaño promedio de los grafos es de 7 en términos de la cantidad de vértices y en términos de la cantidad de aristas su tamaño promedio es de 6.

Las etiquetas de los vértices no pueden ser sustituidas por otras etiquetas sino por

[3]https://sites.google.com/site/xiangbai/

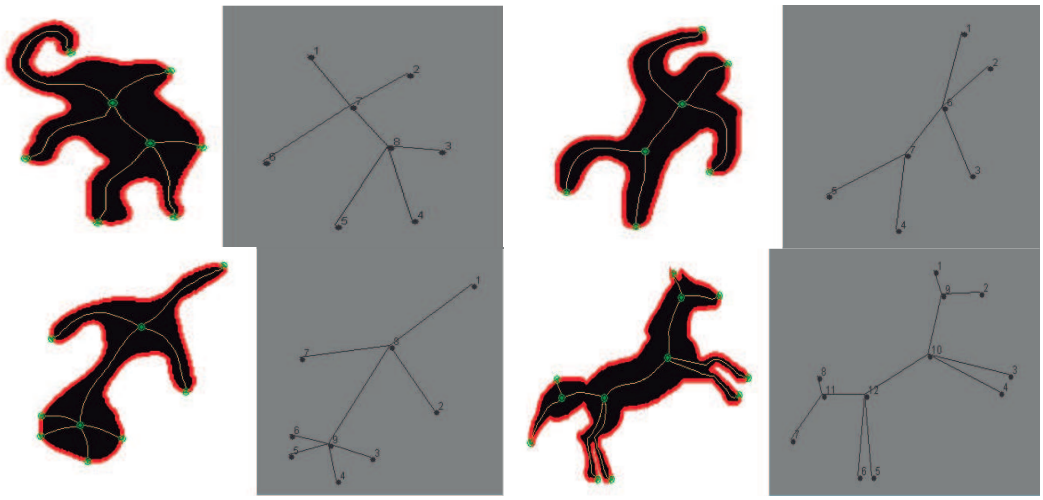

Figura 5.3: Ejemplo de imágenes y sus esqueletos de la base de datos SIS.

ellas mismas, por lo que la matriz de sustitución para los vértices es la matriz identidad (1s en la diagonal y 0s en el resto). Como las etiquetas de las aristas son las distancias entre los vértices que están conectados, entonces si existen posibles sustituciones entre estas etiquetas. Las probabilidades de sustituciones se distribuyen por filas siguiendo la ecuación (5.1):

$$m_{i,j} = \frac{|d_i - d_j|}{|d_0 - d_n|} \tag{5.1}$$

Donde $m_{i,j}$ es la celda en cuestión, d_i y d_j son las distancias de las aristas i y j respectivamente, y d_0 y d_n son la menor y la mayor distancia del conjunto de distancias respectivamente. Luego se normalizan los valores de las celdas de tal manera que la suma de sus valores por fila sea 1 y de esta manera se confecciona la matriz de sustitución para las aristas.

5.2.1. Resultados experimentales

Adicionalmente, se presenta una comparación entre nuestra propuesta utilizando los patrones emergentes en el módulo de reducción de patrones y el método utilizando todos los patrones calculados por VEAM en la colección detallada en la sección 5.2. En esta ocasión se probaron varios valores de γ diferentes a los utilizados en los experimentos de la sección 5.1.1 debido a que con los umbrales utilizados en esa sección no se identificaron patrones. Los valores utilizados para esta base de datos fueron γ =0.5, 0.6, 0.7, 0.8 y 0.9 con varios valores para el umbral de mínimo soporte δ como parámetros para VEAM. Para los experimentos se utilizó validación cruzada con una ventana de 10 (ver figura 4.1).

En la tabla 5.13 se muestra la reducción de la dimensionalidad del conjunto de patrones lograda con el uso de los patrones emergentes. La primera columna de esta tabla muestra los valores del umbral de mínimo soporte usado en VEAM. Las otras seis columnas consecutivas muestran el número de patrones utilizados como atributos en el método de clasificación, donde la primera de estas columnas muestra la cantidad de patrones calculados por VEAM, y las otras cinco muestran el número de patrones emergentes obtenidos utilizando γ =0.9, 0.8, 0.7, 0.6 y 0.5 del conjunto de todos los patrones calculados por VEAM. Nótese que el rango del umbral de mínimo soporte usado está entre 20 % y 40 %, debido a que el conjunto de patrones calculados con umbrales superiores a 40 % es pequeño, no es necesaria la reducción de dimensionalidad, y los patrones obtenidos con umbrales menores que 20 % no resultaron útiles para la clasificación. Estos últimos no son útiles debido a que las distorciones que se permiten por VEAM en los umbrales menores de 20 % introducen ruido en la representación de los datos lo cual afecta en el resultado de la clasificación.

