Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp394-399)

MINING USAGE WEB LOG VIA INDEPENDENT COMPONENT ANALYSIS AND ROUGH FUZZY

SIRIPORN CHIMPHLEE[1], NAOMIE SALIM[2], MOHD SALIHIN BIN NGADIMAN[3],
WITCHA CHIMPHLEE[4], SURAT SRINOY[5]

[1,4]Faculty of Science and Technology
Suan Dusit Rajabhat University, 295 Rajasrima Rd, Dusit, Bangkok, THAILAND
Tel: (+66)-2445675, Fax: (+66) 6687136, http://www.dusit.ac.th
[2,3]Faculty of Computer Science and Information Systems,
University Technology of Malaysia, 81310 Skudai, Johor, MALAYSIA
Tel: (607) - 5532070, Fax: (607) 5565044, http://www.utm.my

Abstract:- In the past few years, web usage mining techniques have grown rapidly together with the explosive growth of the web, both in the research and commercial areas. Web Usage Mining is that area of Web Mining which deals with the extraction of interesting knowledge from logging information produced by Web servers. A challenge in web classification is how to deal with the high dimensionality of the feature space. In this paper we present Independent Component Analysis (ICA) for feature selection and using Rough Fuzzy for clustering web user sessions. Our experiments indicate can improve the predictive performance when the original feature set for representing web log is large and can handling the different groups of uncertainties/impreciseness accuracy.

Keywords:- Web Usage Mining; Web log mining; Independent component analysis; Rough Sets; Fuzzy rough sets;

1. Introduction

With the rapid growth of information on the World Wide Web, automatic classification of web pages has become important for effective indexing and retrieval of web documents. Soft computing is a consortium of methodologies that works synergistically and provides, in one form or another, flexible information processing capability for handling real life ambiguous situations. The principal soft computing tools include fuzzy sets, artificial neural networks, genetic algorithms and rough set theory [1]. Web usage data includes data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse-clicks and scrolls, and any other data generated by the interaction of users and the web. Web using mining (WUM) works on user profiles, user access patterns and mining navigation paths which are being heavily used by e-commerce companies for tracking customer behavior on their sites [1]. Web usage mining of the data generated by the users' interactions with the Web, typically represented as Web server access logs, user profiles, user queries and mouse-clicks. This includes trend analysis (of the Web dynamics information), and Web access association/sequential pattern analysis [2]. The clustering problem is a fundamental problem that frequently arises in a great variety of fields such as pattern recognition, machine learning, and data mining. Pre-processing the trend signals for the purpose of noise removal, qualitative interpretation and dimension reduction is now an important step in developing computer-aided systems for fault detection and diagnosis. There are two aspects in dimension reduction of process dynamic trends: feature extraction from a trend of an individual variable and removal of dependencies among a number of correlated and sometimes redundant variables [3]. Independent component analysis (ICA) aims at extracting unknown hidden factor/components from multivariate data using only the assumption that the unknown factors are mutually independent [4]. Rough sets are a tool to deal with inexact, uncertain or vague knowledge. Specifically, it provides a mechanism to represent the approximations of concepts in terms of overlapping concepts.

The rest of the paper is organized as follows: Section 2 deals with web clustering and features selection. The Rough Sets, Fuzzy Set and Rough Fuzzy are discussed in Section 3. Section 4 provides an experimental design. Sections 5 describe experimental results and discussions and conclusion in Section 6.

2. Web Clustering and Features Selection

2.1. Web clustering

Clustering pertains to unsupervised learning, where no predefined classes are assigned. The key requirement is the need for a good measure of similarity between the instances/patterns. The problem is to group *n* patterns into *c* desired clusters, such that the data points within clusters are more similar than across clusters. Scalable clustering algorithms pertain to working with large volumes of high dimensional data that is inherent to data mining problems [2]. The importance of clustering to Web mining, specifically in the domains of Web Usage mining, make Web clustering an interesting topic of research. This includes clustering of access logs the involves overlapping clusters.

