A Clinical Heart Disease Decision Supportive Optimized Mining Method for Effective Disease Diagnosis

D. Nalini
Assistant professor,
Department of Computer Science,
Kurinji college of Arts and Science,
Tiruchirappalli, Tamilnadu, India.

R. Periasamy, PhD
Research Advisor,
PG and Research Department of Computer Science,
Nehru Memorial College (Autonomous),
Puthanampatti,Trichy(dt), Tamilnadu, India

ABSTRACT
Healthcare industry collects enormous amounts of healthcare data and required to mine and ascertain hidden information for constructive decision making. In recent years, the computer technology and machine learning methods with data mining approach increase techniques in assisting the doctors for productive decision making related to heart disease and stroke identification at an early stage. The needs to reduce the pattern matching loss by performing pattern matching while effectively improving the heart disease identification at much early stage poses severe challenges to the database community. In this paper a method called Clinical Heart Disease-Decision Supportive Optimized Mining (CHD-DSOM) is presented to overcome the pattern matching loss, and perform perfect pattern matching. The method CHD-DSOM categories to enable decision support with multi-dimensional analysis using Mahalanobis distance measure for obtaining dynamic data table information and typically built to support early stage of stroke identification and levels of heart disease. This helps in identifying the solution for different patterns and therefore reducing the

pattern matching loss. With the objective of identifying the level of heart disease in a very accurate manner, the CHD-DSOM method uses the Iterative Dichotomiser 3 based decision tree model.With Iterative Dichotomiser 3 based decision tree model, the stroke is also identified very easily by starting with the original set S as the root node. The entropy value of every attribute is calculated and is placed in the decision tree. Decision tree are constructed from the higher to lower values to perform easy pattern matching, aiming at reducing the processing time for pattern matching. Experiment is conducted with the Cleveland Clinic Foundation Heart disease data set available from UCI repository using the factors such as pattern matching loss rate, accuracy, processing time for pattern matching, computational cost. Experimental analysis shows that the CHD-DSOM method is able to reduce the processing time for pattern matching by 33.23% and reduce the pattern matching loss rate by 29.67% compared to the state-of-the-art works.

Keywords
Machine Learning, Decision Supportive, Optimized Mining, Pattern Matching, Decision Tree IterativeDichotomiser

1. INTRODUCTION
The application of time series data in medical disease diagnosis has been considered one of the most important types of data available in human society. Pattern preserving for time series data to observe the patterns for disease diagnosis was presented in [1]. Query processing was attained by supporting complex queries using novel anonymization model. In [2] probabilistic sequential alignment model was designed with the objective of improving accuracy using time series data. In [3] patterns were efficiently classified for the application of road networks using pattern based classification.

XML-tree based pattern mining [4] was performed using matching cross framework with the objective of improving effectiveness of the pattern being matched. In [5], rough set technique was applied to extract the feature subset at much early stage. However, the above

said methods did not address the accuracy with which the patterns were mined though were applied in various fields. In the proposed CHD-DSOM method, accuracy was improved using Mahalanobis distance measure.

Real world applications in the areas of artificial intelligence, such as pattern recognition not only require minimizing the database dimensionality, but also good classifiers are required to address the classification problems. In [6], fuzzy rough set technique was applied to solve the classification problems for feature reducing minimizing the time for obtaining the required features. Generalized board count approach [7] served not only as a ranking model for election campaigns that extracted the essential features removing the redundant features but also efficiently ranked the features using multi valued objects. Graph similarity search applied in [8] improved the pruning

power with much less query response time using partition based and branch based bounds for extracting features.

Efficient characterization and prediction was made using first order Markov behavior [9] resulting in the feasibility and effectiveness of the method. Extended sub tree [10] for disease diagnosis resulted in the improvement of runtime efficiency using subtree mapping. But the above said methods lack pattern matching loss rate which is address in CHD-DSOM using equality and inequality set.

