Discovery of Interesting Regions in Spatial Data Sets Using Supervised Clustering

Christoph F. Eick, Banafsheh Vaezian, Dan Jiang, and Jing Wang

Department of Computer Science, University of Houston
Houston, Texas 77204-3010, U.S.A
{ceick, bvaezian, djiang, jwang3}@cs.uh.edu

Abstract. The discovery of interesting regions in spatial datasets is an important data mining task. In particular, we are interested in identifying disjoint, contiguous regions that are unusual with respect to the distribution of a given class; i.e. a region that contains an unusually low or high number of instances of a particular class. This paper centers on the discussion of techniques, methodologies, and algorithms to discover such regions. A measure of interestingness and a supervised clustering framework are introduced for this purpose. Moreover, three supervised clustering algorithms are proposed in the paper: an agglomerative hierarchical supervised clustering named SCAH, an agglomerative, grid-based clustering method named SCHG, and lastly an algorithm named SCMRG which searches a multi-resolution grid structure top down for interesting regions. Finally, experimental results of applying the proposed framework and algorithms to the problem of identifying hotspots in spatial datasets are discussed.

1 Introduction

Because of advances in database technologies, data collection techniques, and data gathering devices the amount of spatial data has been growing tremendously in recent years. The goal of spatial data mining is to automate the extraction of interesting and useful patterns that are not explicitly represented in spatial datasets.

This paper centers on discovering interesting regions in spatial datasets; in particular, on identifying disjoint, contiguous regions that are unusual with respect to the distribution of a given class, i.e. a region that contains an unusually low or high number of instances of a particular class. Methodologies, techniques, and algorithms are proposed for this purpose. Challenges that this task faces include the capability to find regions of arbitrary shape and at arbitrary levels of resolution, the definition of suitable parameterized measures of interestingness to instruct discovery algorithms what they are supposed to be looking for, and the need to reduce computational complexity due to the large size of most spatial datasets.

The paper assumes that datasets contain classified examples, and treats region discovery as a clustering problem in which clusters have to be found that maximize an externally given reward scheme. Section 2 proposes reward-based evaluation framework for

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): PKDD 2006, LNAI 4213, pp. 127 – 138, 2006.

region discovery. In sections 3 and 4 three supervised clustering algorithms are introduced and compared. Section 5 discusses related work and section 6 gives a conclusion. Table 1 summarizes the notations used in this paper.

Table 1. Notations used

Notation	Description
$O=\{o_1, \ldots, o_n\}$	Objects in a dataset (or training set)
n	Number of objects in the dataset
$c_i \subset O$	The i-th cluster
$X=\{c_1, \ldots, c_k\}$	A clustering solution consisting of clusters c_1 to c_k
$q(X)$	Fitness function that evaluates a clustering X
C	A class label

2 Measuring the Interestingness of a Set of Regions

As we explained earlier, our approach uses supervised clustering algorithms to identify interesting regions in a dataset. A region, in our approach, is defined as a surface containing a set of spatial objects; e.g. the convex hull of the objects belonging to a cluster. Moreover, we require regions to be disjoint and contiguous; that is, for each pair of objects belonging to a region, there always must be a path within this region that connects the pair of objects. Furthermore, we assume that the number of regions is not known in advance, and therefore finding the best number of regions is one of the objectives of the clustering process. Therefore, our evaluation scheme has to be capable of comparing clusterings that use a different number of clusters.

Our approach employs a reward-based evaluation framework. The quality q(X) of a clustering X is computed as the sum of the rewards obtained for each cluster $c \in X$. Cluster rewards are weighted by the number of objects that belong to a cluster c. In general, we are interested in finding larger clusters if larger clusters are equally interesting as smaller clusters. Consequently, our evaluation scheme uses a parameter β with $\beta>1$ and fitness increases nonlinearly with cluster-size dependent on the value of β, favoring clusters c with more objects.

$$q(X) = \sum_{c \in X} (reward(c) * |c|^{\beta}) \tag{1}$$

Selecting larger values for the parameter β usually results in a smaller number of clusters in the best clustering X. The proposed evaluation scheme is very general; different reward schemes that correspond to different measures of interestingness can easily be supported in this framework, and the supervised clustering algorithm that will be introduced in the second half of the paper can be run with different fitness functions without any need to change the clustering algorithm itself.