2. 2 Independent component analysis (ICA)

In a classification problem, the number of features can be quite large, many of which can be irrelevant or redundant. A relevant feature is defined in [5] as one removal of which deteriorates the performance or accuracy of the classifier, and an irrelevant or redundant feature is not relevant. These irrelevant features could deteriorate the performance of a classifier that uses all features since irrelevant information is included inside the totality of the features. Thus the motivation of a feature selector is (i) *simplifying* the classifier by the selected features; (ii) *improving or not significantly reducing* the accuracy of the classifier; and (iii) *reducing* the dimensionality of the data so that a classifier can handle large values of data [6]. Many approaches as feature selectors have been proposed.

Independent component analysis (ICA) for dimension reduction is to separate these independent components (ICs) from the monitored variables. Introduction of ICA concepts in the early 1980s in the context of neural networks and array signal processing. ICA was originally developed to deal with problems that are closely related to the real world 'cocktail-party' problem. ICA is a method for automatically identifying the underlying factors in a given data set. Dimension reduction using ICA is based on the idea that these measured variables are the mixtures of some independent variables. When given such a mixture, ICA identifies those individual signal components of the mixture that are unrelated. Given that the only unrelated signal components within the signal mixture are the voices of different people. ICA is based on the assumption that source signals are not only uncorrelated, but are also 'statistically independent' [7].

ICA techniques provide statistical signal processing tools for optimal linear transformations in multivariate data and these methods are well-suited for feature extraction, noise reduction, density estimation and regression [8]. The ICA problem can be described as follows, each of h mixture signal $x_1(k), x_2(k),\ldots,x_h(k)$ is a linear combination of q independent components $s_1(k), s_2(k),\ldots,s_h(k)$, that is , X = AS where A is a mixing matrix. Now given X, to compute A and S. Based on the following two statistical assumptions, ICA successfully gains the results: 1) the components are mutual independent; 2) each component observes nongaussian distribution. By X = AS, we have $S = A^{-1}X=WX$ (where $W = A^{-1}$). The take is to select an appropriate W which applied on X to maximize the nongaussianity of components. This can be done in an iteration procedure.

Given a set of *n*-dimensional data vectors $[X^{(1)},X^{(2)},\ldots,X^{(N)}]$, the independent components are the directions (vectors) along which the statistics of projections of the data vectors are independent of each other. Formally, if A is a transformation from the given reference frame to the independent component reference from then

$$X = As$$

Such that

$$p(s) = \prod p_a(s_i),$$

where $p_a(.)$ is the marginal distribution and *p*(s) is the joint distribution over the *n*-dimensional vector *s*.

Usually, the technique for performing independent component analysis is expressed as the technique for deriving one particular W,

$$Y = Wx,$$

Such that each component of y becomes independent of each other. If the individual marginal distributions are non-Gaussian then the derived marginal densities become a scaled permutation of the original density functions if one such *W* can be obtained. One general learning technique [9; 10] for finding one *W* is

$$\Delta W = \eta(I - \phi(y)y^T)W,$$

Where $\phi(y)$ is a nonlinear function of the output vector y (such as a cubic polynomial or a polynomial of odd degree, or a sum of polynomials of odd degrees, or a sigmoidal function) [11].

3. Rough Sets, Fuzzy Set and Rough Fuzzy

3.1 Rough Sets

Rough sets are characterized by their ability for granular computation. In rough set theory a concept *B* is described by its "lower" $(\underline{B})$ and "upper" $(\overline{B})$ approximations defined with respect to some indiscernibility relation. Rough set theory [12] provides an effective means for analysis of data by synthesizing or constructing approximations (upper and lower) of set concepts from the acquired data. The key notions here are those of ''information granule'' and ''reducts''. Information granule formalizes the concept of finite precision representation of objects in real life situation, and reducts represent the core of an information system (both in terms of objects and features) in a granular universe [13].

