Engineering Applications of Artificial Intelligence 50 (2016) 215–225

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Improving graph-based image classification by using emerging patterns as attributes

Niusvel Acosta-Mendoza [a,b,*], Andrés Gago-Alonso [a], Jesús Ariel Carrasco-Ochoa [b], José Francisco Martínez-Trinidad [b], José Eladio Medina-Pagola [a]

[a] Advanced Technologies Application Center (CENATAV), 7a ♯ 21406 e/ 214 and 216, Siboney, Playa, CP: 12200, Havana, Cuba
[b] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro No. 1, Sta. María Tonantzintla, Puebla, CP: 72840, Mexico

ARTICLE INFO

Article history:
Received 17 September 2015
Received in revised form
20 December 2015
Accepted 18 January 2016
Available online 15 February 2016

Keywords:
Approximate graph mining
Approximate graph matching
Emerging patterns
Graph-based classification

ABSTRACT

In recent years, frequent approximate subgraph (FAS) mining has been used for image classification. However, using FASs leads to a high dimensional representation. In order to solve this problem, in this paper, we propose using emerging patterns for reducing the dimensionality of the image representation in this approach. Using our proposal, a dimensionality reduction over 50% of the original patterns is achieved, additionally, better classification results are obtained.

1. Introduction

Several authors have turned their attention to graph-based mining techniques where frequent subgraphs are detected allowing some distortions in the data (Holder et al., 1992; Anchuri et al., 2013; Flores-Garrido et al., 2014b; Acosta-Mendoza et al., 2015). Since in practical applications, it is not common to have two instances exactly equal, several algorithms for frequent approximate subgraph (FAS) mining have been developed (Song and Chen, 2006; Zhang and Yang, 2008; Zou et al., 2009; Jia et al., 2011; Acosta-Mendoza et al., 2012a; Anchuri et al., 2013; Flores-Garrido et al. 2014a,b). The usefulness of the patterns computed by these algorithms has been shown in different classification tasks (Holder et al., 1992; Jia et al., 2011; Acosta-Mendoza et al., 2012c; Gago-Alonso et al., 2013; Morales-González et al., 2014); but only a few of them have been applied to image classification (Acosta-Mendoza et al., 2012a,c, 2013; Morales-González et al., 2014). In these works, the authors have reported good results; however, they use a large number of graphs (patterns) as attributes for describing the images. This fact affects the performance of the classifiers due to the high dimensionality of the representation. Moreover, some of these patterns do not provide useful information for classification. To solve this problem, several strategies to reduce the set of patterns in several contexts have been proposed (Jin and Wang, 2011; Acosta-Mendoza et al., 2013; Kong et al., 2013; Aridhi et al., 2015; Moradi and Rostami, 2015). In this paper, we propose to use emerging approximate graph patterns for image classification. An emerging pattern is a discriminative pattern whose support increases significantly into a class regarding to the remaining classes. Therefore, we will focus on holding only emerging patterns from the set of frequent patterns obtained by a FAS mining algorithm and using them as attributes for image classification. Our proposal reduces the number of patterns to be considered as attributes, and as we will show in our experiments, it allows increasing the efficiency and effectiveness of image classification, compared against the best methods reported in the literature (Acosta-Mendoza et al., 2012a,c, 2013). Furthermore, to the best of our knowledge, this is the first work that uses FAS mining and emerging pattern selection for image classification.

The organization of this paper is the following. In Section 2, some basic concepts are presented. In Section 3, the related work on image classification methods based on FAS, as well as, the FAS mining algorithms are presented. In Section 4, our proposal for reducing the representation of images in the collection through the use of emerging patterns is introduced. Later, in Section 5, through some experiments, the efficiency and effectiveness of our proposal are shown. Finally, our conclusions and future work directions are discussed in Section 6.

* Corresponding author.
E-mail addresses: nacosta@cenatav.co.cu (N. Acosta-Mendoza), agago@cenatav.co.cu (A. Gago-Alonso), ariel@ccc.inaoep.mx (J.A. Carrasco-Ochoa), fmartine@ccc.inaoep.mx (J.F. Martínez-Trinidad), jmedina@cenatav.co.cu (J.E. Medina-Pagola).

http://dx.doi.org/10.1016/j.engappai.2016.01.030
0952-1976/

2. Basic concepts

This work is focused on collections of simple undirected labeled graphs. Henceforth, when we refer to a graph, we are assuming this type of graphs.

A *labeled graph* in the domain of all possible labels $L = L_V \cup L_E$, where L_V and L_E are the label sets for vertices and edges respectively, is a 4-tuple, $G = (V, E, I, J)$, where V is a set whose elements are called *vertices*, $E \subseteq \{\{u, v\} \mid u, v \in V, u \neq v\}$ is a set whose elements are called *edges* (the edge $\{u, v\}$ connects the vertex u with the vertex v), $I : V \rightarrow L_V$ is a *labeling function* for assigning labels to vertices and $J : E \rightarrow L_E$ is a *labeling function* for assigning labels to edges.

Let $G_1 = (V_1, E_1, I_1, J_1)$ and $G_2 = (V_2, E_2, I_2, J_2)$ be two graphs, G_1 is a *subgraph* of G_2 if $V_1 \subseteq V_2$, $E_1 \subseteq E_2$, $\forall u \in V_1, I_1(u) = I_2(u)$, and $\forall e \in E_1, J_1(e) = J_2(e)$. In this case, we use the notation $G_1 \subseteq G_2$ and we also say that G_2 is a *supergraph* of G_1.

