MATHEMATICAL BIOSCIENCES
AND ENGINEERING
Volume **14**, Number **2**, April **2017**

doi:10.3934/mbe.2017031
pp. **511–527**

MACHINE LEARNING OF SWIMMING DATA VIA WISDOM OF CROWD AND REGRESSION ANALYSIS

Jiang Xie and Junfu Xu

School of Computer Engineering and Science
Shanghai University
99 Shangda Road, Shanghai 200444, China

Celine Nie

University High School
4771 Campus Drive, Irvine, CA 92612, USA

Qing Nie*

Department of Mathematics
Center for Mathematical and Computational Biology
University of California, Irvine, CA 92697, USA

(Communicated by Yang Kuang)

Abstract. Every performance, in an officially sanctioned meet, by a registered USA swimmer is recorded into an online database with times dating back to 1980. For the first time, statistical analysis and machine learning methods are systematically applied to 4,022,631 swim records. In this study, we investigate performance features for all strokes as a function of age and gender. The variances in performance of males and females for different ages and strokes were studied, and the correlations of performances for different ages were estimated using the Pearson correlation. Regression analysis show the performance trends for both males and females at different ages and suggest critical ages for peak training. Moreover, we assess twelve popular machine learning methods to predict or classify swimmer performance. Each method exhibited different strengths or weaknesses in different cases, indicating no one method could predict well for all strokes. To address this problem, we propose a new method by combining multiple inference methods to derive Wisdom of Crowd Classifier (WoCC). Our simulation experiments demonstrate that the WoCC is a consistent method with better overall prediction accuracy. Our study reveals several new age-dependent trends in swimming and provides an accurate method for classifying and predicting swimming times.

1. **Introduction.** In competitive swimming, many factors influence the performance and training of swimmers. Previous studies have focused on using devices for swimming performance and technique evaluation [1], studying oxygen uptake

2010 *Mathematics Subject Classification.* 97R40, 92R30, 00A69.

Key words and phrases. Big data, statistical analysis, prediction, swimming model, time series data.

This work was partially supported by the Major Research Plan of NSFC [No. 91330116], and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and National Science Foundation grants DMS1161621 and DMS1562176.

* Corresponding author: qnie@math.uci.edu.

in swimming performance [21], and analyzing age of peak swim speed and gender difference in stroke performance [25].

Recently, a large amount of swimming data has become publicly available. Swimming times by each registered USA swimmer at every USA Swimming sanctioned meet have been recorded during the past thirty years, with all the data available online (`http://www.usaswimming.org`). By tracking each swimmer's time for each stroke, a time series data set could be obtained. As a result, a history of times for one particular stroke at one particular distance for male or female at different ages could be collected. This data provides ample opportunities for data mining and learning that may produce new knowledge in swimming.

In this study, we employ statistical analysis and machine learning tools to analyze USA Swimming data (`http://www.usaswimming.org`). We investigate the performance characteristics of both males and females at different ages for different strokes. In particular, the Pearson correlation is used to investigate the degree of dependency between swimmers performances at age 18 (the typical age for a swimmer to go to college) with their performance at younger ages. Regression analysis is used to approximate the performance curve in terms of age for different strokes.

To classify or predict the swimming times, we utilize twelve different machine learning methods that are based on variations of the following nine approaches: 1) the Support Vector Regression (SVR) method [24, 28, 3], which have been mostly used for pattern recognition; 2) the Artificial Neural Network (ANN) method, which has been widely used in classification [27, 11, 12, 7]; 3) the K-nearest Neighbor (KNN) algorithm in which K training samples of closest distance to the test sample are used to select the label [10, 13]; 4) the Support Vector Machine (SVM) method based on the structural risk minimization principle and the statistical learning theory [23, 4]; 5) the Decision Tree (DT) algorithm, based on a greedy top-down recursive partitioning strategy for tree growth [9, 2]; 6) the Random Forest (RF) approach, which is an ensemble classifier that consists of many decision trees and outputs the class [15, 16]; 7)AdaBoost, which constructs a succession of weak learners by using different training sets that are derived from resampling the original data [20, 6]; 8) the Navie Bayes (NB) classification, which relaxes the restriction of the dependency structures between attributes by simply assuming that attributes are conditionally independent, given the class label [26, 29]; 9) the Latent Dirichlet Allocation (LDA), a supervised method that searches for the project axes on which the data points of different classes are far from each other while requiring data points of the same class to be close to each other [5].