Let $X = \{x_1,\ldots,x_n\}$ be a set of *U* and *R* an equivalence relation on *X*. As usual, *X/R* denotes the quotient set of equivalence classes, which form a partition in *X*, i.e. *xRy* means the *x* and *y* cannot be took apart. The notion of *rough set* [14] born to answer the question of how a subset T of a set X in U can be represented by means of X/R.. It consists of two sets:

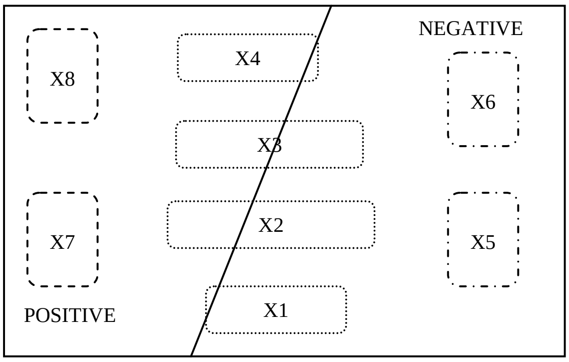

Figure 1: Positive or negative region [27].

G. Chakraborty and B. Chakraborty [27] are proposed POSITIVE and NEGATIVE As shown in Fig.1 are the two partitions induced by the decision attribute, POSITIVE is for "*Decision Yes"* and NEGATIVE is for "*Decision No*". The partitions X5, X6, and X7, X8 induced by condition attributes are clearly positive or negative side of the partition induced by the decision attribute. In addition, partitions X2 and X3 are clear boundary cases, and no decision can be made out of the corresponding partitions. But partition X4 is mostly positive, and partition X1 is mostly negative. Working with different real world data, we have seen that partitions like X1 and X4 are very common.

$$RS^*(T) = \{[x]_R \mid [x]_R \cap T \neq 0\} \tag{1}$$

$$RS_*(T) = \{[x]_R \mid [x]_R \subseteq T\} \tag{2}$$

where $[x]_R$ denotes the class of elements $x, y \in X$ such that *xRy*. $RS^*(T)$ and $RS_*(T)$ are respectively the *upper* and *lower approximation* of T by R, i.e.

$$RS_*(T) \subseteq T \subseteq RS^*(T) \tag{3}$$

Other operations over rough sets include:

- Negative region of $X : U - RS^*(X)$.
- Boundary region of $X : RS^*(X) - RS_*(X)$.
- Quality of approximation of X by RS_* and

$$RS^* : \mu_R S(X) = \frac{card(RS^*(X))}{card(RS_*(X))}$$

3.2. Fuzzy Sets

Fuzzy theory provided a mechanism for measuring the degree to which an object belongs to a set by introducing the "membership degree" as a characteristic function $\mu_A(x)$ which associates with each point x a real number in the range [0,1]. The nearer the value of $\mu_A(x)$ to unity, the larger the membership degree of x in the set A.

Let assume X be a set, then two different *crisp* versions of a fuzzy set A can be define, namely $\overline{A} = \{(x, \mu_{\overline{A}} \mid x \in X\}$ and $\underline{A} = \{(x, \mu_{\underline{A}} \mid x \in X\}$ where

$$\mu\overline{A}(x) = \begin{cases} 1 & \mu_A(x) \geq 0.5 \\ 0 & \mu_A(x) < 0.5 \end{cases} \tag{4}$$

and

$$\mu_{\underline{A}}(x) = \begin{cases} 1 & \mu_A(x) < 0.5 \\ 0 & \mu_A(x) \geq 0.5 \end{cases} \tag{5}$$

Denote $A \subset X$ and $B \subset X$ two fuzzy sets, i.e. $A = \{(x_i, \mu_A(x_i)), i = 1, ..., n\}$ and $B = \{(x_i, \mu_B(x_i)), i = 1, ..., n\}$, the operations on fuzzy sets are extensions of those used for conventional sets (intersection, union, comparison, etc.). The basic operations are the intersection and union as defined as follows:

The membership degree of the *intersection* $A \cap B$ is

$$\mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\} \quad x \in X \tag{6}$$

The membership degree of the *intersection* $A \cup B$ is

$$\mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\} \quad x \in X \tag{7}$$

Furthermore, a common measure of similarity between two fuzzy sets A and B is the l^p-distance, defined as follows [15]. The l^p-distance between two fuzzy sets A and B is given by

$$l^p(A, B) = (\sum_{i=1}^{n} | \mu_A(xi) - \mu_B(xi) |^p)^{\frac{1}{p}} \tag{8}$$

if p=1 the l^p-distance reduces to the fuzzy Hamming distance.