The aforementioned papers discussed about the disease diagnosis using different methods in wide area of applications. In this part, a method has designed to Clinical Heart Disease-Decision Supportive Optimized Mining (CHD-DSOM) to overcome the pattern matching loss, and perform perfect pattern matching. The contributions of CHD-DSOM include the following, (i) to reduce the pattern matching loss by applying Mahalanobis distance measure, (ii) to improve the accuracy of heart disease identification using Iterative Dichotomiser 3 based decision tree model and (iii) to reduce the processing time for pattern matching using decision tree model.

2. CONSTRUCTION OF CLINICAL HEART DISEASE-DECISION SUPPORTIVE OPTIMIZED MINING

In this section, Clinical Heart Disease-Decision Supportive Optimized Mining (CHD-DSOM) method was described. It consists of two parts. First parts present a multi-dimensional analysis of the dynamic data table information for early stage of stroke identification and levels of heart disease that captures the distance measure between two data patterns. It is motivated from the Probabilistic Sequence Translation Alignment Model (PSTAM) [1] aimed for time-series classification, PSTAM extend it considerably to deal with heart and stroke disease diagnosis.

In the second part, CHD-DSOM propose an Iterative Dichotomiser 3 based decision tree model that considers decision tree to perform easy matching. This is done by identifying the entropy value for different data patterns using greedy search model.

The figure 1 given below shows the block diagram of CHD-DSOM method. The CHD-DSOM method is divided into two parts. The first part Mahalanobis Distance Measure in CHD-DSOM efficiently identifies the similar and different patterns using set of equality and inequality aiming at reducing the pattern matching loss. The second part Iterative Dichotomiser 3 based decision tree efficiently diagnosis the stroke and heart disease using the entropy value for each original set using greedy search. This in turn identifies the heart disease in an accurate manner minimizing the processing time for pattern matching.

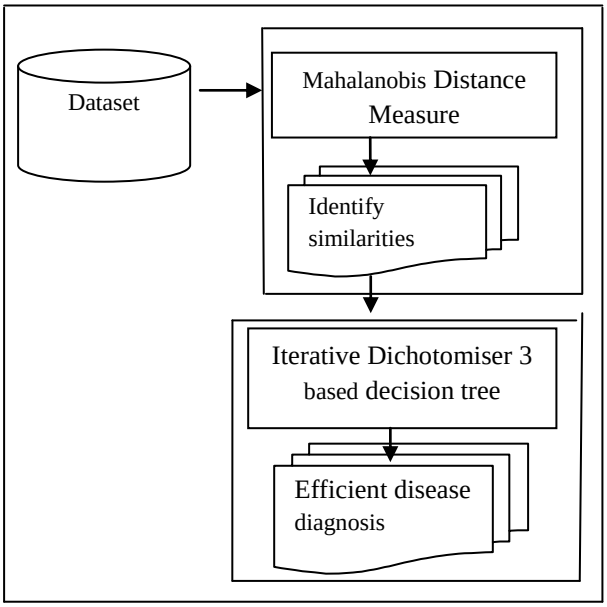

Figure 1 Block diagram of CHD-DSOM method

2.1 Design of Mahalanobis Distance Measure

The first part in the design of CHD-DSOM is to enable decision support with multi-dimensional analysis of the dynamic data table information using Mahalanobis Distance Measure.

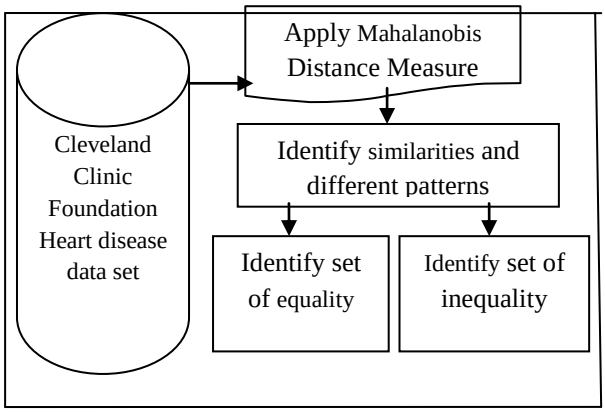

Figure 2 Block diagram of pattern matching using Mahalanobis Distance Measure

Figure 2 shows the block diagram of pattern matching performed using MDM.

