Semi-Supervised Clustering and Aggregation of Relational Data

Hichem Frigui and Cheul Hwang
Multimedia Research Laboratory, CECS dept.,
University of Louisville, USA.

Abstract

We introduce a new semi-supervised approach for Clustering and Aggregating Relational Data (SS-CARD). We assume that data is available in a relational form, where we only have information about the degrees to which pairs of objects in the data are related. Moreover, we assume that the relational information is represented by multiple dissimilarity matrices. These matrices could have been generated using different sensors, features, or mappings. The SS-CARD uses partial supervision information that consists of a small set of must-link *and* cannot-link *constraints. The performance of the proposed algorithm is illustrated by using it to categorize a collection of 500 color images. The results are compared with those obtained by 3 other relational clustering methods.*

Keywords: *Relational Clustering, Semi-supervised clustering, Feature aggregation, Image database categorization.*

1 Introduction

The goal of cluster analysis is to find natural groupings in a set of N objects $O=\{o_1, \ldots, o_N\}$ such that objects in the same cluster are as similar as possible and objects in different clusters are as dissimilar as possible. The set of objects may be described by either object data or relational data. For object data, each object is described by a p-dimensional vector $\mathbf{x_i}=\{x_{i1}, \cdots x_{ip}\} \in \Re^p$, where x_{il} represents the l^{th} feature value. For relational data, only information that represents the degrees to which pairs of objects in the data are related is available. This information is usually stored in an $N \times N$ matrix called the (dis)similarity matrix. Relational clustering is more general in the sense that it is applicable to situations in which the objects to be clustered cannot be represented by numerical features. It is also more practical for situations where the computational complexity of the distance is high, when the distance measure does not have a closed form solution, or when groups of similar objects cannot be represented efficiently by a single prototype.

Clustering of object data has been an active field of research, and several methods have been proposed for this task. However, clustering of relational data has received much less attention. This is despite the fact that several applications would benefit tremendously from relational clustering algorithms. For instance, in content-based image retrieval [18, 17] or web data mining [11], it has been shown that the most effective (dis)similarity measures do not have a closed form. Thus, these measures could not be used in object-based clustering. Another issue in clustering complex data is that the relationship among the objects may be described by *multiple* (dis)similarity matrices. For instance, in image database categorization, we may have one similarity matrix that encodes color information, another matrix for texture information, and another one for structure information. Existing relational clustering algorithms can operate on only one similarity matrix at a time. Thus, one has to partition each matrix independently, or partition a single matrix that combines all matrices in a uniform way. However, the influence of the different similarity matrices is generally not equally important in the definition of the category to which similar patterns belong. Therefore, to obtain meaningful clusters across all similarity matrices, we need to learn *cluster-dependent* relevance weights for each similarity matrix.

Clustering high dimensional data collections is a challenging task. The problem is more complex if, in addition to clustering, one is also interested in learning cluster dependent feature relevance weights. One possible solution to alleviate this problem is to use partial supervision to guide the search process and narrow the space of possible solutions. Recently, semi-supervised learning has emerged as a new research direction in machine learning to improve the performance of unsupervised learning using some supervision information. This additional information is usually available in the form of constraints [15, 9] or labels [12]. Supervision in the form of constraints is more practical than providing class labels. This is because in many real world applications, the true class labels may not be known, and it is much easier to specify whether pairs of points should belong to the same or to the different clusters. In this pa-

978-1-4244-2703-1/08/$25.00 ©2008 IEEE

per, we propose a semi-supervised approach that performs Clustering and Aggregation of Relational Data (SS-CARD). The supervision information consists of a small set of constraints on which instances should or should not reside in the same cluster.

2 Related Work

Relational data consists of an $N \times N$ relational matrix $\mathbf{R}=[r_{jk}]$, where r_{jk} measures the relationship between objects j and k, and satisfy

$$r_{jk} \geq 0; \quad r_{jj} = 0; \quad r_{jk} = r_{kj}. \tag{1}$$

There are several relational clustering algorithms in the literature. One of the most well-known, SAHN [4], is a bottom-up approach that generates clusters by sequentially pairs of similar clusters. PAM [8] is another algorithm that is based on finding k representative objects (medoids) that minimize the sum of the within-cluster dissimilarities. CLARA [8] is a modified version of PAM that can handle very large data sets. Ng and Han [10] proposed a variation of CLARA, called CLARANS, that makes the search for the k-medoids more efficient.

