El impacto de la selección de patrones en la clasificación de imágenes basada en minería de subgrafos frecuentes aproximados

The Impact of the Patterns Selection in Image Classification based on Frequent Approximate Subgraph Mining

Niusvel Acosta-Mendoza[a,b,*], Andrés Gago-Alonso[a], José E. Medina-Pagola[a], Jesús A. Carrasco-Ochoa[b], José Fco. Martínez-Trinidad[b]

[a] *Centro de Aplicación de Tecnologías de Avanzada, La Habana, Cuba.*
[b] *Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México.*
(Minería de Datos y Textos)

Resumen

En la actualidad, varios investigadores se han enfocado en mejorar tareas de clasificación de grafos basadas en minería de subgrafos frecuentes aproximados (SFA). La identificación de este tipo de patrones tiene una amplia variedad de aplicaciones en diferentes dominios de la ciencia, ya que permitir variaciones en los datos es de gran utilidad en varios contextos de aplicación. En este trabajo se presenta un estudio de los métodos aproximados que permiten variaciones semánticas entre etiquetas manteniendo la topología de los grafos. Dicho estudio permitió detectar que, a pesar de que los SFA son de gran utilidad para la clasificación, disminuir el elevado número de patrones que se identifican ha sido un reto para los investigadores. Por tal motivo se han reportado algunos trabajos donde su principal objetivo es disminuir la dimensionalidad del conjunto de patrones que se obtienen al aplicar la minería de SFA. Estos trabajos han logrado mejoras en eficiencia y eficacia obteniendo buenos resultados en clasificación de imágenes. Finalmente, una comparación y un resumen de los resultados alcanzados es presentado también en este trabajo como parte del estudio realizado.
Palabras claves: Clasificación de grafos, classificación de imágenes, selección de patrones, patrones representativos, minería de subgrafos frecuentes aproximados.

Abstract

Currently, several researchers have focused on improving graph classification tasks based on frequent approximate subgraph mining. Frequent graph identification has a wide range of applications in several domains of the science. This is because in this kind of mining, useful data variation for some specific tasks are allowed. In this work, a study of the approximate methods which treat semantic variations between labels keeping the graph topologies is presented. Through this study it was observed that, although the frequent approximate subgraphs are useful for classification, reducing the high number of the identified patterns has been a challenge for researchers. For this reason, several works are reported for reducing the dimensionality of the pattern set computed in the frequent approximate subgraph mining process. These works have achieved improvements in efficacy and efficiency, obtaining good results using this kind of mining for image classification. Finally, a comparison and a summary of these results is also presented, as part of the study, in this paper.
Keywords: Graph classification, image classification, pattern selection, representative patterns, frequent approximate subgraph mining.

[*] Autor correspondiente
Dirección de correo electrónico: nacosta@cenatav.co.cu (Niusvel Acosta-Mendoza)

1. Introducción

La clasificación de grafos basada en minería de subgrafos frecuentes se ha convertido en una tarea principal con una amplia variedad de aplicaciones en varios dominios de la ciencia, como son: biología, química, social, y lingüística, entre otros (Han et al., 2007; Cheng et al., 2010; Jiang et al., 2013). Este tipo de minería puede ser aplicada en todas las áreas donde los datos puedan ser modelados en forma de grafos (Holder et al., 1992; Yan and Huan, 2002; Huan et al., 2003; Nijssen and Kok, 2004; Ketkar et al., 2006; Hossain and Angryk, 2007; Eichinger and Böhm, 2010; Gago-Alonso et al., 2010a,b). Mediante la identificación de este tipo de patrones se han obtenido buenos resultados en diferentes tareas de classificación (Holder et al., 1992; Yan and Huan, 2002; Huan et al., 2003; Nijssen and Kok, 2004; Ketkar et al., 2006; Hossain and Angryk, 2007; Jiang and Coenen, 2008; Elsayed et al., 2010; Gago-Alonso et al., 2010a,b). Sin embargo, se identificaron varios problemas donde este tipo de algoritmos no son de utilidad (Holder et al., 1992; Fellman, 2008), debido a la ausencia de objetos exactamente iguales en la práctica y este tipo de minería no permite variaciones en los datos. Por tal motivo, se comenzaron a desarrollar algoritmos para la minería de subgrafos frecuentes aproximados (SFA), donde se permiten variaciones entre los objetos en el proceso de minería (Holder et al., 1992; Xiao et al., 2007; Song and Chen, 2006; Chen et al., 2007; Jia et al., 2009; Acosta-Mendoza et al., 2012a).

Mediante el uso de los algoritmos para la minería de SFA se han obtenido buenos resultados en diferentes dominios de la ciencia (Holder et al., 1992; Song and Chen, 2006; Xiao et al., 2007; Zhang et al., 2007; Chen et al., 2007; Jia et al., 2009, 2011; Acosta-Mendoza et al., 2012a,c; Morales-González et al., 2014; Acosta-Mendoza et al., 2014b). Sin embargo, Estos algoritmos identifican un elevado número de patrones mediante la minería y no todos estos patrones son de utilidad para la clasificación. Por este motivo surge la necesidad de realizar un proceso de selección de los patrones con un mayor grado de represtación para la clasificación. Para dar solución a este problema se han reportado diversos estrategias para la reducción de la dimensionalidad del conjunto de patrones a utilizar en tareas de clasificación (Acosta-Mendoza, 2013). Ejemplo de estas estrategias son: utilizando los enfoques convencionales de selección de atributos (Acosta-Mendoza et al., 2013, 2014b), y las basadas en el uso de subgrafos frecuentes representativos (Acosta-Mendoza et al., 2014a).

En los enfoques basados en selección de atributos se utilizan diferentes alternativas, como son el uso de: algoritmos de filtrado, algoritmos de envoltura y algoritmos embebidos (Acosta-Mendoza et al., 2013). De estas alternativas, el enfoque de filtrado se consideran como uno de las primeras propuestas de métodos que discriminan los atributos basado en las propiedades de estos y sus relaciones con las clases del conjunto de datos.