Como se puede observar en la tabla 5.13 se logra una notable reducción de la dimensionalidad del conjunto de patrones al utilizar el módulo de reducción de patrones del

Tabla 5.13: Número de patrones utilizados como atributos en el proceso de clasificación sobre la base de datos SIS.

Soporte (δ)	**Todos los patrones**	**Patrones emergentes**				
		γ =0.9	γ =0.8	γ =0.7	γ =0.6	γ =0.5
20 %	12460420	6496182	221961	72676	4104	901
25 %	343670	99472	40113	555	–	–
30 %	108975	135	58	4	–	–
35 %	312	76	29	3	–	–
40 %	163	51	26	–	–	–

método para la clasificación propuesto. Luego, se realizan experimentos utilizando dichos conjuntos de patrones como atributos para la clasificación con el objetivo de evaluar su calidad.

En los experimentos se utilizaron varios clasificadores para evaluar nuestra propuesta. En esta ocación se utilizaron seis clasificadores de Weka v3.6.6 Hall *et al.* (2009) usando los parámetros por defecto de los mismos: red bayesiana (BayesNet), máquina de soporte vectorial (SVM), un clasificador basado en regresión (Regression), uno basado en reglas (Decision-Table), y uno basado en árboles de decisión (J48graft). En estos experimentos se comparan los resultados de la clasificación (accuracy) obtenidos por nuestra propuesta utilizando los patrones emergentes contra el uso de todos los SFAs como atributos para la clasificación. Estos resultados están resumidos en la tabla 5.14.

La primera y segunda columna de la tabla 5.14 muestran el identificador de los clasificadores y los valores del umbral de mínimo soporte, respectivamente. En la tercera columna se muestran los resultados (accuracy) obtenidos por los clasificadores utilizando todos los SFAs como atributos y las otras cinco columnas muestran los resultados obtenidos utilizando solamente los patrones emergentes con γ =0.9, 0.8, 0.7, 0.6 y 0.5, en este mismo orden.

Como se puede observar en la tabla 5.14, los resultados obtenidos por los clasificadores utilizando todos los SFAs como atributos son peores que los obtenidos utilizando

Tabla 5.14: Resultados de la clasificación alcanzados utilizando el método propuesto en la colección de imágenes SIS con y sin el uso de los patrones emergentes como atributos para varios clasificadores con diferentes valores de γ.

Clasificador	δ	**todos los patrones**	**Patrones emergentes**				
			γ =0.9	γ =0.8	γ =0.7	γ =0.6	γ =0.5
NavieBayes	20 %	–	–	**87.10** %	80.65 %	70.97 %	62.50 %
	25 %	–	**77.42** %	74.19 %	73.68 %	–	–
	30 %	65.71 %	**86.96** %	68.42 %	57.89 %	–	–
	35 %	42.86 %	**73.68** %	57.89 %	52.63 %	–	–
	40 %	37.14 %	**68.42** %	52.63 %	–	–	–
SVM	20 %	–	–	**90.32** %	77.42 %	**90.32** %	56.25 %
	25 %	74.29 %	77.42 %	70.97 %	**94.74** %	–	–
	30 %	**82.86** %	82.61 %	68.42 %	26.32 %	–	–
	35 %	37.14 %	**84.21** %	63.16 %	21.05 %	–	–
	40 %	42.86 %	**57.89** %	47.37 %	–	–	–
Regression	20 %	–	–	–	74.19 %	**83.87** %	50.00 %
	25 %	–	74.19 %	64.52 %	**89.47** %	–	–
	30 %	–	**91.30** %	68.42 %	21.05 %	–	–
	35 %	37.14 %	**78.94** %	63.16 %	21.05 %	–	–
	40 %	45.71 %	**63.16** %	**63.16** %	–	–	–
Decision-Table	20 %	–	–	–	–	**67.74** %	62.50 %
	25 %	–	–	61.29 %	**73.68** %	–	–
	30 %	–	**65.22** %	47.37 %	10.53 %	–	–
	35 %	28.57 %	**52.63** %	21.05 %	10.53 %	–	–
	40 %	**25.71** %	21.05 %	21.05 %	–	–	–
J48graft	20 %	–	–	80.65 %	80.65 %	**87.10** %	56.25 %
	25 %	65.71 %	74.19 %	74.19 %	**94.74** %	–	–
	30 %	82.86 %	**86.96** %	57.89 %	42.11 %	–	–
	35 %	34.29 %	**73.68** %	47.37 %	21.05 %	–	–
	40 %	40.00 %	42.11 %	47.37 %	–	–	–