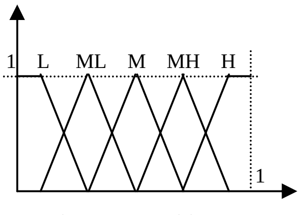

Figure 2: A fuzzy space of five membership function

The fuzzy membership functions corresponding to the informative regions are stored as cases. A collection of fuzzy sets, called fuzzy space, defines the fuzzy linguistic values or fuzzy classes. A sample fuzzy space of five membership function is shown in Figure 2.

3.3. Rough Fuzzy sets

In any classification task the aim is to form various classes where each class contains objects that are not noticeably different. These indiscernible or non-distinguishable objects can be viewed as basic building blocks (concepts) used to build up a knowledge base about the real world [16]. In this paper we propose the rough fuzzy sets, realizing a system capable to efficiently cluster data coming from image analysis tasks. The hybrid notion of rough fuzzy sets comes from the combination of two models of uncertainty like vagueness by handling rough sets and fuzzy sets. Rough sets embody the idea of indiscernibility between objects in a set, while fuzzy sets model the ill-definition of the boundary of a sub-class of this set.

The *rough-fuzzy set* is the generalization of rough set in the sense that here the output class is fuzzy. Let *X* be a set, *R* be an equivalence relation defined on *X*, and the output class $A \subseteq X$ be a fuzzy set. The rough-fuzzy set is a tuple $\langle \underline{R}(A), \overline{R}(A) \rangle$, where the lower approximation $\underline{R}(A)$ and the upper approximation $\overline{R}(A)$ are fuzzy sets of X/R, with membership functions defined in [17, 18] by

$$\mu_{\underline{R}(A)}([x]_R) = \inf\{\mu_A(x) \mid x \in [x]_R\} \qquad \forall x \in X \quad (9)$$

and

$$\mu_{\overline{R}(A)}([x]_R) = \sup\{\mu_A(x) \mid x \in [x]_R\} \qquad \forall x \in X \quad (10)$$

Here, $\mu_{\underline{R}(A)}(x)$ and $\mu_{\overline{R}(A)}(x)$ are the membership values of $[x]_R$ in $\underline{R}(A)$ and $\overline{R}(A)$, respectively.

The rough-fuzzy membership function of a pattern $x \in X$ for the fuzzy output class $A_c \subseteq X$ is defined as

$$l_{Ac}(x) = \frac{\| F \cap A_c \|}{\| F \|}, \qquad (11)$$

Where $F = [x]_R$, and $\| A_c \|$ implies the cardinality of the fuzzy set A_c. Important properties of the rough-fuzzy membership functions that can be exploited in classification task [19].

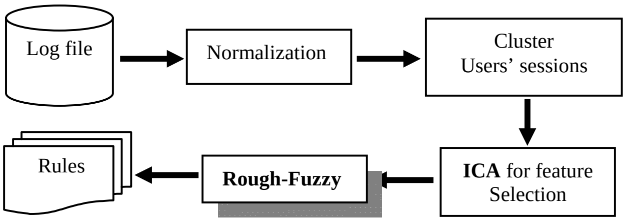

Figure 3: Architecture for web usage mining.

4. Experimental Design

The problem is solved in two steps: 1) feature reduction from measured data; 2) clustering based on selected feature. For the first step, ICA, mostly used in feature reduction from time series data, is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals [8].

5. Experimental Set Up and Results

The prediction models that we build are based on web log data that corresponds with users' behavior. They are used to make prediction for the general user and are not based on the data for a particular client. This prediction requires the discovery of a web users' sequential access patterns and using these patterns to make predictions of users' future access. We will then incorporate these predictions into the web prefetching system in an attempt to enhance the performance.

```
1102801060.863  1897600  172.16.1.98  TCP_IMS_HIT/304 203  GET
http://asclub.net/images/main_r4_c11.jpg - NONE/- image/jpeg
1102801060.863  1933449 172.16.1.183 TCP_MISS/404 526 GET
 http://apl1.sci.kmitl.ac.th/robots.txt -DIRECT/161.246.13.86 text/html
1102801060.863 1933449 172.16.1.183  TCP_REFRESH_HIT/200
3565  GET ttp://apl1.sci.kmitl.ac.th/wichitweb/spibigled/spibigled.html
- DIRECT/161.246.13.86 text/html
```

Figure 4: Sample web log data

The experiment used web data collected from www.dusit.ac.th web server (see example in Figure 4) during 1 December 2004 – 31 December 2004. The total number of web pages with unique URLs is equal to 314 URLs, and there are 13,062 web log records. These records are used to construct the user access sequences (Figure 5). The user session is split into training dataset and testing dataset. The training dataset is mined in order to extract rules, while the testing dataset is considered in order to evaluate the predictions made based on these rules.