En los enfoques basados en el uso de patrones representativos, los patrones emergentes y contrastantes han jugado un importante papel (Acosta-Mendoza et al., 2014a). Un patrón es emergente si aparece mayormente en una clase, mientras que rara vez aparece en el resto de las clases. Un patrón constrastante es un patrón emergente que solamente aparece en una clase, estando ausente en el resto de las clases. Mediante el uso de estos tipos de patrones se reduce el conjunto de SFA manteniendo solamente los más representativos, teniendo en cuenta la información que brindan algunos SFA para la separación de las clases del conjunto de datos.

Teniendo en cuenta estos enfoques se reportan resultados competitivos en la tarea de clasificación sobre diferentes colecciones de imágenes, logrando la disminución de la dimensionalidad del conjunto de patrones utilizados. Estos resultados son presentados como parte del estudio realizado en este artículo.

Este trabajo está organizado de la siguiente manera. En la sección 2 se presentan los conceptos básicos necesarios para la comprensión del resto del artículo. Los trabajos relacionados son presentados en la sección 3. En la sección 4 se detallan los resultados alcanzados en tareas de clasificación de imágenes basadas en la minería de SFA haciendo uso de métodos de selección de patrones. Finalmente, las conclusiones de este artículo y algunas ideas de trabajo futuros son expuestas en la sección 5.

2. Conceptos básicos

En esta sección se presentan los conceptos básicos necesarios para la comprensión del resto del artículo. Estos conceptos se muestran mediante dos secciones: (1) donde se definen los conceptos de la teoría de grafos (ver sección 2.1), y (2) donde se definen los conceptos de aprendizaje automático (ver sección 2.2)

2.1. Teoría de grafos

Este artículo está enfocado al procesamiento de colecciones de grafos simples, etiquetados y no dirigidos. En lo adelante se asumen las propiedades de este tipo de grafos cuando se haga referencia a un grafo.

Definición 1 (Grafo). Un *grafo* es una 5-tupla, $G = (V, E, \phi, I, J)$, en el domino de las posibles etiquetas $L = L_V \cup L_E$, donde L_V y L_E son los conjuntos de etiquetas para los vértices y las aristas respectivamente. En un grafo, V es un conjunto en el cual los elementos son llamados *vértices*, E es un conjunto en el cual los elementos son llamados *aristas*, $\phi : E \rightarrow V \times V$ es la *función de incidencia* (la arista e, mediante la función $\phi(e)$, conecta el vértice u y v si $\phi(e) = \{u, v\}$), $I : V \rightarrow L_V$ es una *función etiquetadora* para a asignación de las etiquetas para los vértices y $J : E \rightarrow L_E$ es una *función etiquetadora* para a asignación de las etiquetas para las aristas.

Definición 2 (Subgrafo y supergrafo). Sean $G_1 = (V_1, E_1, \phi_1, I_1, J_1)$ y $G_2 = (V_2, E_2, \phi_2, I_2, J_2)$ dos grafos, G_1 es un *subgrafo* de G_2 si $V_1 \subseteq V_2$, $E_1 \subseteq E_2$, ϕ_1 es una restricción de ϕ_2 a V_1, I_1 es una restricción de I_2 a V_1, y J_1 es una restricción de J_2 a E_1 (una restricción de una función es el resultado de reducir su dominio). En este caso es usada la notación $G_1 \subseteq G_2$, y se dice que G_2 es un *supergrafo* de G_1.

2.1.1. Enfoques exactos

Definición 3 (Isomorfismo). Dados dos grafos G_1 y G_2, un par de funciones (f, g) es un *isomorfismo* entre estos grafos si $f : V_1 \rightarrow V_2$ y $g : E_1 \rightarrow E_2$ son funciones biyectivas, donde:

1. $\forall u \in V_1 : f(u) \in V_2$ y $I_1(u) = I_2(f(u))$
2. $\forall e_1 \in E_1$, donde $\phi_1(e_1) = \{u, v\}$: $e_2 = g(e_1) \in E_2$, y $\phi_2(e_2) = \{f(u), f(v)\}$ y $J_1(e_1) = J_2(e_2)$.

Si existe un isomorfismo entre G_1 y G_2, se dice que G_1 y G_2 son *isomorfos*.

Definición 4 (Sub-isomorfismo). Dados tres grafos G_1, G_2 y G_3. 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, denotado por $G_1 \subseteq_s G_2$, y se dice también que G_1 es *sub-isomorfo* a G_2.

Definición 5 (Soporte). Sean $D = \{G_1, \ldots, G_{|D|}\}$ una colección de grafos y G un grafo, el *soporte* de G en D se define como la fracción del grafo $G_i \in D$, tal que $G \subseteq_s G_i$. El soporte se obtiene mediante la ecuación (1):

$$supp(G, D) = \frac{|\{G_i \in D : G \subseteq_s G_i\}|}{|D|} \tag{1}$$

Utilizando (1), G es un subgrafo frecuente en D si $supp(G, D) \geq \delta$, para un umbral de soporte dado δ. Los valores de δ están en el rango $[0, 1]$. La *minería de subgrafos frecuentes* consiste en, dado un umbral de soporte δ, encontrar todos los subgrafos frecuentes en una colección de grafos D.

2.1.2. Enfoques aproximados

Las siguientes definiciones establecen un marco útil para describir el proceso de minería de grafos inexacta. En los trabajos con enfoques de minería de grafos (Sanfeliu and Fu, 1983; Messmer and Bunke, 1998; Ambauen et al., 2003; Kuramochi and Karypis, 2004; Neuhaus and Bunke, 2004; Chen et al., 2007; Xiao et al., 2008; Zhang and Yang, 2008; Jia et al., 2009; Zou et al., 2010; Acosta-Mendoza et al., 2012a), los autores primero definen el criterio de comparación entre grafos según el contexto de la aplicación. In general, estos criterios son definidos cumpliendo con la definición 6.