In this paper, due to the lack of space, we only introduce a single measure of interestingness that centers on discovering *hotspots* and *coldspots* in a dataset. The measure is based on a class of interest C, and rewards regions in which the distribution of class C significantly deviates from its prior probability, relying on a reward function τ. τ itself, see Fig. 1, is computed based on p(c,C), prior(C), and based on the following parameters: γ1, γ2, R+, R− with γ1≤1≤γ2; 1≥ R+,R−≥0, η>0 .

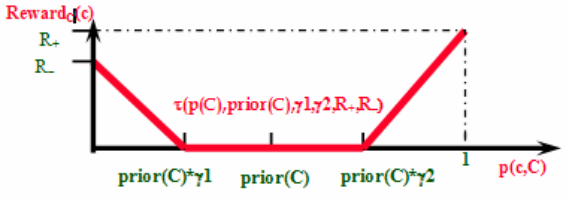


Fig. 1. The reward function τ_C for η=1

The fitness function q(X) is defined as follows:

$$q(X) = \sum_{i=1}^{|X|} \frac{\tau(p(c_i, C), prior(C), \gamma_1, \gamma_2, R+, R-, \eta) * (|c_i|)^{\beta}}{n^{\beta}} \tag{2}$$

with

$$\tau(p(c_i, C), prior(C), \gamma_1, \gamma_2, R+, R-, \eta) = \begin{cases} \left(\frac{((prior(C)*\gamma_1) - P(c_i, C))}{(prior(C)*\gamma_1)} * R-\right)^{\eta} & if \quad P(c_i, C) < prior(C)*\gamma_1 \\ \left(\frac{(P(c_i, C) - (prior(C)*\gamma_2))}{(1-(prior(C)*\gamma_2))} * R+\right)^{\eta} & if \quad P(c_i, C) > prior(C)*\gamma_2 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In the above formula prior(C) denotes the probability of objects in dataset belonging to the class of interest C. The parameter η determines how quickly the reward grows to the maximum reward (either R+ or R−). If η is set to 1 it grows linearly; in general, if we are interested in giving higher rewards to purer clusters, it is desirable to choose larger values for η; e.g. η=8.

Let us assume a clustering X has to be evaluated with respect to a class of interest "Poor" with prior(Poor) = 0.2 in a dataset that contains 1000 examples. Suppose that the generated clustering X subdivides the dataset into five clusters c1, c2, c3, c4, and c5 with the following characteristics: |c1| = 50, |c2| = 200, |c3| = 200, |c4| = 350, |c5| = 200; p(c1, Poor) = 20/50, p(c2, Poor) = 40/200, p(c3, Poor) = 10/200, p(c4, Poor) = 30/350, p(c5, Poor) = 100/200. Moreover, the parameters used in the fitness function are as follows: γ1 = 0.5, γ2 = 1.5, R+ = 1, R− = 1, β = 1.1, η=1. Due to the settings of γ1 = 0.5, γ2 = 1.5, clusters that contain between 0.5 x 0.2 = 10% and 1.5 x 0.2 = 30% instances of the class "Poor" do not receive any reward at all; therefore, no reward is given to cluster c2. The remaining clusters received rewards because the distribution

of class "Poor" in the cluster is significantly higher or lower than its prior distribution. For example, the reward for c1 that contains 50 examples is 1/7 x $(50)^{1.1}$; 1/7 is obtained as follows: $\tau(c1) = ((0.4\text{-}0.3)/(1\text{-}0.3)) * 1 = 1/7$.

$$q(X) = \frac{\frac{1}{7} * 50^{1.1} + 0 + \frac{1}{2} * 200^{1.1} + \frac{1}{7} * 350^{1.1} + \frac{2}{7} * 200^{1.1}}{1000^{1.1}} = 0.129$$

3 Supervised Clustering Algorithms for Region Discovery

As part of our research, we have designed and implemented seven supervised clustering algorithms three of which will be described in this Section.