Based on each individual method, our simulations suggest that when different strokes are taken into account no specific method is superior to the other. It is found, however, each machine learning method contains a method performance ceiling in which the prediction accuracy of certain strokes is approximately 60%. Previously, 29 gene-network-inference methods were investigated, and it was found that the community predictions, by combining 29 methods, are more reliable than individual inference methods [17]. We then propose a Wisdom of Crowd classifier (WoCC) by aggregating the predictions made by each of the 12 methods used in this work. Comparing the accuracy of swim performance predictions using the 12 machine learning methods and the WoCC, we find that the WoCC exhibits more accurate predictions collectively than individual methods.

2. **Data set.** A large observational data set was obtained from USA Swimming, the national governing body for competitive swimming in the United States, on their website http://www.usaswimming.org/DesktopDefault.aspx.

To study the performance of the top swimmers in the USA, we searched for the top swimmers in the 100-meter (100M) freestyle with a time faster than the time standard 'B'. This filter gave us 5512 male and 4199 female swimmers. From there we collected all available times listed on the USA Swimming website dating back to as early as age 10. This process gave us a data set with 4,022,631 times for 9711 swimmers. For most cases, we analyzed performances for swimmers between the ages of 10 and 21.

A vector of five elements is used to describe the record of each data point: stroke, course length, age, time, and power point. The vector takes the following form:

$$record = (stroke, course, age, time, powerpoint). \tag{1}$$

Swimmers compete in four strokes and two different course lengths. The strokes are: freestyle (FR), butterfly (FL), backstroke (BK), and breaststroke (BR). The course lengths are long-course meters (LCM) and short-course yards (SCY). Notation, for example, is as follows: 100 meters as 100M and 100 yards as 100Y. The time for each record is always measured in seconds. The power point is a measurement Hy-Tek value that allows for a comparison of performances across strokes, distances and events, as well as between age groups (For more details please visit-http://www.usaswimming.org/ DesktopDefault.aspx?TabId=757). A sample of the data set is shown in Table 1.

TABLE 1. A sample of the USA Swimming data set

Stroke	Course	Age	Time (sec.)	Power points
100Y_FR	SCY	21	41.12	1053
100M_FL	LCM	24	53.83	926
100M_FR	LCM	25	50.01	930
200Y_FR	SCY	20	96.52	897
400M_IM	LCM	18	273.69	834
800M_FR	LCM	16	520.64	750
...	...	...	...	...

3. **Statistical analysis.**

3.1. **Variance.** The stability of a swimmer's performance may be measured by the variance in time for different strokes [14]. For each swimmer, we denote $mean_x$ and d_x as the average time and variance of times at the age of x. The variance and standard deviation of athlete i are denoted as d_{xi} and std_{xi}. N is defined as the total number of swimmers, and D_x is the average variance for all swimmers at age x:

$$D_x = \frac{1}{N}\sum_{i=1}^{N} d_{xi}. \tag{2}$$

The average coefficient of variation (CV) is a standardized measure of dispersion, probability distribution or frequency distribution.

$$CV = \frac{1}{N}\sum_{i=1}^{N} \frac{std_{xi}}{mean_{xi}}. \tag{3}$$

In principle, a smaller variance indicates a more stable performance at each age. The CV represents the relative variation in performance. Figure 1 and Figure 2 show the average variance and CV for both 100M and 100Y of all four strokes for all male and female athletes in the data set.