Definición 6 (Semejanza). Sea Ω el conjunto de todos los posibles grafos etiquetados en L, la *semejanza* entre dos grafos $G_1, G_2 \in \Omega$ se define como una función $sim : \Omega \times \Omega \rightarrow [0, 1]$. Se puede decir que no existe ningún tipo de isomorfismo entre los subgrafos de G_1 y G_2 si $sim(G_1, G_2) = 0$, mientras mayor sea el valor de $sim(G_1, G_2)$ más semejantes son los grafos, y si $sim(G_1, G_2) = 1$ entonces existe un isomorfismo entre estos grafos.

Cada criterio de comparación (sim) es usado para indicar cuando un grafo está aproximadamente incluido dentro de otro. De esta forma se define el siguiente concepto.

Definición 7 (Isomorfismo aproximado y sub-isomorfismo aproximado). Sean G_1, G_2 y G_3 tres grafos, sea $sim(G_1, G_2)$ una función de semejanza entre grafos, y sea τ un umbral de semejanza, existe un *isomorfismo aproximado* entre G_1 y G_2 si $sim(G_1, G_2) \geq \tau$. Además, si existe un isomorfismo aproximado entre G_1 y G_2, y $G_2 \subseteq G_3$, entonces existe un *sub-isomorfismo aproximado* entre G_1 y G_3, denotado como $G_1 \subseteq_A G_3$.

Definición 8 (Grado de máxima inclusión). Sean G_1 y G_2 dos grafos, sea $sim(G_1, G_2)$ una función de semejanza entre grafos, sea τ un umbral de semejanza, dado que un grafo G_1 puede ser muy semejante a varios subgrafos de otro grafo G_2, el *grado de máxima inclusión* de G_1 en G_2 se define como:

$$maxID(G_1, G_2) = \max_{G \subseteq G_2} sim(G_1, G), \tag{2}$$

donde $maxID(G_1, G_2)$ es el máximo valor de semejanza al comparar G_1 con todos los subgrafos G de $G2$.

Definición 9 (Proyección). Sea $D = \{G_1, \ldots, G_{|D|}\}$ una colección de grafos, sea $sim(G_1, G_2)$ una función de semejanza entre grafos, sea τ un umbral de semejanza y sea G un grafo, la *proyección* de G en D se define como:

$$\Delta(G, D) = \{G_i | G_i \in D, G \subseteq_A G_i\} \tag{3}$$

Utilizando las definiciones 8 y 9, una definición de soporte que permita variaciones en la correspondencia entre grafos puede ser definida.

Definición 10 (Soporte aproximado). Sea $D = \{G_1, \ldots, G_{|D|}\}$ una colección de grafos, sea $sim(G_1, G_2)$ una función de semejanza entre grafos, sea τ un umbral de semejanza y sea G un grafo, el *soporte aproximado* (denotado por *appSupp*) de G en D, en términos del sub-isomorfismo aproximado, se obtiene mediante la ecuación (4):

$$appSupp(G, D) = \frac{\sum_{G_i \in \Delta(G,D)} maxID(G, G_i)}{|D|} \tag{4}$$

Utilizando (4), G es un *SFA* en D se $appSupp(G, D) \geq \delta$, dado un umbral de soporte δ, una función de semejanza entre grafos $sim(G_1, G_2)$ un umbral de semejanza τ. Los valores de δ y τ están en el rango $[0, 1]$, ya que la semejanza está definida en $[0, 1]$. La *minería de SFA* consiste en, dado un umbral de soporte δ y un umbral de semejanza τ, encuentra todos los SFA en una colección de grafos D, usando una función de semejanza *sim* determinada.

Definición 11 (Patrón emergente y patrón contrastante). Sea $D = \{G_1, \ldots, G_{|D|}\}$ una colección de grafos, sea C un conjunto de clases $C = \{c_1, \ldots, c_{|C|}\}$, donde $c_i \neq c_j$, $U_{c_i} = D$ y sea G un grafo de D, se dice que G es un *patrón emergente* (Dong and Li, 1999; Li et al., 2000; Acosta-Mendoza, 2013; 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 (Zhao et al., 2011; Acosta-Mendoza, 2013; Acosta-Mendoza et al., 2014a).

2.2. Aprendizaje automático

La *clasificación supervisada* y la *clasificación no supervisada* son las dos categorías que agrupan los tipos de técnicas de clasificación dentro del aprendizaje automático. Mediante estas técnicas se logra asignar una categoría o clase a un objeto o fenómeno físico del conjunto de clases o categorías 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 pertenece. Mientras que la clasificación supervisada si cuenta con el conocimiento a priori o modelos ya clasificados de antemano (Acosta-Mendoza, 2013). En este artículo se realiza un estudio de varios trabajos enfocados en la clasificación supervisada de imágenes.

Definición 12 (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 contienen. Discriminar los atributos y objetos más relevantes de un conjunto de datos es de gran utilidad para mejorar la eficiencia en esta tarea. Por lo que, en minería de datos, aprendizaje automático, reconocimiento de patrones y estadística es común el uso de métodos para la selección de atributos con el objetivo de reducir la dimensionalidad basados en la utilidad y presición de la clasificación.

Definición 13 (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 Z tal que $R(Z) = max_{X \subseteq A} R(X)$.

Donde una búsqueda exhaustiva no es viable para explorar el espacio de búsqueda, ya que en el peor de los casos se realizan 2^n comparaciones. Para responder a esta problemática se han propuesto alternativas como algoritmos de filtrado, los cuales intentan descartar los atributos que no son relevantes para la clasificación. Para lograr esto, dichos algoritmos se basan únicamente en la evaluación de las propiedades intrínsecas de los atributos y sus relaciones con las clases del conjunto de datos, manteniendo un bajo costo computacional debido a los criterios utilizados para dicha evaluación.