los patrones emergentes, los cuales utilizan un conjunto reducido de atributos para la clasificación. Una mejora se puede ver con $\delta = 25\,\%$ usando los clasificadores SVM y J48graft, donde estos clasificadores son capaces de obtener todos los resultados. Con estos clasificadores nuestra propuesta, utilizando un 0,17 % de los patrones, alcanza una mejora mayor del 20 % del accuracy obtenido con todos los SFAs como atributos para la clasificación. Otra mejora significativa se evidencia con $\delta = 20\,\%$ donde ningún clasificador fue capaz de obtener resultados si no se utiliza nuestro método, ocurriendo de igual manera en algunos casos donde $\delta = 25\,\%$ y 30 %. Estos resultados muestran que el uso de los patrones emergentes como atributos para la clasificación permite obtener mejores resultados de clasificación y una reducción significativa de la dimensionalidad

del problema.

En general, los resutados alcanzados en este experimento, sobre la base de datos SIS, muestran que el método propuesto no depende de la representación en árboles de cuadrantes (*quad-trees*). Este hecho nos permite validar que el uso de los patrones emergentes, como atributos para representar las imágenes de la colección, proporciona mejores resultados que el uso de todos los patrones calculados por VEAM.

5.3. Síntesis y conclusiones

En este capítulo se presentó una comparación entre el método de clasificación propuesto en (Acosta-Mendoza *et al.*, 2012a) y el método propuesto en esta tesis. En nuestro método se utilizaron los patrones emergentes y varios algoritmos de selección de atributos tipo filtrado en el módulo de reducción de patrones. Para hacer este tipo de comparaciones se utilizaron colecciones de imágenes sintéticas de diferentes tamaños las cuales han sido previamente utilizadas en la literatura. Además, se utilizó una colección de esqueletos estructurales de siluetas de imágenes reales similar a una usada en la iteratura.

En nuestros experimentos se pudo observar que el uso de los patrones emergentes como atributos en la clasificación impactó positivamente el desempeño de la clasificación. El efecto positivo del uso de los patrones emergentes se evaluó mediante tres pruebas de significancia estadística. Dicha evaluación indica que el método propuesto utilizando patrones emergentes supera al método que utiliza todos los SFAs como atributos para la clasificación. De esta manera se logra cumplir parcialmente el objetivo específico 4 de esta investigación.

En los experimentos con algoritmos de selección de atributos tipo filtrado, se puede observar una mejora en la reducción del número de atributos y se logra mejorar significativamente el resultado de la clasificación comparado con el uso de todos los atributos. Se

comprobó mediante pruebas de significancia estadística que el uso de los algoritmos de selección de atributos utilizados tuvo un impacto satisfactorio en la clasificación. De este modo se logra el cumplimiento del objetivo específico 4 y mediante el uso del módulo de reducción de patrones se logra dar solución a los problemas 1 y 2 mencionados en la sección 1.1.

Finalmente, se realizó un estudio comparativo entre el uso de los patrones emergentes y el uso de los algoritmos de selección de atributos. Para dicha comparación se escogieron las opciones identificadas como las mejores de las secciones 5.1.1 y 5.1.1. Como resultado de esta comparación se comprobó que el uso del algoritmo para la selección de atributos GRAE logró impactar en mayor medida que el uso de los patrones emergentes E(0.4) en la clasificación. Dicho impacto fue comprobado mediante pruebas de significancia estadística. Sin embargo, mediante el uso de E(0.4) se logra una mejor reducción del conjunto de atributos obteniendo resultados cercanos a los obtenidos mediante GRAE. Por este motivo, se llega a la conclusión de que el uso de GRAE y de E(0.4) son viables dado que con ambos se mejora la calidad de la clasificación reduciendo la dimensionalidad del conjunto de atributos. De esta manera se da cumplimiento al objetivo específico 5 de esta tesis.

