A SUPPORT VECTOR CLUSTERING BASED APPROACH FOR SPATIOTEMPORAL ANALYSIS IN SECURITY INFORMATICS

Jon Devine [a, *], Anthony Stefanidis [b]

[a] Dept. of Spatial Information Science and Engineering, University of Maine, Orono, Maine 04469 - jdevine@umit.maine.edu

[b] Dept. of Earth Systems and Geoinformation Sciences, George Mason University, Fairfax, VA 22030 - astefani@gmu.edu

Commission II, WG II/1

KEY WORDS: Spatiotemporal, clustering, kernel methods, modeling spatiotemporal helix

ABSTRACT:

Security informatics involves the application of information technology to protect public health and security. A common research topic in security informatics is the identification and representation of clusters of events (e.g., a disease cluster or a crime hot-spot). Understanding why clusters change shape and move over time would be valuable to researchers in security informatics, providing them with greater means for discovering causes as well as examining the effectiveness of mitigation efforts. A first step towards this type of understanding necessitates the establishment of methods for the description of how clustering events evolve over time. However, existing approaches for the analysis of clusters are limited in their ability to describe spatiotemporal behaviour such as movement and deformation. This research presents a framework for facilitating such spatiotemporal descriptions based on support vector clustering and the spatiotemporal helix. Benefits of this approach include the absence of bias a prioi regarding the shape or number of clusters and the ability to describe spatiotemporal behaviour in terms of both changes in shape and movement. Results based on simulated data suggest the effectiveness of this approach for spatiotemporal analysis in a range of application domains in security informatics.

1. INTRODUCTION

Recent technological development has produced important new sources for generating spatiotemporal data and has significantly enhanced the accuracy of existing data collection techniques. The existence of these new data offer real opportunity to advance a range of fields. One such application area where these data can offer potential for analysis is in security informatics.

Security informatics is an umbrella term describing a diverse collection of research domains including homeland security, law enforcement, and public health among others. Broadly, it can be defined as the application of information technology for the maintenance of public safety and well-being. A common analytical problem in security informatics is the identification of regions with elevated concentrations of events. Illustrative examples include the identification of disease clusters and hot spots for particular crimes. Within the realm of security informatics, research objectives such as these have been characterized by three principal questions (Zeng, Chang et al. 2004; Chang, Zeng et al. 2005): 1) How to identify regions within the study area having high or low concentrations of events? 2) How to determine if any areas of variant concentration are the result of random variation or are statistically significant, and if the variation is not random are there explanatory variables that can explain this deviation? 3) How to identify significant changes in the distribution of events (Zeng, Chang et al. 2004; Chang, Zeng et al. 2005)?

The first two of these questions are traditional research objectives in several research domains in security informatics, notably in epidemiology. Among the methods available to identify and delineate potential clusters, the scan statistic has recently emerged as popular. This technique is effective in identifying areas with clustering, but has a major drawback in terms of its reliance on scanning windows of fixed shape (i.e., circular or elliptical) which implies bias a priori regarding the shape of clustering events and limits ability to describe changes in the spatiotemporal behaviour such as change in cluster shape with much detail.

To address these shortcomings, Zeng, Chang et al. (2004) and Chang, Zeng et al.(2005) proposed methodologies based on support vector clustering (SVC). SVC is a kernel method and, as with all kernel methods, relies on kernel transformation to high-dimensional feature space to make non-linear learning tractable. In this feature space a relatively simple decision function, the minimum bounding hypersphere, is applied. When this representation of cluster boundaries is mapped back to the initial input data space the cluster boundaries can be complex in shape and composed of multiple polygons.

Products of the SVC algorithm include labels for points distinguishing clusters from outliers as well as representation of cluster boundaries in input space. For analysis in security informatics, Zeng, Chang et al. (2004) investigated application potential for SVC-produced representations of cluster boundaries through comparison against the scan statistic and hierarchical clustering results. Conclusions suggested further

* Corresponding author.

consideration of SVC for spatiotemporal applications due to SVC's capability to produce more complex representations of clusters. Later consideration of SVC for security informatics applications was conducted in spatiotemporal context (Chang, Zeng et al. 2005). In this research the authors incorporated time directly into the SVC algorithm and examined the potential of SVC-produced point labels for spatiotemporal analysis. Results from this analysis identified points both in space and time that were either inside or outside of clusters. However, with the output being clouds of clustered points, this approach offers limited ability to describe the spatiotemporal behaviour.