Existen varios criterios para la evaluación de las propiedades de los atributos y sus relaciones con las clases de los datos, siendo la *ganancia de información*, *chi-cuadrado* y el *cociente de la ganancia de información* tres de los más usados en la literatura (Acosta-Mendoza, 2013).

Definición 14 (Ganancia de información). La *ganancia de información* consiste en calcular la información mutua de un conjunto de atributos X relativa al conjunto de clases C, definida como:

$$IG(X, C) = H(X) - H(X|C) \tag{5}$$

$$H(X) = -\sum_{x \in X} p(x) log\, p(x) \tag{6}$$

$$H(X|C) = -\sum_{x \in X, c \in C} p(x, c) log \frac{p(x, c)}{p(c)} \tag{7}$$

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

Definición 15 (Cociente de la ganancia de información). El *cociente de la ganancia de información* consiste en calcular la razón de beneficio de un conjunto de atributos X respecto a las clases, definida como:

$$GRAE(X, C) = \frac{IG(X, C)}{H(C)} \tag{8}$$

Definición 16 (Chi-cuadrado). El criterio *chi-cuadrado* consiste en calcular el valor estadístico chi-cuadrado de cada atributo respecto a la clase. De esta manera se obtiene el nivel de correlación entre la clase y cada atributo:

$$CHI = \sum_{i} \frac{(v0_i - ve_i)^2}{ve_i} \tag{9}$$

donde $v0_i$ 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.

3. Trabajos relacionados

En la literatura se han reportado varios algoritmos para la minería de SFA en colecciones de grafos, los cuales usan diferentes funciones de semejanza para el cálculo de la correspondencia entre grafos. Existen varios enfoques para la este tipo de minería, por ejemplo: (1) algoritmos basados en distancia de edición (Holder et al., 1992; Song and 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 and Chen, 2006) no busca el camino de menor costo, solamente busca uno que satisfaga la inexactitud especificada; (2) algoritmos basados en β-arista sub-isomorfismo (Zhang et al., 2007; Zhang and Yang, 2008), el cual solamente permite variaciones entre aristas y etiquetas de aristas; (3) algoritmos basados en sub-homeomorfismo con vértices/aristas disjuntas (Xiao et al., 2007, 2008), los cuales calculan estructuras aproximadas con topología invariante; (4) algoritmos basados en sub-isomorfismo entre grafos inciertos (Zou et al., 2009; Li et al., 2012), donde el soporte esperado para cada candidato es calculado sobre una colección de subgrafos construida utilizando las probabilidades de que no ocurran en la colección original; y (5) algoritmos basados en probabilidades de sustitución (Jia et al., 2009, 2011; Acosta-Mendoza et al., 2012a), donde no siempre una etiqueta de vértice o una etiqueta de arista puede reemplazar o ser reemplazada por otra. Los algoritmos *VEAM* (Acosta-Mendoza et al., 2012a) y *APGM* (Jia et al., 2009, 2011) utilizan matrices de sustitución para realizar la minería de SFA 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 and Yang, 2008; Zou et al., 2009; Jia et al., 2009, 2011; Li et al., 2012), análisis de redes genéticas regulatorias (Song and Chen, 2006), análisis de redes sociales y de vínculos (Holder et al., 1992), entre otros. Sin embargo, a pesar de que los patrones calculados por estos algoritmos han mostrado ser útiles para varias tareas de clasificación, estos calculan un gran número de patrones en el proceso de la minería. El número de estos patrones crece a medida que disminuye el umbral de soporte y/o el umbral de semejanza, lo cual afecta negativamente en el rendimiento de los clasificadores. Además, muchos de estos patrones no son de utilidad como atributos para la clasificación ya que no son representativos para alguna clase en específico. Por este motivo se han reportado trabajos que incluyen métodos de selección de patrones en los esquemas de clasificación propuestos (Acosta-Mendoza, 2013; Acosta-Mendoza et al., 2013, 2014a,b). Dichos métodos de selección de patrones están basados en el uso de los enfoques convencionales de selección de atributos (Acosta-Mendoza et al., 2013, 2014b) y en la identificación de SFA emergentes (Acosta-Mendoza et al., 2014a). Estos últimos trabajos han reportado resultados competitivos en tareas de clasificación de imágenes, logrando una considerable reducción de la dimensionalidad del conjunto de atributos a utilizar. Además, de esta manera se lograron mejoras en los resultados de la clasificación.

4. Resultados alcanzados utilizando la minería de SFA

En esta sección se muestran los resultados obtenidos utilizando la minería de SFA en tareas de clasificación de imágenes, así como las mejoras alcanzadas al aplicar métodos de selección de patrones en estas tareas. Dichos resultados se muestran sobre varias colecciones de imágenes sintéticas y reales representadas en forma de grafos.

4.1. Colecciones de grafos utilizadas

En la literatura se han utilizado varias colecciones de imágenes con el objetivo de mostrar la utilidad de los patrones calculados por los algoritmos para la minería de SFA. Algunas de estas colecciones serán utilizadas en este artículo para mostrar una comparación entre los aportes reportados. Cada colección utilizada en este artículo está dividida de forma aleatoria en dos sub-colecciones (entrenamiento y prueba):

- *COIL* (Nene et al., 2008), donde se tienen imágenes de objetos reales tomados desde diferentes puntos de vistas. En este caso se utilizan 25 objetos aleatorios de 100 que posee la colección y está dividida en 198 (11 %) imágenes para el entrenamiento y 1602 para prueba. Las imágenes son representadas en forma de grafo haciendo uso de las pirámides irregulares de grafos de cada imagen que proveen una jerarquía de las particiones a diferentes niveles de resolución (Brun and Kropatsch, 2001; Kropatsch et al., 2005) y seleccionando el grafo de mejor calidad utilizando una la medida propuesta en (Morales-González and García-Reyes, 2011).

