An Improved Classification Technique Using Adaptive Gaussian Mixture Model for User Behavior Modeling

M.Rekha Sundari[1], Prasad Reddy Pvgd[2], Y.Srinivas[3]

[1]Asst. Professor, Dept. of CSE, GITAM UNIVERSITY, Visakhapatnam
[2]Professor, Dept. of CS&SE, ANDHRA UNIVERSITY, Visakhapatnam
[3]Professor, Dept. of IT, GITAM UNIVERSITY, Visakhapatnam
rekhasundari@yahoo.com, prasadreddy.vizag@gmail.com, ysrinivasit@rediffmail.com

Abstract: The purpose of Web usage mining is to extract the knowledge that is concealed in the web log files of a Web server. By implementing statistical and data mining techniques to the Web log data, curious patterns regarding the users' navigational behaviour can be grabbed, such as user and page clusters, as well as possible correlations between Web pages and user groups. In this paper, we propose a Classification method based on Adaptive Gaussian Mixture Model (AGMM). Under the AGMM, user's navigational behaviour in a Web site can be seamlessly classified to discover the cluster of likeminded users. Our experimental results show that our method can achieve better classification accuracy when compared to standard classification approaches based on Gaussian Mixture Model.

Keywords: Clustering, Classification, Gaussian Mixture Model.

1. Introduction

Web usage mining (WUM) [2] is the mechanism of discovering user navigation patterns from web data. This can be considered as a three-phase process, consisting of data training, pattern detection and pattern analysis phases. In order to spot the users, sessions, page views etc, Web log data are pre-processed. In order to explore impressive patterns statistical methods, as well as data mining methods are applied second phase. In the third phase of the Web usage mining process the patterns explored are further scrutinized.

Web usage mining, can be employed on various data sources such as web access logs, cookies, data tags, login information, client or server side scripts, packet sniffing, etc. All these e-resources have its benefits and ill effects, but important sources for web usage mining are web access logs. Reports on content accessibility and visitors navigation paths are outcomes of WUM. Application designer cannot predict beforehand the behaviour of the visitors that can be discovered by WUM by exercising data mining techniques [7] on the vast data of web access logs.

The formal model currently in use based on web access logs for personalisation of web pages is coined by cooley et.al (1999). This process involves three layers. The first step is aimed at processing the log data and classifying the pages based on the user characteristics. The second step aims at transforming sessions into transactions by labelling into clusters based on user preferences. In the final step the data is pre-processed and meaningful data is extracted by which the web logs can be transferred in to knowledge sources.

Association rule mining [5] is a technique for determining frequent patterns, associations and correlations among sets of items. To determine the correlations between pages accessed in sync during a session server Association rules are used. These rules can be used to spot all the direct and indirect relationships between pages that are viewed together .These relationships contain groups of users with specific interests. Web site restructuring, improving the system’s performance through pre-fetching Web data are some of the after-effects of ARM beside business applications.

Sequential pattern discovery [13] is an extension of association rules mining. This technique incorporates the notion of time sequence to divulge patterns of co-occurrence. Predictions related to visit patterns of users can be determined and new directions of navigation can be suggested to the users using this approach.

The two forms of data analysis models that can be used to extract models describing important data classes are Clustering and Classification [3].Clustering is a mechanism of arranging groups with identical data. Classification is a data mining (machine learning) technique used to predict group membership for data instances.

In the context of Web mining, user clusters and page clusters are the two different types of clusters that can be formed. Page clustering [15] considers the user impression and groups the pages that are conceptually related...
Clustering is a discovery tool. Data can be quickly summarized using cluster analysis, especially if the objects are scattered into many groups. Clustering is an unsupervised algorithm, this statement clearly elucidates that no prior knowledge about the clusters to be formed or about the data members of each cluster is required. Previously uneventful associations and structure of data can be identified using clustering. The users are classified in to n number of groups based on some similarity. Clustering finds groups of data that can build a model of the problem based on those groupings. Model construction is the basic advantage of using clustering beside many benefits like data organization, and categorization and data compression. Basically there are two types of clustering algorithms, partitional clustering and hierarchical clustering. Partitional clustering divides data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. Hierarchical clustering forms a set of nested clusters organized as a hierarchical tree. The partitional clustering algorithm K-Means is used for clustering the users in to similar groups. A specific number of disjoint, flat (non-hierarchical) clusters are generated by K-Means clustering. The K-Means method is numerical, unsupervised, non-deterministic and iterative algorithm that produces fairly higher accuracy, lower RMSE and requires less computation time. Performance of K-Means is better than hierarchical clustering and techniques like fuzzy C-means and Self organized maps [1], and is recommended for huge dataset while hierarchical clustering algorithms are preferred for small data set.