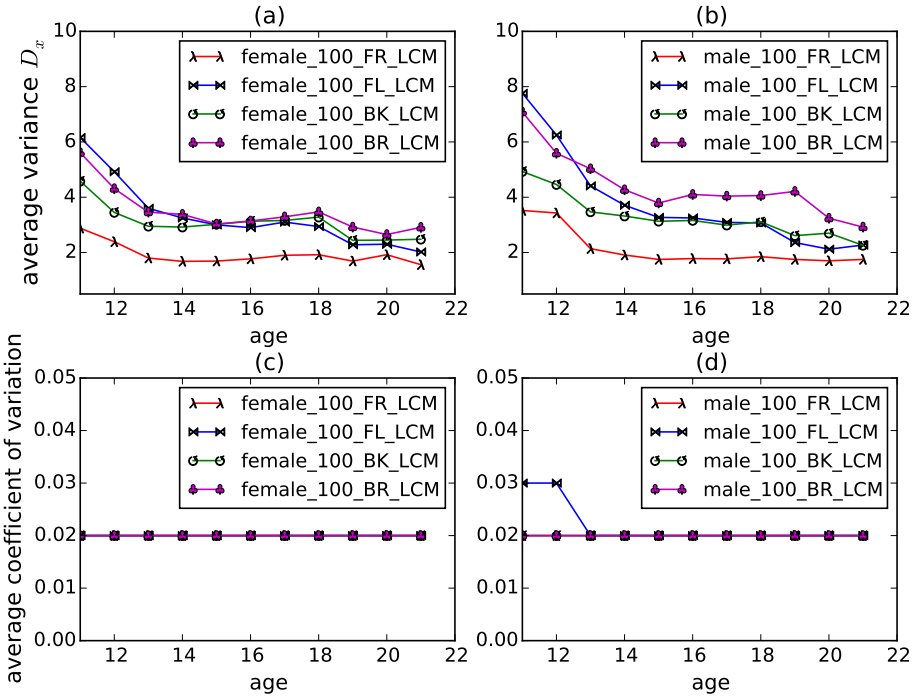

FIGURE 1. The average variance and CV as a function of age for different strokes in the LCM.

Based on Figure 1 and Figure 2, we made the following observations:

1. In Figure 1(a,b) and Figure 2(a,b), the performance variance of both male and female athletes decreased with increasing age, indicating a more stable performance as swimmers became older. After the age of 14, greater stability was suggested as the variances slowly decreased and became flat.
2. The variances ($age > 12$) in the LCM of Figure 1 were slightly larger than those for the SCY of Figure 1 in the same stroke, which is likely due to the longer distance in the meter course (one lap in SCY is 25 yards, while one lap in LCM is 50 meters).
3. For females in Figure 1(a), the variances for each age ($age > 13$) were very close to each other in butterfly, backstroke, and breaststroke. For males in Figure 1(b), the variance in breaststroke was greater than the others.
4. Among older male and female swimmers ($age > 13$), the 100M breaststroke (with the exception of Figure 1(a)) exhibited the largest variances. From this it is suggested that breaststroke, a stroke largely based off of timing and tempo, is the most difficult to maintain performance. The 100M freestyle has the smallest variances, most likely due to freestyle being the foundation stroke for most swimmers. Due to butterflys need for body fluidity and upper body

