An Efficient Ensemble Sequence Classifier

I-Hui Li[1], I-En Liao[2*], Jin-Han Lin[2], Jyun-Yao Huang[2]
[1] Ling Tung University, Taichung, Taiwan.
[2] National Chung Hsing University, Taichung, Taiwan.

* Corresponding author. Tel.: +886-4-22854146; email: ieliao@nchu.edu.tw
Manuscript submitted September 29, 2015; accepted December 3, 2015.
doi: 10.17706/jsw.11.2.133-147

Abstract: The techniques of classification are through learning historical data to help people to predict the class label of data, and they have been applied to solve many problems. In the real world, there exists many sequence data, such as genome sequences, those should be learned and analyzed for predicting class labels. The traditional classification methods are unsuitable for sequence data. This study proposes an Ensemble Sequence Classifier (ESC). The ESC consists of two stages. The first stage generates a Sequence Classifier based on Pattern Coverage Rate (SC-PCR) in two phases. The first phase mines sequential patterns and builds the features of each class, whereas the second phase classifies sequences based on class scores using a pattern coverage rate. The second stage creates an ensemble classifier by some classifiers built from the first stage, to improve the prediction accuracy. The experimental results confirm that the SC-PCR and ESC schemes achieve high classification accuracies for both synthetic and medical sequence datasets, even when the training set contained only a limited number of sequential patterns. The average and worst accuracies of SC-PCR are 95.8% and 80.3%, respectively. The average accuracy of ESC is 96.97%, and the worst accuracy is 87%.

Key words: Classification, ensemble classifiers, ensemble sequence classifiers, sequence classification.

1. Introduction

Classification is an essential technique to help people for exploring the hidden knowledge within large-scale datasets and then using this knowledge to make informed decisions [1]. In recent years, classification methods are widely used in biomedical research and applications. However, in many cases, the data within real-world datasets are ordered in accordance with their timestamps, e.g., customer consumption records, patient treatment records, biomedical data [2], and so on. For such data sequences, traditional classification methods yield a poor prediction accuracy since the data structure of sequence is more complex than that of data processed in traditional classification. The literature contains various studies in which specifically-designed sequence classifiers have been applied to the processing of sequential data patterns. For example, Tuzun and Dalmau [3] used a sequence data analysis technique to diagnose limbic encephalitis. Bazinet and Cummings [4] evaluated the performance of several sequence classification schemes in taxonomically or functionally classifying DNA sequence fragments. Kalitzin *et al*. [5] proposed a novel remote-sensing paradigm for detection of a particular class of motor seizures in Video Sequences. Stavri and Michie [6] surveyed current classification systems in behavioural science from the natural, medical and social sciences. Consequently, developing more effective techniques for analyzing data sequences is essential in the data science domain.

Accordingly, this study proposes an enhanced-performance Ensemble Sequence Classifier (ESC) for biomedical sequence classification purposes. The ESC scheme comprises two stages. The first stage constructs a Sequence Classifier based on Pattern Coverage Rate (SC-PCR) scheme, whereas the second stage uses several classifiers built in the first stage to construct an ensemble classifier with an improved prediction accuracy. The experimental results show that the proposed SC-PCR and ESC schemes achieve high prediction accuracies even for datasets with only a limited number of sequential patterns. The main contributions of this study can be summarized as follows:

1) The construction of an Ensemble Sequence Classifier (ESC) integrating sequential pattern mining and a classification architecture.
2) A SC-PCR classifier based on a sophisticated evaluation formula comprising four separate items of information, namely the length rate, the support weight, the similarity degree, and the pattern coverage rate (The details are shown in Section 3.2.2). Such an approach enables a robust classification performance even if the data are skewed.
3) The use of an ensemble classifier concept to improve the prediction accuracy.

The remainder of this paper is organized as follows. Section 2 reviews previous proposals for sequence classification methods. Section 3 describes the system architectures and detailed methods of the proposed SC-PCR and ESC schemes. Section 4 presents and discusses the experimental results. Finally, Section 5 summarizes the major contributions of the present study.

2. Related Work

Tseng and Lee [7] proposed two methods for creating a sequence classification model. The first method was designated as CBS-ALL (i.e., Classify-By-Sequence ALL) and was designed to find sequential pattern rules, referred to as CSP (Candidate Sequential Pattern) rules, from the entire dataset by means of class support and transaction support calculation. The CSP rules were then used to construct a classifier, in which the scores of the unknown sequences were calculated and then compared with stored score values in order to predict the probable class of the sequence. The second method, designated as CBS-CLASS, was designed to find the CSP rules from each class rather than from the entire dataset. The experimental results showed that CBS-CLASS achieved higher prediction accuracy than CBS-ALL. However, it has a poor robustness toward skewed and / or missing data. Moreover, in trimming the CSP rules, CBS-CLASS was required to scan the data incessantly, and therefore has a high computational cost and a low efficiency. Accordingly, in a later study, Tseng and Lee [8] proposed a third sequence classification scheme, referred to as CBS (Classify-By-Sequence). CBS combines sequential pattern mining with a mathematical induction probability-based approach. CBS was similar to CBS-ALL, but used a CSP-Miner technique to prune the CSP-tree in order to generate the CSP rules. CBS had good accuracy and stability and could reduce the impact of the minimum support. However, as with CBS-CLASS, CBS needed to scan the data incessantly. Furthermore, several of the CBS parameters must be set in advance, which is problematic in practice.

Exarchos *et al*. [9] proposed a sequence classification model comprised of two phases. In the first phase, referred to as CBS_CLASS, cSPADE [10] was used to mine the sequential patterns associated with each class. In the testing stage, the model computed the score of each sequence and then compared this score with that of each sequential pattern associated with each class by means of a score matrix. Finally, the sequence was then assigned to the class with the highest score. However, the prediction result may not be ideal, therefore, in the second phase of the proposed model, optimization software [11] was used to assign appropriate weights to each sequential pattern and class in order to improve the classification accuracy. The experimental results confirmed that the optimization procedure yielded a significant improvement in the classification performance. However, the optimization process requires a time-consuming and complex

iterative procedure. As a result, the proposed scheme is poorly suited to the real-time processing of modern large-scale, complex datasets.

Chen and Chen [12] proposed a classification algorithm designated as FESP (Frequent Emerging Sequence Patterns, in which new support and growth rate of support measures were used to find frequently occurring patterns in DNA sequence databases. Notably, these patterns not only retained the information provided by the order of the bases in the gene sequences, but also captured the interactions among these bases. Having identified the frequently emerging sequence patterns, these patterns were then used to construct classification rules with which to classify new DNA sequences.

