ISSN:2231-0711

V. Chitraa | IJCSET(www.ijcset.net) | January 2016 | Vol 6, Issue 1, 14-18

Feature Selection in Session Matrix Using Enhanced ICA

Dr. V. Chitraa
Associate Professor,
CMS College of Science & Commerce, Coimbatore, Tamil Nadu, India.

***Abstract*-User behavior analysis helps the visitors browsing work more effective and simple and is an important research area in this internet scenario. It is used to analyze a new user and study their present behavior and classify them in a group of similar users. The group consists of user navigation patterns which are used for many applications like personalization, site improvement, business intelligence, and web page recommender system. The navigation pattern consists of features which are both relevant and irrelevant for mining process. To improve the overall clustering and classification results only important features have to be discovered in user navigation matrix. This work analyzes each and every feature and selects the relevant and important features for mining by using enhanced Independent Component Analysis. The experimental results shows that the proposed method selects more relevant features in lesser time.**

***Keywords* - Feature Selection, ICA, Navigation Patterns, Preprocessing, Web Usage Mining**

I. INTRODUCTION

Web mining, is the art of discovering useful information from the World Wide Web by employing techniques of data mining on web data. Web mining is categorized into three types based on the data used to mine information. They are web content mining, web usage mining and web structure mining [1]. Web content mining discovers useful information from contents, data and documents. Web structure mining uses models of link structures and topology of hyperlinks for extracting useful information. Web usage mining mines data derived from the interaction between users and the web applications. The data is collected in the form of server access logs that are generated automatically for each and every interaction made by the users with the website and are stored in the server in the form of transaction logs, commonly called as web log data. The data collected at a Web server includes Internet Protocol addresses, requested page, reference page, status, size of the file being transferred and access time.

A web user navigation pattern analysis system is used to analyze a new user and study their present behavior and classify them in a group. This classification results are used for many applications *viz.*, personalization, website redesign or site improvement, business intelligence, navigation behavior prediction and web page recommender system. The system consists of three phases such as preprocessing, grouping and classifying users and analyzing the resultant users behavior.

The first phase of the study focuses on the conversion of raw web log data into a form that is easily accessible by the machine learning algorithms. This step performs four major tasks, namely, data cleaning, user and session identification, navigation path filling and selecting features. The main aim of data cleaning is to remove unwanted or irrelevant details from the raw web log data so as to avoid processing time on unnecessary details. Users are categorized as those with same IPaddress and browsing agent. Session identification is carried out using the assumption that if a certain predefined period of time between two accesses is exceeded, a new session starts at that point. A user session is a delimited set of pages visited by the same user within the duration of one particular visit to a website. Each session is assumed to be of duration 30 minutes and a new session starts after each 30 minutes [2].

The third task of preprocessing is path completion. Path completion is process of adding the page accesses that are not in the web log but those which actually have occurred. The path filling method uses a simple procedure that analyzes previous page accessed and referral page fields of web log data. The method also uses a post processing step that adjusts the reference length of adjacent pages [3]. The result of this task will produce a web log data with no missing values.

Using the cleaned and complete web log data, the next step of preprocessing is the formation of user session matrix which acts as the input for the proposed work. Each row in the matrix is the navigation patterns of the visitors in the log. Web Pages acts as columns or features and rows are user sessions. For each visit one entry is added in the matrix cell and for revisit each cell value is incremented [4].

Not all the features are relevant for the mining process. All the attributes in the mining procedures not only increase the complexity of the analysis, but also decrease the accuracy of the result and give unnecessary load to the systems. Moreover the presence of redundant and irrelevant attributes could mislead the analysis. When the size increases, the storage, transmission and processing of data becomes difficult. The cost involved in mining algorithms is calculated on the basis of time required for the algorithm to run and the size of the dataset. To reduce processing time and to alleviate the limitations of large data during mining process data reduction is needed. Performance is measured in terms of accuracy of the learning model. There are several issues which motivated to carry out Feature selection process in navigation pattern analysis. Some of the major issues are:

- If many users visit some pages repeatedly the other pages may not be considered as most important pages.

- In clustering, due to large dimensionality, all the patterns in the sessions become equidistant and increase the complexity.
- When user navigation patterns contain large number of attributes, a problem of over-fitting occurs at the classification stage to classify similar user navigation patterns from test data.

