SUPPORT VECTOR CLUSTERING FOR WEB USAGE MINING

WEI SHUNG CHUNG
School of Computer Science
The University of Oklahoma
Norman, Oklahoma

THEODORE B. TRAFALIS
School of Industrial Engineering
The University of Oklahoma
Norman, Oklahoma

LE GRUENWALD
School of Computer Science
The University of Oklahoma
Norman, Oklahoma

ABSTRACT
This paper applies the use of support vector clustering (SVC) in the domain of web usage mining. In this method, the data points are transformed to a high dimensional space called the feature space, where support vectors are used to define a smallest sphere enclosing the data. A soft-margin constant is used to handle outliers. The paper then performs experiments to compare SVC and the K-Means algorithm using a web server log obtained from a real life educational web site. The experimental results show that SVC provides better user session clusters than the K-Means algorithm in term of intra-cluster scatter measure but perform much worst in term of time. SVC does not require the number of clusters to be determined a priori and it can handle outliers and any shape of clusters.

INTRODUCTION

Web data can be used to understand customers' browsing behaviors and to gain a strong competitive advantage in e-business. Web usage mining analyzes browsing patterns from Web log data in order to group users having similar browsing patterns. The discovered knowledge is critical for e-commerce businesses in order to derive better business intelligence. Web Usage Mining analyzes Web log usage patterns to cluster users of similar interest. After the users clusters are identified, a new visitor with a limited navigation history can be categorized into one of the clusters. By knowing the new user's predicted interest, marketer can send the right promotion information to the new user.

(Ben-Hur et al., 2001) presents a novel method using the support vector machines approach for clustering called Support Vector Clustering (SVC). SVC does not have a bias of the number or the shape of clusters and it can handle noise or outliers.

The objective of this paper is to present a comparison study of SVC and K-means for Web usage data mining using real-life Web logs. (http://www.hippocrates.ouhsc.edu) We chose K-means for comparison because it is a widely used algorithm for clustering. (Shahabi et al., 1997) and

(Mobasher et al., 2000) used the technique to perform clustering in Web usage mining. Section 2 describes SVC and K-means. Section 3 and 4 present the comparison studies and experimental results, respectively. Section 5 gives the conclusions and future research.

SUPPORT VECTOR CLUSTERING

Support Vector Clustering (Ben-Hur et al., 2001) is a non-parametric clustering algorithm based on the support vector machine approach (Vapnik 1995). A Gaussian kernel function is used to map the data points to a high dimensional space called the feature space. The objective is to search for the minimal enclosing sphere. The mapping of the sphere back to the data space gives rise to the formation of clusters. Increasing the width parameter of the Gaussian kernel will increase the number of clusters. A soft margin constant is used to handle outliers by allowing the sphere in the feature space not to enclose all points. In the first stage of the Support Vector Clustering, the sphere with the minimal radius, which encloses the data points in feature space, is computed. In the second stage, a cluster assignment based on a geometric approach is used. If a pair of data points belongs to different clusters, then any line that connects them must pass beyond the minimal enclosing sphere in the feature space. In other words, there exists a point y on the line with the radius larger than the radius of the sphere in the feature space, R_{sphere}.

Next, we explain the basic ideas of SVC formulation as described in (Ben-Hur et al., 2001). Let $\{\mathbf{x_i}\}_{1\le i\le N} \subseteq X$ be a data set of N points, with $x_i \subseteq IR^d$, the input space. We search for the smallest sphere enclosing the data of radius R after applying the nonlinear transformation Φ from the data space X to the feature space.

The optimization problem can be written as follows:

Min R^2

s.t.

$$\left\|\Phi(\mathbf{x_j}) - \mathbf{a}\right\|^2 \le R^2 + \xi_j \qquad \text{j = 1,….., N}$$

$$\xi_j \ge 0$$

The constraints of the optimization problem are as follows:

$$\left\|\Phi(\mathbf{x_j}) - \mathbf{a}\right\|^2 \le R^2 + \xi_j \qquad (1)$$

where a is the center of the sphere and $\xi_j \ge 0$ are slack variables that are used to handle outliers.