Our previous study [13] proposed a sequence classification method designated as SCM (Sequential Pattern Length Based Sequence Classifier Model) based on the model proposed by Exarchos *et al*. [9], but without the optimization process. SCM consisted of two phases. In the first phase, PrefixSpan [14] was used to identify the sequential patterns in the dataset and to delete repeated sequential patterns among different classes. In the second phase, the score matrices of each class were calculated and the sequence was assigned to the class with the highest score. SCM used an evaluation formula based on three specific items of information, namely the pattern length, the support count and the class weight. The experimental results showed that SCM achieves a good classification accuracy despite the absence of an optimization phase. For example, the best, average and worst classification accuracies were found to be 95%, 88% and 66.7%, respectively. However, the act of deleting the repeated sequential patterns among different classes reduces the features and / or information available for each class and therefore increases the likelihood of classification errors; particularly when some (or all) of the classes have only a small number of sequential patterns.

An enhanced version of the SCM scheme designated as Pattern Coverage Rate-based Sequence Classification Model (PCRSCM) [15] was proposed later. In the proposed model, the sequential patterns were mined to determine the characteristics of each class, and the pattern coverage rates and class scores were then calculated in order to predict the class of each sequence. Notably, the sequence classification process was performed using an estimation method based on a Pattern Coverage Rate (PCR) metric. The experimental results showed that PCRSCM was not only faster than SCM due to its use of the PCR estimation method, but was also better able to deal with skewed data through its use of a more sophisticated evaluation formula comprising four items of in-formation: namely a similarity comparison measure, the pattern length, the pattern support rate, and the pattern rate of each class. The experimental results showed that PCRSCM achieved an excellent prediction performance for both synthetic and real sequence datasets and was robust toward various parameters, such as the average trans-action length, the average sequence length, and so on. Overall, the maximum, mean and minimum classification accuracies of the proposed scheme were found to be 97%, 92% and 70% respectively.

3. Ensemble Sequence Classifier

Assume that a sequence is composed of one or more elements. For example, in sequence <{B, C, D} {A, D} {B}>, each {} is an element, {B} is a single item within the element, and {B, C, D} and {A, D} are itemsets. Furthermore, the sequential order of the elements represents the occurrence order, and each sequence is associated with a particular class. The problem considered in the present study is that of analyzing the relationships among the sequence data so as to identify the characteristics of each class, and to construct a classifier for predicting the class label of each sequence.

Problem Definition: Assume that there exists a sequence dataset SD and that L is a set of class labels. Assume further that each sequence $S \in SD$ is associated with a particular class label, $cl \in L$. The sequence classification task involves building a sequence classifier C; namely a function which maps any sequence S

to a class label $cl \in L$, i.e., $C:S->cl$; $cl \in L$. In the second stage of the classification model proposed in the present study, the mapping process is performed using an ensemble classifier, written as $ESC(C_1, ..., C_n):S->cl$; $cl \in L$, where n is the number of classifiers.

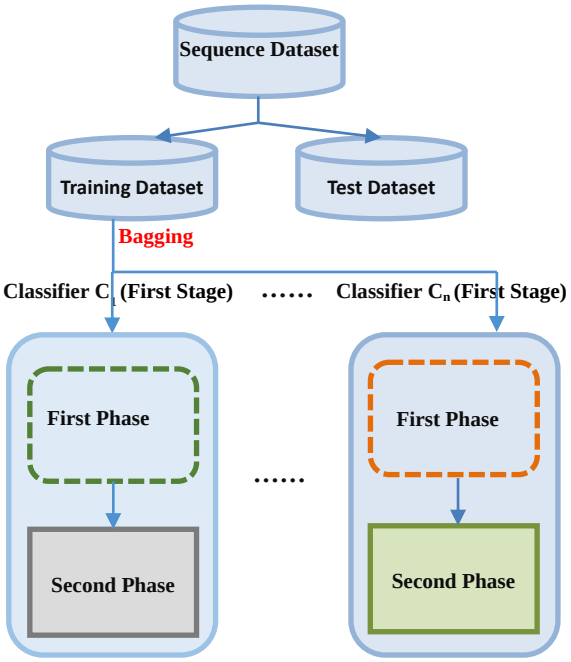

Fig. 1. System architecture of first stage of ESC.

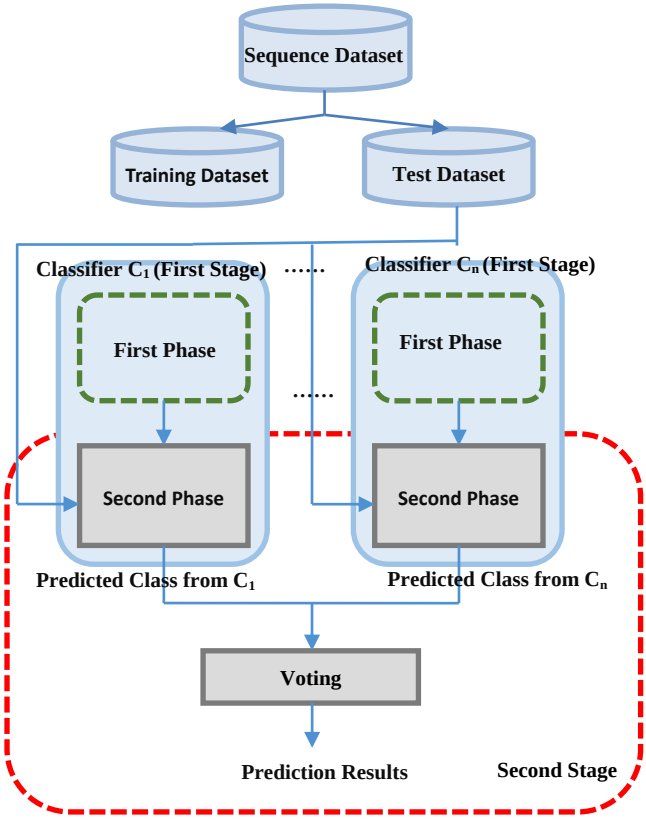

Fig. 3. Ensemble sequence classification.

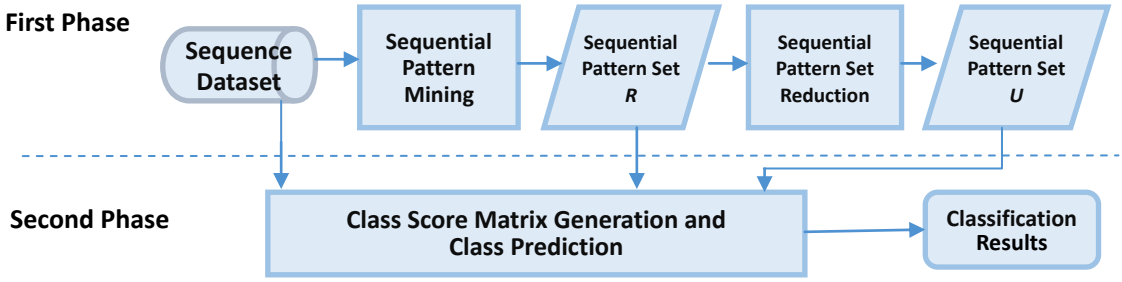

Fig. 2. Architecture of sequence classifier based on pattern coverage rate.