Thus to improve the overall clustering and classification results only important features have to be discovered in user navigation matrix. In the proposed work only relevant features are selected to enhance further process.

There are three types of Feature selection algorithms in machine learning literature: filter methods, wrapper methods, and embedded methods [5]. The wrapper methods assess subsets of variables according to their usefulness to a given predictor by conducting a search for a good subset using the learning algorithm itself as part of the evaluation function. Embedded methods perform variable selection as part of the learning procedure and are less computationally intensive than wrapper method but specific to a learning machine. Filters perform feature selection independently of any particular classifier, being motivated by the properties of the data distribution itself. In this method optimal subset is searched by using an empirical risk estimate for a particular classifier. It easily scales very high-dimensional datasets, computationally simple and fast, and independent of the classification algorithm. This paper mainly focuses on selection of features by enhancing Independent Component Analysis which follows filter method.

This research paper is organized into five sections and structured as follows. Section two analyzes the works with related problem and section three describes the methodology in detail. Section four presents' experimental results and section five concludes the whole process and scope for further work.

2. Related Work

Dimension reduction is important in cluster analysis, which not only makes the high dimensional data addressable and reduces the computational cost, but can also provide users with a clear picture and visual examination of the interesting data [6]. Many emerging dimensionality reduction techniques have been proposed in the literature. For example, Local Dimensionality Reduction (LDR) approach tries to find local correlations in the data, and performs dimensionality reduction on the locally correlated clusters of individual data and dimensionality reduction adaptively adjusted and integrated with the clustering process [7].

Several dimensionality reduction methods are available in the literature which has been used in several applications [8]. Many methods use the second order statistics, and the variance or covariance of the data in the optimization of the objective function. Some of the methods are Principal Component Analysis, Multivariate Linear Regression, Partial Least Squares, Factor Analysis and Canonical Correlation Analysis. Second-order methods reduce the redundancy of the data by using a simpler representation but do not give a better representation since the latent variables extracted by them are not independent. Rather, they are correlated.

Factor Analysis is a linear method and its goal is to uncover common factors to reduce the dimension of datasets following the factor model [7]. It holds orthogonal relation of factors and does not depend on scale of the variables. Projection pursuit is another linear reduction technique, which finds the interesting projections in the high dimensional data and is used for optimal visualization of the clustered structure of the data. It also incorporates second-order information.

In recent years, principal component analysis (PCA) and Independent component analysis (ICA) are widely used in various applications due to their significant performance [9]. Both are linear dependent methods which takes only the linear dependencies between variables. Principal Component Analysis is a filter method widely employed for feature selection using the linear dimension technique. The transformation takes place linearly by retaining the characteristics of the original data set. It is useful in image processing applications [10] for numeric attributes. It performs an orthogonal transformation on the input space, to produce a lower dimensional space in which the main variations are maintained.

Blind Source Separation (BSS) problem is an area where ICA found its use initially to find some hidden sources from the observed mixture data. These hidden sources are assumed to be mutually independent and may give a simpler and better representation of the data. They are called the independent components.

In web usage mining research, not much work has been done to reduce features. A novel approach for feature selection based on rough set theory for Web usage mining is proposed by Inbarani et al., [11]. The idea of Independent Component Analysis to remove irrelevant features to enhance the rough fuzzy based clustering to cluster navigation patterns is given [12].

3. Methodology

Independent Component Analysis (ICA) is a rich statistical technique developed for digital signal processing applications for revealing hidden factors for measurement of signals. It also acts as a powerful tool for analyzing text document data if the documents are presented in a suitable numerical form. ICA has received considerable interest in recent years because of its versatile applications such as source separation, channel equalization, speech recognition and functional magnetic resonance imaging, face recognition, telecommunication, predicting stock market place and financial market data mining [13]. It has been used for dimension reduction and representation of word histograms [14].

Independence is defined by the probability densities in mathematics between two scalar-valued random variables x and y [13]. The variables x and y are said to be independent if information about x does not give any information on y and vice versa. Let us denote by p1 (y1, y2) the joint probability density function of y1 and y2.

$$p1(y1) = p(y1, y2)\, dy2$$

and p2 follows the same. Then we define that y1 and y2 are independent if and only if the joint probability density function is featured in the following way.

P (y1, y2) = p1 (y1) p2 (y2)

This definition extends naturally for any number, 'n'of random variables, in which case the joint density must be a product of n terms.