The Lagrangian L is formulated as follows to solve the above problem:

$$L(R,\mathbf{a},\beta,\mu,\xi) = R^2 - \sum_j (R^2 + \xi_j - \left\|\Phi(\mathbf{x_j}) - \mathbf{a}\right\|^2)\beta_j - \sum_j \xi_j\mu_j + +C\sum_j \xi_j \qquad (2)$$

where

$\beta_j \geq 0$ and $\mu_j \geq 0$ are Lagrange multipliers.

C is a constant and

$\sum_j \xi_j$ is a penalty term.

Differentiating with respect to R, ξ_j and **a** leads to

$$\frac{\partial L(R,\mathbf{a},\beta,\mu,\xi)}{\partial R} = 0 \Rightarrow \sum_j \beta_j = 1 \quad (3)$$

$$\frac{\partial L(R,\mathbf{a},\beta,\mu,\xi)}{\partial \xi_j} = 0 \Rightarrow \beta_j = C - \mu_j \Rightarrow \mu_j = C - \beta_j \quad (4)$$

$$\frac{\partial L(R,\mathbf{a},\beta,\mu,\xi)}{\partial a} = 0 \Rightarrow a = \sum_j \beta_j \Phi(\mathbf{x_j}) \quad (5)$$

We can then write the Karush-Kuhn-Tucker complementary conditions as follows:

$$\xi_j \mu_j = 0 \quad (6)$$

$$(R^2 + \xi_j - \left\|\Phi(\mathbf{x_j}) - \mathbf{a}\right\|^2)\beta_j = 0 \quad (7)$$

We turn the Lagrangian into the dual form by eliminating R, a and μ_j with the substitution of Eq. (3), (4) and (5) :

$$W(\beta) = \sum_j \Phi(\mathbf{x_j})^2 \beta_j - \sum_{ij} \beta_i \beta_j \Phi(\mathbf{x_j}).\Phi(\mathbf{x_i}) \quad (8)$$

with the constraints $0 \leq \beta_j \leq C$

Maximizing the dual form is the SV problem that we are solving. As in the Support Vector Machine method, the dot products can be replaced by a Mercer kernel $K(x_i,x_j)$. In this paper, we use the Gaussian kernel,

$K(\mathbf{x_i}, \mathbf{x_j}) = e^{-q\left\|\mathbf{x_i} - \mathbf{x_j}\right\|^2}$ where q is the width parameter.

The dual W can be rewritten after replacing the dot products with the Gaussian kernel as follows:

$$W = \sum_j K(\mathbf{x_j}, \mathbf{x_j})\beta_j - \sum_{i,j} \beta_i \beta_j K(\mathbf{x_i}, \mathbf{x_j}) \quad (9)$$

This problem is equivalent to the standard SV optimization problem when the Gaussian kernel is used.

The distance (square of radius) of each point's feature space image from the center of the sphere can be defined as:

$$R^2(\mathbf{x}) = \left\|\Phi(\mathbf{x}) - \mathbf{a}\right\|^2 \quad (10)$$

Substituting the center of the sphere with equation (5) and the kernel, the above can be rewritten as follows:

$$R^2(\mathbf{x}) = K(\mathbf{x},\mathbf{x}) - 2\sum_j \beta_j K(\mathbf{x_j},\mathbf{x}) + \sum_{i,j} \beta_i \beta_j K(\mathbf{x_i},\mathbf{x_j}) \quad (11)$$

The sphere in the feature space has the radius of R as defined below.

$$R_{sphere} = \{R(\mathbf{x_i}) \mid \mathbf{x_i} \text{ is a support vector}\} \quad (12)$$

The average of radius of all support vectors can be used as the radius of the sphere.