3.1. System Architecture

The Ensemble Sequence Classifier (ESC) proposed in this study is based on the simple two-stage architecture proposed by Exarchos *et al.* [9]. As shown in Fig. 1, the sequence dataset is divided into a training dataset and a test dataset. ESC consists of two stages. The first stage generates several Sequence Classifiers based on the Pattern Coverage Rate (SC-PCR) using the sequences in the bagged training dataset. As shown, the first stage comprises two phases. In the first phase, the classifier mines the sequential patterns and constructs the features of each class. In the second phase, the sequences are classified in accordance with their class scores using a Pattern Coverage Rate (PCR) approach. In the second stage of ESC, an ensemble classifier is built using several classifiers constructed in the first stage in order to further

improve the prediction accuracy.

3.2. Sequence Classifier Based on Pattern Coverage Rate (SC-PCR)

Fig. 2 illustrates the architecture of the proposed SC-PCR classifier. As described above, SC-PCR consists of two phases. In the first phase, SC-PCR mines the sequential patterns of each class (i.e., set R in Fig. 2) and then deletes any sequential patterns which are repeated among the different classes to produce a reduced sequential pattern set (designated as set U in Fig. 2). In the second phase, a class score matrix is generated in order to predict the class label of each sequence.

3.2.1. Sequential pattern mining

Sequential pattern mining provides the ability to detect frequently occurring patterns and to detect (or predict) anomalies in the data [16]. A sequence s can be denoted as $<s_1\ s_2\ \dots\ s_n>$ is an ordered set of n elements [17]. For example, a frequent sequential pattern $<(ab)(c)>$ indicates that items a and b often occur at the same time, and are then frequently followed by item c. In the first phase of SC-PCR, the sequential patterns are mined for each class in order to identify the class features. Suppose there exists a sequence dataset SD, in which each sequence belongs to a particular class i, i = 1, 2, ..., N, where N is the total number of classes. Let all of the sequence data belonging to the same class be grouped as N sub-datasets, i.e., $SD=\{SD_1 \cup SD_2 \cup SD_3 \cup \dots \cup SD_N\}$. Given a minimum support, the PrefixSpan algorithm is then used to find the sequential pattern set R, i.e., $R=\{R_1 \cup R_2 \cup R_3 \cup \dots \cup R_N\}$, where R_i represents the sequential patterns belonging to class i. Assume that $R_i=\{P_{i1}, P_{i2}, P_{i3}, ..., P_{imi}\}$, where P_{ik} is the k^{th} sequential pattern in class i, $1 \leqq k \leqq m_i$. That is, R_i has m_i sequential patterns, i.e., $|R_i|=m_i$. Having constructed the sequential pattern set R, SC-PCR produces the reduced sequential pattern set U by deleting any sequential patterns in R which are repeated among the different classes. Let the reduced sequential pattern set be denoted as $U=\{U_1 \cup U_2 \cup U_3 \cup \dots \cup U_N\}$, where $U_1 \subseteq R_1$, $U_2 \subseteq R_2$, ..., $U_i \subseteq R_i$.

3.2.2. Class score matrix generation and class prediction

The second phase of SC-PCR uses a PCR-based approach to improve the efficiency of the prediction process. The details are presented in the following.

1) **Pattern Coverage Rate Calculation:** Assume that there exist g sequence data whose classes are to be predicted. SC-PCR compares each sequence S_j with U and computes the corresponding PCR, i.e., $1 \leqq j \leqq g$. In other words, the PCR is defined as the proportion of each sequential pattern in U_i which is a sub-sequence within every sequence S_j. For example, suppose that there exist sequential patterns $P_1=<(c)(k)>$ and $P_2=<(cd)(k)>$, and sequences $S_1=<(a)(cdf)(k)>$ and $S_2=<(a)(c)(d)(k)>$. Since P_1 and P_2 are both sub-sequences of S_1, the sub-sequence of S_2 is only P_1. Consequently, the pattern coverage rate of S_1 is equal to 100%, while that of S_2 is equal to 50%.

Having calculated the PCR between S_j and U_i, the coverage rate of each sequence S_j in every class, i.e., T_{j1}, T_{j2}, ..., T_{jN} can be found.

2) **Similarity Degree Calculation:** The sequential pattern similarity degree calculation performed in SC-PCR utilizes the **S2MP** (Similarity Measure for Sequential Patterns) method proposed by Saneifar *et al*. [18]. Notably, S2MP not only takes into consideration the levels and locations of the itemsets in the two sequences being compared, but also considers the levels of each item in the two itemsets.

3) **Pattern Score Matrix Generation:** SC-PCR generates a **P**attern **S**core **M**atrix (*PSM*) for each class, in which the rows of the matrix represent the sequential patterns belonging to the class; while the columns represent the sequence data. Specifically, the values stored in the matrix are the scores where each sequence S_j corresponds to every sequential pattern. In calculating the pattern scores, SC-PCR considers four different items of information, namely the length rate, the support weight, the similarity degree, and the pattern coverage rate. Let P_S_{ikj} be the pattern score of the j^{th} sequence of

the k^{th} sequential pattern in the i^{th} class, and there are m_i sequential patterns in class i, i.e., $1 \leqq k \leqq m_i$. Furthermore, let P_S_{ikj} be computed as

$$P\_S_{ikj} = P_{ik}\_len\_rate \times sup\_wg_{ik} \times SimDeg_{ikj} \times T_{ji}$$
$$k = 1,2, \dots, mi$$
$$P_{ik}: The\ k^{th}\ sequential\ pattern\ in\ i^{th}\ class$$
$$P_{ik}\_len\_rate: P_{ik}'s\ length/The\ max\ length\ of\ sequential\ pattern\ in\ i^{th}\ class$$
$$sup\_wg_{ik}: P_{ik}'s\ support/\sum_{k=1}^{mi} P_{ik}'s\ support \quad (1)$$
$$SimDeg_{ikj}: The\ similarity\ Degree\ between\ P_{ik}\ and\ S_j$$
$$T_{ji}: The\ pattern\ coverage\ rate\ of\ S_j\ in\ i^{th}\ class$$

4) **Class Score Matrix Generation:** Having calculated the *PSM* for each class, SC-PCR produces a Class Score Matrix (*CSM)* [9], in which the rows of the matrix represent the classes, while the columns represent the **sequences**. The values in the matrix indicate the score for each sequence S_j for each class (i.e., the degree of belonging of each sequence S_j to each class). More specifically, the class score C_S_{ij} is the score of the j^{th} sequence with regard to the i^{th} class, i.e.,

$$C\_S_{ij} = \sum_{k=1}^{mi} P\_S_{ikj} \quad (2)$$

5) **Class Prediction:** In accordance with the values stored in the *CSM*, SC-PCR predicts the class of each sequence as follows:

$$predicted_{class(S_j)} = arg_i \max\left\{C_{S_{ij}}\right\} \quad (3)$$

In the event that the number of maximum C_S_{ij} for a sequence (j^{th}) is greater than one, SC-PCR simply picks one of the corresponding classes at random.