The objective of this research is to take a different approach to spatiotemporal analysis with SVC that can describe changes in shape and movement of clustering events. Like Zeng, Chang et al. (2004), this approach is based on SVC-produced representations of cluster boundaries, defining clusters in terms of regions with high concentrations of event instances rather than as point clouds. To incorporate a temporal dimension, these derivations are repeated through time resulting in data not unlike video sequences of image data. Given this similarity, a method already proven capable of describing event evolution in video data, the spatiotemporal helix (ST helix), is proposed. Developed as a means for the summarization of event behaviour in image data, the ST helix readily incorporates SVC-produced representations of cluster boundaries and can be used to support spatiotemporal queries regarding cluster behaviour over time. To demonstrate how SVC can be coupled with the ST helix for spatiotemporal analysis of event behaviour, a simulation based example is described which highlights topics for future research.

2. SUPPORT VECTOR CLUSTERING

Support vector clustering (SVC) is a non-parametric kernel-based approach to the problem of describing clustering in data. Key advantages of this over popular approaches is the absence of any assumptions regarding the number or shapes of clusters. Distinguishing SVC from other kernel methods is its use of a minimum bounding hypersphere decision function in feature space written

$$\left\|\Phi(x_i)-a\right\|^2 \le R^2+\xi_i \qquad \forall i \tag{1}$$

where x_i are event instances
Φ is a non-linear transformation
a is the center of the hypersphere
R the radius of the min. bounding hypersphere
ξ_i is a slack variable for the soft constraint

To solve for the minimum bounding hypersphere, the following Lagrangian is used

$$L = R^2 - \sum_{i=1}^{N}\left(R^2+\xi_i\left\|\Phi(x_i)-a\right\|^2\right)\alpha_i - \sum_{i=1}^{N}\xi_i\mu_i + C\sum_{i=1}^{N}\xi_i \tag{2}$$

which describes a convex cost problem. By imposing stationarity (derivatives equal to zero) the following first order conditions

$$\sum_i \alpha_i = 1 \tag{3}$$

$$a = \sum_i \alpha_i \Phi(x_i) \tag{4}$$

$$\alpha_i = C - \mu_i \cdot \tag{5}$$

can be derived. The inequality constraint (Eq.) implies Karush Kuhn Tucker (KKT) conditions

$$\xi_i \mu_i = 0 \tag{6}$$

$$\alpha_i (R^2 + \xi_i - \left\|\Phi(x_i)-a\right\|^2) = 0 \tag{7}$$

With both the first order conditions and the KKT conditions, Eq. 1 can be written entirely in terms of the parameters α_i so that

$$W = \sum_j \Phi(x_j \cdot x_j)\alpha_j - \sum_{i,j}\alpha_i\alpha_j\Phi(x_i)\cdot\Phi(x_j) \tag{8}$$

Since the μ_i do not appear in Eq. 7, they can be replaced by the constraint

$$0 \le \alpha_i \le C\,,\; i = 1,\ldots,N \tag{9}$$

where C is a user-defined parameter controlling the influence of outliers. With the interpretation of the hypersphere radius delimits clusters, Ben-Hur, Horn, et al. (2001) termed the instances located a distance less than the radius from the center a as being *interior points*, those located beyond the radius as *bounded support vectors*, and those on the surface as *support vectors*. In other words, the hypersphere is a representation of the cluster boundary and support vectors appear along the boundary. From Eq. 3 and Eq. 9, it follows that the number of bounded support vectors, or outliers, is limited by C so that the maximum number of outliers is less that 1/C. Therefore 1/NC where N is the number of event instances can be interpreted as an upper bound on the percentage of outliers accepted by the cluster boundary.