The above algorithms generate a crisp (or hard) partition where each object belongs to only one cluster. In most real applications, categories are rarely well separated (i.e. clusters overlap). In this case, the partition is best described by fuzzy memberships [1], particularly, along the overlapping boundaries. Fuzzy clustering techniques allow the user to quantitatively distinguish between objects which are strongly associated with particular clusters from those that have only a marginal association with multiple clusters. Since relational clustering methods do not use the notion of prototypes, fuzzy memberships could be used to identify few representatives to summarize the clusters' content.

Some of the early fuzzy relational clustering methods include the algorithms proposed by Ruspini [14] and Rouben [13]. Other notable fuzzy techniques for clustering relational data include FANNY[8], and RFCM[7]. More recently, several improvements to the RFCM and FANNY have been proposed. For instance, the NERF [6] generalizes the RFCM to the case of arbitrary dissimilarity data. The competitive agglomeration of relational data [11] extends the NERF to find the optimal number of clusters through a competitive agglomeration process. A robust RFCM was proposed in [2].

Existing work in relational clustering addresses the issues of optimal partition, finding the optimal number of clusters, and robustness to noise and outliers. It assumes that the pairwise similarity information is available in one *global* matrix. However, in many real applications that involve clustering complex objects, multiple sources of information may generate multiple similarity matrices. In this case, the traditional approach is to cluster one global matrix that combines the partial similarities in a uniform way. Unfortunately, this approach may not be effective since the influence of the different similarity matrices is generally not equally important in the definition of the category to which similar patterns belong. Therefore, to obtain meaningful clusters across all similarity matrices, we need to learn cluster-dependent relevance weights for each similarity matrix. The CARD algorithm [3] addresses this issue. However, learning using CARD, like other unsupervised learning methods, may lead to sub-optimal solutions depending on the complexity of the data. In fact, CARD is more prone to local minimum since it attempts to learn the optimal partition and the optimal feature relevance weights simultaneously. To overcome this potential drawback, we propose a semi-supervised version of CARD, called SS-CARD. The supervision information consists of pairwise *must-link* and *cannot-link* constraints between few data samples. This type of supervision constraints have been used recently in object based clustering [15, 9], but not in relational based clustering.

3 Semi Supervised Clustering and Aggregation of Relational Data

We assume that we have S dissimilarity matrices $\mathbf{R}^1, \cdots, \mathbf{R}^S$. Each $\mathbf{R}^s=[r^s_{ij}]$ satisfies (1). The different dissimilarities could have been generated using different sensors, different sets of features, or a different mapping. For instance, in image database categorization, we could have one matrix that encodes the pairwise color dissimilarity, a second matrix for the pairwise texture dissimilarities, and a third one for the pairwise structure dissimilarities. Let $\mathbf{W}=[w_{is}]$, where $w_{is} \in [0,1]$ is the relevance weight for dissimilarity matrix $\mathbf{R}^s$ with respect to cluster i. A high (low) value of w_{is} indicates that the relational information in matrix $\mathbf{R}^s$ is relevant (irrelevant) for the definition of cluster i, and that this matrix should (should not) have a significant impact on the creation of this cluster. The global dissimilarity between objects j and k, $\bar{\mathbf{R}} = [\bar{r}_{jk}]$ is computed by aggregating the partial dissimilarities and their relevance weights, i.e.,

$$\bar{r}_{jk} = \sum_{s=1}^{S} w_{is}^q r_{jk}^s, \tag{2}$$

where $q \in (1, \infty)$ is a discriminant exponent.

Let $\mathcal{M}_l$ be the set of *must-link* pairs such as $(x_j, x_k) \in \mathcal{M}_l$ means that x_j and x_k should be assigned to the same cluster. Similarly, let $\mathcal{C}_l$ be the set of *cannot-link* pairs such as $(j,k) \in \mathcal{C}_l$ means that x_j and x_k should not be assigned to the same cluster. The SS-CARD minimizes the following

objective function:

$$J = \sum_{i=1}^{C} \frac{\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^2 u_{ik}^2 (\sum_{s=1}^{S} w_{is}^q r_{jk})}{2\sum_{k=1}^{N} u_{ik}^2} + \alpha\Big[\sum_{(j,k)\in\mathcal{M}_l}\sum_{i=1}^{C}\sum_{l=1,l\neq i}^{C} u_{ij}u_{lk} + \sum_{(j,k)\in\mathcal{C}_l}\sum_{i=1}^{C} u_{ij}u_{ik}\Big] \quad (3)$$

subject to

$$u_{ij} \in [0,1]\ \forall\, i,\ j; \quad \text{and} \quad \sum_{i=1}^{C} u_{ij} = 1,\ \forall\, i, \quad (4)$$

and

$$w_{is} \in [0,1]\ \forall\, i,\ s; \quad \text{and} \quad \sum_{s=1}^{S} w_{is} = 1,\ \forall\, i. \quad (5)$$