3.3. Ensemble Sequence Classifier (ESC)

Ensemble classification is a technique in which multiple classifiers are combined in order to obtain an enhanced prediction accuracy. A good ensemble classifier with an equal voting mechanism is able to both reduce the biased prediction of risk (from each single classifier) and to improve the forecasting accuracy of the data categories. In the present study, the individual SC-PCR classifiers are trained using the Bagging method proposed by Breiman [19]. Bagging is an approach for increasing the stability or accuracy of a classifier. When using the Bagging method for ensemble classification, each classifier generates a predicted class for every sequence in the test dataset using the method described above, and a majority voting process is then applied to assign a final class to each sequence, as shown in Fig. 3. In the present study, each classifier in ESC is assumed to have the same weight, and thus the number of ensemble classifiers is set as odd in order to ensure a unique prediction outcome.

Bagging uses the take-out back method [19]. In each iteration, the Bagging method uniformly samples data from the training dataset and uses this data to replace the original data; thereby generating new, resampled training datasets. In these newly-built training datasets, some of the original data may be repeated many times, while some of the data may no longer occur. Thus, the Bagging method generates a training dataset with diverse characteristics for each individual classifier. By training each classifier with the characteristics of different datasets, ESC can train the differences of each classifier. Moreover, each classifier has its own innate degree of accuracy as a result of the proposed SC-PCR approach. Consequently,

integrating multiple classifiers in order to obtain the final decision-making outcome yields a significant improvement in the prediction accuracy.

4. Experiments and Results

The prediction accuracies of SC-PCR and ESC were evaluated as follows:

$$accuracy = (Number\ of\ correctly\ predicted\ sequences)/(Number\ of\ sequences) \quad (4)$$

The performance of the SCM [13] and PCRSCM [15] classification methods proposed by the current group in previous studies was evaluated using the same dataset as that considered in [9]. The corresponding results are shown in Figs. 4 and 5 [15]. Accordingly, in the present study, the prediction performance of the SC-PCR and ESC schemes was compared with that of SCM and PCRSCM. To ensure a robust investigation, the comparison process was performed using both synthetic datasets and biomedical sequence datasets. Moreover, the classification accuracies of SC-PCR and ESC were also evaluated for the HS3D (Homo Sapiens Splice Sites Dataset) DNA sequence dataset [20], and compared with that of the FESP scheme [12].

4.1. Synthetic Datasets

The prediction performance of the SC-PCR and ESC classifiers was evaluated using synthetic datasets obtained by adjusting the values of the notations in the IBM Quest Synthetic Data Generator [21] (see Table 1). Initially, four experiments were conducted using datasets produced by modifying just three of the notations, i.e., T, S and N. In every case, the sequences in the synthetic dataset were associated with just two classes. Each experiment involved 30000 training data and 7500 test data. Moreover, the number of ensemble classifiers was set as 5 in every case. Two further experiments were then performed using synthetic datasets with different numbers of classes and different distributions of the sequences over these classes. The detailed experimental designs and results are presented in the following.

Table 1. Notations in IBM Quest Synthetic Data Generator.

	number of thousands of customers (sequences) within dataset
	average transaction (itemset) length
	average sequence length (i.e., average number of itemsets)
	number of distinct items

Experiment (1)

The first experiment was performed using a synthetic dataset designated as C30T2S10N50, and was repeated for four different values of the minimum support in a class, namely 0.4, 0.5, 0.6 and 0.7. Fig. 6 presents the corresponding prediction accuracies of the SCM, PCRSCM, SC-PCR and ESC schemes, respectively. Since the average itemset length is equal to just 2 (i.e., T=2), the number of mined sequential patterns is very small in every case. However, it is seen that for each of the considered support values, SC-PCR and ESC still yield an accurate prediction of the sequence class. From inspection, the maximum, mean and minimum accuracies of the SC-PCR scheme are 97.13%, 91.6% and 80.3%, respectively. Furthermore, the maximum, mean and minimum accuracies of the ESC scheme are 97.2%, 92% and 87%, respectively.

Experiment (2)

The predictive ability of the proposed schemes given a longer average itemset length (i.e., T=4) was evaluated using a synthetic dataset designated as C30T4S10N50. As in the first experiment, four different values of the minimum support in a class were considered, namely 0.4, 0.5, 0.6 and 0.7. The corresponding results are presented in Fig. 7. It is observed that for the SCM and PCRSCM schemes, the prediction

accuracies fall significantly to 83% and 86%, respectively, given a minimum support in a class equal to 0.6. This is because the simple pattern score calculation of these two methods. However, the corresponding accuracies of the SC-PCR and ESC schemes reduce only very slightly. Overall, the maximum, mean and minimum accuracies of the SC-PCR scheme and ESC scheme are 99.35%, 98.7% and 97.87%, respectively, and 99.5%, 99.1% and 98.7%, respectively.

Fig. 8 shows the prediction results obtained by the four schemes for a synthetic dataset with an average itemset length of T=10. There are obvious differences between these four methods, for the distinct pattern score calculation and ensemble scheme. For the SC-PCR scheme, the maximum, mean and minimum accuracies are 98.8%, 98.4% and 97.87%, respectively, while for the ESC scheme, the corresponding accuracies are 99.8%, 99.8% and 99.7%, respectively.

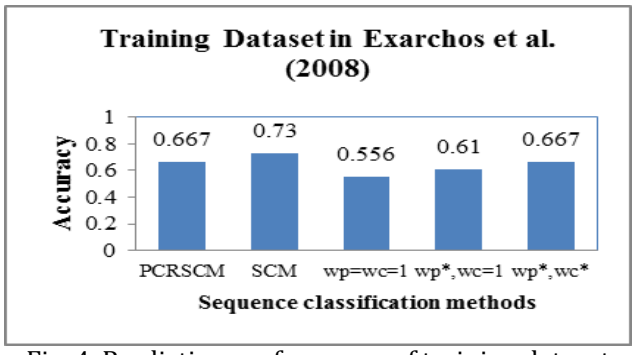

Fig. 4. Prediction performance of training dataset with different methods [15].

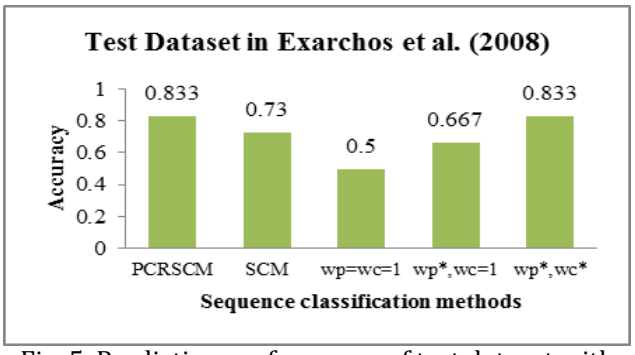

Fig. 5. Prediction performance of test dataset with different methods [15].