With the input data x_i appearing as a dot product the advantages of the"kernel trick" become accessible, as dot product can be replaced by an appropriate kernel function. The sole constraint placed upon functions to be considered as a kernel for this substitution is that their Gram matrix be symmetric and positive semi-definite to guarantee convexity and a unique solution (kernels must be a Mercer kernel). This research uses Gaussian kernels of the form

$$K(x_i, x_j) = e^{-q\left\|x_i - x_j\right\|^2} \tag{10}$$

where $$q = 1/2\sigma \tag{11}$$

because they are not sensitive to outliers (Tax and Duin 1999) and because of their previous application in spatial analysis with kernel density estimation. Following kernel substitution with a Gaussian kernel, the problem in Eq. can be written as

$$W = 1 - \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \tag{12}$$

From the above formulations, it is evident that the hypersphere is determined by the (unbounded) support vectors alone. As a result, these points, with $0 < \alpha_i < C$, alone are used to map the hypersphere back to input space and produce a representation of the cluster boundary in that space. The process first involves the derivation of an input space value R corresponding to the radius in feature space. This value is used to test calculated values from a mesh of input space points $z \in Z$. Those locations where the values match R are interpreted as an input space representation of the cluster boundary. R can be written:

$$R^2(z) = \left\|\Phi(z) - a\right\|^2 \tag{13}$$

With $a = \sum_i \alpha_i \Phi(x_i)$ (Equation 3.34) and kernel substitution for Φ, this formulation can then be rewritten

$$R^2(z) = K(z,z) - 2\sum_i \alpha_i K(z, x_{SV_i}) + \sum_{i,j} \alpha_i \alpha_j K(x_{SV_i}, x_{SV_j}) \tag{14}$$

$$R^2(z) = 1 - 2\sum_i \alpha_i K(z, x_{SV_i}) + \sum_{i,j} \alpha_i \alpha_j K(x_{SV_i}, x_{SV_j}) \tag{15}$$

$$R^2(z) = 1 - 2\sum_i \alpha_i K(z, x_{SV_i}) + S_X \tag{16}$$

where x_{SV_i}, x_{SV_j} are support vectors and S_X is the constant sum of the product of the kernel-based representation of the inner product of support vectors with their Lagrangian multipliers α_i where $0 < \alpha_i < C$. With the Gaussian kernel in the first term equal to one and S_X, the only term that varies in Eqs. 14-16 with each novel point z is the second one.

Using these formulations, the radius of the minimum bounding hypersphere can be obtained through the consideration of the points $x_i \in Z$ where the x_i are the support vectors, or equivalently

$$R = \{R(x_i) \mid x_i \text{ is a support vector}\}. \tag{17}$$

This value, generated by the points lying on the hypersphere, can then be used to compare against the values generated through the expression in Eq. 17 and by identifying those points z with values that are equal, therefore determining contours representing the extent of the clusters in input space. This can be written

$$\{z \mid R(z) = R\}. \tag{18}$$

Motivated by its ability to provide complex representations of clustering without bias in regards to shape or the number of clusters, SVC has already been twice examined for its application potential in security informatics. The first of these investigations examined SVC in a purely spatial context, comparing SVC generated representation of cluster boundaries against those produced by the scan statistic and hierarchical clustering (Zeng, Chang et al. 2004). The second of these investigations considered SVC for spatiotemporal analysis, exploiting the ability of kernel methods to handle high-dimensional data by incorporating time directly into the SVC algorithm (Chang, Zeng et al. 2005). Output from this algorithm was successful in identifying clustered points, but by consisting of clouds of labelled points this method does not allow for direct description of clusters behaviour over time.

To address this inability to describe spatiotemporal behaviour of clustering phenomena, this research proposes a new approach. Like the work by Zeng, Chang et al. (2004), this approach is based on SVC-produced representations of cluster boundaries. Given that these boundaries appear in raster-type format and that they can be produced through time, there is a resemblance of these data to those that occur in video sequences. Correspondingly, these SVC-generated results are inputted into an existing framework for spatiotemporal analysis of areal events in image data, the ST helix.