5.1 Web log pre-processing

Web log files contain a large amount of erroneous, misleading, and incomplete information. This step is to filter out irrelevant data and noisy log entries. Elimination of the items deemed irrelevant by checking the suffix of the URL name such as gif, jpeg, GIF, JPEG, jpg, JPG. Since every time a Web browser downloads an HTML document on the Internet, several log entries such as graphics and script are downloaded too. In general, a user does not explicitly request all of the graphics that are in the web page, they are automatically down-loaded due to the HTML tags. Since web usage mining is interested in studying the user's behavior, it does not make sense to include file requests that a user does not explicitly request. The HTTP status code returned in unsuccessful requests because there may be bad links, missing or temporality inaccessible pages, or unauthorized request etc: 3xx, 4xx, and 5xx. Executions of CGI script, Applet, and other script codes. We also eliminated enough Meta data to map these requests into semantically meaningful actions, as these records are often too dynamic and contain insufficient information make sense to decision makers.

5.2 Session identification

After the pre-processing, the log data are partitioned into user sessions based on IP and duration. Most users visit the web site more than once. The goal of session identification is to divide the page accesses of each user into individual sessions. The individual pages are grouped into semantically similar groups. A user session is defined as a relatively independent sequence of web requests accessed by the same user [20]. Fu et al. [21] identify a session by using a threshold idle time. If a user stays inactive for a period longer than the identified max_idle_time, subsequent page requests are considered to be in another episode, thus another session. Most researchers use heuristic methods to identify the Web access sessions [22] based on IP address and a time-out not exceeding 30 minutes for the same IP Address. A new session is created when a new IP address is encountered after a timeout. Catledge and Pitkow [23] established a timeout of 25.5 minutes based on empirical data. In this research, we use IP address time-out of 30 minutes to generate a new session (Figure 5).

Session 1 : 900, 586, 594, 618
Session 2 : 900, 868, 586
Session 3 : 868, 586, 594, 618
Session 4 : 594, 618, 619
Session 5 : 868, 586, 618, 900

Figure. 5: User session from data set.

We assume the access pattern of a certain type of user can be characterized by a certain minimum length of a users transaction, and that the corresponding future access path is not only related to the last accessed URL. Therefore, users with relatively short transactions (e.g. 2-3 accesses per transaction) should be handled in a different way from users with long transactions (e.g. 10-15 accesses per transaction) [24]. In this study, we proposed a case definition design based on the transaction length. User transactions with lengths of less than 3 are removed because it is too short to provide sufficient information for access path prediction [24].

6. Conclusion

Soft computing methodologies, involving fuzzy sets, neural networks, genetic algorithms, rough sets, and their hybridizations, have recently been used to solve data mining problems. They strive to provide approximate solutions at low cost, thereby speeding up the process. A categorization has been provided based on the different soft computing tools and their hybridizations used, the mining function implemented, and the preference criterion selected by the model [25]. A novel approach using independent component analysis for dimension reduction from dynamic trend signals has been presented. However, ICA has been restricted to unsupervised cases [26]. In this paper we have presented soft computing on Web mining by using ICA for feature selection and Rough Fuzzy to handle clustering web user sessions. This approach is useful to cluster web usage access patterns in web log.

7. Acknowledgements

This research was supported in part by grants from Suan Dusit Rajabhat University www.dusit.ac.th at Bangkok, Thailand.

Reference:-

[1] S.K.Pal, V.Talwar and P.Mitra, Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, *IEEE Transactions on neural network*, 13 (5), 2002.