The first term in (3) is used to seek compact clusters and their partial dissimilarity relevance weights. The second term consists of the cost of violating the pairwise *must-link* and *cannot-link* constraints. The penalty terms are weighted by the membership values of the points that violate the constraints. In other words, the penalty term is larger when the points are at the core of the cluster (high membership). When both terms are combined, SS-CARD will seek compact clusters and their relevance weights while minimizing the number of violated constraints. The value of α in (3) controls the importance of the supervision compared to the sum of intra-cluster distances.

Minimization of J with respect to the cluster membership $\mathbf{U}$ yields

$$u_{vt} = u_{vt}^{RFCM} + u_{vt}^{Const} \quad (6)$$

where

$$u_{vt} = \frac{1}{\sum_{i=1}^{C}\left(d_{vt}^2/d_{vt}^2\right)}, \quad (7)$$

and

$$u_{vt}^{Const} = \frac{\alpha}{2d_{vt}^2}(\overline{C_t} - C_{vt}). \quad (8)$$

In (8), where C_{vt} and $\overline{C_t}$ are defined as

$$C_{vt} = \sum_{(x_t,x_k)\in\mathcal{M}_l}\sum_{l=1,l\neq v}^{C} u_{lk} + \sum_{(x_t,x_k)\in\mathcal{C}_l} u_{vk},$$

and

$$\overline{C_t} = \frac{\sum_{i=1}^{C}\left(\frac{\sum_{(x_t,x_k)\in\mathcal{M}_l}\sum_{l=1,l\neq i}^{C} u_{lk} + \sum_{(x_t,x_k)\in\mathcal{C}_l} u_{ik}}{d_{it}}\right)}{\sum_{i=1}^{C}\left(\frac{1}{d_{it}}\right)}.$$

In other words, the membership of a point in a given cluster depends on its relative proximity to that cluster (u^{RFCM} term) and the cost of constraints violations incurred by that cluster assignment(u^{Const} term). In (7), d_{ik} is the distance between sample $\mathbf{x}_k$ and the membership vector

$$\mathbf{v}_i = \frac{(u_{i1}^m, \cdots, u_{iN}^m)^t}{\sum_{j=1}^{N} u_{ij}^m}, \quad (9)$$

defined in terms of the relational matrix $\bar{\mathbf{R}}$ as

$$d_{ik}^2 = (\bar{\mathbf{R}}\mathbf{v}_i)_k - \frac{\mathbf{v}_i^t\bar{\mathbf{R}}\mathbf{v}_i}{2}, \quad (10)$$

Optimization of (3) with respect to w yields

$$w_{is} = \frac{1}{\sum_{p=1}^{S}\left(\bar{D}_{is}/\bar{D}_{ip}\right)^{1/(q-1)}}, \quad (11)$$

where

$$\bar{D}_{is} = \sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^2 u_{ik}^2 r_{jk}^s. \quad (12)$$

The SS-CARD algorithm is summarized below.

The U-CARD Algorithm

Fix number of clusters C, $m \in [1,\infty)$, and $q \in [1,\infty)$;
Initialize the fuzzy partition matrix $\mathbf{U}$;
Initialize relevance weights w_{is} to $1/S\ \ \forall\, i, s$;
REPEAT
Compute total dissimilarities $\bar{\mathbf{R}}$ using (2);
Compute membership vectors $\mathbf{v}_i$ using (9);
Compute distances using (10);
Update fuzzy memberships using (6);
Update relevance weights using (12);
UNTIL (*fuzzy memberships do not change*);

4 Application: Image Database Categorization

To illustrate the ability of SS-CARD to aggregate and cluster multiple dissimilarity measures derived from real data sets, we use them to categorize an image database. We use a subset of 500 color images from the COREL image collection. This subset includes 10 categories with 50 images in each one. Fig. 1 displays a sample image from each category. The images within most categories are selected to have high intra-group variations. For instances, the *"roses"* category includes images of roses of different colors (red, yellow, white, pink, etc.). Similarly, the *"guns"* category includes guns of different shapes and at different orientations.