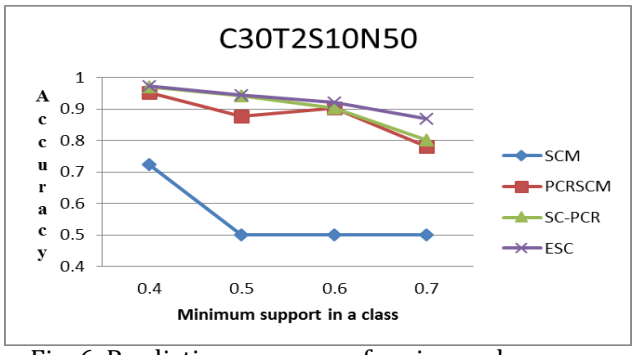

Fig. 6. Prediction accuracy of various schemes as function of minimum support in a class.

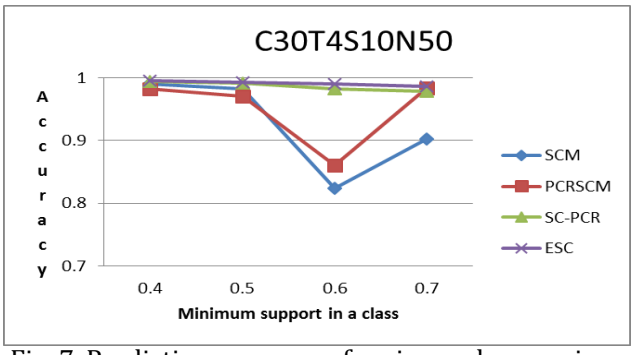

Fig. 7. Prediction accuracy of various schemes given T=4.

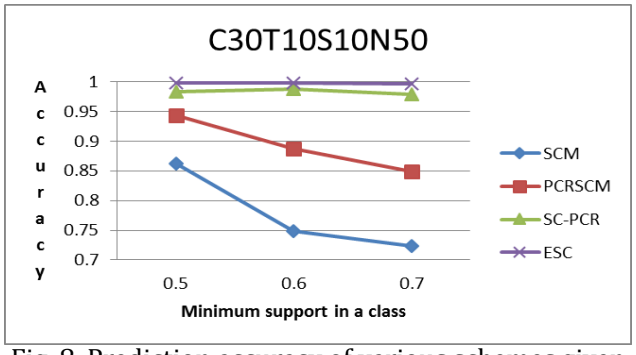

Fig. 8. Prediction accuracy of various schemes given T=10.

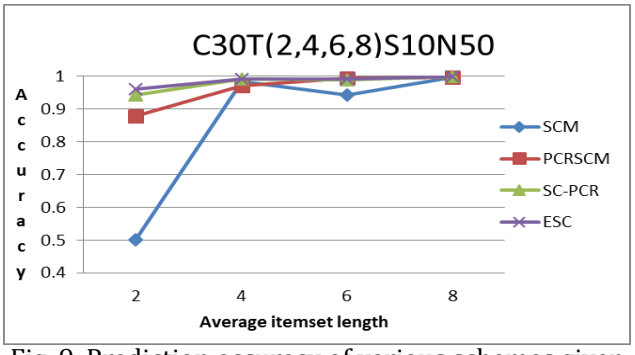

Fig. 9. Prediction accuracy of various schemes given T=2, 4, 6 and 8.

Fig. 9 shows the classification performance of the various schemes given synthetic datasets with four different average itemset lengths, namely T=2, 4, 6 and 8, respectively (i.e., C30T(2,4,6,8)S10N50). Note that the minimum support in a class is equal to 0.5 in every case. For the case of T=2, the prediction task is more

complex since the number of sequential patterns is very limited. However, the SC-PCR and ESC schemes still achieve prediction accuracies of 94% and 96%, respectively. As the average itemset length increases, the number of sequential patterns also increases. Consequently, the prediction performance of all four schemes improves. However, that of the SCM method reduces significantly for the largest considered mean itemset length (i.e., T=2) since the number of mined sequential patterns is very small. From inspection, the maximum, mean and minimum accuracies of the SCM method are found to be 99.7%, 85.63% and 50%, respectively. From inspection, the best, average and worst accuracies of SC-PCR are 99.8%, 98.1% and 94.3%, respectively, while the best, average and worst accuracies of ESC are 99.8%, 98.5% and 96%, respectively.

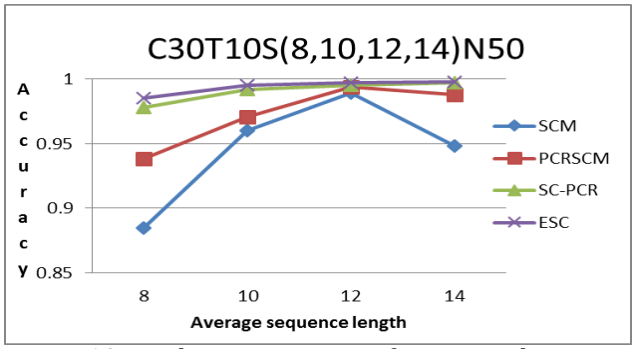

Fig. 10. Prediction accuracy of various schemes given S=8, 10, 12 and 14.

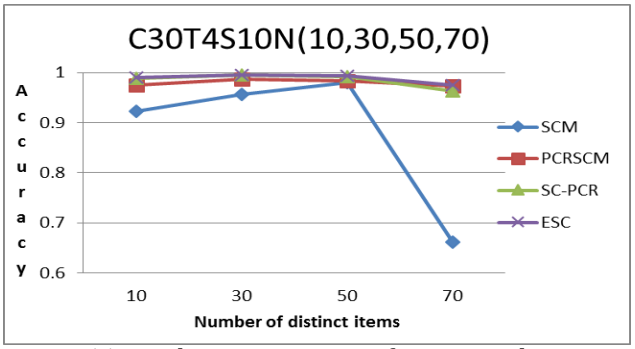

Fig. 11. Prediction accuracy of various schemes given N=10, 30, 50 and 70.

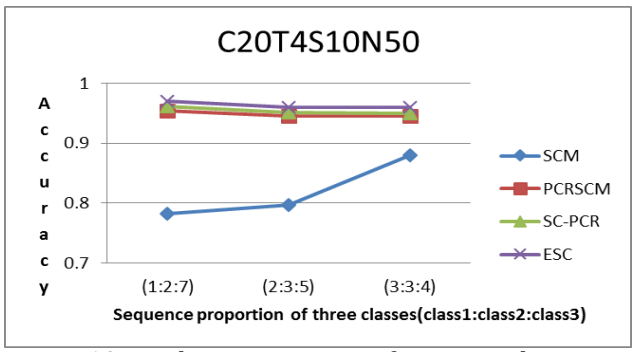

Fig. 12. Prediction accuracy of various schemes given different assignments of sequences to different classes.

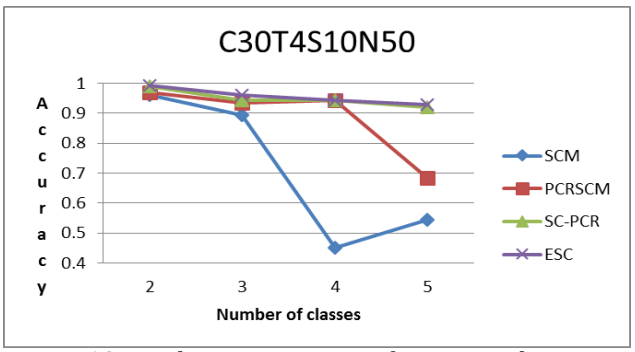

Fig. 13. Prediction accuracy of various schemes given different numbers of classes.