Scaling a data item into one of several rooted classes is the intention of Classification technique. To construct a concise model of the distribution of class labels in terms of predictor features is the objective of supervised learning. The resulting model is then used to designate class labels to the new instances. A designed classifier is said to generalize well if the classifier achieves similar classification accuracy to both training samples and real world samples. There are different approaches for modelling a classifier: fixed models, parametric models and nonparametric models. Fixed model is used when the exact input-output relationship is known. Parametric model is used when its parametric mathematical form can be obtained. Nonparametric model is used when the relationships between input vectors and their associated classes are not well understood.

Similarly, mixture models that are parametric or non-parametric can be applied for classification, in which probability density function (PDF) is used to classify the data among various clusters [16][12][10]. In this paper, we applied Gaussian Mixture model (GMM) - a parametric classification model to classify the user navigational behavior and to discover the cluster of likeminded users as considering the practical constraints like memory usage, training time ,classification time GMM outperforms when compared to K nearest neighbour, back propagation and decision tree algorithms.

For better classification accuracy, we further expanded the GMM to Adaptive Gaussian Mixture model (AGMM). The developed models are tested under msnbc dataset [6]. In order to have a comprehensive study of data, the data is to be clustered such that structured information can be pooled among the clusters and based on these clusters of users behaviors, the data can be classified appropriately. Of all the popular clustering techniques available, we have chosen K-means clustering algorithm in this paper in order to cluster the data.

The rest of the paper is organized as follows. Section 2 of the paper presents motivation and section 3 highlights the GMM. In section 4, AGMM is presented. The methodology along with experimentation results is presented in section 5 and finally section 6 of the paper summarizes the results with conclusion.

2. Motivation
Model based classification techniques are very much useful, among the model based techniques, Gaussian mixture model based techniques are mostly preferred, the main advantage of using the model based approaches is that, the data is considered as a mixture of probabilistic distributions, and the data objects that follow the same distribution can be regarded as a class. However, mixture models such as a Mixture of Gaussians still suffer parameter over-fitting problems, when large dimensional data is considered, they cannot interpret the relationships among the data, which plays an important role in web mining.

To overcome these disadvantages of the traditional GMM, in this paper Adaptive Gaussian Mixture model is preferred. The main advantage of using Adaptive GMM is that, in the Adaptive Gaussian Mixture Model algorithm, each data value can be modelled by a mixture of Gaussian distributions. Each Gaussian represents a feature or attribute of the user. The mean and standard deviations of each of the distributions can be updated over time, enabling the algorithm to adapt to estimate the relationships between the users more appropriately. The weight and standard deviation of each distribution determine whether a feature/attribute corresponds to a particular user or to a different user.
3. Classification Using GMM

GMM Classifier role is to assign the input data represented by their features to a number of different categories. GMM is an unsupervised classifier i.e. the training samples of a classifier are not labelled to show their category membership.

In the GMM classifier, the conditional PDF of the observation vector with respect to the different classes is modelled as a linear combination of multivariate Gaussian pdf’s. Each of them has the following general form:

\[ f_\text{GMM}(x) = \sum_{i} \pi_i \frac{1}{\sqrt{2\pi \sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \]  \hspace{1cm} (1)

Where \( x \) is a d- component feature vector, \( \mu \) is the d- component vector containing the mean of each feature; \( \sigma \) is the d- component vector containing the standard deviation of each feature. The assumption is made that the features are independent. Thus \( p(x) \) can actually be written as the product of the univariate probability densities for the elements of \( x \).

4. Adaptive Gaussian Mixture Model

AGMM is an improvised version of GMM in which the probability density is a function of \( x, \mu, \sigma \) as equivalent to GMM and with two additional parameters \( n \) and \( N \) where \( N \) is total number of samples present in the data and \( n \) is number of samples in each cluster.

The probability density function of Adaptive GMM is given by:

\[ f_\text{AGMM}(x, \mu, \sigma, n, N) = \frac{1}{\sqrt{2\pi \sigma}} \left( \frac{1}{N} \sum_{i=1}^{n} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}} + \frac{1}{N} \sum_{j=n+1}^{N} e^{-\frac{(x-\mu_j)^2}{2\sigma^2}} \right) \]  \hspace{1cm} (2)

Assume that each sample \( x \) is a d dimensional vector. Let \( x = [x_1, x_2, \ldots, x_d] \). As the features are independent, the mean and standard deviation are also calculated independently. For a cluster with \( n \) samples, the mean \( \mu \) and standard deviation \( \sigma \) of each feature \( x_i \) is calculated by taking the \( x_i \)'s of all the samples in that particular cluster. So the mean and the standard deviation are given by the equations 3 and 4.

\[ \mu = \frac{1}{n} \sum_{j=1}^{n} x_{ij} \]  \hspace{1cm} (3)

\[ \sigma = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \mu)^2} \]  \hspace{1cm} (4)

5. Methodology and Experimental Results

5.1 Dataset

We used the msnbc.com anonymous web data as test data. This data describes the page visits of users who visited msnbc.com.