3.1 Supervised Clustering Using Agglomerative Hierarchical Techniques (SCAH)

SCAH is an agglomerative, hierarchical supervised clustering algorithm. Initially, it forms single object clusters, and then greedily merges clusters as long as the clustering quality improves. In more detail, a pair of clusters (c_i, c_j) is considered to be a merge candidate if c_i is the closest cluster to c_j or c_j is the closest cluster to c_j. Distances between clusters are measured by using the average distance between the objects belonging to the two clusters. The pseudo code of the SCAH algorithm is given in Fig. 2.

Inputs:
A dataset O={o_1,...,o_n}
A dissimilarity *Matrix* $D = \{d(o_i,o_j) \mid o_i,o_j \in O\}$,
Output:
Clustering $X=\{c_1, c_2, \ldots, c_k\}$

Algorithm:
1) **Initialize**:
   Create single object clusters: $c_i = \{o_i\}, 1 \le i \le n;$
   Compute merge candidates
2) **DO FOREVER**
   a) Find the pair (c_i, c_j) of merge candidates that improves q(X) the most
   b) If no such pair exist terminate, returning $X = \{c_1, c_2, \ldots c_k\}$
   c) Delete the two clusters c_i and c_j from X and add the cluster $c_i \cup c_j$ to X
   d) Update merge candidates

Fig. 2. The SCAH Algorithm

In general, SCAH differs from traditional hierarchical clustering algorithms which merge the two clusters that are closest to each other in that it considers more alternatives for merging clusters. This is important for supervised clustering because merging two regions that are closest to each other will frequently not lead to a better clustering, especially if the two regions to be merged are dominated by instances belonging to different classes.

3.2 Supervised Clustering Using Hierarchical Grid-Based Techniques (SCHG)

Grid-based clustering methods are designed to deal with the large number of data objects in a high dimensional attribute space. A grid structure is used to quantize the space into a finite number of cells on which all clustering operations are performed. The main advantage of this approach is its fast processing time which is typically independent of the number of data objects, and only depends on the number of occupied cells in the quantized space.

SCHG is an agglomerative, grid-based clustering method. Initially, each occupied grid cell is considered to be a cluster. Next, SCHG tries to improve the quality of the clustering by greedily merging two clusters that share a common boundary. The algorithm terminates if q(X) cannot be improved by further merging.

3.3 Supervised Clustering Using Multi-Resolution Grids (SCMRG)

Supervised Clustering using Multi-Resolution Grids (SCMRG) is a hierarchical grid based method that utilizes a divisive, top-down search: each cell at a higher level is partitioned further into a number of smaller cells in the next lower level, but this process only continues if the sum of the rewards of the lower level cells is higher than the obtained reward for the cell at the higher level. The returned cells usually have different sizes, because they were obtained at different level of resolution. The algorithm starts at a user defined level of resolution, and considers three cases when processing a cell:

1. If a cell receives a reward, and its reward is larger than the sum of the rewards associated of its children and larger than the sum of rewards of its grandchildren, this cell is returned as a cluster by the algorithm.
2. If a cell does not receive a reward and its children and grandchildren do not receive a reward, neither the cell nor any of its descendents will be included in the result.
3. Otherwise, all the children cells of the cell are put into a queue for further processing.

The algorithm also uses a user-defined cell size as a depth bound; cells that are smaller than this cell size will not be split any further. The employed framework has some similarity with the framework introduced in the STING Algorithm [13] except that our version centers on finding interesting cells instead of cells that contain answers to a given query, and only computes cell statistics when needed and not in advance as STING does.