Capítulo 6

Conclusiones, aportaciones y trabajo futuro

6.1. Conclusiones

En esta tesis se propone un módulo de reducción de patrones con el fin de reducir la dimensionalidad de los vectores de atributos utilizados en la clasificación de imágenes. El método propuesto en esta tesis es comparado con un método que utiliza todos los patrones como atributos en la clasificación. Teniendo en cuenta los experimentos reportados se pudo llegar a las siguientes conclusiones:

- Se recomienda usar el método propuesto en esta tesis para la clasificación de imágenes representadas en forma de grafos, y que contengan un elevado número de patrones frecuentes aproximados.
- Mediante el uso de los selectores de atributos (tipo filtrado) se obtienen mejores resultados de clasificación respecto al uso de los patrones emergentes y al uso de todos los atributos en la clasificación.

- Los patrones emergentes reducen la dimensionalidad del conjunto de atributos en mayor porción que los algoritmos de filtrado obteniendo buenos resultados de clasificación comparables a los logrados mediante estos selectores.

- Se recomienda el uso de los patrones emergentes como atributos para la clasificación cuando se desee reducir el tamaño de la representación, logrando un ahorro considerable de memoria y tiempo.

- El uso de los patrones emergentes y de la selección de atributos no solo permite disminuir la dimensionalidad del problema, sino que también permite mejorar los resultados de la clasificación.

Finalmente, se consideran cumplidos los objetivos de esta tesis con la propuesta del método para la clasificación de imágenes. Los problemas planteados en esta investigación fueron solucionados, aunque todavía queda trabajo por hacer para mejorar los resultados alcanzados. Sin embargo, se logró una mejora significativa en el desempeño de la clasificación reduciendo, a su vez, la dimensionalidad del conjunto de patrones (atributos) utilizados para representar las imágenes.

6.2. Aportaciones del trabajo de investigación

Las aportaciones de esta investigación son las siguientes:

1. Método para la selección de un subconjunto de patrones (SFA) que permita reducir la dimensionalidad de los vectores de atributos para la clasificación de imágenes.

2. Método de clasificación de imágenes utilizando un subconjunto de SFA sin comprometer la eficacia de la clasificación.

6.3. Trabajo futuro

Los resultados obtenidos en esta tesis nos motivan a realizar, como trabajo futuro, un análisis más profundo de nuestro método en distintas circunstancias y bases de datos. Esto se realizará con el objetivo de respondernos algunas interrogantes que surgen a raíz de la solución propuesta:

- ¿Qué tan sencible es nuestro método a la selección de γ?
- ¿Qué tanto afecta el desbalance de las clases al rendimiento del método propuesto?
- ¿Cuál es la cantidad de elementos necesarios para que nuestro método obtenga buenos resultados de clasificación?
- ¿Cómo impactaría en los resultados de la clasificación la asignación de pesos a los patrones emergentes teniendo en cuenta que tan representativos a una clase sean?

El módulo de reducción de patrones propuesto en esta tesis está enfocado en el post-procesamiento de los resultados de los algoritmos para la MSFA. Los patrones que calculan estos algoritmos no toman en cuenta las ocurrencias por clases de estos patrones. Pudiera ser ventajoso la inclusión de las propiedades de los patrones emergentes en el proceso de búsqueda de la MSFA.

Por otra parte, se podría disminuir la redundancia de los patrones calculados por los algoritmos para la MSFA si se calcularan los patrones cerrados. Si se tienen en cuenta solamente los patrones cerrados y emergentes como atributos para la clasificación, se pudiera disminuir aún más la dimensionalidad del problema, evitando la redundancia en la información.

Anexos

Notaciones

(V, E, I, J)	Grafo etiquetado con sus cuatro componentes
V	Conjunto de vértices de un grafo
E	Conjunto de aristas de un grafo
I	Función que asigna etiquetas a los vértices de un grafo
J	Función que asigna etiquetas a las aristas de un grafo
u, v	Vértices de un grafo
$\{u, v\}$, e	Aristas de un grafo (Si u y v son vértices de del grafo para el caso $\{u, v\}$)
L_V, L_E	Conjuntos de etiquetas para los vértices y aristas de un grafo respectivamente
L	Conjunto de todas las posibles etiquetas para un grafo ($L = L_V \cup L_E$)
G, T, G_1, G_2	Grafos etiquetados
Ω	Conjunto de todos los posibles grafos etiquetados
$G_1 \subseteq G_2$	G_1 es subgrafo de G_2 (Si G_1 y G_2 son grafos)
i, j, n, m	Subíndices (números enteros no negativos)