A. General Procedure of ICA

The process of extraction of the independent components from the observed mixture data is called independent component analysis. To extract new patterns or best patterns for the purpose of reducing the dimensions of patterns space and to achieve better performances ICA is used. Each and every feature in the navigation patterns is normalized by calculating mean and standard deviation. Absolute mean is calculated for each and every row. Independent matrix is created by comparing and shrinking the values. Finally the independent matrix is multiplied with original matrix and the mean is calculated for all attributes. The mean is compared with a threshold and attributes which are less than threshold are selected.

B. Feature Selection in User Sessions Matrix

Feature selection in this context is the measurement and elimination of irrelevant session matrix attributes that are generally not used by the users. The user navigation pattern differs from one user to another user in the user session matrix. For feature selection, component analysis techniques are implemented. Principal Component Analysis is used to select features but it is too expensive for multivariate data and in Independent Component Analysis the weight matrix is randomly generated which affects the performance and does not give accurate estimation. To solve these issues, the present research work proposes an extended ICA. A Quantum α -Skew Divergence is integrated into the ICA to build a new reduced matrix in the proposed method.

C. Quantum Skew Divergence based ICA method (QSDICA)

The input for this method is the navigation patterns and features of patterns from the user session matrix. The features are to be evaluated and eliminated for the next phase of mining in this process. The pattern matrix consists of 'n' features which are the web pages in the web site. The patterns of the users are sessions and are represented by T. The navigation pattern is denoted by $x_t(j)$ where j = 1…n and t = 1…T. 'I' acts as the index of the patterns. Absolute mean is found for all I where 1 < I < T. The user navigation pattern matrix is normalized by using $(np_i - m_i) / 2\sigma_i$ where m_i and σ_i are the mean and the standard deviation of np_i respectively. Weight matrix W for normalized matrix is generated randomly by using normrnd $(\mu, \sigma,$ m, n) function in the toolbox where$'\mu'$ is the mean, $'\sigma'$ is the standard deviation, m and n are the number of rows and columns of input matrix. Then, the different independent component weight matrix is computed for the generated weight matrix by using Cartesian product between two attributes. Different independent component analysis weights are denoted as $f_I(w)$.

$$f_I(w) = \prod_{j=1}^{m} f(w_j) \quad (1)$$

The new quantum α-skew divergence matrix is calculated from the weighted independent component matrix and normalized user pattern matrix. The asymmetric skew divergence simply smoothens one of the distributions by mixing it to a degree. Skew divergence is observed to achieve lesser error rates [15]. For any fixed α ∈ (0,1) where 'α'is a scalar and is called as a skewing parameter. The quantum α-skew divergence between states of the $f(nx_j)$ and f_I (w) is given by

$$QSKD_{\alpha}(f(x_j)||f_I(w)) = \frac{1}{-\log(\alpha)} QS(f(x_j)||\alpha f_I(w) + (1-\alpha)f_I(w)) \quad (2)$$

where

$$QS(f(nx_j)||f_I(w)) = Trf(nx_j)(1ogf(nx_j) - \log f_I(w) \quad (3)$$

The absolute mean value for each of the user navigation patterns n_p with coefficients $a_{ij,}$ is calculated from quantum α-skew divergence weight matrix 'QW' and the value in 'QW' is qw_{ij}

$$a_{ij} = \frac{1}{T}\sum_{j=1}^{T} |qw_{ij}| \quad (4)$$

The weight value in QW is compared with a_{ij}, and minimum values are shrunk to zero and another weight matrix W' is generated. Then the new weight matrix W' is multiplied with normalized input matrix and it is defined as $y_i(t)$.

$y_i(t) = \sum_j qw_{ij}x_j(t)$for i = 1,,..T, j = 1,....n (5)

For j = 1,2…n, let U denote the two-sided n-value associated to the j^{th} feature.

Absolute mean for each feature is calculated. As a final step in the algorithm, features with absolute mean values greater than threshold value are selected.

D. Pseudocode - QSDICA

- Input data :' m ' input navigation patterns matrix X =[x_1(t)……x_n(t)]
- Normalize each user navigation patterns np_i related features by $\frac{(npi - mi)}{2\sigma i}$
- Generate weight matrix W
- Find different independent components column wise $f(w_j)$ using eqn 1.
- Perform quantum divergence for weight matrix and normalized matrix
- For each row in QW calculate absolute mean
- For all qw_{ij} in QW
  if $|qw_{ij}| < a_{ij}$, then $|qw_{ij}|$ shrinks to zero and a new W' is generated.
  Else weight =qw_{ij}
- Multiply new weight matrix W' of dimension T x n to the normalized matrix, it is specified as $y_i(t)$
- Calculate the absolute mean for each column
- If ya_{ij} > mean select feature else go to next column
- End the process.