The contour that encloses the points in the data space is the set

$$\{\mathbf{x} \mid R(\mathbf{x}) = R_{sphere}\}$$

Parameters q and C are used to control the shape of the enclosing contours in the input space. These parameters affect the number of support vectors. For fixed q, as C is decreased, the number of SVs decreases since some of them turn into bounded SVs and the resulting shapes of the contours become smoother. Bounded SV is data point that has the radius, calculated using Eq. (11), larger than R_{sphere}. After the support vectors are identified, the next step is to label/define the clusters. The clusters are defined as the connected components of the graph induced by A. A is an adjacency matrix where A_{ij} between pairs of points x_i and x_j:

A_{ij} = 1 if for all y on the line segment connecting x_i and x_j , $R(\mathbf{y}) \le R_{sphere}$ (13)

= 0 otherwise

This labeling method can be used because the line segment connecting points in different components contain points outside the sphere whereas the line connecting "close neighbors" in the same component lies inside the sphere.

COMPARISON STUDIES

The server logs are obtained from a centralized educational learning Web site of OU Health Science Center (http://www.hippocrates.ouhsc.edu). Three sets of user sessions are studied. The first experiment is run using the first 50 user sessions and the second experiment is run using the next 50 user sessions. The third experiment is run using the 200 user sessions. Since K-Means requires users to provide the K parameter that is the number of clusters, an intra-cluster scatter measure is used to determine K. We measure the intra-cluster scatter for K starting from 2 to 13 and choose the K that gives the least scatter. Intra-cluster scatter measure is used to measure the quality of the clustering results. The definition of intra-cluster scatter is defined as follows:

$$J_e = \sum_{i=1}^{C} \left(\sum_{x_j \in G_i} \left\| x_j - m_i \right\|^2 \right)$$ where C is the number of clusters,

$x_j \in G_i$ means data point x_j is in cluster i, m_i is the center of G_i , defined as

$m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j$ for i = 1,2,….., C. n_i is the number of data points in cluster i.

Intra-cluster scatter measures the compactness of the clusters. The smaller the J_e is, the more compact the clusters are.

Using the heuristics proposed in (Cooley 1999), we define a unique user based on IP address. Based on (Mobasher et. al.1996), the individual log entries are grouped into user sessions. A user session is a sequence of accesses by a user. The duration of elapse time between any two consecutive accesses in the session is limited to a specified threshold, 30 minutes. The features used are the valid url pages of the site. A unique number i is assigned to each URL, $i \in \{1,.....,m\}$ where m is the total number of URLs. To facilitate various data mining tasks, the jth user session is formulated as an m-dimensional binary attribute vector s_{ij} with the property

s_{ij} = 1 if the user accessed the ith URL during the jth session

= 0 otherwise

EXPERIMENTAL RESULTS

From the three experimental studies, we found that SVC provides better clustering results than K-Means algorithm. For the first experiment, K-Means yields the least scatter, 84.662 when K = 8. For SVC, the intra-cluster scatter measure is zero when q is increased to 3 and C = 1 since it gives the perfect partition. SVC gives identical users sessions in each cluster when q is 3. The clustering results are better than K-Means even when q is lower than 3. When q = 1, the intra-cluster scatter measure is 39.578, lower than the best result of K-Means, 84.662 by 53 %. In other words, the clusters given by SVC are 53% more compact than that of the K-Means. When q = 2, SVC clustering result is almost 70% more compact than K-Means.

Table 1: The values of Intra-Cluster Scatter measures for K-Means and SVC using the first data set. (First 50 Users)

Clustering algorithm	Intra-Cluster Scatter
K-Means (K=8)	84.662
SVC (q=3, C=1)	0

In the second experiment, every data point becomes a single cluster when q is increased from 1 to 2 with intra-cluster scatter value of 46.349. C has to be decreased in order to further refine the clustering solution. When C is decreased from 1, SVC starts to detect outliers and perform clustering without considering the outliers. This improves the effectiveness of the clustering algorithm. The intra-cluster scatter for SVC in the second experiment is shown in Table 2. The best result from SVC is 16.692 when q = 1 and C = 0.05. In this case, there are 24 outliers detected. K-Means gives the intra-cluster scatter value of 52.014 with

K = 12. Clusters provided by SVC are about 68% more compact than that of the K-Means.