The Mahalanobis Distance Measure in CHD-DSOM supports the solution with early stage of stroke identification and levels of heart disease. As shown in figure 2, similar and different patterns are identified from the input Cleveland Clinic Foundation Heart disease dataset. The multi-dimensional analysis helps in identifying which patterns (i.e. patterns observed from stroke and heart disease) are more similar and which patterns are more different. With this similarities and different patterns using Mahalanobis Distance Measure, the rate of pattern matching is improved or reduces the patterns matching loss. With the solution obtained, the CHD-DSOM method identify the distance 'Dis' in such a way that similar patterns are closer to each other than different patterns. The CHD-DSOM method uses Mahalanobis Distance Measure (MDM) to derive the similar or different patterns.

Let us consider a collection of samples '$S = p_1, p_2, .., p_n$' where 'n' is the number of samples in the Cleveland Clinic Foundation Heart disease data set. Here '$p_i \in DV^m$' where 'm' is the number of features in the sample 'S'. Let the set of equality (i.e. belonging to same class) be denoted as given below,

$$Set_E = \{(p_i, p_j), \quad (1)$$

$$where\ p_i, p_j\ belong\ to\ the\ same\ class\}$$

From (1), the set of equality 'Set_E' is formulated where 'p_i, p_j' belong to the same class and have same patterns. Let the set of inequality (i.e. belonging to different class) be denoted as given below

$$Set_{IE} = \{(p_i, p_j), \quad (2)$$

$$where\ p_i, p_j\ do\ not\ belong\ to\ the\ same\ class\}$$

From (2), the set of inequality 'Set_{IE}' is formulated where 'p_i, p_j' does not belong to the same class and have different patterns. Then, the method CHD-DSOM is formulated by an observed set of data patterns as given below

$$DP = \{(p_i, q_i, t_i), \ldots, (p_n, q_n, t_n) \in (P, Q, T) \quad (3)$$

From (3), '$p_i, .., p_n$' represents the 'n' different data patterns and '$q_i, .., q_n$' denotes the binary values '$[0, 1]$' to represent the diagnosis results with '$t_i, .., t_n$' representing the trust value of several patterns. In the CHD-DSOM method, '$p_i, .., p_n$' represents the information of different patients obtained from Cleveland Clinic Foundation Heart disease data

set. Then, the similarity using Mahalanobis Distance Measure is as given below

$$Dis_M(p_i, p_j) = (p_i, p_j)^T - (p_i, p_j) \quad (4)$$

From (4), the distance measured 'Dis_M' between two patterns '(p_i, p_j)' is obtained. Then, the similarity between two data patterns is measured as given below.

$$Sim(p_i, p_j) = Dis_M(p_i, p_j) \quad (5)$$

From (5), the similarity 'Sim' between two data patterns 'p_i, p_j' measures how similar the two data patterns are. The resultant value is assigned with either '$0\ or\ 1$'. If the distance of similarity between '$p_i, p_j\ is\ 0$', then the resultant value is assigned with '1', otherwise the resultant value is assigned with '0'. This in turn reduces the pattern matching loss

Input: Samples '$S = p_1, p_2, .., p_n$',
Output: Reduced pattern matching loss with varied samples provided as input
Step 1: **Begin** Step 2: **For** each sample S Step 3: Measure the set of equality using (1) Step 4: Measure the set of inequality using (2) Step 5: Evaluate the observed set of data patterns using (3) Step 6: Evaluate Mahalanobis Distance Measure using (4) Step 7: Measure the similarity between two patterns using (5) Step 8: **End for** Step 9: **End**

Figure 3 MDM algorithmic description

Figure 3 shows the algorithmic description using MDM measure. As shown in the figure 3, for each sample, to start with, the set of equality and inequality is measured to identify whether the patterns belong to the same class or different class. Next, based on the equality and inequality results, the observed set of data patterns are evaluated that represents the diagnosis results with the base trust values given as threshold. Finally, the Mahalanobis Distance Measure is formulated to find the level of similarity between two data patterns. This in turn supports in minimizing the pattern matching loss.