3. SPATIOTEMPORAL HELIX

The spatiotemporal helix is a framework for summarizing the evolution of spatiotemporal phenomena. Designed to allow efficient querying of data and to support intuitive visual representations of event evolution, the primary strength of the ST helix include its ability to facilitate complex description and query of event evolution in terms of both the event's trajectory and its deformation.

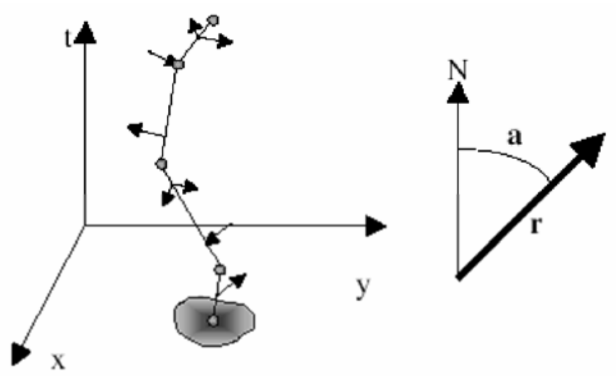

Figure 2. A spatiotemporal helix.

The gray spheres depict nodes and define the helix's spine. Prongs are represented by arrows. Outward facing arrows denote expansion and inward facing arrows indicate contraction. Arrow length reflects the magnitude of deformation while their angle reflects the azimuth range over which the deformation occurred.

In order to determine which changes in velocity and shape are significant, and consequently which nodes and prongs constitute a ST helix, self-organizing map (SOM) and differential snakes techniques are applied. The derivation of appropriate nodes is complex in that significant change in trajectory implies consideration not only of the distance traveled over time, but also of the direction. For this reason, a sophisticated approach involving a geometric adaptation of self-organizing maps is used which assigns more nodes to time periods of intense change and fewer nodes to periods of stability (Partsinevelos, Stefanidis et al. 2001).

To develop the prongs an adaptation of deformable contour models, differential snakes, is used. This method considers changes in an events shape as a function of differences in the distance from an event's center of mass to points on its boundary from time *t* to time *t+dt*. The percentage of change in these distances is successively compared against a user defined threshold to identify where significant changes have occurred (Agouris, Stefanidis et al. 2001). These significant changes and their sign, negative change implying contraction and positive change indicating expansion, are recorded as prongs (Agouris and Stefanidis 2003).

With selection of appropriate parameter values for determining significance, a concise signature of the evolution of event occurring over the time period frame *t1* to *t2* can be captured by a ST helix and written

$$Helix_{t1,t2}^{objid} = \{node_1,...node_n; prong_1,...prong_m\} \qquad (19)$$

(Stefanidis, Eickhorst et al. 2003). This signature is the basis for the development of similarity metrics outlined by Croitoru, Agouris et al. (2005) which demonstrated the ability of the node and prong data stored in ST helix as capable of differentiating the evolution of 25 different hurricanes and facilitated discussion regarding the similarity of their evolution. These results suggest that the ST helix could also be used for spatiotemporal analysis involving the description of clustering.

4. EXPERIMENTS

To illustrate how SVC can be used to describe clustering that can be incorporated into the ST helix for modeling spatiotemporal behavior, this section presents a simulation-based example. The simulation consists of 6 frames each representing approximately 200 instances and with varying amounts of clustering and randomly generated noise. SVC was conducted on the data in each of the 6 frames to generate areal representations of clusters which were then inputted into the ST helix modeling framework. Two selected frames appear in Figure 3.

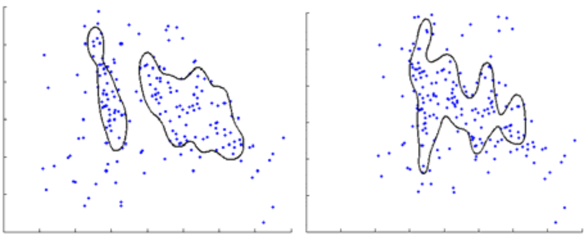

Figure 3. The two images above depict SVC results occurring at different time frames of the simulation (q=12, C=0.0077).