- *GREC* (Riesen and Bunke, 2008), donde las imágenes representan símbolos de los planos arquitectónicos o electrónicos. Compuesta por 1100 imágenes y está dividida en 572 (52 %) imágenes para el entrenamiento y 528 para prueba. Las imágenes se representan en forma de grafos utilizando los puntos críticos en las imágenes seleccionados de forma semi-automática presentado en (Riesen and Bunke, 2008).

- Imágenes sintéticas (*CoenenDB*) obtenidas mediante el Generador aleatorio de imágenes de Coenen[1], donde las imágenes representan dos tipos de vistas (marítimas y terrestres). Esta colección está compuesta por 2000 imágenes y está dividida en 1200 (60 %) imágenes para el entrenamiento y 800 para prueba. Las imágenes se representaron en forma de grafos utilizando la información de las hojas del árbol generado mediante la técnica de quad-tree (Finkel and Bentley, 1974).

Las características específicas de cada colección se muestran en la tabla 1.

Cuadro 1: Colecciones de imágenes utilizadas.

Colección	**COIL**	**GREC**	**CoenenDB**
Cantidad grafos	1800	1100	2000
Cantidad etiquetas de vértices	152	4	18
Cantidad etiquetas de aristas	27	24	24
Tamaño promedio de los grafos	135	11	49
Cantidad clases	25	22	2

4.2. Impacto de la selección de patrones en clasificación de imágenes basada en la minería de SFA

En los esquemas convencionales de clasificación de imágenes basados en minería de SFA lo primero que se realiza es la construcción de la colección de grafos que representan las imágenes dadas (Jiang and Coenen, 2008; Acosta-Mendoza et al., 2012a,b,c; Morales-González et al., 2014). Luego, se aplica un algoritmo para la minería de SFA y con los patrones obtenidos se crean los vectores de atributos. Estos vectores se le pasan a un algoritmos de clasificación para que se entrene y construya un modelo que se utiliza para etiquetar (clasificar) las imágenes del conjunto de prueba. Mediante este esquema se han reportado buenos resultados de clasificación; sin embargo, el número de patrones calculados se hace improcesable bajo algunas condiciones como: decremento de los umbrales de soporte y semejanza. Debido a que muchos de estos patrones no son de utilidad para la clasificación, se han propuesto módulos de selección de patrones para reducir la dimensionalidad de los vectores de atributos a utilizar en la clasificación (Acosta-Mendoza et al., 2013, 2014a).

En el trabajo presentado por Acosta-Mendoza *et al.* (Acosta-Mendoza et al., 2013) se reportan buenos resultados haciendo uso de los algoritmos de selección de atributos: ganancia de información (***I**nformation **G**ain IG*), chi-cuadrado (*CHI*) y el cociente de evaluación de la ganancia de información (***G**ain **R**atio **A**ttribute **E**valuation GRAE*). Por otro lado, en el trabajo (Acosta-Mendoza et al., 2014a) se propone el uso de un subconjunto de SFA que cumplen con las propiedades de patrones emergentes en las colecciones utilizadas. Para esto se necesita especificar el valor de un umbral γ que delimita los SFA emergentes del resto.

En esta sección se muestra una comparación entre el método convencional de clasificación de imágenes que usa todos los SFA y los métodos que utilizan solo un subconjunto de estos patrones como atributos para la clasificación. Primero se compara la reducción de dimensionalidad alcanzada. En la tabla 2 se muestra la cantidad de atributos utilizados para la clasificación de estos métodos en cada colección utilizada en este artículo. Esta tabla está compuesta por cuatro conjuntos de cinco columnas que especifican las colecciones de imágenes y una columna final que especifica un clasificador por fila. Las cinco colunmas de cada colección indican la cantidad de atributos que se utilizaron para la clasificación con su correspondiente clasificador especificado. El número de atributos seleccionado por los algoritmos de filtrado (IG, CHI y GRAE) fue obtenido experimentalmente en los rangos: [50,300] para CoenenDB, [50,600] para GREC y [50,1500] para COIL. Para la obtención de forma experimental del valor del umbral para los emergentes (γ) se utilizó el rango [0.2,0.8], siendo $\gamma = 0{,}4$ el mejor valor para este umbral en las colecciones CoenenDB y GREC, y $\gamma = 0{,}6$ para la colección COIL.

La dimensionalidad fue reducida notablemente utilizando las estrategias de selección de patrones (ver tabla 2). Esta reducción está sobre el 50 % en el 70 % de los experimentos realizados, mejorando la eficiencia de los algoritmos de clasificación.

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

Cuadro 2: Cantidad de atributos usados en el proceso de clasificación.

CoenenDB (δ = 20 %)					GREC (δ = 3 %)					COIL (δ = 30 %)					
Todos	**CHI-Q**	**IG**	**GRAE**	**PE**	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**	**Clasificador**
	125	275	250			425	400	525			1500	1400	1200		SVM
	100	100	100			450	500	500			1450	1500	1000		BayesNet
745	125	125	125	132	715	50	50	50	109	2668	50	50	50	208	AdaBoost
	250	250	300			325	275	200			1500	400	375		Reg.
	150	150	150			450	300	550			1400	50	50		D-Table.
	275	175	300			550	550	200			950	1450	550		J48graft

Por otro lado, en la tabla 3 se muestran los resultados de la clasificación obtenidos en las colecciones de imágenes utilizando los esquemas de clasificación con y sin la selección de patrones. En esta tabla se muestra una comparación entre el uso o no del módulo de selección. Esta comparación se muestra con el objetivo de analizar la utilidad de los métodos de clasificación basado en la minería SFA que utilizan la selección de patrones.