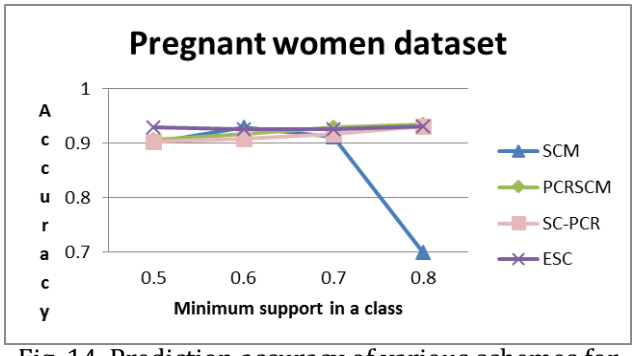

Fig. 14. Prediction accuracy of various schemes for pregnant women dataset.

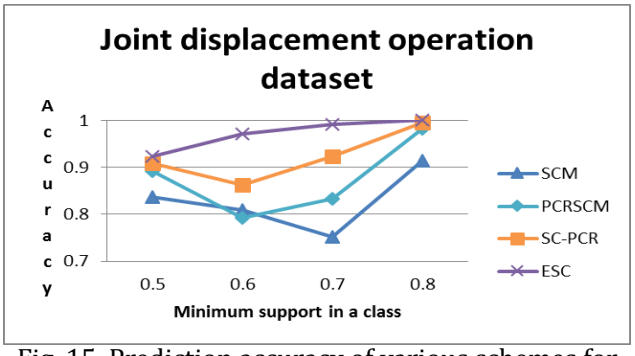

Fig. 15. Prediction accuracy of various schemes for joint displacement operation dataset.

Experiment (3)

The classification performance of the proposed schemes was further evaluated using four synthetic datasets with average sequence lengths of S=8, 10, 12 and 14, respectively (i.e., C30T10S(8,10,12,14)N50).

Note that the minimum support in a class was set equal to 0.5 in every case. Notably, as the average sequence length increases, the number of mined sequential patterns also increases. Consequently, the sequence classification task becomes more complex. As shown in Fig. 10, the prediction accuracy of the SCM scheme drops significantly given an average sequence length of S=14. However, the prediction accuracies of SC-PCR and ESC retain a high and relatively stable value as the average sequence length is increased from S=12 to S=14. For the SC-PCR and ESC methods, the best, average and worst accuracies are 99.7%, 99.1% and 97.8%, respectively, and 99.8%, 99.3% and 98.5%, respectively.

Experiment (4)

The classification experiment was repeated using four datasets in which the number of distinct items in each sequence was set equal to N=10, 30, 50 and 70, respectively (i.e., C30T4S10N(10,30,50,70)). In each case, the minimum support in a class was set as 0.5. For a given dataset size, the number of mined sequential patterns reduces as the number of distinct items increases, and thus the classification problem becomes more challenging. As shown in Fig. 11, the accuracy of SCM reduces significantly given the maximum considered number of distinct items, i.e., N=70. However, the classification performance of SC-PCR and ESC reduces only very slightly. For the SC-PCR scheme, the best, average and worst accuracies are 99.5%, 98.5% and 96.4%, respectively. Finally, for the ESC scheme, the best, average and worst accuracies are 99.6%, 98.9% and 97.5%, respectively.

Experiment (5)

A further series of experiments was performed using a C20T4S10N50 dataset in which the sequences were assigned to three different classes with three different ratios, namely 1:2:7, 2:3:5 and 3:3:4. Each experiment involved 20000 training data and 5000 test data. Furthermore, the minimum support in a class was set equal to 0.5 in every case. The corresponding results are presented in Fig. 12. It is seen that irrespective of the distribution of the sequences over the classes, the accuracies of the SC-PCR and ESC schemes remain largely unaffected. And the poor pattern score scheme in SCM results in the lower accuracy. For the SC-PCR scheme, the maximum, mean and minimum accuracies are 96.2%, 95.4% and 95%, respectively. Finally, for the ESC scheme, the maximum, mean and minimum accuracies are 97%, 96.3% and 96%, respectively.

Experiment (6)

The final experiment was performed using a C30T4S10N50 dataset with the minimum support in a class set as 0.5 and the number of classes set as 2, 3, 4 and 5, respectively. In every case, the sequences were distributed uniformly over the various classes. Given a larger number of classes, the prediction problem becomes more complex. Thus, as shown in Fig. 13, the accuracy of SCM reduces to 56% given five classes in the dataset, while that of PCRSCM falls to 68.5%. However, the accuracies of SC-PCR and ESC remain close to 92% and 93%, respectively. The best, average and worst accuracies for the SC-PCR and ESC schemes are 99.2%, 95% and 92%, respectively, and 99.2%, 95.7% and 93%, respectively.

4.2. Medical Sequence Datasets

The performance of the SC-PCR and ESC classification schemes was further evaluated using two medical datasets containing patient information obtained from a hospital in Taichung, Taiwan in 2009. The first dataset related to the patient histories of pregnant women, whereas the second related to patients scheduled for joint displacement operations. In both cases, the dataset comprised three attributes (i.e., the patient ID, the physical examination item, and the date). Moreover, the data records were sorted in date order.

Experiment (7)

The first experiment was performed using the pregnant women dataset containing 395 sequence data. The dataset contained just two classes, namely “natural childbirth” and “caesarean birth”. Classification

experiments were performed using four different values of the minimum support in a class, namely 0.5, 0.6, 0.7 and 0.8. In every case, the dataset was divided into 295 training data and 100 test data. Moreover, the average itemset length was 10, the average sequence length was 5, and the number of ensemble SC-PCR classifiers was set as 5.

Since the dataset contained relatively few sequential patterns, the accuracy of the SCM scheme falls to around 70% given the highest considered value of the minimum support in a class, i.e., 0.8, as shown in Fig. 14. However, for the SC-PCR and ESC schemes, the prediction accuracies retain a high and approximately constant value irrespective of the minimum support in a class. For the SC-PCR and ESC schemes, the best, average and worst accuracies are 93%, 91% and 90%, respectively, and 93%, 92.8% and 92.7%, respectively.

Experiment (8)

The joint displacement operation dataset comprised 246 sequences. As in the previous case, the dataset contained just two classes, namely "hip joint displacement" and "knee joint displacement". The dataset was divided into a training dataset with 186 sequences and a test dataset with 60 sequences. Experiments were performed using minimum support in a class values of 0.5, 0.6, 0.7 and 0.8, respectively. In every case, the average itemset length was 25, the average sequence length was 6, and the number of ensemble SC-PCR classifiers was set as 5.