Each image in the collection is characterized by six feature subsets: HSV color histogram, HSV color moments, Dominant colors, wavelet and Gabor texture, and edge histogram. For the wavelet and color moments features, we use

Figure 1: The 10 categories used in the experiment. Each category contains 50 images.

the Euclidean distance to compute the pairwise image dissimilarities. The high-dimensionality of Edge Histogram features (150-Dim) makes the Euclidean distance not the best choice. Instead, we use the histogram intersection distance [16]. For the dominant color features, computing the distances is not trivial. This is because each image can have a different number of dominant colors, and a dominant color from one image can be matched to multiple dominant colors in another image. One such distance that could be used is the Earth Mover's Distance (EMD) [18]. The EMD has attractive properties for content-based image retrieval and it was shown that it matches perceptual similarity better than other distances [18]. For the Garbor filtered feature, we also use the EMD. This is because an image filtered with a particular scale and a particular orientation may be partially matched at different scales and orientations [18]. For HSV Color histogram, we use Quadratic distance [5] which was shown to outperform the Euclidian distance. We will use $\mathbf{R}^{wave}$, $\mathbf{R}^{mmt}$, $\mathbf{R}^{EHD}$, $\mathbf{R}^{Dom}$, $\mathbf{R}^{Gab}$, and $\mathbf{R}^{Quad}$ to refer to these dissimilarity matrices.

We should note here that the histogram intersection distance, the EMD, and the Quadratic distance which are commonly used in content-based image retrieval, could not be used in an object-based clustering algorithm. For the EMD, it is not practical to solve the optimization problem for every data point in every cluster and in every iteration. Also, for all distances, the concept of cluster centroid (averaging the features) is not suitable for representing the elements of the cluster.

The SS-CARD was used to categorize the 500 images using the 6 partial dissimilarities. All dissimilarity matrices were normalized to have the same dynamic range. The results were compared with those obtained by the basic relational FCM (RFCM)[7], the SAHN [4], and FANNY [2]. Since these algorithms are not designed to aggregate multiple dissimilarity matrices, we use $\mathbf{R}^{Tot} = \mathbf{R}^{Gab}+\mathbf{R}^{Quad}+\mathbf{R}^{mmt}+\mathbf{R}^{EHD}+\mathbf{R}^{Dom}+\mathbf{R}^{wave}$ as input to these algorithms. We also compare the results of SS-CARD with those obtained by the basic CARD (no supervision) [3]. For all algorithms, we fix the number of clusters C=10 , m=2, and use the same initialization. For the SS-CARD, we first perform few iterations of CARD (i.e., no supervision), then to simulate user's supervision (i.e., feedback), we identify few samples on the clusters' boundary and use the ground truth to create must-link and cannot-link constraints.

The overall performance of the different algorithms is shown in Table 1, where the ground truth was used to measure the percentage of correctly partitioned images. As it can be seen, CARD and SS-CARD outperform the other 3 algorithms significantly. This is because these algorithms, while partitioning the data, learn cluster dependent feature relevance weights for each dissimilarity matrix. Moreover, as it can be seen, the semi-supervision information improves the results significantly.

Table 1: Comparison of 5 different algorithms.

SAHN	**FANNY**	**RFCM**	**CARD**	**SS – CARD**
41.40%	52.60%	61.60%	78.60%	90.00%

To analyze the results further, we display 20 representative images from sample clusters generated by the different algorithms. Figures 2 displays the results for the RFCM. As it can be seen, the two clusters do not match the user intuition of a cluster. This is because the RFCM does not have a provision for cluster dependent dissimilarity relevance weights. Thus, it tends to group images that are similar with respect to *all* partial dissimilarities. For instance, there are several other images of *"doors"* that have different colors in our collection. These images were not lumped into the cluster shown in Fig. 2(a) because the contribution of the color dissimilarity is too large.