[2] D.Arotaritei, S.Mitra, Web mining: a survey in the fuzzy framework, *Journal of Fuzzy Sets and Systems* 148, 2004, 5-19.

[3] R. F. Li, X.Z Wang, Dimension reduction of process dynamic trends using independent component analysis, *Computers and chemical engineering*, 26 (2002) 467-473.

[4] Editorial, Independent component analysis and beyond, *Journal of Signal Processing*, 84, 2004, 215-216.

[5] M. Dash and H. Liu, Consistency-based search in feature selection. *Journal of Artificial Intelligence*, 151, 2003, 155–176.

[6] T.Wakaki, H.Itakura, and M.Tamura, Rough Set-Aided Feature Selection for Automatic Web-Page Classification, *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, 2004

[7] J.V.Stone, Independent component analysis: an introduction, *TRENDS in Cognitive Sciences*, 6(2), 2002, 59-64.

[8] H.Song, L.Zhong, and B.Han, Structural Damage Detection by Integrating Independent Component Analysis and Support Vector Machine, *ADMA*, LNAI 3584, 2005,670-677.

[9] S. Amari, Natural gradient works effciently in learning, Neural Computing. 10 (1998) 251–276.

[10] H.-H. Yang, S. Amari, Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information, Neural Computing. 9 (1997) 1457–1482.

[11] J.Basak, Weather Data Mining Using Independent Component Analysis, *Journal of Machine Learning Research*, 5, 2004, 239-253.

[12] Z.Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic, Dordrecht, 1991.

[13] S.K.Pal, Soft data mining, computational theory of perceptions, and rough-fuzzy approach, *International Journal of information science*, 163, 2004 5-12.

[14] Z.Pawlak, Rough sets, *International Journal of Information and Computer Science*, 11(5), 1982, 341–356.

[15] A. Petrosino, G. Salvi, Rough Fuzzy set based scale space transforms and their use in image analysis, *International Journal of Approximate Reasoning*, 2005.

[16] D.Vijay Rao, V.V.S. Sarma, A rough-fuzzy approach for retrieval of candidate components for software reuse, *Journal of Pattern Recognition Letter*, 24, 2003, 875-886.

[17] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, *International Journal of General Systems* 17 (2–3), 1990, 191–209.

[18] D. Dubois, H. Prade, *Putting rough sets and fuzzy sets together*, Intelligent Decision Support, Handbook of Applications and Advances of the Rough Set Theory, Kluwer Academic Publishers, Dordrecht, 1992.

[19] M.Sarkar, Rough-fuzzy functions in classification, *Journal of Fuzzy Sets and Systems*, 132, 2002, 353-369.

[20] R. Cooley, P-N. Tan, J. Srivastava, Discovery of Interesting Usage Patterns from Web Data, *In Springer-Verlag LNCS/LNAI series*, 2000.

[21] Y. Fu, K. Sandhu, and M.Y. Shih, Clustering of Web Users Based on Access Patterns, *Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, San Diego: Springer, 1999.

[22] G. Pallis, L. Angelis, and A. Vakali, Model-based cluster analysis for web users sessions, *Springer-Verlag Berlin Heideberg*, 2005, 219-227.

[23] L.Catledge, and J.E. Pitkow, Characterizing Browsing Behaviors on The World Wide Web, *Computer networks and ISDN Systems*, 27(6), 1995.

[24] C. Wong, S. Shiu, and S. Pal, Mining Fuzzy Association Rules for Web Access Case Adaptation, *Proc. of the workshop Programme at the fourth International Conference on Case-Based Reasoning, Harbor Center in Vancouver*, British Columbia, Canada, 2001.

[25] S. Mitra, Data Mining in Soft Computing Framework: A survey, *IEEE Transactions on neural networks*, 13(1), 2002.

[26] S.Akaho, Conditionally independent component analysis for supervised feature extraction, *Journal of Neurocomputing*, 49, 2002, 139-150.

[27] G. Chakraborty, B.Chakraborty, “A Rough-GA Hybrid Algorithm for Rule Extraction from Large Data”, IEEE International conference on Computational Intelligence on Measurement Systems and Applications (CIMSA 2004), 14-16 July, Boston, USA, 2004, pp. 85-90.