2.2 Design of Iterative Dichotomiser 3 based decision tree model

The second part in the design of the CHD-DSOM method is to use the Iterative Dichotomiser 3 based decision tree model to identify the level of the heart disease and stroke very accurately and therefore diagnose the disease at an early stage. Let us consider an original set 'R'. The Iterative Dichotomiser 3 based decision tree model efficiently generates a decision tree from the original set 'R'. The objective behind using Iterative Dichotomiser 3 algorithm is that is creates decision tree model for dichotomizing data patterns or efficiently classifies the data patterns until a leaf node is arrived. Figure 4 shows the Iterative Dichotomiser 3 based decision tree model to diagnose stroke and heart disease using Cleveland Heart Disease dataset extracted from UCI repository.

In order to obtain the ideal attribute values, the CHD-DSOM method measures the entropy value. The entropy value for every attribute is evaluated and is placed in the decision tree. The entropy for CHD-DSOM is mathematically formulated as given below.

$$E(R) = -\sum r(p) \log 2\ r(p) \quad (6)$$

From (6), the entropy 'E' of the original set 'R' is obtained using the set of classes 'p' in 'R' and using the ratio 'r' of number of similar data patterns in class 'p' to the overall data patterns in the original set 'R'. Decision tree are constructed from the higher to lower values for the easy way of the pattern matching. The CHD-DSOM method employs the decision tree model so that the path pattern matching is performed with minimal processing time. Figure 4 shows the algorithmic description for ID3 decision tree model using the original set or root node 'R'. As shown in the decision tree model using ID3, for each subset, the entropy value of each attribute is measured. Followed by this, the original set is split into subsets with the aid of the attributes in the decision tree and the entropy value which has the minimum is selected as the continuum attribute. The algorithmic description of ID3 decision tree model is given below.

Input: Original set 'R', set of classes 'p'
Output: Pattern matching with minimal processing time
Step 1: **Begin** Step 2: Let 'R' be the original set Step 3: Let 'R' denotes the root node Step 4: **Repeat** for each subset Step 5: **For** each iteration through unused attribute Step 6: Measure the entropy value of that attribute using (6) Step 7: Select the smallest entropy value of that attribute Step 8: **End for** Step 9: **Until** (all subset is processed) Step 10: **End**

Figure 4 Algorithm for ID3 decision tree model

The continuum attribute is further used to construct the decision tree. Similar process is repeated with all the other attributes. In this way, attributes that are not selected are only considered to obtain the data patterns, minimizing the processing time for pattern matching.

3. EXPERIMENTAL SETUP

Performance experiments of CHD-DSOM method are conducted with various conditions using JAVA platform. Cleveland Heart disease dataset from UCI repository is taken for the experimental work to categorize the rich features. Cleveland Heart disease dataset contains 76 attributes, but experiments are conducted using only 14 of them. Experiments with the Cleveland database using CHD-DSOM method concentrated on attempting to differentiate between the presence of disease by the values 1, 2, 3, 4 and the absence of disease by the value 0. The dataset description of Cleveland Heart disease dataset from UCI repository is provided

4. SIMULATION RESULTS AND DISCUSSION

The performance of Clinical Heart Disease-Decision Supportive Optimized Mining (CHD-DSOM) method is compared with the existing Pattern Preserving Anonymization for Time Series Data (PPA-TSD) [1] and Probabilistic

Sequence Translation Alignment Model (PSTAM) [1]. The performance is evaluated according to the following metrics.

4.1 Impact of Pattern matching loss rate

The pattern matching loss rate measures the data pattern match loss rate with respect to different patients. The pattern match loss rate is the ratio of difference between the number of patients and the number of data pattern match rate. It is measured in terms of percentage (%) and is formulated as given below.

$$PMLR = \left(\frac{n - DP_{matc\ h}}{n}\right) * 100 \qquad (7)$$

From (7), the pattern matching loss rate '$PMLR$' is obtained according to the number of patients 'n' provided as input using Cleveland heart disease dataset. Lower the pattern matching loss rate, more efficient the method is said to be.