Figure 3 displays sample images generated by CARD. These images are not the most representatives ones. Several incorrectly partitioned images, with low memberships, were selected to outline the hard to cluster images. The 3 shown clusters are the closest (share many images) with those in figure 2. As it can be seen, these clusters are more compatible with the user's notion of clusters. This improvement in performance is due mainly to the learned cluster dependent relevance weights. These relevance weights are shown in table 2. For instance, the cluster in figure 3 (c) does not mix images from the classes *"beach" "gun" "bus"* etc. as in Fig. 2(c). For this cluster,$\mathbf{R}^{Gab}$ is the most relevant feature. In other words, texture features from these images is more effective to represent this cluster. These cluster dependent relevance weights enables images of *"doors"* (with different color but similar texture) to be lumped into one group. However, this cluster is still not perfect as it includes few incorrect images. That is without partial supervision, the learning ability of the CARD is limited. The SS-CARD, On the other hand, uses few pairwise constraints to guide the clustering process. As a result, the performance improves significantly (from 78% to 90%). For instance, for the third

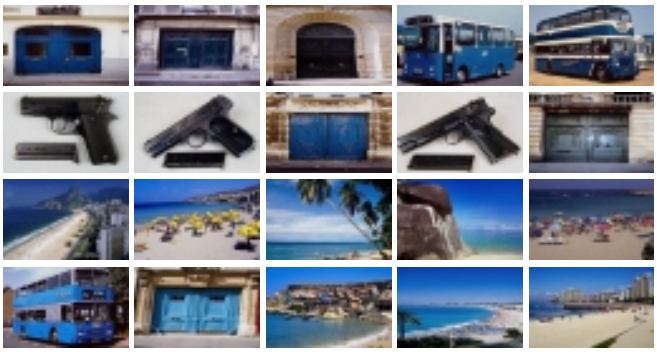
Table 2: Relevance weights for the three clusters displayed in Fig. 3.

Cluster	$\mathbf{R}^{Gab}$	$\mathbf{R}^{Quad}$	$\mathbf{R}^{mmt}$	$\mathbf{R}^{EHD}$	$\mathbf{R}^{Dom}$	$\mathbf{R}^{wav}$
Cluster 1	0.16	0.07	0.12	**0.19**	0.07	**0.38**
Cluster 2	**0.17**	0.12	**0.26**	0.10	0.14	**0.20**
Cluster 3	**0.28**	**0.18**	0.13	0.13	0.13	0.16

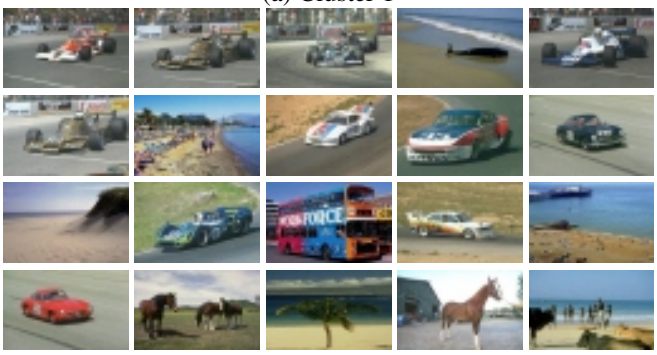
Table 3: Relevance weights for the three clusters displayed in Fig. 4.

Cluster	$\mathbf{R}^{Gab}$	$\mathbf{R}^{Quad}$	$\mathbf{R}^{mmt}$	$\mathbf{R}^{EHD}$	$\mathbf{R}^{Dom}$	$\mathbf{R}^{wav}$
Cluster 1	0.14	0.11	**0.17**	**0.19**	0.10	**0.29**
Cluster 2	**0.21**	0.10	**0.23**	0.14	0.14	**0.18**
Cluster 3	**0.26**	**0.17**	0.12	0.15	0.09	**0.21**

cluster described above, the SS-CARD adjusted the weights such that $\mathbf{R}^{EHD}$ and $\mathbf{R}^{Wav}$ became more relevant and the color features became less relevant. Figure 4 displays representative images of the same three clusters shown in 3. The feature relevance weights of all 10 clusters is shown in 3.