4 Experimental Evaluation

4.1 Datasets

In order to study the performance of the clustering algorithms presented in section 3, we conducted experiments on a benchmark consisting of 6 spatial datasets. Table 2

gives a summary for the datasets used. Objects belonging to those data sets consist of longitude and latitude and a non-spatial part which is the class label associated with that object.

Table 2. Datasets used in the benchmark

	Dataset Name	# of objects	# of classes
1	B-Complex9	3,031	2
2	Volcano	1,533	2
3	Earthquake-1	3,161	3
4	Earthquake-10	31,614	3
5	Earthquake-100	316,148	3
6	Wyoming-Poverty	493,781	2

B-Complex9 is a two dimensional synthetic spatial dataset whose examples are distributed having different, well-separated shapes. Earthquake and Volcano are spatial datasets containing the longitude and latitude of earthquakes and volcano eruptions, that are classified based on their severity of the event. Earthquake-1 and Earthquake-10 are smaller datasets contains 1% and 10% of the original data respectively. Wyoming-Poverty is a two dimensional spatial dataset indicating the poverty status of residents of the state of Wyoming based on 2000 census data. For more details about the datasets see [12].

4.2 Experimental Evaluation

The proposed algorithms have been evaluated on a benchmark consisting of the datasets described in section 4.1. We tested the algorithms' capability to identify very pure, potentially very small regions (β=1.01, η=6, γ1=0.5, γ2=1.5, R+=1, R−=1) and to identify larger regions (β=3, η=1, γ1=0.5, γ2=1.5, R+=1, R−=1). Table 3 summarizes the results and Fig. 6 and 7 visualize the result of SCAH, SCHG and SCMRG for the B-Complex9 and Volcano datasets. The result of SCAH on B-Complex9 and Volcano is identical for both set of parameters, so we visualized them only once. Purity and quality of final clustering and the number of clusters obtained are reported for each algorithm-dataset pair. Quality is measured using q(X) that was introduced in Section 2. Purity of a clustering is defined as the number of examples that belong to the most frequent class of a cluster over the total number of examples belonging to clusters. All Experiments were conducted on a DELL D600 workstation with an Intel Pentium M processor 1.80GHz and 1 GB of main memory. SCAH did not succeed in producing results (indicated by DNF in Table 3) within 6 hours for the Earthquake-10 dataset and ran out of storage for the Wyoming and Earthquake-100 datasets when creating the initial clustering.

Table 3. Experimental Results for $\beta = 1.01/\eta = 6$ and $\beta = 3/\eta = 1$

Dataset	Algorithms	SCAH	SCHG	SCMRG	SCAH	SCHG	SCMRG
	Parameters	$\beta = 1.01, \eta = 6$			$\beta = 3, \eta = 1$		
B-Complex9	Purity	1	0.998	1	1	0.997	0.863
	Quality	0.974	0.974	0.957	0.008	0.044	0.002
	Clusters	17	15	132	17	9	22
Volcano	Purity	1	0.692	0.979	1	0.692	0.885
	Quality	0.940	0.091	0.822	1E-5	7E-4	1E-4
	Clusters	639	56	311	639	31	221
Earthquake-1	Purity	1	0.844	0.938	0.853	0.840	0.814
	Quality	0.952	0.399	0.795	0.004	0.086	0.006
	Clusters	479	33	380	161	10	93
Earthquake-10	Purity	DNF	0.840	0.912	DNF	0.834	0.807
	Quality	DNF	0.398	0.658	DNF	0.077	0.006
	Clusters	DNF	37	506	DNF	12	153
Earthquake-100	Purity	DNF	0.842	0.909	DNF	0.837	0.808
	Quality	DNF	0.389	0.560	DNF	0.083	0.006
	Clusters	DNF	38	780	DNF	9	191
Wyoming	Purity	DNF	0.772	0.721	DNF	0.769	0.661
	Quality	DNF	0.027	0.227	DNF	0	0.001
	Clusters	DNF	489	89	DNF	391	78

