Two Levels Similarity Modelling: a Novel Content Based Image Clustering Concept

Amar Djouak[1] and Hichem Maaref [2]

1 Agriculture high institute (ISA) (computer science and statistics laboratory)- Catholic Lille University. 48, boulevard Vauban. 59000. Lille. France.

E-mail: a.djouak@isa-lille.fr

2 IBISC Laboratory (CNRS FRE 3190) 40 Rue du Pelvoux, 91025 EVRY Cedex. France. Phone: +33169477555, Fax: +33169470306,

E-mail: hmaaref@univ-evry.fr

Abstract In this work, we applied a co-clustering concept in content based image recognition field. In this aim, we introduced a two levels similarity modelling (TLSM) concept. This approach is based on a new images similarity formulation using obtained co-clusters. The obtained results show a real improvement of image recognition accuracy in comparison with obtained accuracy obtained using one of classical co-clustering systems.

1 Introduction

The aim of any unsupervised classification method applied to contend based image retrieval is to gather considered images to be similar. In this case, algorithms treat the data in only one direction: lines or columns, but not both at the same time. Contrary to the one-dimensional clustering, co-clustering (also called bi-clustering) proposes to process the data tables by taking of account the lines and the columns in a simultaneous way. That implies the consideration of the existing correlation between the data expressed in lines and columns. Thus, co-clustering is a more complete data view since it includes a new concept “the mutual information” which is a bond between the random variables representing the clusters. An optimal coclustering is that which minimize the difference (the loss) of mutual information between the original random variables and mutual information between the clusters random variables. Several co-clustering structures are proposed in the literature [1],[2],[3]. The choice of one of these structures is directly related on the considered application. That is also related to the relational complexity between the lines and the columns elements. Among used co-clustering approaches, in [4], Qiu proposed a new approach dedicated to content based image categorization us-

ing the bipartite graphs to simultaneously model the images and their features. Indeed, the first bipartite graph set is associated to the images and the second unit is associated to their features. The bonds between the two sets characterize the existing degrees of correspondence between the images and their features. This method is very promising and could open the way for many possibilities to adapt co-clustering techniques to images recognition and retrieval applications. In this work, to propose one of the first coclustering based content based images recognition and search, our images are described by features vectors.

They form a two-dimensional table (lines/columns) such as the lines are the DataBase images and the columns are the features which describe them. Then, we propose in this paper a new co-clustering modelling which introduce a two levels similarity concept. The aim of this method is to improve the image retrieval accuracy and to optimize time processing. In other hand, we use one of classical co-clustering approaches to comparing its performances with those obtained with our method. Moreover, the choice of BIVISU system [5] applied initially to manage gene expression data is justified by its great conceptual simplicity.

This work is organized as follows: section 2 is devoted to the BIVISU co-clustering system. Section 3 introduce the two levels association concept and model the general architecture of the proposed approach. In section 4, different experimental results are presented and commented. Finally, one synthesizes presented work and exposes the future tracks.

2 Used Co-clustering Algorithm: BIVISU

The used co-clustering algorithm (BIVISU system) can detect several co-clusters forms (constant, constant rows, constant columns, additive and multiplicative co-clusters as defined in [5]). Also, this method uses parallel coordinate (PC) plots for visualize high dimensional data in a 2D plane. Besides visualization, it has been exploited to re-formulate the co-clustering problem [5].

To cluster the rows and columns simultaneously, clustering of rows is first performed for each pair of columns in used algorithm. Further columns are then merged to form a big co-cluster. The approach “merge and split”, i.e. merging paired columns and splitting in rows is performed then for obtaining final co-clusters. Actually, the "merge and split" process is repeated for each column-pair. Since there is either no significantly large co-cluster found or the same co-clusters are detected, only the three co-clusters are obtained for our example. This algorithm is based on a clear formalism and provides good quality co-clustering results. In the following paragraph, one extends it with introducing the two levels similarity concept.

3 Two Levels Similarity Model

The difficulty in choosing co-clustering algorithm parameters generates some search errors. To attenuate these errors, one tries to introduce a method which consider the similarity described by the obtained co-clusters in a different way from that usually used..