The method is implemented in the modeled session matrix. For comparison existing methods like PCA, ICA was also implemented.

IV. Experimental Results

This work has been implemented in MATLAB r2011a. The data sets used in this work are three real time datasets. The experiments were undertaken in three web log data sets which are collected from ecommerce web server, academic institution web server and a research journal web server. User's requests are accumulated in web server and it is an automatic process performed by the web server. It typically stores information like IP address, user id and password, date and time stamp, status field indicating whether a request is successful or not, size of the file being transferred, referring URL and finally, the name and version of the browser being used.All three are IIS web servers and the version is Log 3.0. To prove the applicability of web log analysis in different domains, the data sets are collected from three different web logs. Initially the e-commerce log file consists of 23141 raw log entries, institution log file consists of 56596 entries and journal's log file consists of 6697. All the datasets are experimented with the existing and proposed methods and evaluated.

A. Data Cleaning

Data cleaning phase is performed and irrelevant entries are removed which are not necessary for mining process. Entries with graphics and videos format such as gif, JPEG, etc., are removed. The status code field of log entry indicates whether the request was successfully fulfilled or the details of the status of the request from server etc., Robots or agents also acts like browser. All these details are not necessary for the mining process. So in this step all these details are removed. The entries with URL with gif, JPEG, etc and the status code < 200 and > 299 are removed. The resultant log entries after data cleaning is tabulated:

Table 1: Data Cleaning Results

Data Sets	No of Entries	No of Entries after Cleaning
E-commerce	23141	9687
Institution	56596	51148
Journal	6697	6005

B. Session Matrix Construction

Initially unique pages are selected from the cleaned log and these pages are the attributes. Sessions are identified by Navigation oriented method. Totally 999 sessions were identified in e-commerce log file and in the institutions' log there are 1400 sessions and there are 180 sessions in the researchers journal log file. In the matrix constructed the sessions are rows and web pages are attributes otherwise called as features. This matrix is the input for the proposed work.

C. Feature Selection

Feature selection in this work is the measurement and elimination of the unimportant web log attributes in the session matrix that are not mostly used by the users and are irrelevant. The number of web pages is selected by retrieving the unique pages from the log file. In the ecommerce web log there are 14 unique web pages navigated by users and they are the attributes for user session matrix. The number of attributes selected from e-commerce log is 14, in academic institution web log there are 13 attributes and in researchers journal the number of attributes selected are only four since the website is new and has less number of active pages and researchers repeatedly browse only few pages. The existing methods such as PCA, ICA and proposed method Quantum Skew Divergence based Independent Component Analysis are implemented and relevant features are selected. The performance of these approaches has been evaluated using the parameters such as number of features and time taken for executing the different algorithms in the three data sets taken.

D. Time taken for Feature Selection

All the three data sets are executed with PCA, ICA and proposed QSDICA in Matlab. The number of features selected in each method is considered. The time taken for three algorithms are taken and compared. It has been noted that the proposed method gives results at a minimum time than the existing methods. Number of features selected, Time taken for each method for three data sets are tabulated below and the comparison is depicted in figure 1.

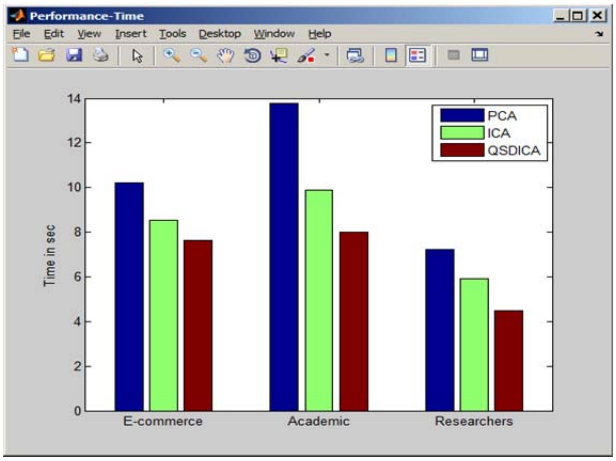

FIG 1. COMPARISON OF EXECUTION TIME

TABLE 2: FEATURE SELECTION RESULTS

Samples (after cleaning)	E-Commerce			Institution			Research		
Number of samples	19239			51148			6005		
Methods	PCA	ICA	QSDICA	PCA	ICA	QSDICA	PCA	ICA	QSDICA
No of features selected	8	6	5	11	8	6	4	4	3
Time in seconds	10.22	8.54	7.62	13.78	9.9	8	7.22	5.9	4.5