As can be seen in Table 3, SCAH outperforms SCHG and SCMRG for β=1.01/ η=6, and, in general, performs quite well for small β values, such as 1.01. In general, we observed for these and other datasets that SCAH merges pure clusters that share the same majority class initially. Consequently, it does a quite good job, if the task is to identify small regions that are pure. However, when SCAH reaches the point when it runs out of pure clusters to merge, it has the tendency to terminate prematurely with too small regions. It should be noted that the penalty for having more clusters is quite small if β is 1.01 and the penalty for losing purity when merging clusters is quite high when η is 6. However, for β=3/η=1, the reward gains from having larger clusters are quite significant. This explains why SCAH is outperformed by the other two algorithms for several datasets.

Why does SCAH terminate prematurely when we are interested in obtaining large clusters? The first reason is that SCAH only considers a single merge candidate per cluster whereas SCHG considers four merge candidates per cluster initially; therefore, SCHG has more options for merge, and therefore is less likely to terminate prematurely.

The second reason is that SCAH's look-ahead horizon is too limited. For example, when running SCAH for the B-Complex9 dataset for β=3, the algorithm terminates with 17 clusters, seven of which are depicted in Fig. 3: the outside elliptical shape belongs to class 1 and the two inside spots belong to class 0. Based on average

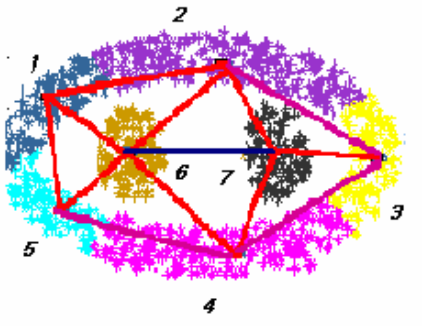
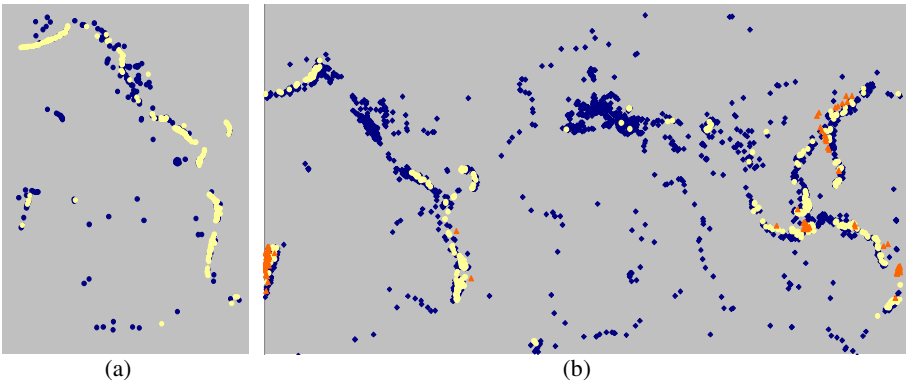


Fig. 3. A part of the B-Complex9 Dataset

distance, the merge candidates considered by SCAH are: 1 and 6, 5 and 6, 3 and 7, 2 and 7, 4 and 7. SCAH will stop here since no improvement made using a single merge. However, for β=3 a higher reward can be obtained by merging all 7 clusters, but SCAH fails to do so, because it terminates if q(X) cannot be improved by a single merge.

Another interesting observation is that the SCHG algorithm outperforms the SCMRG for β=3 for all datasets tested. On the other hand, SCMRG algorithm performs better, compared to SCHG, on the datasets containing chain patterns such as volcano, as depicted in Fig. 4, for β=1.01. This can be attributed to the fact that SCHG is limited to the predefined size of the cells, whereas SCMRG uses grid-cells of different size. As we see in Fig. 6, regions discovered for the volcano dataset are purer than those discovered by the SCHG algorithm.