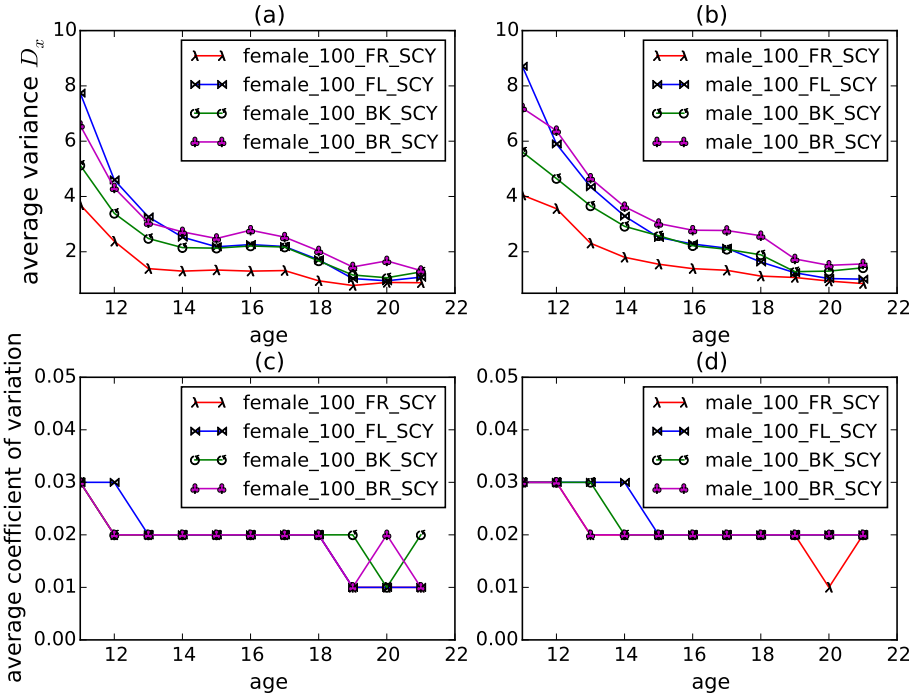

FIGURE 2. The average variance and CV as a function of age for different strokes in the SCY.

strength, it was not surprised to observe that at younger ages ($age < 12$) the 100M butterfly was the least stable stroke.

5. In Figure 1(c,d) and 2(c,d), the CVs are small for both the LCM and SCY, indicating small variability relative to the mean times.

Next, we analyzed the variances of different distances for freestyle in the 100M, 200M, 400M, and 800M. Usually, the mean times for events of different distances varied greatly. For example, the mean time of 100M is around 50 seconds with a variance of 1 to 8 seconds, the mean time of 200M is around 100 seconds with the variance of 6 to 12 seconds, and the mean time of 400M is around 270 seconds with the variance of 19 to 28 seconds. We found it difficult to compare the variances of events in different distances without normalization. For better comparison, we normalized each distance by a corresponding factor to measure the variances relatively in the 100M; i.e., we halved the time in the 200M, divided it by four in the 400M, and divided it by eight in the 800M.

The variances in performance of female swimmers at different distances were clustered together, suggesting that stability of performance has no significant relationship with distance, as shown in Figure 3(a). However, the male 100M freestyle showed to be significantly more stable than at other distances, as seen in Figure 3(b). One observation was that the variance curves of females at the four distances were smaller compared with curves of males. The male 800M FR exhibited an oscillatory behavior at the age of 20, which was partly due to the lack of a large data sample.

In Figure 4, both the average variance for 100Y freestyle and 200Y freestyle decreased with increasing age, and the average variance for 200Y was larger. In addition, the CVs for both the LCM and SCY were smaller, indicating small variabilities relative to the mean times.