La tabla 3 está compuesta por dos subtablas que muestran los resultados del (a) accuracy y (b) F-measure, respectivamente. La primera y segunda columna de estas subtablas especifican la colección utilizada y el valor del umbral de soporte, respectivamente. Las subtablas se dividen en tres conjuntos de columnas que representan los resultados alcanzados con los clasificadores especificados en la parte superior de estas. Cada conjunto de columnas está compuesto por cinco columnas que indican los resultados de la clasificación utilizando: todos los atributos, los atributos seleccionados con chi-cuadrado (CHI), los seleccionados con ganancia de información (IG), los seleccionados según el cociente de la ganancia de información (GRAE), y los patrones emergentes (PE), respectivamente. Finalmente se muestran los promedios de los resultados de varios clasificadores respecto a las diferentes colecciones.

Cuadro 3: Resultados (%) de la clasificación utilizando varios clasificadores sobre diferentes colecciones de grafos con y sin el uso de varios algoritmos de selección de patrones.

(a) Accuracy

		J48graft					D-Table					Reg.				
Colección	δ	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**
CoenenDB	20 %	97.25	97.50	97.50	**97.75**	97.25	94.38	**95.88**	94.00	95.25	93.00	96.25	96.75	96.75	**96.88**	96.75
GREC	3 %	**82.20**	81.63	81.63	**82.20**	81.59	65.72	65.72	**66.48**	65.72	65.43	83.14	**85.61**	83.52	82.39	82.17
COIL	30 %	79.96	82.95	79.21	82.33	**86.27**	52.06	55.12	58.43	**63.17**	61.17	74.34	79.56	81.71	85.27	**87.89**
Promedio		86.47	87.36	86.11	87.43	**88.37**	70.72	72.24	72.97	**74.71**	73.21	84.58	87.31	87.33	88.18	**88.94**

		AdaBoost					BayesNet					SVM				
Colección	δ	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**
CoenenDB	20 %	**94.00**	**94.00**	**94.00**	**94.00**	92.75	90.38	92.75	92.75	**93.25**	90.00	95.38	95.75	95.75	96.25	**96.88**
GREC	3 %	–	–	–	–	–	87.88	**88.07**	87.88	**88.07**	87.30	**94.51**	92.42	92.61	93.37	92.22
COIL	30 %	–	–	–	–	–	**90.51**	90.07	90.13	89.70	86.50	90.20	89.45	89.26	**91.14**	90.28
Promedio		31.33	31.33	31.33	31.33	30.92	89.59	90.30	90.25	**90.34**	87.93	93.36	92.54	92.54	**93.59**	93.13

(b) F-measure

		J48graft					D-Table					Reg.				
Colección	δ	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**
CoenenDB	20 %	97.23	97.50	97.50	**97.76**	97.24	94.49	**95.94**	94.13	95.31	92.96	96.21	96.73	96.72	**96.86**	96.73
GREC	3 %	**86.96**	**86.96**	**86.96**	85.11	81.70	28.13	28.13	**30.51**	28.13	27.86	78.43	80.00	85.11	**85.71**	78.80
COIL	30 %	**91.18**	82.89	84.56	**91.18**	85.90	48.80	58.00	54.20	**63.60**	61.90	67.06	78.08	81.82	80.77	**87.60**
Promedio		**91.79**	89.12	89.67	91.35	88.28	57.14	60.69	60.61	**62.35**	60.91	80.57	84.94	**87.88**	87.78	87.71

		AdaBoost					BayesNet					SVM				
Colección	δ	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**	**Todos**	**CHI-Q**	**IG**	**GRAE**	**PE**
CoenenDB	20 %	**93.89**	**93.89**	**93.89**	**93.89**	93.36	90.29	92.37	92.37	**93.02**	90.90	95.39	95.74	95.71	96.22	**96.89**
GREC	3 %	14.50	14.50	14.50	14.50	13.84	86.96	86.96	86.96	86.96	86.63	89.36	88.89	88.89	**91.30**	86.60
COIL	30 %	15.54	15.63	15.63	**18.31**	16.00	**87.32**	**87.32**	84.35	85.14	85.70	**92.19**	90.37	86.52	89.71	90.00
Promedio		41.31	41.31	41.31	**42.23**	41.07	88.19	**88.88**	87.89	88.37	87,74	92.31	91.67	91.66	**92.41**	91,16

Los resultados de la clasificación alcanzados utilizando los método de clasificación utilizando un subconjunto de patrones son competitivos con los resultados logrados utilizando todos los patrones. También, es importante señalar que al utilizar el módulo de selección se usa un número mucho menor de atributos en la clasificación.

Adicionalmente, en la tabla 4 se presentan comparaciones de significancia estadística entre los clasificadores utilizando todos los atributos y utilizando solo un subconjunto de atributos. Para esta comparación se utiliza una prueba de significancia estadística reportada en (García and Herrera, 2008) (*Bergmann* (Bergmann and Hommel, 1988)) con 0.5 como valores de α. En la primera columna de la tabla 4, "Todos" representa el método que utiliza todos los atributos calculados por VEAM (Acosta-Mendoza et al., 2012a), mientras "IG", "CHI", "GRAE" y "PE" representan los métodos que incluyen el módulo de selección mediante ganancia de información, chi-cuadrado, cociente de evaluación de la ganancia de información y las propiedades de los patrones emergentes, respectivamente. Las columnas

restantes muestran el enfoque que se identificó como mejor opción según la prueba de significancia estadística. El símbolo "–" indica que no existe diferencias estadísticamente significativas entre los diferentes enfoques.

Cuadro 4: Pruebas de significancia estadística para diferentes clasificadores en varias colecciones de grafos (imágenes) utilizando todos los atributos y usando los atributos seleccionados por diferentes algoritmos de selección, donde α =0.05.