Data Characteristics: The data is collected from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msnbc.com for the entire day. Each single line in the dataset designates page views of a user during a day. Each entry in the line corresponds to a user’s request for a page. Table 1 below shows the data format of the data set used. Entries are recorded at the level of page category but not at the level of URL. The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports". All page requests addressed via a caching mechanism were not entered in the server logs and, hence, not present in the data.

Each category is associated--in order--with an integer starting with "1". For example, "frontpage" is associated with 1, "news" with 2, and "tech" with 3. Each describes the hits--in order--of a single user. For example, the third user hits "news" twice, and the "tech" thrice and "local" once.
Table 1 Data Format

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 1</td>
<td>2</td>
<td>3, 2, 4, 2, 3, 3</td>
<td>2, 9, 9, 12</td>
</tr>
<tr>
<td>1, 2, 11, 15, 8</td>
<td>1, 12, 12, 8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.2 Pre-processing Dataset

We assume that there is a set of $n$ unique URLs appearing in the pre-processed log, $U = \{url_1, url_2, url_3, ..., url_{17}\}$ and a set of $m$ user transactions: $T = \{t_1, t_2, ..., t_{500}\}$. We represent the transactions as a bit vector $t=<u_{t1}, u_{t2}, ..., u_{t17}>$. Where

$$u_{tj} = \begin{cases} 1, & \text{if } url_j \in t \\ 0, & \text{otherwise} \end{cases}$$

The data format in table 1 is pre-processed and each entry in the processed data gives the count of user visits to that particular page. If the user doesn’t visit that particular page entry will be treated as 0. The sample of this form is illustrated in table 2. The number of user transactions $m = 500$ and the number of URLs $n = 17$.

Table 2 Format of data grouped by transactions

<table>
<thead>
<tr>
<th></th>
<th>URL1</th>
<th>URL2</th>
<th>URL3</th>
<th>...</th>
<th>URL17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tnx 1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Tnx 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Tnx 3</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Tnx 500</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>

5.3 Clustering transactions into user groups by using k-means algorithm

The K-means approach of clustering is a traditional one [9], where the process begins with a random selection of $K$ objects each objects representing a cluster mean. The distance of the objects to the centre mean are calculated using Euclidean distance method, and these objects are transformed to the clusters having similar means. This process is continued till a convergence is attained [8][11].

The above methodology is utilized for our purpose, where the clustering of the web pages is carried out by initializing a random value of $K=10$. The data considered is a 17-dimensional binary transaction, and the data is clustered in to vectors of 10 groups, basing on similarity.
5.4 Classification using GMM and AGMM

We consider a set X of m observations of d features $x = [x_1, x_2, x_i, \ldots, x_m]$. Assuming that the observations are independent and identically distributed, the likelihood that the entire set of observations has been produced by a class $C_i$ is:

$$p(x | c_i) = \prod_{t=1}^{m} p(x_t | c_i) \quad \ldots (5)$$

We assume that $p(x_t | C_i)$ is a mixture of m multiple univariate gaussians. At this stage, for all the classes of users the parameters of the GMM and AGMM are estimated: $\mu_{t,i}$ and $\sigma_{t,i}$ with $t = 1 \ldots m$ and $I = 1 \ldots I_0$ (in our case) using equations 3 and 4.

5.5 Classification Test

Assuming all is well and that we have managed to train the GMM and AGMM, we can proceed to the classification test. A feature vector $x$ is said to belong to a user class if it maximizes $p(C_i | x) = p(x | C_i).p(C_i)$. In the case where we assume that all the classes can occur with the same probability, we are actually concerned by maximizing $p(x | C_i)$ for every possible class of users.

5.6 Results

Figure 1 shows the classification accuracy of our experiment with msnbc dataset. The result is tested by varying the number of users randomly and the results clearly prove that if the number of users increases, AGMM classifies more accurately compared to GMM. The results show that AGMM Model reaches a maximum accuracy of 74.5% at 400 users.

![Classification Accuracy of GMM and AGMM](image)

6. Conclusion

In this paper, we reported on a Classification Model with the Adaptive GMM. We proposed both GMM and AGMM for classifying user's navigational behaviour in a Web site and to discover the cluster of likeminded users. We found that Adaptive GMM performs better than the GMM. Since our experiment used only one dataset, it is important to adopt other type of dataset to verify our methodology as a future work.

References