Given two graphs G_1 and G_2, a function f is an *isomorphism* between these graphs if $f : V_1 \rightarrow V_2$ is a bijective function, where:

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

If there is an isomorphism between G_1 and G_2, we say that G_1 and G_2 are *isomorphic*. If G_1 is isomorphic to G_3 and $G_3 \subseteq G_2$, then we say that there is a *sub-isomorphism* between G_1 and G_2, and we also say that G_1 is sub-isomorphic to G_2.

Let $D = \{G_1, \ldots, G_{|D|}\}$ be a collection of graphs and G be a labeled graph in L, the *support* value of G in D is defined as the fraction of graphs $G_i \in D$, such that there is a sub-isomorphism between G and G_i. This value of support is obtained through the following equation:

$$supp(G, D) = \frac{|\{G_i \in D : G \text{ is sub-isomorphic to } G_i\}|}{|D|} \quad (1)$$

Let Ω be the set of all possible labeled graphs in L, the *similarity* between two graphs $G_1, G_2 \in \Omega$ is defined as a function $sim : \Omega \times \Omega \rightarrow [0, 1]$. We say that the graphs are very different if $sim(G_1, G_2) = 0$, the higher the value of $sim(G_1, G_2)$ the more similar the graphs are, and if $sim(G_1, G_2) = 1$ then there is an isomorphism between these graphs.

In Fig. 1, an example of the similarity evaluation between two graphs is shown. For our example, it is supposed that the similarity function is based on the product of the similarities between vertices and between edges of two graphs with the same topology. Also, it is assumed that the vertices with labels "A" and "C" could substitute the vertices with labels "D" and "A" with similarity 0.667 and 0.2, respectively (represented by dashed lines in Fig. 1). Moreover, it is assumed that edges with label "Y" could substitute edges with label "Z" with similarity 0.667 (represented by dotted lines in Fig. 1). Finally, the vertices with label "B" and edges with labels "W" cannot substitute any other vertex or edge except themselves with a similarity of 1.0 (represented by continuous lines in Fig. 1). Then, the graph G_1 is similar to G_2 with a similarity of 0.09.

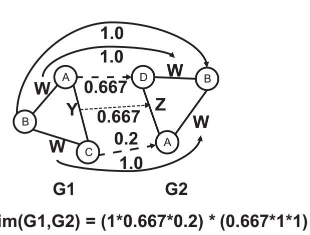

Fig. 1. Example of similarity between two graphs.

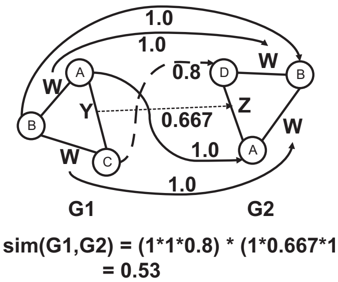

Fig. 2. Example of an alternative similarity between two graphs.

Between two labeled graphs, there could be more than one inexact correspondence among its vertices and edges. In Fig. 2, a different vertex correspondence between the graphs G_1 and G_2, of the previous example, is shown. This correspondence is achieved supposing that label "C" could substitute label "D" with a similarity 0.8. Then, using the same similarity function (the product of the similarities between vertices and between edges) a similarity value of 0.53 between G_1 and G_2 is obtained.

As we can see in Figs. 1 and 2, there could be several correspondences between two graphs. Therefore, $sim_{max}(G_1, G_2)$ is defined as the highest similarity value among all the possible correspondences between G_1 and G_2.

Using this definition of similarity between two graphs, a definition of support that allows inexact graph matching can be defined. Let $D = \{G_1, \ldots, G_{|D|}\}$ be a graph collection and G be a labeled graph in L, the support (denoted by *appSupp*) value of G in D, in terms of the similarity, is obtained through the following equation:

$$appSupp(G, D) = \frac{\sum_{G_i \in D} sim_{max}(G, G_i)}{|D|} \quad (2)$$

Using (2), when $appSupp(G, D) \geq \delta$, for a given support threshold δ, then G is a *frequent approximate subgraph* (FAS) in D. The value of the support threshold δ is in $[0, 1]$ because the similarity is defined in $[0, 1]$. *Frequent approximate subgraph mining* consists in finding all the FASs in a graph collection D, using a similarity function *sim* and a support threshold δ.

3. Related work

In the literature, only one work which aims to avoid using graph patterns that do not positively contribute to the classification task under the image classification approach based on FAS mining has been proposed (Acosta-Mendoza et al., 2013). This approach combines FAS mining with attribute selection algorithms for reducing the attribute dimensionality in the image classification tasks. In this work, the authors represented an image collection based on an n-dimensional vector approach using graph patterns, where the attributes of the vector are the patterns. Next, they apply an attribute selection algorithm for reducing the dimensionality of the vector; finally the classification task is performed over the reduced representation. In Acosta-Mendoza et al. (2013) a good dimensionality reduction was achieved, and at the same time the classification results were also improved.

On the other hand, an important tool for performing the aforementioned image classification task is a FAS mining algorithm. Several algorithms for FAS mining in graph collections have been developed based on different similarity functions for graph matching (Holder et al., 1992; Song and Chen, 2006; Xiao et al., 2008; Zhang and Yang, 2008; Zou et al., 2009; Jia et al., 2011; Acosta-Mendoza et al., 2012a; Gago-Alonso et al., 2013), but only two of them are based on substitution probabilities keeping the