A major criticism of SVC is related to the difficulty of selecting appropriate parameter values. With SVC with Gaussian kernels, these parameters are the bandwidth σ and the parameter C that controls the percentage of outliers. To generate the results produced in this research, a variety of parameter settings were examined and values were adjusted according to degree of clustering and noise in each frame. Guidelines for the selection of appropriate parameter values is an on-going topic of research that will need to be addressed before SVC can be implemented. The effects on the shape of the boundaries produced by SVC can be seen in Figures 4 and 5. In Figure 4 the bandwidth σ is manipulated indirectly by varying the values of $q = 1/\sigma$. As σ becomes smaller (as q is increased) the number of support vectors increases and boundary shape becomes more complex. Likewise, varying C also has a strong effect on cluster shape (Figure 5). Given the dramatic variation in the results shown in Figures 3-5, before SVC can become available for widespread in security informatics applications, research regarding the selection of appropriate parameter values will have to deliver suggestions for the 'best' parameter values for various datasets.

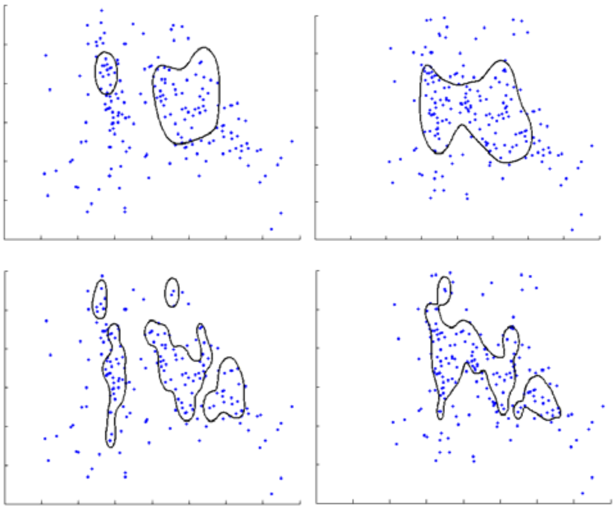

Figure 4. Illustration of the effect of changing kernel bandwidth on cluster representation. Both rows of images are derived from the same data with the same value for C as in Figure 3 (C=0.0077). The different boundary representations result from different values for q (q=6 in top row, q=24 in bottom row).

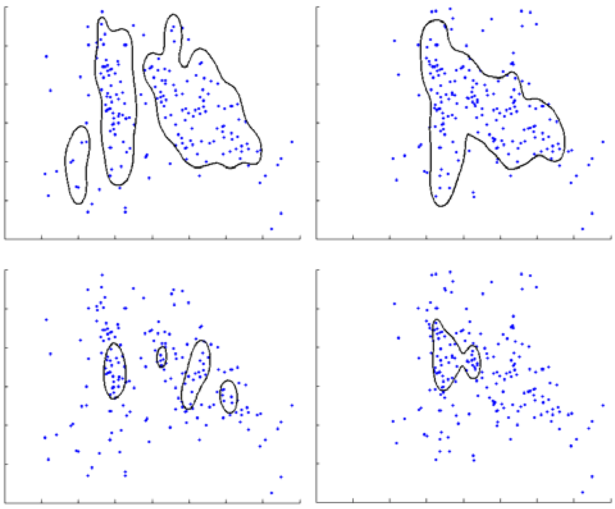

Figure 5. Illustration of the effect of changing C on cluster representation. Both rows of images are derived from the same data with the same bandwidth (q=12) as in Figure 3. The different boundary representations result from different values for C (C=0.01 in top row, C=0.0059 in bottom row).