(a) (b)

Fig. 4. Original Volcano and Earthquake Datasets, a) Part of Volcano b) Earthquake

The average running time of each algorithm is depicted in Fig. 5. The SCAH algorithm is less efficient compared to SCHG and SCMRG. The slowness SCAH can be

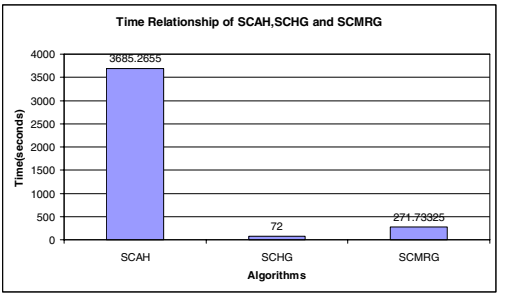


Fig. 5. Average running time of SCAH, SCHG and SCMRG

attributed to its time consuming distance computations: SCAH is required to compute distances between pairs of clusters in order to make the decision of which two clusters have to be merged in the next step, whereas SCHG and SCMRG do not compute any distances at all, and determine merge/split candidates quickly from the underlying grid structure. Moreover, the average running time of SCMRG is higher than SCHG because SCMRG looks for the regions of interest at different levels of resolution, whereas SCHG searches for the interesting clusters using grid cells of fixed size.

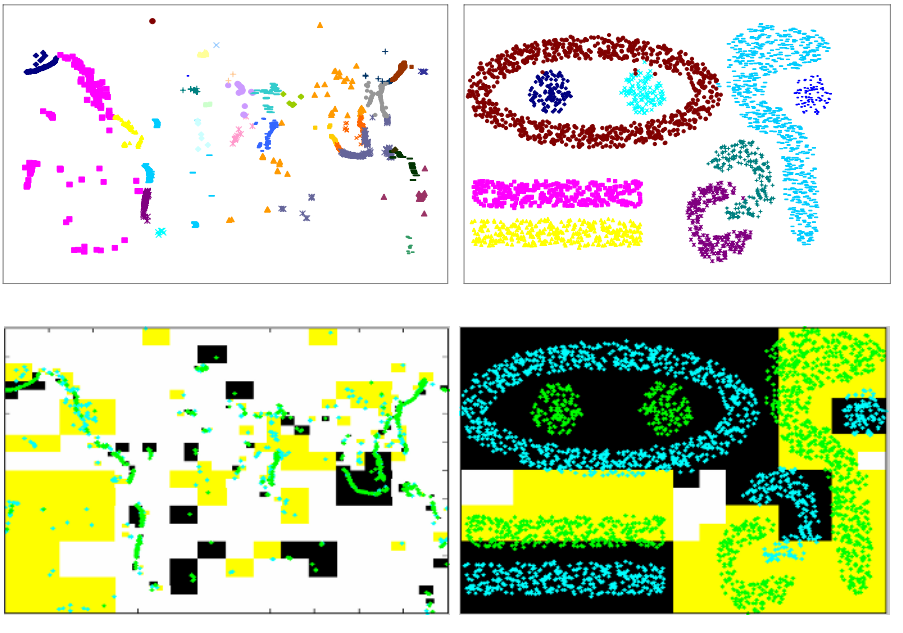


Fig. 6. The result of running supervised algorithms, from top to bottom SCAH, SCHG and SCMRG on two datasets for parameters $\beta = 3/\eta = 1$

