Customer Behavior Clustering Using SVM*

Zhongying Yang, Xiaolong Su

Dept of Compute Science & Technology China University of Mining & Technology
Xuzhou Jiangsu 221000, China
yzy@xzmu.edu.cn, suxl@cumt.edu.cn

Abstract

In order to supply better service for network customers, deeply analyzing customers’ behavior is required. This paper extracts three features from customers’ network behavior which is divided into different categories, such as browsing news, downloading shared resources and real-time communications etc. Support vector machine is used to perform clustering, thanks to its fast and valid executing, especially in the situation of small datasets. Using the analysis results, we can make our applications and services more personalized and easier to be used.

1. Introduction

Support vector machine is a new machine learning technology invented by Vapnik together with his team in Bell lab. It is developed from optimal separating hyperplane that can be divided linearly. SVM has some special advantages in pattern recognition area, especially for small samples, non-linear and high-dimensional problem. The objective of SVM is to obtain the best solution using current information rather than get the best solution when the sample tends to be infinite. The algorithm will transform to a quadratic type optimization problem which can gets global optimal solution theoretically. SVM doesn’t have the dimension problem, so its time complexity is independent of sample dimension. It can realize different learning algorithm using different kernel functions such as polynomial approximation, radial basic function and hierarchical sensor network etc. As SVM adopts structure risk minimization principle, it has strong generalizing ability. Now, the most popular SVM models contain C-SVM, v-SVM, RSVM (Reduced SVM), WSVM (weighted SVM) and LS-SVM. All of them are developed from the original SVM by changing the objective function of the optimizing issue. C-SVM is the widely applicable model. With the rapid development of network, modern data communication network reach an unprecedented
scale. Network plays an important role in peoples’ routine lives. Many kinds of applications based internet become out. At the same time, the telecom companies confront many new challenges because they lack analysis tools facing the sharp increase of information. They need to supply better services by finding the customer concerned information rightly and rapidly. In order to make more profit depend on the applications running in the network, they need to mine the customers’ potential demand, enlarge their market by doing user behavior analysis. Using data mining technology to find user behavior and improving the website services has become an important research topic. The traditional classification of user behavior analysis is a Web-based classification which can determine the type and class of Web pages. These methods use static content, so it is difficult to achieve universal analytical tools and need a lot of manpower to ensure that the accuracy. This paper proposes a support vector machine based on clustering of user behavior.

2. Support Vector Model:

For a given training set:

\[ T = \{(x_i, y_i), \ldots, (x_j, y_j)\} \quad (2-1) \]

Where, \( x_i \in \mathbb{R}^n, \quad y_i \in \{-1, 1\}, \quad i = 1, 2, \ldots, l \)

If we can find an decision function on \( C \) like \( f(x): C \rightarrow Y \), we can infer the value of \( y \) corresponding to any arbitrary \( x \) using \( f(x) \).

The C-SVC model is:

\[
\begin{align*}
\min_{\alpha} & \quad \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{l} \alpha_i \\
\text{subject to} & \quad \sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C \\
& \quad i = 1, \ldots, l
\end{align*}
\]

(2-2)

The model is essentially a nonlinear optimization problem, and \( K(x_i, x_j) \) is called kernel function. This paper adopts the Gaussian radial equilibrium kernel function. \( C \) is the compromise coefficient (If \( C \) increases, we get smaller empirical risk, otherwise, the confidence risk gets lower).

Propose \( \alpha^* \) is a optimal solution of C-SVC, the classification function is [6]:

\[ f(x) = \text{sgn} \left( \sum_{\alpha_i > 0} y_i \alpha_i^* K(x_i, x) + b^* \right) \]

(1-3)

\( b^* \) is calculated by equation 2-4:

\[ b^* = y_j - \sum_{\alpha_i > 0} y_i \alpha_i^* k(x_i, x_j) \quad (\alpha_i^* > 0) \]

(2-4)

The KKT condition of C-SVM model is:

\[ \alpha_i = 0 \iff y_i u_i \geq 1 \]

\[ 0 < \alpha_i < C \iff y_i u_i = 1 \]

\[ \alpha_i = C \iff y_i u_i \leq 1 \]

(2-5)

where \( u_i = \sum_{\alpha_j > 0} y_j \alpha_j k(x_j, x_i) - b \)

The classification result of C-SVM model is showing like figure 1:
Because the C-SVM model is non-linear model, so it is better than other models in classification accuracy and structural risk minimization.