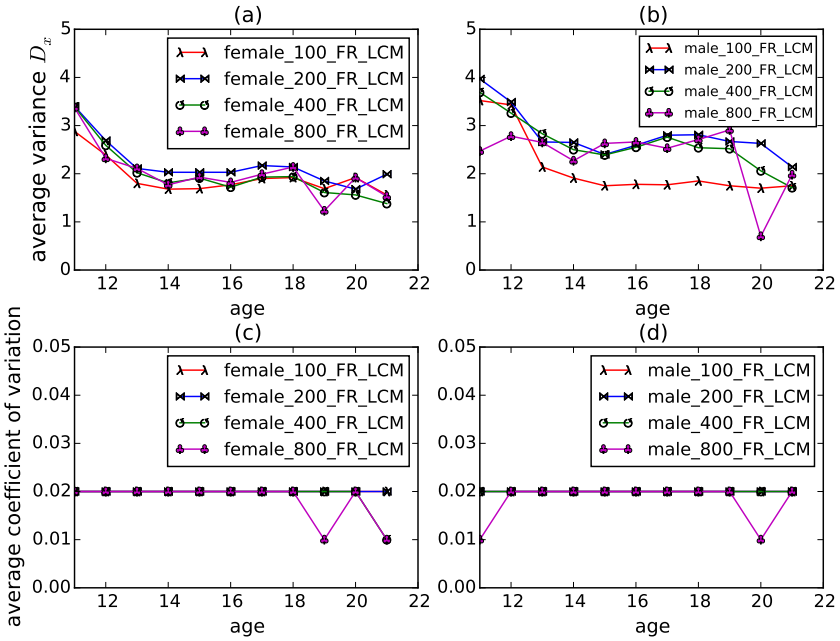

FIGURE 3. The average variance and CV in time for different distances (LCM).

3.2. **Pearson correlation.** The Pearson product-moment coefficient is a measure of the linear correlation between two variables X and Y and takes values between -1 and +1, where +1 represents the strongest positive correlation, 0 is no correlation, and -1 represents the strongest negative correlation. This quantity is widely used as a measure of linear dependence or association between two variables [19]. Here we use this quantity to study the potential correlation between swimmers' performances at age 18 and at younger ages. For a given swimmer, let $X = [x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{18}]$ where x_a is the average swimming time at the age of a. We use the Pearson correlation coefficient to measure the degree of swimming performance correlation between the swimmer's time at $x_a (a \in [10, 16])$ and x_{18}, defined as

$$\rho_{x_a, x_{18}} = \frac{cov(x_a, x_{18})}{\sigma_{x_a} \sigma_{x_{18}}} \tag{4}$$

where cov is the covariance and σ_x is the standard deviation of x. Typically, a graded interpretation of the correlation strength (Dancey & Reidy's 2004 categorization) is based on the following characterization:

1. 0.0-0.2 = weak or zero
2. 0.2-0.4 = modest
3. 0.4-0.6 = moderate
4. 0.6-0.8 = strong
5. 0.8-1.0 = very strong

Figure 5 shows the Pearson correlation between the swimmer's performances at age 18 and those at younger ages. For all male LCM events in Figure 5(b), age 16 showed the strongest correlation with age 18. The Pearson correlation coefficient is almost steady in the 100M breaststroke, 100M freestyle, and 100M backstroke, indicating that the developing male body has little to do with performance at young ages (ages 10 to 13). After age 13, for all four strokes, the Pearson correlation coefficient steadily increases. Coaches should enhance training for male swimmers at the age of 13. Surprisingly, 100M breaststroke and 100M backstroke follow the same trend in correlation: performances at younger ages in these two strokes have a strong correlation with performances at age 18. Training in breaststroke and backstroke should be technical and enhanced at a younger age due to the direct correlation of swims at an older age. Freestyle and butterfly training do not have to be stressed until the age of 14, at which point the coefficient is higher.

The female LCM events exhibit different trends. Unlike the trends in the coefficients between the males?100M breaststroke and 100M backstroke, the female coefficients show an increasing trend from age 10 to 13. Based on the increasing trend, coaches for female swimmers should enhance training of backstroke and breaststroke during these ages. Like the male coefficient trends for breaststroke and backstroke, the female coefficient shows similarities in trend but is larger than those of males. For both males and females, all strokes exhibit the greatest coefficient at age 16. This coefficient indicates the need for peak training at that age. The 100M freestyle has the lowest correlation with performance compared with other strokes.