c_1, c_2, ..., c_n	Clases
$C = \{c_1, \ldots, c_{\|C\|}\}$	Conjunto de clases
M, MV, ME	Matrices de sustitución
$m_{i,j}$	Una celda de una matriz de sustitución
$O(G_1, G_2)$	Conjunto de ocurrencias del grafo G_1 en el grafo G_2
$ExtSet(G)$	Conjunto de extensiones del grafo G
D	Colección de grafos sin especificar sus elementos
$\{G_1, G_2, \ldots, G_m\}$	Colección de grafos especificando sus elementos
G_i	Grafo de una colección
$supp(G, D)$	Soporte de G en D
$appSupp(G, D)$	Soporte aproximado de G en D
δ	Umbral de frecuencia
τ	Umbral mínimo de isomorfismo aproximado
γ	Umbral de soporte para la identificación de un patrón emergente
F	Resultado de la minería (Conjunto de todos los SFA)
$\diamond$	Operador de extensión en el paradigma crecimiento de patrones

Acrónimos

MSFA	*Minería de Subgrafos Frecuentes Aproximados*
SFA	*Subgrafo Frecuente Aproximado*
IG	*Ganancia de Información*
CHI-Q	*Chi-Cuadrado*
GRAE	*Cociente de evaluación de la ganancia de información de los atributos*
LS	*Puntuación laplaciana*
NW	*Noroeste*
NE	*Noreste*
SW	*Suroeste*
SE	*Sureste*
N	*Norte*
E	*Este*
W	*Oeste*
S	*Sur*
SUBDUE	*Algoritmo para la MSFA*
RNGV	*Algoritmo para la MSFA (**R**egulator **N**etwork **G**eneration **V**ariation)*
VEAM	*Algoritmo para la MSFA (**V**ertex and **E**dge **A**pproximate graph **M**iner)*
APGM	*Algoritmo para la MSFA (**AP**promimate **G**raph **M**ining)*
gApprox	*Algoritmo para la MSFA*

Referencias

N. Acosta-Mendoza, A. Gago-Alonso, & J. E. Medina-Pagola. Frequent approximate subgraphs as features for graph-based image classification. *Knowledge-Based Systems*, 27:381–392, March 2012a.

N. Acosta-Mendoza, A. Gago-Alonso, & J. E. Medina-Pagola. On speeding up frequent approximate subgraph mining. In *Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP'12)*, páginas 316–323, Buenos Aires, Argentina, 2012b. Springer-Verlag Berlin Heidelberg.

N. Acosta-Mendoza, A. Morales-González, A. Gago-Alonso, E. B. García-Reyes, & J. E. Medina-Pagola. Classification using frequent approximate subgraphs. In *Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP'12)*, páginas 292–299, Buenos Aires, Argentina, 2012c. Springer-Verlag Berlin Heidelberg.

R. Alves, D. S. Rodríguez-Baena, & J. S. Aguilar-Ruiz. Gene association analysis: a survey of frequent pattern mining from gene expression data. *Briefings in Bioinformatics*, 11(2):210–224, 2010.

O. Bahadir & A. Selim. Image classification using subgraph histogram representation. In *Proceedings of the 20th International Conference on Pattern Recognition*, páginas 1112–1115, Washington, DC, USA, 2010a. IEEE Computer Society.

O. Bahadir & A. Selim. Image classification using subgraph histogram representation. In *Proceedings of the 20th International Conference on Pattern Recognition*, páginas 1112–1115, Washington, DC, USA, 2010b. IEEE Computer Society.

X. Bai & L.J. Latecki. Discrete skeleton evolution. In *Energy Minimization Methods in Computer Vision and Pattern Recognition, 6th International Conference, EMMCVPR'07*, páginas 362–374. Ezhou, China, Springer, 2007.

P. Bermejo, L. de la Ossa, J.A. Gámez, & J. Miguel-Puerta. Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. *Knowledge-Based Systems*, 25(1):35–44, 2012.

V. Bolón-Canedo, N. Sánchez-Maroño, & A. Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34(3):483–519, 2013.

C. Borgelt & M. R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. *IEEE International Conference on Data Mining (ICDM'02)*, páginas 51–58, 2002.