Even given a minimum support in a class value equal to 0.8, the dataset still contains a large number of sequential patterns. Therefore, as shown in Fig. 15, all four methods achieve a relatively high classification performance. The maximum, mean and minimum accuracies of SC-PCR are 96.6%, 92.2% and 86.2%, respectively. Finally, the maximum, mean and minimum accuracies of ESC are 100%, 97.26% and 92.4%, respectively.

4.3. DNA Sequence Dataset — HS3D

The classification performance of SC-PCR and ESC was also further evaluated using the HS3D DNA sequence dataset. Moreover, the classification performance of the two schemes was compared with that of the FESP method proposed by Chen and Chen [12]. The HS3D dataset marks out the splice sites Exon-Intron and Intron-Exon associated with protein production in the homo sapiens gene. In general, the ability to pinpoint the splice sites of a gene is beneficial in understanding the protein production behavior of genes. A splice site identification model will offer great help for biomedical and the gene and in facilitating disease prevention. Thus, if SC-PCR and ESC can achieve a high classification accuracy for the HS3D dataset, they can serve as a useful tool in constructing splice site identification models for all manner of DNA sequence datasets.

Experiment (9)

The HS3D dataset contained two class labels, namely EI (exon intron) and IE (intron exon). The EI and IE data sequences had the forms shown in Figs. 16 and 17, respectively. The sequence length was around 140 in both cases. The experiments were performed using 5576 training samples (EI: 2796, IE: 2780) and 100 test samples (EI: 50, IE: 50). An inspection of the samples revealed that GT was located at nucleotide 71 in the EI sequences and AG was located at nucleotide 69 in the IE sequence. Accordingly, only nucleotides 69->72 in the DNA sequence were analyzed in the experiments. Three specific outcomes were identified: (1) nucleotide 69->70 is AG and nucleotide 70->71 is not GT, and thus the class label is IE; (2) nucleotide 71->72 is GT and nucleotide 69->70 is not AG, and thus the class label is EI; (3) nucleotide 69->70 is AG and nucleotide 70->71 is also AG, and thus the sequence (IE or EI) should be predicted using the proposed SC-PCR and ESC schemes.

The results obtained in Experiment (9) showed that for a sequence length of 14, the SC-PCR and ESC schemes both yield a high prediction accuracy. Additionally, nucleotide 69->72 in the present DNA

sequences has more features. Therefore, for each DNA sequence in the training data, based on nucleotide 68->73, the experiments considered 15 items in the forward direction and 15 items in the backward direction. That is, for each sample, two subsequences comprising nucleotides 54->68 and 73→87, respectively, were generated. A Pattern Score Matrix (*PSM*) was generated for each subsequence using the method described in Section 3.2, and the *CSM* was then calculated by summing up the two *PSM*s. The *CSM* was then used to predict the class of the original sample.

The experimental results are presented in Fig. 18. The best, average and worst accuracies of SC-PCR are 93%, 92% and 91%, respectively. Meanwhile, the best, average and worst accuracies of ESC are 93%, 92.3% and 92%, respectively. Table 2 compares the prediction accuracies of SC-PCR and ESC with that of FESP.

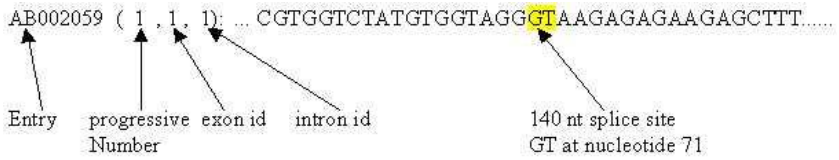

Fig. 16. DNA sequence format with EI (Exon Intron) splice site [20].

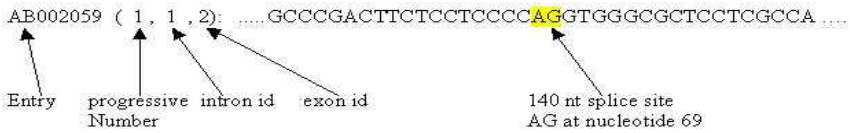

Fig. 17. DNA sequence format with IE (Intron Exon) splice site [20].

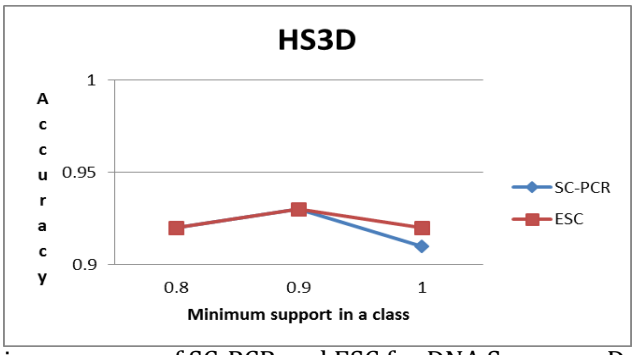

Fig. 18. Prediction accuracy of SC-PCR and ESC for DNA Sequence Dataset — HS3D.

Table 2. Prediction Accuracy of ESC, SC-PCR and FESP for DNA Sequence Dataset — HS3D

Method	Best Accuracy	Average Accuracy	Worst Accuracy
SC-PCR	93%	92%	91%
ESC	93%	92.3%	92%
FESP (k=300 in [12])	71%	70%	69%

4.4. Experiment Analysis

The SC-PCR classifier proposed in this study utilizes an enhanced sequential pattern reduction approach, which minimizes the number of patterns to be considered in the classification process while retaining all of the information required to ensure an accurate prediction outcome. Moreover, the proposed score matching method based on a PCR approach focuses on the key characteristics of the sequences. The string similarity comparison method makes possible a detailed examination of the similarities among the itemsets in different sequences. In addition, the quantified score calculation formula prevents one parameter from dominating the classification outcome, while the detailed score formula is designed in such a way as to prevent the same score being generated for more than one sequence. Moreover, the ensemble classifiers

concept provides a simple yet effective means of improving the prediction accuracy without the need for an optimization process.

SC-PCR utilizes PrefixSpan to identify the classification features of the training data. As a result, SC-PCR is suitable for the classification of sequences with a high degree of similarity among their itemsets; particularly given an abundance of classification features and a suitable sequence length. However, even when only relatively few classifications features are available and the sequences are distributed non-uniformly over the various classes, SC-PCR still achieves a higher prediction accuracy than SCM or PCRSCM. The experimental results have shown that SC-PCR consistently achieves a good classification performance irrespective of the number of sequential patterns (C), the length of the sequential patterns (S), the length of the itemsets (T), the number of items (N), or the distribution of the sequences over the classes. Overall, it has been shown that the average and worst accuracies of SC-PCR are 95.8% and 80.3%, respectively. In other words, SC-PCR is both highly stable and highly accurate.

The experimental results have shown that the classification performance of SC-PCR can be further enhanced by adopting an ensemble classifiers concept. The ESC scheme proposed in this study, based on the Bagging method ensures that each classifier has a certain repetitive training data. Such an approach prevents any single classifier from biasing the prediction outcome when the training dataset is very large. Thus, ESC can make each SC-PCR classifier with a variety of features to improve global accuracy. The experimental results have shown that ESC has mean and minimum prediction accuracies of 96.97% and 87%, respectively.