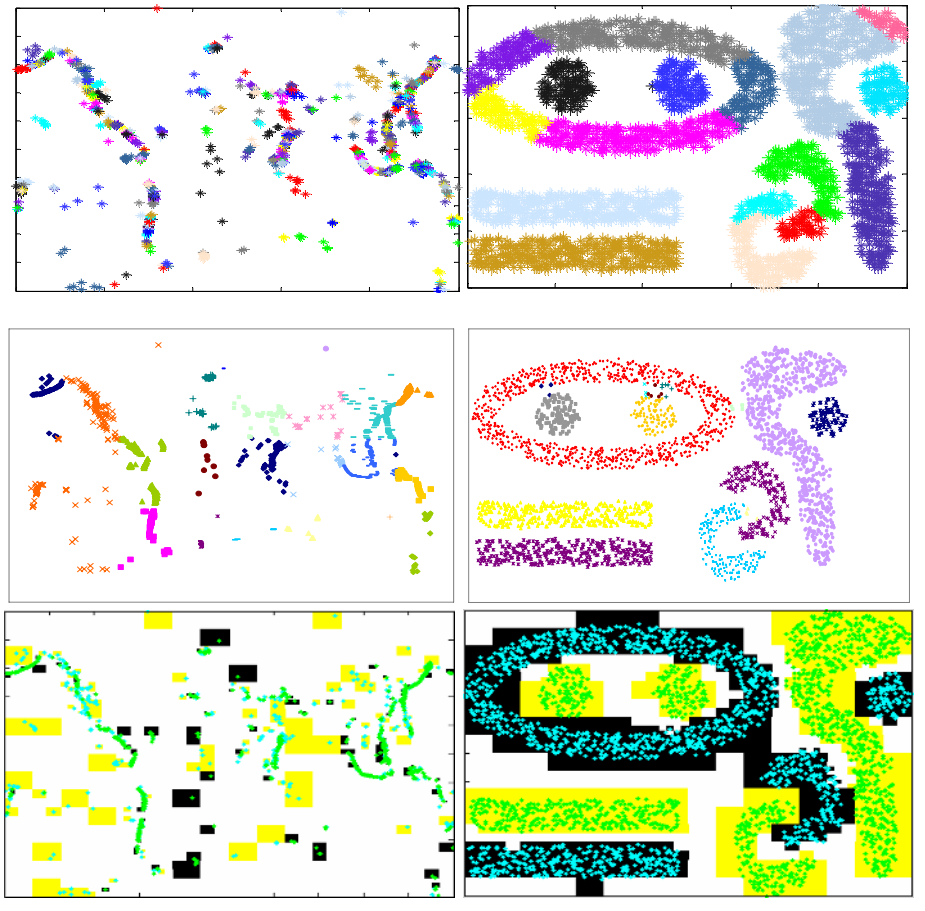


Fig. 7. The result of running SCAH, SCHG and SCMRG on two datasets for $\beta = 1.01/\eta = 6$

5 Related Work

Supervised Clustering [7] centers on partitioning classified examples, maximizing cluster purity while keeping the number of clusters low. Supervised clustering algorithms have been originally proposed to enhance classification algorithms [6]. Our work also has some similarity with work in search-based semi-supervised clustering (also see [1]); for example, Demiriz et al. [5] propose an evolutionary clustering algorithm in which solutions consist of k centroids and the objective of the search process is to obtain clusters that minimize (the sum of) cluster dispersion and cluster impurity. This paper centers on the application of supervised clustering to a new problem: region discovery in spatial datasets containing classified examples.

There also has been some work on co-location rule discovery whose relationship to our work is worth discussing. Co-location rule discovery centers on finding subsets of spatial features frequently located together. Approaches to discover co-location rules

in the literature can be categorized into three groups: spatial statistics [3, 4], association rules [9] and event centric spatial co-location [10]. It should be noted that all mentioned approaches center on finding frequent, global patterns that characterize the complete dataset, whereas our approach centers on finding local regions that are unusual or unexpected with respect to the global characteristics of the dataset.

Hotspot discovery has also been investigated by past research. Williams [14] proposes an evolutionary hotspot discovery architecture that uses traditional clustering, rule induction, and a domain specific fitness function. Tay & Lim et al. [11] describes a region growing method for hotspot discovery, which selects seed points first and then grows clusters from seed points by adding neighbor points as long as a density threshold condition is satisfied. Brimicombe proposes the Geo-ProZone algorithm [2] for hotspot discovery that employs adaptive recursive tessellations. This algorithm supports different level of resolution and recursively decomposes the space with variable decomposition ratios using rectangular grid cells. Finally, Klösgen and M. May [8] propose a multi-relational framework for subgroup discovery within a spatial database system.