3. Feature Extraction

In order to clustering the users’ behaviors, we should extract the feature of them which can reflect their behaviors. Using these features as input, the C-SVM model can cluster the users’ behaviors into different classes. According to document [7], the variable rate of URL suffixes (including the access path & file name) can be an index feature of news browsing behaviors, the access dispersion rate can be an index feature of resources sharing behaviors and the quantity of post request can be an index feature of real-time communication behaviors. This paper uses these index features as the input parameters of the SVM model.

Def 2.1 $T_d$: The variable rate of URL suffixes, it can be calculate by formula:

$$T_{d(i,j)} = | \text{Suffix of } D_g \cap \text{Suffix of } D_{i\leadsto j} | \quad (3-1)$$

When $T_d(i,j)$ is getting smaller, it indicate the variable rate of URL suffixes is greater. The characteristic of news website is fast information renew speed, numerous information quantity and frequently accessing.

Generally, user will access similar news at a small time interval. So $T_d(i,j)$ is an excellent index to distinguish it from others.

Def 2.2 $T_{ip}$: Access dispersion rate, it can be calculate by formula:

$$T_{ip(i,j)} = \frac{d_{ip(i,j)}}{|D_y|} \quad (3-2)$$

$d_{ip}(i,j)$ is the distinct IP address quantity in $D_{ij}$. The shared resources contain files, music such as advertisement, product introduction etc. Usually users will forward to other address once they get the resources needed. So the access dispersion rate of this behavior is greater than others.

Def 2.3 $T_p$: The quantity of post request in users’ access records. It can be calculate by formula:

$$T_{p(i,j)} = \frac{N_{p(i,j)}}{|D_y|} \quad (3-3)$$

$N_p(i,j)$ is the record quantity of post method in $D_y$. Because SPs and terminal users need to interact with each other, there are many upload actions using post method in this process. There are more post requests in real-time communication behaviors.
4. Structure of SMO

The training algorithm of SVM is actually solving a convex quadratic programming problem. The classical algorithms to solve it contain Newton method, quasi Newton method, gradient projection method and interior point method. These algorithms need to update the parameters of the kernel matrix in each iteration. When the dataset is large, the memory required of $Q$ will exceed the memory of any ordinary computer. Osuna putted forward a decomposing algorithm in 1997 which broke a large QP problem down into a sequence comprised of small QP problems. During each iteration, the algorithm adds a KKT condition violated point to last QP sub-problems; the value of objective function always satisfies the constraint which enables the convergence of the sequence of sub-problems until the last point of all the KKT conditions are satisfied, resulting in the solution of the original problem. Sequential Minimal Optimization is a special case of decomposing algorithm. Its working set is firmly set to 2 and there is no iteration in SMO. The objective function of SMO is:

$$
\begin{align*}
\min_{\alpha_i, \alpha_j} & \frac{1}{2} k_{ii} \alpha_i^2 + \frac{1}{2} k_{jj} \alpha_j^2 + y_i y_j k_{ij} \alpha_i \alpha_j - (\alpha_i + \alpha_j) \\
+ & y_i \alpha_i v_i + y_j \alpha_j v_j + \frac{\tau - \alpha_j}{4} ((\alpha_i - \alpha_i^k)^2 + (\alpha_j - \alpha_j^k)^2) \\
S.T. & \quad y_i \alpha_i + y_j \alpha_j = \sum_{i=1}^l y_i \alpha_i = \text{Constant} \\
& \quad 0 \leq \alpha_i, \alpha_j \leq C
\end{align*}
$$

As $y_i \alpha_i + y_j \alpha_j = \text{Constant}$, we can get a convex quadratic programming by eliminating $\alpha_i$. If we neglect $0 \leq \alpha_2 \leq C$, the solution will be $\alpha_2^{new} = \alpha_2 + \frac{y_2 (E_i - E_2)}{\eta}.

$$
E_i = y_i - y_j, \eta = k(x_i, x_j) + k(x_2, x_2) - 2k(x_i, x_2)
$$

If the constrain is concerned, the solution of $\alpha_2^{new}$ is:

$$
\alpha_2^{new, clipped} = \begin{cases} 
H & \text{if } \alpha_2^{new} \geq H \\
\alpha_2^{new} & \text{if } L < \alpha_2^{new} < H \\
L & \text{if } \alpha_2^{new} \leq L
\end{cases}
$$

By decomposing the original problem into a series of small sub-problems, SMO doesn’t need any extra memory matrix or iteration. Although iteration counts increases for there are more work sets, the training time decreases because the computing time of each iteration is very short.