Table 2: The values of Intra-Cluster Scatter measures for K-Means and SVC using the second data set. (Second 50 Users)

Clustering algorithm	Intra-Cluster Scatter
K-Means (K=12)	52.014
SVC (q=1, C=0.05)	16.692

In the third experiment, SVC gives the best intra-cluster scatter of 206.550 when q = 1.3 and C =1 as shown in Table 3 below. In this case, there are 20 clusters found. However, only two clusters have a cardinality larger than 2. K-Means gives the best intra-cluster scatter of 224.584 when K = 18. The clusters found using SVC are about 8% more compact than that of the K-Means.

Table 3: The values of Intra-Cluster Scatter measures for K-Means and SVC using the third data set. (200 Users)

Clustering algorithm	Intra-Cluster Scatter
K-Means (K=18)	224.584
SVC (q=1.3, C=1)	206.550

We also measure the running time of both algorithms. In the first experiment, SVC requires 10 minutes to cluster 50 user sessions compared to 5 seconds as required by K-Means. When the number of user sessions increases to 200, SVC takes about 120 minutes compared to about 1 minute as required by K-Means.

Table 4: The running time of K-Means and SVC with varying number of users

Number of Users	Running time for K-Means (second)	Running time for SVC (second)
50	5	600
200	60	7200

CONCLUSIONS AND FUTURE RESEARCH

The comparison shows that SVC provides better clusters in term of intra-cluster scatter especially when the data set is small but perform much worst in term of time when the number of users sessions increase. On average, SVC gives smaller intra-cluster scatter measure compared to K-Means as shown in the above experimental results. SVC provides more compact clusters compared to K-Means. The clustering results can also be improved when outliers are handled in SVC. It has been shown in (Ben-Hur et al., 2001), SVC can handle clusters with arbitrary shape and handle noisy data. SVC's ability to handle outliers improves the clustering results. However, a more efficient way to determine the optimal parameters q and C for SVC has to be derived. In addition, a scalable SVC is also desired to handle large data sets.

REFERENCES

(Ben-Hur et al., 2001) A. Ben-Hur, D. Horn, H.T. Siegelmann, and V.Vapnik. "Support Vector Clustering," Journal of Machine Learning Research 2 pp. 125-137 December 2001

(Cooley et al., 1999) R. Cooley, B. Mobasher and J. Srivastava, "Data preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, 5-32 1999

(Fu et al., 1999) Y. Fu, K. Sandhu, and M. Shih, "Clustering of Web Users Based on Access Patterns," International Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), San Diego, CA, 1999.

(Shahabi et al., 1997) C. Shahabi, A.M. Shakesh, J. Adibi, V. Shah, "Knowledge Discovery from Users Web-Page Navigation ," In Proceedings of the IEEE RIDE97 Workshop, April 1997.

(Mobasher et al., 1996) B. Mobasher, N. Jain, E-H. Han, and J. Srivastava "Web Mining: Pattern discovery from world wide web transactions," Technical Report 96-050, University of Minnesota, Sept, 1996

(Vapnik 1995) Vladimir N. Vapnik. "The Nature of Statistical Learning Theory." Springer, N.Y., 1995

(Mobasher et al., 2000) B. Mobasher, R. Cooley, and J. Srivastava. "Automatic personalization based on Web usage mining," In Communications of the ACM, (43) 8, Aug, 2000.

(Jain and Dubes) A. K. Jain and R. C. Dubes. "Algorithms for clustering Data." Prentice Hall, 1988.