For SCY events, the male and female Pearson correlation coefficient steadily increases for all strokes. However in male backstroke, there is a decrease in correlation from ages 11 to 13 as shown in Figure 5(d). This is most likely because males, on

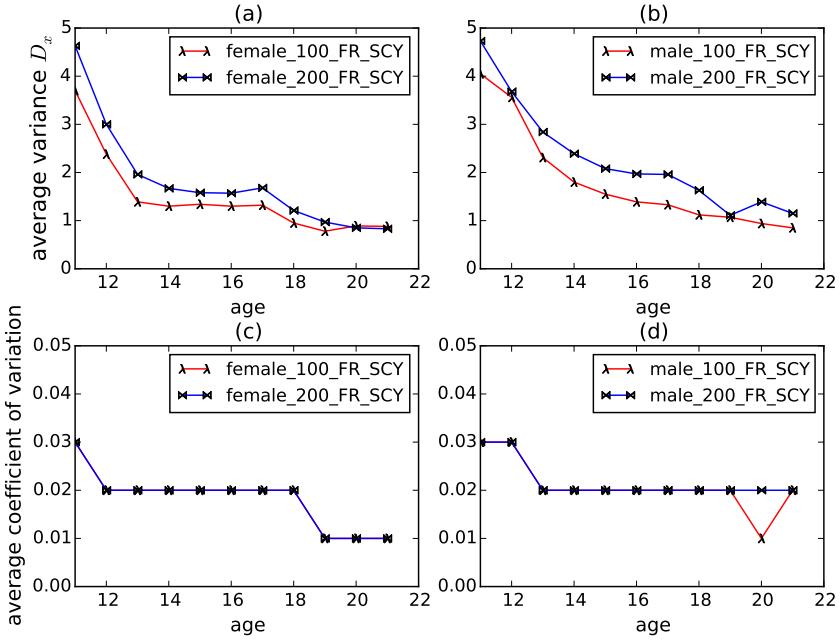

FIGURE 4. The average variance and CV in time for different distances (SCY).

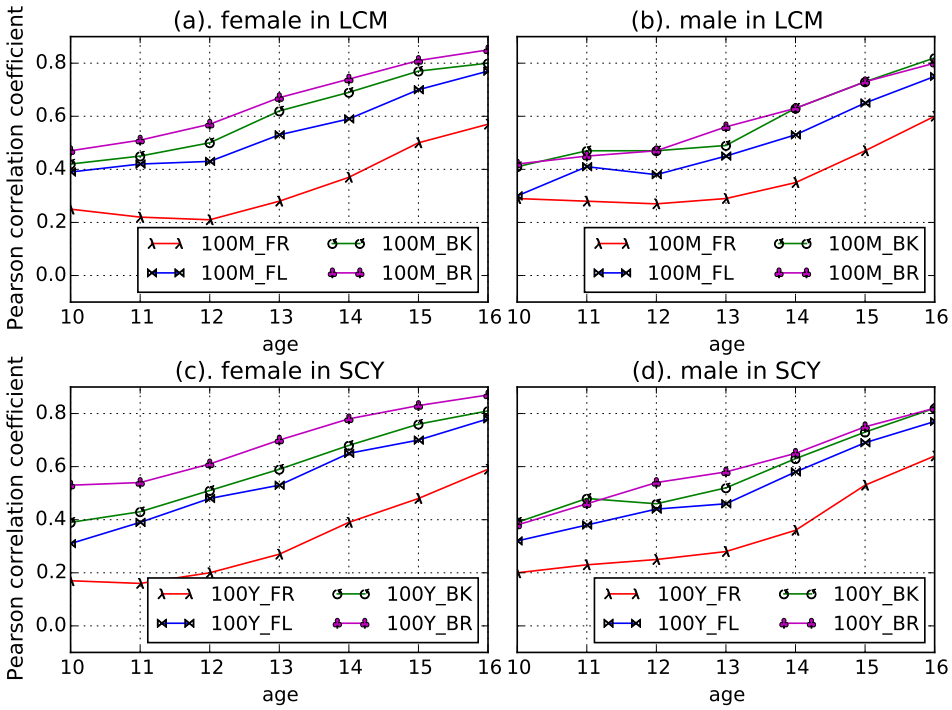

FIGURE 5. Pearson Correlation coefficient between younger ages and age 18.

average, start puberty at age 12, and many go through growth spurts and physical changes during that period of time. Once again, the 100Y freestyle shows low correlation coefficients throughout the younger ages.