In SMO algorithm, it is important to choose the working set $B$ of each sub-problem. It seriously affects the convergence and effectiveness. SMO uses a heuristic approach rather than the traditional steepest descent method in determining the working set. The outer loop traverse the training samples to find a violation sample of KKT condition; the inner loop scans the non-boundary support vectors to find a violation sample of KKT condition which matched the outer loop optimization. This article takes the second-order information and any use of the working set of nuclear matrix selection rules (WSS3). WSS3.
The selection algorithm procedure is as follows:

1) Calculate the parameters according to input feature data:
\[
\alpha_{ij} = k_{ii} + k_{jj} - 2k_{ij},
\]
\[
b_{ij} = -y_i \nabla p(\alpha^k_i) + y_j \nabla p(\alpha^k_j),
\]
\[
\gamma_{ij} = \begin{cases} \alpha_{ij}, & \alpha_{ij} > 0 \\ t, & \text{other} \end{cases}
\]
\[
y_i\alpha_i + y_j\alpha_j = -\sum_{k \neq i, j} y_i\alpha_k = \text{constant}
\]
\[
0 \leq \alpha_i, \alpha_j \leq C
\]

2) Choose
\[
i \in \arg \max \{-y_i \nabla p(\alpha)|t \in I_u(\alpha)\}
\]
\[
j \in \arg \min \{-\frac{\gamma_{ij}}{\alpha_j} | t \in I_{su}(\alpha), \gamma_i\nabla p(\alpha), \gamma_j \nabla p(\alpha) \}
\]

3) Format working set user i,j {, }

The structure of SMO using WSS3 working set selection is:

a) Set initial feasible solution \( \alpha^1 \), set k=1
b) If \( \alpha^k \) satisfies the KKT constrain, the algorithm stops. Otherwise, select working set B using WSS3.
c) Calculate \( \gamma_{ij} \), if \( \gamma_{ij} > 0 \), solve the sub-problem:
\[
\text{sub}(B) = \min_{d_B} \nabla p(\alpha^k) B d_B + \frac{1}{2} d_B^T B \nabla^2 p(\alpha^k) B d_B
\]
Subject to:
\[
y_B d_B = 0
\]
\[
d_t \geq 0, \text{ if } \alpha^k_t = 0, t \in B
\]
\[
d_t \leq 0, \text{ if } \alpha^k_t = C, t \in B
\]
If \( \gamma_{ij} < 0 \), solve two Lagrange multiplier problem and assign the optimal result to \( \alpha_{ij}^{k+1} \)
d) Set \( \alpha_N^{k+1} = \alpha_N^k, k + 1 \rightarrow k \). Go to step 2.

5. Simulation

Currently, the classical SVM algorithm is only support two types classification. For multi-class, there are mainly two categories. One is combining tow class classifiers to complete multi-class target recognition, and the other is calculating the parameters of multiple classification plane into one optimization problem. Although the second approach looks simpler, its training speed is slower than the first one for it has more parameters to calculate. Moreover, the classification accuracy is not always better than the first method. So, this paper adopts the second method. The algorithm architecture is showing in
We use network flow mirror technology to get data that come from the CUMT campus network. We use the hostname as the index while doing data statistic. The category information is shown in Table 1.

Figure 2. Multi-Target clustering architecture

Table 1 Category information of different Behavior

<table>
<thead>
<tr>
<th>Category</th>
<th>Distinct Hostnames</th>
<th>Web Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>1680</td>
<td>168865</td>
</tr>
<tr>
<td>Shared Resources</td>
<td>1290</td>
<td>2355</td>
</tr>
<tr>
<td>Real-time Communication</td>
<td>2070</td>
<td>98653</td>
</tr>
<tr>
<td>Other</td>
<td>2356</td>
<td>105623</td>
</tr>
</tbody>
</table>

By extracting variable rate of URL suffixes, access dispersion rate and the quantity of post request in users’ access records, we can cluster the user behavior into three classes.

Figure 3. The clustering result

Figure 3 is the cluster result using WSS3 working set selection method. The dark part represents the news browsing behavior while the light represents others. So the support vectors can distinguish these tow behaviors.
According to figure 4, the feature of users’ behaviors satisfy the expect result. For example, the quantity of post method in real-time communication is greater than others.

6. Conclusion

This paper puts forward a cluster method of users’ behaviors based SVM. By extracting the features of these behaviors and then use them as model input. The experiment proved that the method is feasible and effective. But deeper analysis of user behavior is also needed for further study.

References