H. Bunke & K. Riesen. Towards the unification of structural and statistical pattern recognition. *Pattern Recognition Letters*, 33:208–297, 2012.

C. Chen, X. Yan, F. Zhu, & J. Han. gapprox: Mining frequent approximate patterns from a massive network. In *IEEE International Conference on Data Mining (ICDM'07)*, páginas 445–450, 2007.

D. Conte, P. Foggia, C. Sansone, & M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 18(3):265–298, 2004.

W. Dhifli, R. Saidi, & E. M. Nguifo. A novel approach for mining representative spatial motifs of proteins. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB)*, páginas 506–508, Orlando, Florida, New York, NY, USA, 2012.

W. Dhifli, R. Saidi, & E. M. Nguifo. Mining representative unsubstituted graph patterns using prior similarity matrix. *Machine Learning in Computational Biology (MLCB'12) (NIPS workshop)*, March 8 2013.

G. Dong & J. Bailey. Overview of contrast data mining as a field and preview of an upcoming book. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference*, páginas 1141–1146, Vancouver, BC, Canada, December 11 2011.

G. Dong & J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 43–52. San Diego, Caliofornia, United States, 1999.

R. Duin & E. Pekalska. The dissimilarity representations for pattern recognition: Foundations and applications. *World Scientific*, 2005.

B. Duval, J.K. Hao, & J.C. Hernandez. A memetic algorithm for gene selection and molecular classification of cancer. In *Genetic and Evolutionary Computation Conference (GECCO'09)*, páginas 201–208. ACM, Montreal, Québec, Canada, 2009.

F. Eichinger & K. Böhm. Software-Bug Localization with Graph Mining. In Charu C. Aggarwal & Haixun Wang, editors, *Managing and Mining Graph Data*, volume 40. Springer-Verlag New York, 2010.

A. Elsayed, F. Coenen, C. Jiang, M. García-Fiñana, & V. Sluming. Region of interest based image categorization. In *Proceedings of the 12th international conference on Data warehousing and knowledge discovery*, páginas 239–250, Berlin, Heidelberg, 2010a. Springer-Verlag.

A. Elsayed, F. Coenen, C. Jiang, M. García-Fiñana, & V. Sluming. Corpus callosum mr image classification. *Knowledge-Based Systems*, 23(4):330–336, 2010b.

G. Fang, W. Wang, B. Oatley, B. V. Ness, M. Steinbach, & V. Kumar. Characterizing discriminative patterns. *Computing Research Repository (CoRR)*, abs1102.4, 2011.

A.J. Ferreira & M.A.T. Figueiredo. Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters*, 33(13):1794–1804, 2012.

R. A. Finkel & J. L. Bentley. Quad trees: A data structure for retrieval on composite keys. *Acta Informatica*, 4:1–9, 1974.

A. Gago-Alonso, N. Acosta-Mendoza, & A. Muñoz-Briseño. A new proposal for graph classification using frequent geometric subgraphs. *To appear in Data and Knowledge Engineering*, August 2013. doi: 10.1016/j.datak.2013.04.001.

A. Gago-Alonso, J. A. Carrasco-Ochoa, J. E. Medina-Pagola, & J. Fco. Martínez-Trinidad. Duplicate candidate elimination and fast support calculation for frequent subgraph mining. In *Proceedings of the 10th international conference on Intelligent data engineering and automated learning*, páginas 292–299, Berlin Heidelberg, 2009. Springer-Verlag.

S. García & F. Herrera. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, páginas 2677–2694, 2008.

M. Garcia-Borroto. Searching extended emerging patterns for supervised classification. In *Proceedings of the 10th international conference on Intelligent data engineering and automated learning*, Computer Science Department, National Institute for Astrophysics Optics and Electronics, Puebla, Mexico, 2010.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, & I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(Issue 1), 2009.

X. He, D. Cai, & P. Niyogi. Laplacian score for feature selection. *Advances in Neural Information Processing Systems*, 18:507–514, 2006.

L.B. Holder, D.J. Cook, & H. Bunke. Fuzzy substructure discovery. In *Proceedings of the 9th International Workshop on Machine Learning*, páginas 218–223, San Francisco, CA, USA, 1992.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

G. Hommel. A stagewise rejective multiple test procedure. In *Biometrika*, volume 75, páginas 383–386, 1988.

M. S. Hossain & R. A. Angryk. Gdclust: A graph-based document clustering technique. In *Proceedings of the 7th IEEE International Conference on Data Mining Workshops*, páginas 417–422, Washington, DC, USA, 2007.