V. CONCLUSION

Analysis of navigation patterns is a new and useful research area due to its applicability in various domains. Due to the information explosion and tremendous increase in the number of users for websites, the knowledge of user's interests from web logs helps web masters and growth of business. The research work focused on the selection of relevant features from user session matrix. The feature selection method selects relevant features using Independent Component Analysis (ICA) which eliminates the unimportant attributes to enhance the accuracy of classification and to speed up mining process. To get more relevant features a Quantum α-Skew Divergence is integrated with ICA in weight matrix generation. This proposed QSDICA approach gives more comprehensive results than the existing ICA and PCA methods due to the skew divergence factor. The resultant features are mined into resultant groups and a whenever a new user's entry arrives they are analyzed and classified in the further process.

REFERENCES

[1]. Junior C.G.M.,Gong Z.," Web Structure Mining: An Introduction",IEEE International Conference on Information Acquisition, 2005.

[2]. Robert. Cooley, Bamshed Mobasher and Jaideep Srinivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", journal of knowledge and Information Systems,1999.

[3]. V.Chitraa , Antony SelvadossThanamani, "An Efficient Path Completion Technique for Web Log Mining", IEEE International Conference On Computational Intelligence and Computing Research,TamilNadu CollegeOf Engineering,Coimbatore,Dec 2010.

[4]. Chitraa.V., and Dr.Antony Selvadoss Thanamani, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing". International Journal of Computer Applications 34(9), November 2011. Published by Foundation of Computer Science, New York, USA.

[5]. Maldonado, Sebastián, Richard Weber, and Jayanta Basak. "Simultaneous feature selection and classification using kernel-penalized support vector machines." Information Sciences 181.1 (2011): 115-128.

[6]. Büchner, M. Baumgarten, S.S. Anand, M.D. Mulvenna, J. G. Hughes, Navigation Pattern Discovery from Internet Data, in WEBKDD, San Diego, CA 1999.

[7]. Fodor, I.K: "A Survey of Dimension Reduction Techniques". LLNL Technical Report, UCRL-ID-148494", 2002.

[8]. Ding, Chris, Tao Li: "Adaptive Dimension Reduction Using Discriminant Analysis and K-means Clustering", International Conference on Machine Learning, Corvallis, OR, 2007 .

[9]. Kolenda, L.K. Hansen, and S. Sigurdsson. "Independent components in text", In M. Girolami, editor, Advances in Independent Component Analysis. Springer- Verlag, 2000.

[10]. Nedevschi S., Bota S. and Tomiuc C., Stereo-Based Pedestrian Detection for Collision-Avoidance Applications. IEEE Transactions on Intelligent Transportation Systems, vol. 10, no. 3, 2009.

[11]. Inbarani, H. Hannah, K. Thangavel, and A. Pethalakshmi. "Rough set based feature selection for web usage mining." Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on. Vol. 1. IEEE, 2007.

[12] Chimphlee, Siriporn, et al. "Mining Usage Web Log Via Independent Component Analysis And Rough Fuzzy." Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases. World Scientific and Engineering Academy and Society (WSEAS), 2006.

[13] Hyvarinen, J. Karhunen, and E. Oja, "Independent Component Analysis". New York: Wiley, 2001.

[14]. Kolenda, L.K. Hansen, and S. Sigurdsson. "Independent components in text", In M. Girolami, editor, Advances in Independent Component Analysis. Springer- Verlag, 2000

[15]. Lillian Lee, "On the effectiveness of the skew divergence for statistical language analysis", Proceedings of Artificial Intelligence and Statistics (AISTATS), 2001.

AUTHOR

Dr V. Chitraa is working as an Associate Professor in CMS College of Science & Commerce, Coimbatore. Her research interest lies in Web Mining, Data Mining, Database Concepts and knowledge mining. She has published many papers in reputed international journals and conferences. She is a member of IEEE student chapter