When modeling SVC-produced results for clustering, other challenges materialized. Principal among these is that existing ST helix methodologies were designed for analysis of events composed of single polygons. Meanwhile, SVC-generated representations of clustering, as in reality, may be represented by multiple polygons at any given time period. Interaction among these polygons implies more complex spatiotemporal behavior such as merging/splitting and appearing/disappearing. Accurately describing these behaviors could be important to many applications in security informatics (e.g., why did two criminal hot spots merge?). The simulation presented here was explicitly designed to highlight the types of obstacles presented when modeling spatiotemporal phenomena. In terms of the ST helix, modeling of this type of behavior has implications in assignment of trajectories (i.e., start/end a trajectory, merge/split with an existing trajectory) and is a topic of current and future research. For the helix depicted in Figure 6 a simple framework, adapted from Devine and Stefanidis (2008), involving trajectory and changes in area over successive frames were used to allocate trajectories.

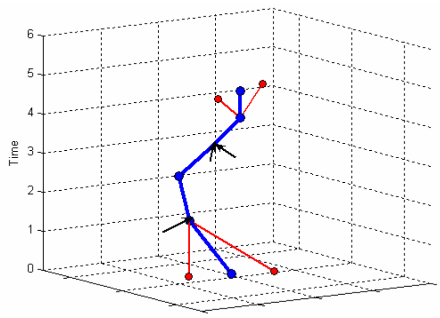

Figure 6. Spatiotemporal helix representing the spatiotemporal in the simulation.

5. CONCLUSIONS AND FUTURE WORK

A limitation of many existing techniques for the description of the behavior of clustering events is bias in the shape or number of clusters in a study area at a given time. Therefore, these methods may be unable to describe spatiotemporal behavior such as movement and deformation. Understanding why clustering events change shape and move over time would be valuable to researchers in security informatics, providing them with another means of discovering causes these clustering events as well as monitoring the effectiveness of mitigation efforts to control them. A first step towards such understanding necessitates description of how clustering events deform and move over time. This research presents a framework for facilitating such descriptions based on SVC-derived representations of clustering and the ST helix. Results based on simulated data suggest the effectiveness of this approach to spatiotemporal analysis in security informatics, but also highlighted necessity for future research addressing the assignment of parameter values and an expansion of ST helix methodology for the incorporation of complex spatiotemporal cluster behavior such as merging/splitting and appearance/disappearance.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under Grant No. 0429644 through the Science and Engineering Informatics and Information Integration Program (SEIII)..

REFERENCES

Agouris, P. and A. Stefanidis (2003). Efficient Summarization of Spatiotemporal Events. *Communications of the ACM*, 46(1), pp. 65-66.

Agouris, P., A. Stefanidis, et al. (2001). Differential Snakes for Change Detection in Road Segments. *Photogrammetric Engineering & Remote Sensing*, 67(12), pp. 1391-1399.

Ben-Hur, A., D. Horn, et al. (2001). Support Vector Clustering. *Journal of Machine Learning Research*, 2, pp. 125-137.

Chang, W., D. Zeng, et al. (2005). Prospective Spatiotemporal Data Analysis for Security Informatics. *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria, pp. 1120-1124.

Croitoru, A., P. Agouris, et al. (2005). Rotation, Translation, and Scale Invariant 3D Trajectory Matching by Pose Normalization. *ACM-GIS'05*, Bremen, ACM Press.

Devine, J. and A. Stefanidis (2008). Support Vector Machines for Spatiotemporal Analysis in Geosensor Networks. *Advances Spatial Data Handling*. In press.

Partsinevelos, P., A. Stefanidis, et al. (2001). Automated Spatiotemporal Scaling for Video Generalization. *IEEE International Conference on Image Processing*, Thessaloniki, Greece.

Stefanidis, A., K. Eickhorst, et al. (2003). Modeling and Comparing Change Using Spatiotemporal Helixes. *ACM-GIS'03*, New Orleans, ACM Press.

Tax, D. M. J. and R. P. W. Duin (1999). Support Vector Domain Description. *Pattern Recognition Letters*, 20, pp. 1991-1999.

Zeng, D., W. Chang, et al. (2004). A Comparative Study of Spatio-Temporal Hotspot Analysis Techniques in Security Informatics. *2004 IEEE Intelligent Transportation Systems Conference*, Washington D.C., IEEE.