(a) Cluster 1



(b) Cluster 2

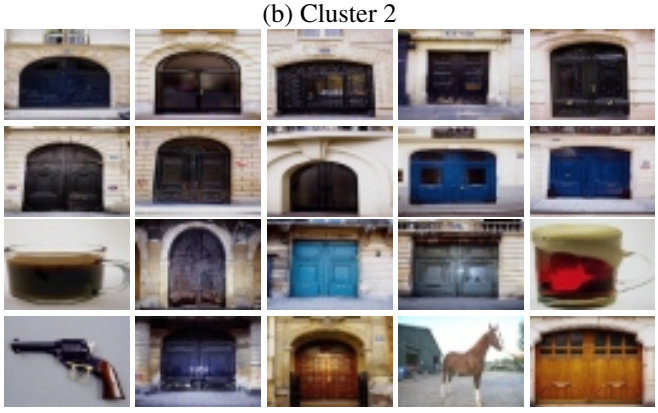
Figure 2: Top 20 representative images from two typical clusters generated by RFCM

5 Conclusions

In this paper, we have presented an approach that performs fuzzy clustering and aggregation of multiple relational data guided by a small set of pairwise constraints. SS-CARD minimizes one objective function for both the optimal partition and for the relevance weight of each partial dissimilarity matrix in each cluster. The supervision information improves the performance of CARD and leads to better feature relevance weights. SS-CARD is based on the well-known relational FCM algorithm. Thus, it can inherit most of the advantages of FCM-type clustering algorithms. For instance, techniques that were used to extend the RFCM to find the optimal number of clusters [11], and to reduce the effect of noise and outliers [2] could be adapted to SS-CARD. We are currently investigating these extensions.

(a) Cluster 1

(b) Cluster 2



(c) Cluster 3

Figure 3: 20 sample images from the first 3 clusters generated by CARD

(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

Figure 4: 20 sample images from the first 3 clusters generated by SS-RFCM

We have illustrated the performance of SS-CARD by using it to categorize a collection of color images. Categorization of generic images is a difficult task, mainly because different feature sets are suitable for different subsets of images. We have shown that using multiple dissimilarity matrices, SS-CARD can learn optimal relevance weights for each dissimilarity in the different image categories.

acknowledgment

This work was supported in part by NSF Awards IIS-0514319 and CBET-0730802, and by an Office of Naval Research award number N00014-05-10788. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the U. S. Government.

References

[1] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, 1981.

[2] R. N. Dave and S. Sen. Robust fuzzy clustering of relational data. In *proc. IEEE Conf. on Fuzzy Systems*, pages 713–727, 2002.

[3] H. Frigui and C. Hwang. Clustering and aggregation of relational data with applications to image database categorization. In *IASTED IMSA*, Hawaii, 2006.

[4] S. P. H. and S. R. R. *Numerical taxonomy*. Freeman, San Fransisco, 1973.

[5] J. Hafner, H. Sawhney, W. Flickner, and W. Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17:729–736, 1995.

[6] R. J. Hathaway and J. C. Bezdek. Nerf c-means: Non-euclidean relational fuzzy clustering. *Pattern Recognition*, 27(3):429–437, 1994.

[7] R. J. Hathaway, J. W. Davenport, and J. C. Bezdek. Relational duals of the c-means algorithms. *Pattern Recognition*, 22:205–212, 1989.

[8] L. Kaufman and P. J. Rousseeuw. *Finding groups in data*. John Wiley & Sons, New York, 1990.

[9] M. C. N. Grira and N. Boujemaa. Semi-supervised fuzzy clustering with pairwise-constrained competitive agglomeration. In *proc. IEEE Conf. on Fuzzy Systems*, pages 867–872, 2005.

[10] R. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *proc. 20th VLDB conference*, pages 144–155, 1994.

[11] N. O., F. H., K. R., and J. A. Extracting web user profiles using relational competitive fuzzy clustering. *International Journal on Artificial Intelligence Tools*, 9(4):509–526, 2000.

[12] W. Pedrycz. Algorithms of fuzzy clustering with partial supervision. *Pattern Recognition Letters*, 3:13–20, 1985.

[13] M. Roubens. Pattern classification problems and fuzzy sets. *Fuzzy Sets and Systems*, 1:239–253, 1978.

[14] E. Ruspini. Numerical methods for fuzzy clustering. *Information Science*, 12:319–350, 1970.

[15] A. B. S. Basu and R. Mooney. Active semi supervision for pairwise constrained clustering. In *proc. the SIAM int. Conf. on Data mining*, pages 333–344, 2004.

[16] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[17] J. Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture Libraries. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.

[18] R. Y., T. C., and G. L.J. A metric for distributions with applications to image databases. In *proc. 6th international conf. on Computer Vision*, pages 59–66, 1998.