Initially, the features are extracted from the query image and from all the DataBase images. The extracted features can varying with application context. Two tables are built. One gathers the DataBase images represented by all their features. The other contains the same table with the addition of query image and its features.

The second stage consists in applying – separately- the co-clustering algorithm which allows to obtain a relatively small co-clusters number (a number preliminary chosen $\in[1,20]$). This stage allows to obtain a two-dimensional grouping based on the lines/columns interaction. The choice of a sufficiently small initial co-clusters number has justified by keeping a certain generalization capacity and that by merging the co-clusters wich are sufficiently dependent. After, this number can vary in an iterative way according to desired search quality.

Obtaining a lines/columns grouping generates some associations between these lines and columns. Indeed, co-clustering allows to associate the images to the features according to the obtained co-clusters scheme. In this case, on can obtain two associations sets : a unit to represent the co-clusters relating to the table “DataBase images /features” and another one for the co-clusters relating to the table “DataBase images + query image /features”. Each co-cluster is characterized by some associations which one can merge in the columns (images in our case) direction. These associations can be formalized by equation 1.

$$\text{Image x associated to feature y} \tag{1}$$

x vary between 1 and N for the first table (images DataBase before query image introduction, N is the images number in the DataBase) and x vary between 1 and N+1 for the second table (images DataBase after query image introduction,). y vary between 1 and M (M is the features number).

The following stage allows the hierarchical strategy construction which gives the images considered similar to the query image. This strategy is based on a associations combination extracted from the two tables at the same time (before and after query image introduction). In fact, the query image addition generates local lines/columns interactions modifications, which implies a change of the co-clustering diagram obtained before introduction of this new image. The idea is to exploit this variation effect to elaborate a comparative strategy of the interactions between the query image and the DataBase images and also the internal interactions beetwen DataBase images before query image introduction.

Indeed, two similarity levels are proposed: the level relating to associations between the query image and the DataBase images and the level relating to the DataBase images associations between them before query image introduction. That

allows after appropriate associations weighting to merge the strongest similarities in the two levels and to exclude the too weak judged bonds progressively and thus consequently to exclude the images considered not sufficiently similar to the query image. For weighting strategy, it is done by calculating the features number relating to each association. Then, a large number implies a strong association and vice versa. Thus, according to user and the application context, a qualitative thresholding is possible by fixing the minimal acceptable associations weights numbers. Figure 1 gives the algorithmic scheme of the two levels association method.

Start_Algorithm

Inputs : *Images DataBase/features table*
Images DataBase+ query image/features table

Co_clustering algorithm application

Co-clusters sets obtention

Direct similarity detection

Indirect similarity detection

Combination of tow similarities into the same set

Outputs : *Similar images appear in the order imposed by the associated features number*

End_Algorithm

Fig. 1. Two levels similarity method algorithm

Finally, one could introduce a precision/recall test which would validate (or not) the obtained result. A result evaluation would be made and that by comparing the obtained precision/recall values with minimal pre fixed thresholds according to application requirements. If the result is judged sufficiently good (stability of obtained error between two successive iterations), the processing operation would stop. If not, one would introduce a co-clusters number gradual increase to obtain a larger precision on the level of the associations development and thus a more precise result. Then, the co-clusters number would be increased in an iterative way until obtaining desired result.

4 Experimental Results

In this section, one experimented two level similarity concept and one compared co-clustering results with those obtained with BIVISU system (initially used for the expression gene data applications).

The tests carried out consist in introducing a features table (26 columns representing 26 features [6] : classical low level features , color histograms features, wavelet transform features and finally rotation translation and scaling invariance by Trace transform) of processed images (200 lines representing the images) and then introducing features vector for each query image and retaining the associations generated by the obtained co-clusters to determine the images subset as being most similar to each query image. Figure 2 shows a sample of the used DataBase images. Thus and for each introduced query image, the initially obtained co-clusters structure is modified, and that modify its local similarity with one or more co-clusters.