Table 3. Prediction Accuracies of SCM, PCRSCM, SC-PCR and ESC

Method	Average Accuracy	Worst Accuracy
SCM (published in 2011 [13])	81.63%	45.2%
PCRSCM (published in 2013 [15])	92.62%	68.5%
SC-PCR (proposed method)	95.8%	80.3%
ESC (proposed method)	96.97%	87%

Table 3 compares the average and worst prediction accuracies of the SCM, PCRSCM, SC-PCR and ESC schemes, respectively, as evaluated over Experiments (1)~(8). It is seen that PCRSCM outperforms SCM due to its pattern coverage rate calculation; while SC-PCR outperforms PCRSCM through its use of a score function with a greater ability to learn the different features of the sequences. Finally, ESC outperforms SC-PCR through it use of an ensemble classifiers approach.

5. Conclusions

In today's era of data repositories containing huge volumes of information, efficient techniques are required for analyzing this information in order to make better decisions regarding the future. Accordingly, this study has proposed two classifiers, namely a Sequence Classifier based on a Pattern Coverage Rate (SC-PCR) and an Ensemble Sequence Classifier (ESC), for effectively learning the characteristic features of a set of training sequences such that the classes of new sequences can be reliably predicted. Importantly, the SC-PCR and ESC classifiers are both robust to the effects of various cases as shown in previous experimental results. Furthermore, they require no optimization process and consider a large number of sequence features in the score calculation formula used as the basis for sequence classification. The experimental results have shown that SC-PCR has mean and minimum accuracies of 95.8% and 80.3%, respectively, while ESC has mean and minimum accuracies of 96.97% and 87%, respectively.

Modern hospitals and biomedical research centers require accurate prediction applications in predictive medicine. The proposed SC-PCR and ESC schemes can achieve high classification accuracies for both synthetic and biomedical sequence datasets. As a result, SC-PCR and ESC both provide a feasible solution for

sequence classification applications in the biomedical engineering domain.

Acknowledgment

This study is sponsored by the Ministry of Science and Technology in Taiwan under the contracts no. MOST 103-2221-E-005 -052 -MY2.

References

[1] Birlutiu, A., Groot, P., & Heskes, T. (2013). Efficiently learning the preferences of people. *Machine Learning*, *90*, 1-28.

[2] Vijayarani, S., & Deepa, S. (2014). Protein sequence classification in data mining — A study. *International Journal of Information Technology, Modeling and Computing*, *2(2)*, 1-8.

[3] Tuzun, E., & Dalmau, J. (2007). Limbic encephalitis and variants: classification, diagnosis and treatment. *The neurologist*, *13(5)*, 61-271.

[4] Bazinet, A. L. (2012). Cummings MP. A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, *13(1)*, 92.

[5] Kalitzin, S., Petkov, G., Velis, D., Vledder, B., & Lopes da Silva, F. (2012). Automatic segmentation of episodes containing epileptic clonic seizures in video sequences. *IEEE Transactions on Biomedical Engineering, 59(12)*, 3379-3385.

[6] Stavri, Z., & Michie, S. (2012). Classification systems in behavioural science: Current systems and lessons from the natural, medical and social sciences. *Health Psychology Review*, *6(1)*, 113-140.

[7] Tseng, V. S., & Lee, C. H. (2005). CBS: A new classification method by using sequential patterns. *Proceedings of SIAM International Conference on Data Mining* (pp. 596-600).

[8] Tseng, V. S., & Lee C. H. (2009). Effective temporal data classification by integrating sequential pattern mining and probabilistic induction. *Expert Systems with Applications*, *36*, 9524-9532.

[9] Exarchos, T. P., Tsipouras, M. G., Papaloukas, C., & Fotiadis, D. I. (2008). A two-stage methodology for sequence classification based on sequential pattern mining and optimization. *Data & Knowledge Engineering*, *66(3)*, 467-487.

[10] Zaki, M. J. (2000). Sequence mining in categorical domains: incorporating constraints. *Proceedings of the* 9*th International Conference on Information and Knowledge Management* (pp. 422-429).

[11] Papageorgiou, D. G., Demetropoulos, I. N., & Lagaris, I. E. (2004). MERLIN-3.1.1. A new version of the Merlin optimization environment. *Computer Physics Communications*, *159(1)*, 70-71.

[12] Chen, X., & Chen, J. (2011). Emerging patterns and classification algorithms for DNA sequence. *Journal of Software*, *6(6)*, 985-992.

[13] Li, I H., Lin, M. C., & Liao, I E. (2011). A sequential pattern length based sequence classifier model. *Proceedings of International Conference on Information Management* (p. 95).

[14] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., *et al*. (2004). Mining sequential patterns by pattern-growth: The PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering*, *16(11)*, 424-1440.

[15] Li, I H., Huang, J. Y., Liao, I E., & Lin, J. H. (2013). A sequence classification model based on pattern coverage rate. *Proceedings of the 8th International Conference on Grid and Pervasive Computing*, *LNCS* 7861 (pp. 737-745).

[16] Li, I H., Huang, J. Y., & Liao, I E. (2014). Mining sequential pattern changes. *Journal of Information Science and Engineering*, *30(4)*, 973-990.

[17] Lin, M. Y., & Lee, S. Y. (2005). Fast discovery of sequential patterns through memory indexing and database partitioning. *Journal of Information Science and Engineering*, *21*, 109-128.

[18] Saneifar, H., Bringay, S., Laurent, A., & Teisseire, M. (2008). S2MP: Similarity measure for sequential patterns. *Proceedings of the 7th Australasian Data Mining Conference* (pp. 95-104).
[19] Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123-140.
[20] Pollastro, P., & Rampone, S. (2013). HS3D: Homo sapiens splice sites dataset. Retrieved June 25, 2013, from the Nucleic Acids Research, Annual Database Issue. http://www.sci.unisannio.it/docenti/rampone/
[21] IBM Quest Market-Basket Synthetic Data Generator. Retrieved June 25, 2013, from the http://www.cs.rpi.edu/ ~zaki /software/IBM-datagen.tar.gz

I-Hui Li received the PhD degree in computer science and engineering from National Chung Hsing University, Taiwan. She is an assistant professor in Ling Tung University. Her research interests are in data mining, algorithms and wireless networks.

I-En Liao received the PhD degree from the Ohio State University. He is a professor and the chairman of the Department of Computer Science and Engineering of National Chung Hsing University, Taiwan. His research interests are in data mining, XML database, and wireless networks. He is a member of the ACM and the IEEE Computer Society.

Jin-Han Lin received the MS degree in computer science and engineering from National Chung Hsing University. He is a substitute service in National Chung-Shan Institute of Science and Technology, Taiwan. His research interests are in the areas of data mining and algorithms.

Jyun-Yao Huang is currently a PhD student in National Chung Hsing University. His research interests are in the areas of applied cryptography, network security and semantic web, with current focus on secure data services in cloud computing.