3.3. **Regression analysis.** We then performed curve fitting [8] to describe the nonlinear relationship between swimming time and age. To reduce the variability in performance of the athletes in the data set, we divided them into four groups based on their fastest time in the 100M freestyle at age 18. The top 25% is called $Group1$, $25-50\%$ is $Group2$, etc. Figure 6 and Figure 7 show scatter plots for all male and female performances from ages 10 to 18 in each group. The average time at different ages is best fitted with a quadratic polynomial for each group.

The fitting curves and coefficients of the quadratic polynomial ($y = ax^2 + bx + c$) shown in Figure 6 and Figure 7 are shown in Table 2.

First, we observed that for females, the slopes of the four curves, yet again, slowly decreased with increasing age and then became flat after the age of 14 for both LCM and SCY of all groups as seen in Figures 6 and 7. This suggests a clear trend in which 14-years-old is a turning point for females, and the improvement in time, before the age of 14, from age to age is much more significant than after the age of 14.

Second, for males in both LCM and SCY in all groups, the coefficients are very close to each other for the 100M and 100Y freestyle. Although athletes may move from one group to another group with increasing age, they make similar progress as a group. Performance after the age of 14 continues to show major improvement in both courses. The overall slopes for males from ages 10 to 14 are similar to those

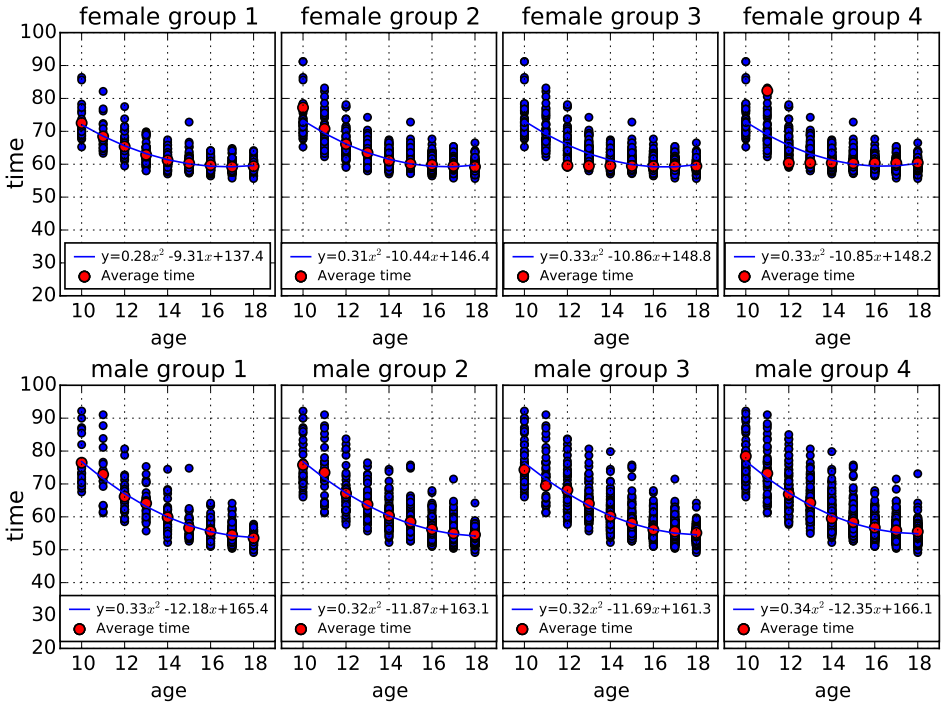

FIGURE 6. 100M freestyle performance regression analysis.