Classifier	J48graft	Decision Table	Regression	AdaBoost	BayesNet	SVM
Todos vs. GRAE	–	–	GRAE	–	–	–
Todos vs. CHI-Q	–	–	–	–	–	–
Todos vs. IG	–	–	IG	–	–	–
Todos vs. PE	–	PE	PE	–	–	–
IG vs. GRAE	–	–	–	–	–	–
IG vs. CHI	–	–	–	–	–	–
CHI vs. GRAE	–	–	–	–	–	–
IG vs. PE	–	–	PE	IG	–	–
CHI vs. PE	PE	–	PE	–	CHI	–
GRAE vs. PE	GRAE	–	–	–	–	–

A partir de la información mostrada en las tablas 2, 3 y 4, se puede concluir que el uso de algoritmos de selección es de ayuda para el desarrollo de un mejor método de clasificación de imágenes basado en la minería de SFA. El algoritmo de selección GRAE es la mejor opción. Esto se debe a los resultados globales que se mostraron en las tablas mencionadas, donde se logra una considerable reducción de atributos y una buena calidad en los resultados de la clasificación.

5. Conclusiones y trabajo futuro

Muchos han sido los esfuerzos orientados al mejoramiento de la clasificación tanto en eficiencia como en eficacia de los métodos basados en la minería de SFA. Se han reportado trabajos que han tenido un impacto favorable en las tareas de clasificación de imágenes. Con los aportes realizados por varios autores se ha logrado la disminución de la dimensionalidad de los atributos a utilizar en la clasificación con mejoras relevantes en sus resultados. Estos aportes son comparados en este trabajo, en tareas de clasificación de imágenes. Dichas comparaciones permitieron mostrar la utilidad de la selección de patrones y atributos en este tipo de tareas. Finalmente, Al aplicar los métodos reportados para la clasificación de imágenes, que incluyen el módulo de selección, se logró una reducción de más del 50 % de la dimensionalidad de los vectores de atributos en el 70 % de los experimentos realizados.

Como trabajo futuro, se desarrollarán algoritmos que identifiquen patrones representativos directamente en el proceso de la minería. Esto permitirá disminuir el tiempo de procesamiento de los métodos para la clasificación al eliminar el módulo de selección como post-procesamiento intentando mantener los resultados alcanzados hasta el momento.

Referencias

Acosta-Mendoza, N., July 2013. Clasificación de imágenes basada en subconjunto de subgrafos frecuentes aproximados. Master's thesis, The National Institute of Astrophysics, Optics and Electronics of Mexico (INAOE).

Acosta-Mendoza, N., Gago-Alonso, A., Carrasco-Ochoa, J., Martínez-Trinidad, J., Medina-Pagola, J., november 2013. Feature Space Reduction for Graph-Based Image Classification. In: Proceedings of the 18th Iberoamerican Congress on Pattern Recognition (CIARP'13). Vol. Part I, LNCS 8258. Springer-Verlag Berlin Heidelberg, Havana, Cuba, pp. 246–253.

Acosta-Mendoza, N., Gago-Alonso, A., Carrasco-Ochoa, J., Martínez-Trinidad, J., Medina-Pagola, J., 2014a. Improving Graph-Based Image Classification by using Emerging Patterns as Attributes. Currentrly with minor revision status in Knowledge-Based Systems.

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

Acosta-Mendoza, N., Gago-Alonso, A., Medina-Pagola, J., Carrasco-Ochoa, J., Martínez-Trinidad, J., 2014b. La minería de subgrafos frecuentes aproximados para la clasificación de imágenes. Tech. rep., accepted in Serie Gris, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV).

Acosta-Mendoza, N., Gago-Alonso, A., Medina-Pagola, J. E., 2012b. Clasificaión de imágenes utilizando minería de subgrafos frecuentes aproximados. Revista Cubana de Ciencias Informáticas (RCCI) 5 (4), 1–10.

Acosta-Mendoza, N., Morales-González, A., Gago-Alonso, A., García-Reyes, E., Medina-Pagola, J., 2012c. Classification using frequent approximate subgraphs. In: Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP'12). Vol. LNCS 7441. Buenos Aires, Argentina, Springer-Verlag Berlin Heidelberg, pp. 292–299.

Ambauen, R., Fischer, S., Bunke, H., 2003. Graph edit distance with node splitting and merging, and its application to diatom identification. In: Proceedings of the 4th IAPR international conference on Graph based representations in pattern recognition. GbRPR'03. Springer-Verlag, Berlin, Heidelberg, pp. 95–106.

Bergmann, G., Hommel, G., 1988. Improvements of general multiple test procedures for redundant systems of hypotheses. In: P.Bauer, G. Hommel, and E. Sonnemann, editors, Multiple Hypotheses Testing. Springer, Berlin, pp. 100–115.

Brun, L., Kropatsch, W., 2001. Introduction to combinatorial pyramids. Digital and image geometry: advanced lectures, 108–128.

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

Cheng, H., Yan, X., Han, J., 2010. Mining graph patterns. In: Aggarwal, C., Wang, H. (Eds.), Managing and Mining Graphs Data. Vol. 40. Advances in Database Systems. Springer., pp. 365–392.

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

Eichinger, F., Böhm, K., 2010. Software-Bug Localization with Graph Mining. In: Aggarwal, C. C., Wang, H. (Eds.), Managing and Mining Graph Data. Vol. 40 of Advances in Database Systems. Springer-Verlag New York.

Elsayed, A., Coenen, F., Jiang, C., nana, F. G.-F., Sluming, V., 2010. Corpus Callosum MR Image Classification. Knowledge-Based Systems 23, 330–336.

Fellman, P., November 2008. Modeling terrorist networks-complex systems at the mid-range. In: Downloaded from the internet.

Finkel, R., Bentley, J., 1974. Quad Trees: A Data Structure for Retrieval on Composite Keys. Acta Informatica 4, 1–9.