Y. Jia, J. Huan, V. Buhr, J. Zhang, & L. Carayannopoulos. Towards comprehensive structural motif mining for better fold annotation in the "twilight zone" of sequence dissimilarity. *BMC Bioinformatics*, 10(S-1), 2009.

Y. Jia, J. Zhang, & J. Huan. An efficient graph-mining method for complicated and noisy data with real-world applications. *Knowledge Information Systems*, 28(2):423–447, 2011.

C. Jiang, F. Coenen, & M. Zito. A survey of frequent subgraph mining algorithms. 2012. Knowledge Engineering Review.

C. Jiang & F. Coennen. Graph-based image classification by weighting scheme. In *Proceedings of the Artificial Intelligence*, páginas 63–76. Springer, Heidelberg, 2008.

C. Jiang, F. Coennen, & M. Zito. Frequent sub-graph mining on edge weighted graphs. In *Proceedings of the 12th International Conference on Data Warehousing and Knowledge Discovery*, páginas 77–88, Berlin, Heidelberg, 2010. Springer-Verlag.

A. Jiménez, F. Berzal, & J. C. Cubero-Talavera. Frequent tree pattern mining: A survey. *Intelligence Data Analysis*, 14(6):603–622, 2010.

N. Jin & W. Wang. Lts: Discriminative subgraph mining by learning from search history. In *Proceedings of the 27th International Conference on Data Engineering (ICDE)*, páginas 207–218, Hannover, Germany, 2011.

Y. Keneshloo & S. Yasdani. A relative feature selection algorithm for graph classification. *Advances in Databases and Information Systems*, 186:137–148, 2013.

N. Ketkar, L. Holder, & D. Cook. Mining in the proximity of subgraphs. In *ACM KDD Workshop on Link Analysis: Dynamics and Statics of Large Networks*, 2006.

X. Kong, P. S. Yu, X. Wang, & A. B. Ragin. Discriminative feature selection for uncertain graph classification. In *Proceedings of Computing Research Repository (CoRR)*, volume abs1301.6626, 2013.

M. Koyutürk, A. Grama, & W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, páginas 200–207, 2004.

J. Li, G. Dong, & K. Ramamohanarao. Instance-based classification by emerging patterns. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, páginas 191–200. Springer-Verlag, 2000.

H. Liu, J. Li, & L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, 13:51–60, 2002.

A. Morales-González & E. B. García-Reyes. Assessing the role of spatial relations for the object recognition task. *The 15th Iberoamerican Congress on Pattern Recognition*, páginas 549–556, 2010.

A. Morales-González & E. B. García-Reyes. Simple object recognition based on spatial relations and visual features represented using irregular pyramids. *Multimedia Tools and Applications*, páginas 1–23, 2011.

O. B. Norshafarina, J. B. Fantimatufaridah, O. B. Mohd-Shahizan, & I. B. Roliana. Review of feature selection for solving classification problems. *Journal of Research and Innovation in Information Systems*, páginas 54–60, 2013.

P. K. Novak, N. Lavrac, & G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009.

K. Ohara, M. Hara, K. Takabayashi, H. Motoda, & T. Washio. Pruning strategies based on the upper bound of information gain for discriminative subgraph mining. *Knowledge Acquisition: Approaches, Algorithms and Applications*, páginas 50–60, 2009.

O. Papapetrou, E. Ioanno, & D. Skoutas. Efficient discovery of frequent subgraph patterns in uncertain graph databases. In *Proceedings of the 14th International Conference on Extending Database Technology*, páginas 355–366, New York, NY, USA, 2011. ISBN 978-1-4503-0528-0.

E. Pekalska, R. Duin, & P. Paclik. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2):189–208, 2006.

B.B. Pineda-Bautista, J.A. Carrasco-Ochoa, & J.Fco. Martínez-Trinidad. General framework for class-specific feature selection. *Experts Systems with Applications*, 38(8): 10018–10024, 2011.

G. Poezevara, B. Cuissart, & B. Crémilleux. Extracting and summarizing the frequent emerging graph patterns from a dataset of graphs. *J. Intell. Inf. Syst.*, 37(3):333–353, 2011.