Figure 5 shows the result of pattern matching loss rate versus the varying number of patients. To better perceive the efficacy of the proposed CHD-DSOM method substantial experimental results are illustrated in Figure 5. The CHD-DSOM method is compared against the existing PPA-TSD [1] and PSTAM [2].

Figure 5, the proposed CHD-DSOM method performs relatively well when compared to two other methods PPA-TSD [1] and PSTAM [2]. The CDH-DSOM method had better changes using the extensive Mahalanobis distance measure

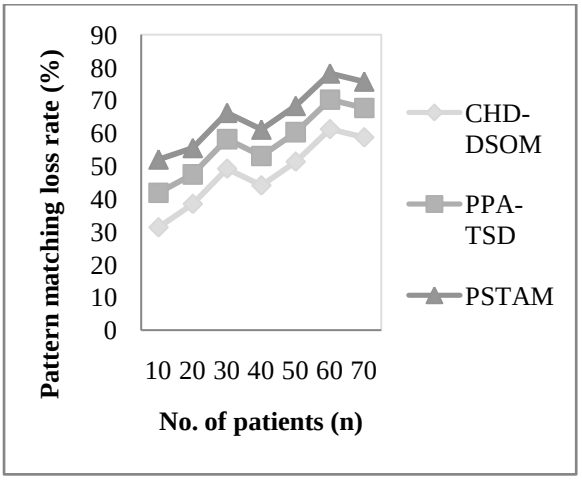

Figure 5 Impact of pattern matching loss rate with respect to patients

This is because in order to obtain the distance measure using multi-dimensional analysis, the stroke identification and level of heart disease is obtained based on the identified set of equality and inequality. This in turn decreases the pattern matching loss rate by 20.39% compared to PPA-TSD and 38.95% compared to PSTAM.

4.1 Impact of accuracy

Accuracy is used as the statistical measure of how well the diagnosis of a disease is made. Accuracy therefore measures the correct diagnosis of disease with respect to the number of patients. It is mathematically formulated as given below

$$A = \left(\frac{\text{Correct diagnosis of disease}}{n}\right) * 100 \qquad (8)$$

From (8), the accuracy 'A' is evaluated in term of percentage (%). Higher, the accuracy, more efficient the method is said to be. The targeting results of identifying the disease accuracy using CHD-DSOM method is compared with two state-of-the-art methods [1], [2] in figure 6 is presented for visual comparison based on the number of patients.

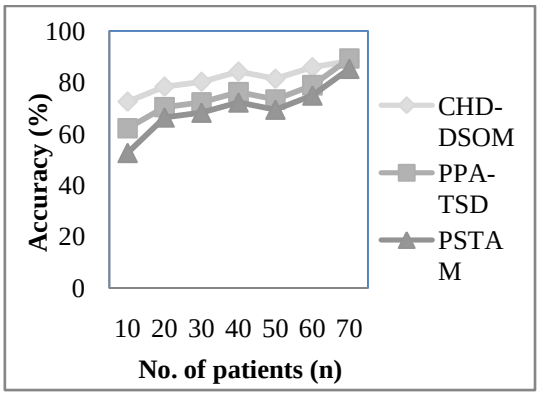

Figure 6 Impact of accuracy

This CHD-DSOM method differs from the PPA-TSD [1] and PSTAM [2] in that incorporated Iterative Dichotomiser 3 based decision tree model. By applying an Iterative Dichotomiser 3 based decision tree model, the entropy value of every attribute is measured and is placed in the decision tree. This dichotomized principle in CHD-DSOM method efficiently classifies the data patterns until a leaf node is arrived helps in the disease pattern and its level. Therefore the accuracy of disease being identified is improved by 8.72% compared to PPA-TSD and 14.77% compared to PSTAM.