6 Summary

In this paper, we introduced a supervised clustering approach and a reward-based evaluation framework for region discovery. Finding interesting regions in spatial datasets is viewed as a clustering problem, in which the sum of rewards for the obtained clusters is maximized, and where the reward associated with a cluster reflects its degree of interestingness for the problem at hand. We explained that this approach is quite different from most other work in spatial data mining that mostly uses association rules. Different measures of interestingness can easily be supported in the proposed framework by designing different reward-based fitness functions; in this case, neither the supervised clustering algorithm itself nor our general evaluation framework has to be modified. We also discussed how hierarchical, and grid-based clustering algorithms can be adapted for supervised clustering in general, and for region discovery in particular, and presented evidence concerning the usefulness of the proposed framework for hotspot discovery problems. Finally, the paper identified some shortcomings of agglomerative clustering algorithms, such as SCAH. Our current work explores the use preprocessing techniques to speed up SCAH, and on the use of proximity graphs to merge clusters more intelligently.

References

1. Basu, S., Bilenko,M., Mooney, R.: Comparing and Unifying Search-based and Similarity-Based Approaches to Semi-Supervised Clustering. in Proc. ICML03 Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, Washington DC (2003) 42-29.
2. Brimicombe, A.J.: Cluster Detection in Point Event Data Having Tendency Towards Spatially Repetitive Events, 8th International Conference on GeoComputation, Ann Arbor, Michigan (2005).

3. Chou, Y.: In Exploring Spatial Analysis in Geographic Information System, Onward Press, (ISBN: 1-56690-119-7), (1997).
4. Cressie, N.: In Statistics for Spatial Data, Wiley-Interscience, (ISBN: 0-471-00255-0), (1993).
5. Demiriz, A., Benett, K.-P., and Embrechts, M.J.: Semi-supervised Clustering using Genetic Algorithms, in Proc. ANNIE, St. Louis, Missouri (1999).
6. Eick, C., Zeidat, N.: Using Supervised Clustering to Enhance Classifiers, in Proc. 15th International Symposium on Methodologies for Intelligent Systems, Saratoga Springs, New York (2005).
7. Eick, C., Zeidat, N., Zhao, Z.: Supervised Clustering - Algorithms and Benefits, UH-Technical Report UH-CS-05-10; short version appeared in Proc. International Conference on Tools with AI (ICTAI), Boca Raton, Florida (2004) 774-776.
8. Klösgen, W. and May, M.: Spatial Subgroup Mining Integrated in an Object-Relational Spatial Database. In Proc. Principles of Data Mining and Knowledge Discovery Conference (PKDD), Springer-Verlag, Berlin (2002) 275-286.
9. Koperski, K., Han, J.: Discovery of Spatial Association Rules in Geographic Spatial Data bases, in Proc. 4th Int. Symp. Advances in Spatial Databases, Maine (1995) 47-66.
10. Shekhar, S., Huang, Y.: Discovering Spatial Co-location Patterns: A Summary of Results, in Proc. of 7th International Symposium on Spatial and Temporal Databases (SSTD01), L.A., CA (2001).
11. Tay S.C., Lim, K.H., Yap, L.C.: Spatial Data Mining: Clustering of Hot Spots and Pattern Recognition, IEEE (2003).
12. UH Data Mining & Machine Learning Group: http://www.tlc2.uh.edu/dmmlg/Datasets.
13. Wang, W., Yang, J., Muntz, R.: STING: A Statistical Information Grid Approach to Spatial Data Mining, Proceedings of the 23rd VLDB Conference, Athens, Greece (1997).
14. Williams, G.J.: Evolutionary Hot Spots Data Mining, An Architecture for Exploring for Interesting Discoveries, Pacific Asia Conference on Knowledge Discovery and Data Mining, Beijing, China (1999).