P. Pudil, J. Novovicova, & J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, 1994.

K. Riesen & H. Bunke. Iam graph database repository for graph based pattern recognition and machine learning. páginas 208–297. Orlando, USA, 2008.

K. Riesen, M. Neuhaus, & H. Bunke. Graph embedding in vector spaces by means of prototype selection. *GbRPR, LNCS 4538*, páginas 383–393, 2007.

G. Rodríguez-Bermúdez, P.J.García-Laencina, J. Roca-González, & J. Roca-Dorda. Efficient feature selection and linear discrimination of (eeg) signals. *Neurocomputing*, 115(4):161—165, September 2013.

W. Shen, Y. Wang, X. Bai, H. Wang, & L.J. Latecki. Shape clustering: Common structure discovery. *Pattern Recognition*, 46(2):539–550, 2013.

R.J. Simes. An improved bonferroni procedure for multiple tests of significance. In *Biometrika*, volume 73, páginas 751–754, 1986.

S. Solorio-Fernández, J.A. Carrasco-Ochoa, & J.Fco. Martínez-Trinidad. Hybrid feature selection method for supervised classification based on laplacian score ranking. *Lecture Notes in Computer Science, Advances in Pattern Recognition*, 6256(3):260–269, 2010.

Y. Song & S. Chen. Item sets based graph mining algorithm and application in genetic regulatory networks. In *Proceedings of the IEEE International Conference on Granular Computing*, páginas 337–340, Atlanta, GA, USA, 2006.

B. Spillmann, H. Bunke, E. Pekalska, & R. Duin. Transforming strings to vector spaces using prototype selection. *Structural, Syntactic, and Statistical Pattern Recognition LNCS*, 4109:287–296, 2006.

D.W. Tan, W. Yeoh, Y.L. Boo, & S.Y. Liew. The impact of feature selection: a data-mining application in direct marketing. *Int. Syst. in Accounting, Finance and Management*, 20(1):23–38, 2013.

Y. Xiao, W. Wang, & W. Wu. Mining conserved topological structures from large protein-protein interaction networks. In *Proceedings of the 18th IEICE data engineering workshop / 5th DBSJ annual meeting*, Hiroshima, Japan, 2007. DEWS'2007.

Y. Xiao, W. Wu, W. Wang, & Z. He. Efficient algorithms for node disjoint subgraph homeomorphism determination. In *Proceedings of the 13th International Conference on Database Systems for Advanced Applications*, páginas 452–460, New Delhi, India, 2008.

C. Xue-wen. An improved branch and bound algorithm for feature selection. *Pattern Recognition Letters*, 24(12):1925–1933, 2003.

X. Yan & J. Huan. gspan: Graph-based substructure pattern mining. In *Proceedings International Conference on Data Mining*, páginas 721–724. Maebashi, Japan, 2002.

Y. Ye, Q. Wu, J.Z. Huang, M.K. Ng, & X. Li. Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognition*, 46(3): 769–787, 2013.

S. Zhang & J. Yang. Ram: Randomized approximate graph mining. In *Proceedings of the 20th International Conference on Scientific and Statistical Database Management*, páginas 187–203, Hong Kong, China, 2008.

S. Zhang, J. Yang, & V. Cheedella. Monkey: Approximate graph mining based on spanning trees. In *International Conference on Data Engineering*, páginas 1247–1249, Los Alamitos, CA, USA, 2007. IEEE ICDE.

Y. Zhao, G. Wang, Y. Li, & Z. Wang. Finding novel diagnostic gene patterns based on interesting non-redundant contrast sequence rules. *IEEE International Conference on Data Mining (ICDM)*, páginas 972–981, 2011.

Z. Zhao, L. Wang, H. Liu, & J. Ye. On similarity preserving feature selection. *IEEE Trans. on Knowl. and Data Eng.*, 25(3):619–632, 2013.

Z. Zou, J. Li, H. Gao, & S. Zhang. Frequent subgraph pattern mining on uncertain graph data. In *Proceeding of the 18th ACM conference on Information and knowledge management*, páginas 583–592, New York, NY, USA, 2009. ACM.

Z. Zou, J. Li, H. Gao, & S. Zhang. Mining frequent subgraph patterns from uncertain graph data. *IEEE Transactions on Knowledge and Data Engineering*, 22(9):1203–1218, 2010.