4.2 Impact of processing time for pattern matching

The processing time for pattern matching is the product of data patterns to the time taken to process the data patterns. The processing time for pattern matching is measured in terms of milliseconds (ms) and is formulated as given below.

$$PT_{pm} = DP * Time\ (DP) \qquad (9)$$

From (9), the processing time for pattern matching is obtained for different set of data patterns. Lower the processing time for pattern matching, more efficient the method is said to be. Figure 7 given below shows the processing time for pattern matching for CHD-DSOM method, PPA-TSD [1] and PSTAM [2] versus increasing number of data patterns in the range of 5 to 35. The processing time improvement returned over PPA-TSD and PSTAM increase gradually as the number of data patterns obtained gets increased though not linear because of the changes observed in the fasting blood sugar level.

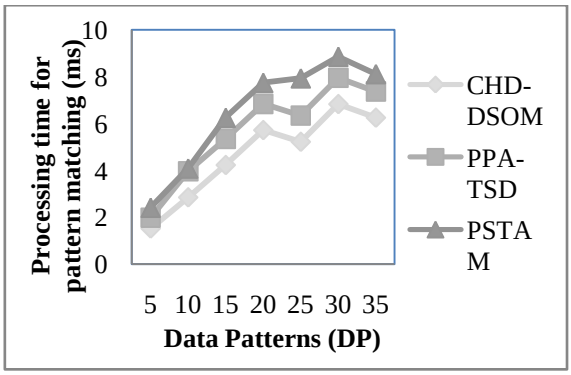

Figure 7 Impact of processing time for pattern matching

From figure 7, it is illustrative that the processing time for pattern matching is reduced using the proposed method CHD-DSOM. For example when the number of data patterns is 30, the percentage improvement of CHD-DSOM method compared to PPA-TSD is 19.40 percent and compared to PSTAM is 35.13 percent respectively. This is because by constructing the decision tree from higher to lower values, from the original set, based on the entropy value, the processing time is reduced by 24.20% compared to PPA-TSD. Similarly, the processing time for pattern matching is reducing in the CHD-DSOM method using ID3 decision tree algorithm that efficiently detects the correct data pattern structure by 42.16% compared to PSTAM.

4.3 Impact of computational cost for disease diagnosis

The computational cost for disease diagnosis for CHD-DSOM method is elaborated and comparison made with two other methods PPA-TSD and PSTAM respectively. With consider the method with 70 patients in the age group 30 – 59 for experimental purpose using JAVA.

Figure 8 shows the disease diagnosis rate taken to perform computational cost for disease diagnosis with respect to 70 patients. As shown in the figure with the increase in the number of patients and the data patterns being observed for several patients, the computational cost is increased using all the methods, but comparatively lesser using the proposed CHD-DSOM method.

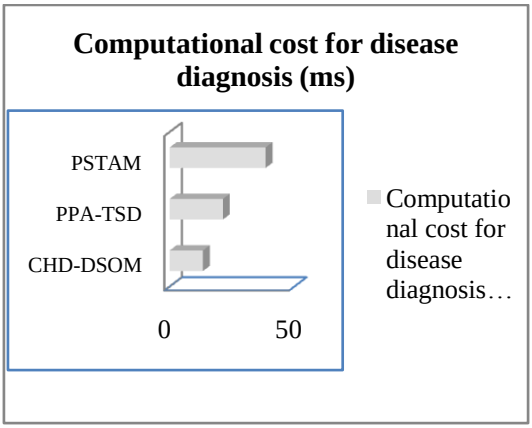

Figure 8 Impact of computational cost for disease diagnosis

The computational cost for disease diagnosis using CHD-DSOM method is reduced owing to the fact that the proposed method uses Mahalanobis Distance Measure. With this Mahalanobis Distance Measure, the similarity between two data patterns is measured in an efficient manner. Moreover, by applying the decision tree model, efficient classification of data patterns is performed through Iterative Dichotomiser 3 algorithm. With this, the computational cost is reduced in CHD-DSOM by 59.85% compared to PPA-TSD and 80.17% compared to PSTAM.