Gago-Alonso, A., Carrasco-Ochoa, J. A., Medina-Pagola, J. E., Martínez-Trinidad, J. F., August 2010a. Full Duplicate Candidate Pruning for Frequent Connected Subgraph Mining. Integrated Computer-Aided Engineering 17, 211–225.

Gago-Alonso, A., Puentes-Luberta, A., Carrasco-Ochoa, J. A., Medina-Pagola, J. E., Martínez-Trinidad, J. F., 2010b. A new algorithm for mining frequent connected subgraphs based on adjacency matrices. Intelligent Data Analysis 14, 385–403.

García, S., Herrera, F., 2008. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research 9, 2677–2694.

Han, J., Cheng, H., Xin, D., Yan, X., November 2007. Frequent pattern mining: Current status and future directions. In: Data Mining and Knowledge Discovery. Vol. 15. 10th Anniversary Issue, pp. 55–86.

Holder, L., Cook, D., Bunke, H., 1992. Fuzzy substructure discovery. In: ML92: Proceedings of the ninth international workshop on Machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 218–223.

Hossain, M. S., Angryk, R. A., 2007. Gdclust: A graph-based document clustering technique. In: ICDMW '07: Proceedings of the Seventh IEEE International Conference on Data Mining Workshops. IEEE Computer Society, Washington, DC, USA, pp. 417–422.

Huan, J., Wang, W., Prins, J., 2003. Efficient mining of frequent subgraphs in the presence of isomorphism. In: The 3rd IEEE International Conference on Data Mining. Melbourne, FL, pp. 549–552.

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

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

Jiang, C., Coenen, F., 2008. Graph-based Image Classification by Weighting Scheme. In: Proceedings of the Artificial Intelligence. Springer, Heidelberg, pp. 63–76.

Jiang, C., Coenen, F., Zito, M., 2013. A survey of frequent subgraph mining algorithms. Knowledge Engineering Review 23 (4), 302–308.

Ketkar, N., Holder, L., Cook, D., August 2006. Mining in the proximity of subgraphs. Analysis and Group Detection KDD Workshop on Link Analysis: Dynamics and Statics of Large Networks.

Kong, X., Yu, P., Wang, X., Ragin, A., 2013. Discriminative feature selection for uncertain graph classification. In: Proceedings of Computing Research Repository. Vol. abs1301.6626. CoRR.

Kropatsch, W., Haxhimusa, Y., Pizlo, Z., Langs, G., 2005. Vision pyramids that do not grow too high. Pattern Recognition Letters 26, 319–337.

Kuramochi, M., Karypis, G., 2004. GREW: A Scalable Frequent Subgraph Discovery Algorithm. Technical Report, University of Minnesota, Department of Computer Science 04-024.

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

Li, J., Zou, Z., Gao, H., 2012. Mining frequent subgraphs over uncertain graph databases under probabilistic semantics. VLDB J. 21 (6), 753–777.

Messmer, B. T., Bunke, H., 1998. A new algorithm for error-tolerant subgraph isomorphism detection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 20(5):493–504.

Morales-González, A., Acosta-Mendoza, N., Gago-Alonso, A., García-Reyes, E., Medina-Pagola, J., 2014. A new proposal for graph-based image classification using frequent approximate subgraphs. Pattern Recognition 47 (1), 169–177.

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

Nene, S., Nayar, S., Murase, H., 2008. Columbia object image library (coil-100). Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshop, SSPR & SPR 2008.

Neuhaus, M., Bunke, H., 2004. A probabilistic approach to learning costs for graph edit distance. In: J. Kittler, M. Petrou, and M. Nixon, eds. Proceedings 17th International Conference on Pattern Recognition, Cambridge, United Kingdom. pp. Vol. 3, pp. 389–393.

Nijssen, S., Kok, J., 2004. A Quickstart in Frequent Structure Mining can make a Difference. In: The 10th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA, pp. 647–652.

Riesen, K., Bunke, H., 2008. IAM Graph Database Repository for Graph Based Pattern Recognition and Machine Learning. Orlando, USA, pp.

208–297.

Sanfeliu, A., Fu, K. S., 1983. A distance measure between attributed relational graphs for pattern recognition. In: IEEE Transactions on Systems, Man, and Cybernetics (Part B). pp. 13(3):353–363.

Song, Y., Chen, S.-S., 2006. Item sets based graph mining algorithm and application in genetic regulatory networks. Data Mining, IEEE International Conference on Volume, Issue, 337–340.

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

Xiao, Y., Wu, W., Wang, W., He, Z., 2008. Efficient Algorithms for Node Disjoint Subgraph Homeomorphism Determination. In: Proceedings of the 13th international conference on Database systems for advanced applications. Springer-Verlag, Berlin, Heidelberg, New Delhi, India, pp. 452–460.

Yan, X., Huan, J., 2002. gSpan: Graph-Based Substructure Pattern Mining. In: International Conference on Data Mining. Maebashi, Japan.

Zhang, S., Yang, J., 2008. RAM: Randomized Approximate Graph Mining. In: The 20th International Conference on Scientific and Statistical Database Management. Hong Kong, China, pp. 187–203.

Zhang, S., Yang, J., Cheedella, V., 2007. Monkey: Approximate Graph Mining Based on Spanning Trees. In: International Conference on Data Engineering. IEEE ICDE, Los Alamitos, CA, USA, pp. 1247–1249.

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

Zou, Z., Li, J., Gao, H., Zhang, S., 2009. Frequent subgraph pattern mining on uncertain graph data. In: CIKM'09: Proceeding of the 18th ACM conference on Information and knowledge management. ACM, New York, NY, USA, pp. 583–592.

Zou, Z., Li, J., Gao, H., Zhang, S., 2010. Mining frequent subgraph patterns from uncertain graph data. IEEE Trans. on Knowl. and Data Eng. 22 (9), 1203–1218.