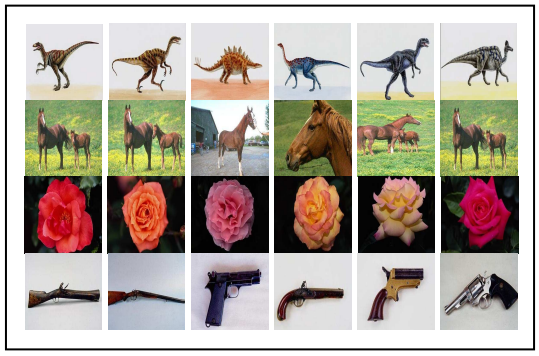

Fig. 2. Used images DataBase sample

Then, the images into the co-clusters associated to query image are turned over in an order which is based on the present features number in each associated co-cluster.

Generally, one notices an acceptable image search quality according to the images heterogeneity and the difficulty in formalizing the BIVISU system parameters adjustments (the maximum lines and columns dimensions per co-cluster...). This difficulty generates some coarse image search errors what leads us to say that an interactive use of this system in the content based image recognition field will be more beneficial in precision term.

Precision/recall diagrams for a sample of 24 images (of figure 2) are given in figure 3 for classical co-clustering method and for TLSM method. We can observe easely the added value and the superiority of our approach for the choosen images.

Finally, one notes that in spite of the first encouraging results, a thorough experimentation will allows to confirm the two levels similarity model potential and thus to give a solid experimental validation of this method.

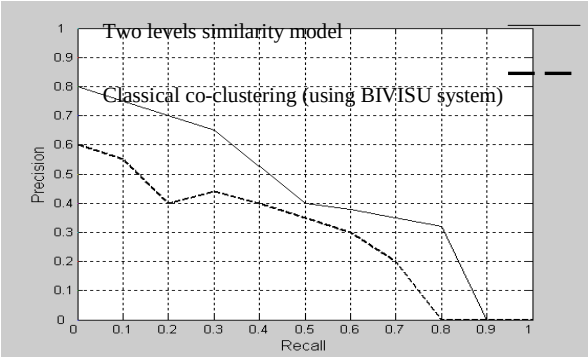

Fig. 3. Pecision/recall diagrams

5 Conclusion

In this work, one interested in an unsupervised classification tool : co-clustering, which is increasingly used in several applications. For this purpose, , one proposed a new co-clustering adaptation by introducing the two levels similarity model (T.L.S.M). This new approach takes into account the query image effect on the initial similarities between used DataBase images. To compare this method with classical co-clustering approaches, one studied a recent co-clustering system initially dedicated to the expression gene data classification and one applied it to the content based images recognition. The preliminary obtained results show the great potential of our method to improve classical co-clustering precision. However, a more solid experimental study allows to validate our method and to generalize its use in various applications.

References

1. I. S. Dhillon, S. Mallela et D. S. Modha: Information theoretic Coclustering, International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2003.
2. A.L. Oliveira, S.C. Madeira : Biclustering algorithms for biological data analysis: a syrvey, EEE/ACM Trans. Computational Biology and Bioinformatics , vol. 1, no. 1, pp. 24-45, 2004.
3. Y. Klugar, R. Basri, J. T. Chang, et M. Gerstein: Spectral biclustering of microarray data: co-clustering genes and conditions. In Genome Research, volume 13, pages 703–716, 2003.
4. G. Qiu : Bipartite graph partitioning and content-based image clustering, CVMP 2004, 1st European Conference on Visual Media Production (CVMP), The IEE, Savoy Place, London, 15-16 March 2004
5. K.O. Cheng, N.F. Law, W.C. Siu, T.H. Lau. : BiVisu: Software Tool for Bicluster Detection and Visualization. Bioinformatics Advance Access published June 22, 2007
6. A.Djouak, K, Djemal and H.Maaref : Modular statistical optimization and VQ method for image recognition – 2nd international workshop on artificial neural networks and intelligent information processing. ANNIP 2006. Portugal. Aout 2006.