5. CONCLUSION

In this work, an effective Clinical Heart Disease-Decision Supportive Optimized Mining (CHD-DSOM) method is presented. The method reduces the computational cost for disease diagnosis with minimum processing time for pattern matching and therefore provides accuracy of disease diagnosis rate for stroke and heart disease. The goal of medical image disease diagnosis is to improve the pattern matching accuracy using the training and test images obtained from Cleveland Heart disease dataset which significantly contribute to the relevance. To do this, the first method designed Mahalanobis Distance Measure to determine the set of equality and set of inequality of patient information to reduce the pattern matching loss rate. Then, based on this measure, CHD-DSOM proposed an Iterative Dichotomiser 3 based decision tree model that efficiently generates a decision tree until a leaf node is arrived reducing the computational cost for disease diagnosis. With the leaf node generated and measured entropy value for each attribute, similar data patterns are observed with the objective of improving the accuracy. In addition, an ID3 decision tree algorithm based on the entropy value and distance measure ensures efficient pattern matching for varied training and test images. Through the experiments using Cleveland Clinic Foundation Heart disease data set from UCI repository that observed medical image diagnosis provided more accurate results compared to existing methods. The results show that CHD-DSOM method offers better performance with an improvement of matching of disease being diagnosed by 11.74% and reduces the computational cost by 70.01% compared to PPA-TSD and PSTAM respectively.

6. REFERENCES

[1] Lidan Shou, Xuan Shang, Ke Chen, Gang Chen and Chao Zhang, "Supporting Pattern-Preserving Anonymization for Time-Series Data", IEEE Transactions on Knowledge and Data Engineering, Volume 25, Issue 4, April 2013, Pages 877 – 892.

[2] Minyoung Kim, "Probabilistic Sequence Translation-Alignment Model for Time-Series Classification", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 2, February 2014, Pages 426 – 437.

[3] Jae-Gil Lee, Jiawei Han, Xiaolei Li and Hong Cheng, "Mining Discriminative Patterns for Classifying Trajectories on Road Networks", IEEE Transactions on Knowledge and Data Engineering, Volume 23, Issue 5, May 2011, Pages 713 – 726.

[4] Jiaheng Lu, Tok Wang Ling, Zhifeng Bao and Chen Wang, "Extended XML Tree Pattern Matching: Theories and Algorithms", IEEE Transactions on Knowledge and Data Engineering, Volume 23, Issue 3, March 2011, Pages 402 – 416.

[5] Jiye Liang, Feng Wang, Chuangyin Dang and Yuhua Qian, "A Group Incremental Approach to Feature Selection Applying Rough Set Technique", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 2, February 2014, Pages 294 – 308.

[6] Suyun Zhao, Eric C.C. Tsang, Degang Chen and XiZhao Wang, "Building a Rule-Based Classifier—A Fuzzy-Rough Set Approach", IEEE Transactions on Knowledge and Data Engineering, Volume 22, Issue 5, May 2010, Pages 624 – 638.

[7] Ying Zhang, Wenjie Zhang, Jian Pei, Xuemin Lin, Qianlu Lin and Aiping Li, "Consensus-Based Ranking of Multivalued Objects: A Generalized Borda Count Approach", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 1, January 2014, Pages 83 – 96.

[8] Weiguo Zheng, Lei Zou, Xiang Lian, Dong Wang and Dongyan Zhao, "Efficient Graph Similarity Search Over Large Graph Databases", IEEE Transactions on

Knowledge and Data Engineering, Volume 27, Issue 4, April 2015, Pages 964 – 978.

[9] Wenjing Zhang and Xin Feng, "Event Characterization and Prediction Based on Temporal Patterns in Dynamic Data System", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 1, January 2014, Pages 144 – 156.

[10] Ali Shahbazi and James Miller, "Extended Subtree: A New Similarity Function for Tree Structured Data", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 4, April 2014, Pages 864 – 877.