TABLE 2. The coefficients of the prediction equation (LCM and SCY)

course	LCM			SCY		
Female Group 1	0.28	-9.31	137.4	0.31	-10.55	140.4
Female Group 2	0.31	-10.44	146.4	0.34	-11.22	145.2
Female Group 3	0.33	-10.86	148.8	0.35	-11.48	146.6
Female Group 4	0.33	-10.85	148.2	0.35	-11.37	145.2
Male Group 1	0.33	-12.18	165.4	0.20	-7.99	126.5
Male Group 2	0.32	-11.87	163.1	0.24	-9.36	136.6
Male Group 3	0.32	-11.69	161.3	0.27	-10.24	143.2
Male Group 4	0.34	12.35	166.1	0.29	-10.78	147.3

after age 14, suggesting that males improve consistently between the ages of 10 and 18.

Finally, for both the male and female analysis seen in Figure 6 and Figure 7, at each age, the times for *Group*1 are consistently clustered together compared to other groups. As the swimming times become slower, the range of times grows smaller at each age.

4. **Machine learning.** We first used the nine different machine learning methods, KNN, Linear-SVM, RBF-SVM , DT, RF, AdaBoost, NB, LDA and Quadratic Discriminant Analysis(QDA) [22], to classify the level of swimming performance. In addition, Quadratic Polynomial Regression(QPR), ANN, and Support Vector Regression(SVR) models are applied to predict times. The above methods are used

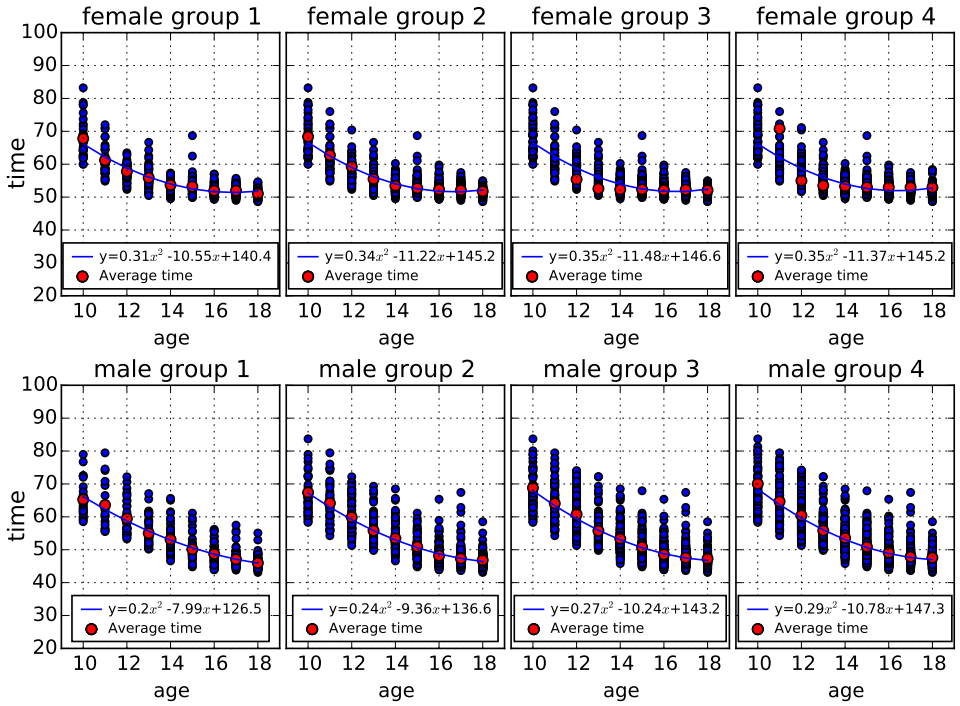

FIGURE 7. 100Y freestyle performance regression analysis.

to recognize patterns to generate classifications of data or, in this case, to predict swim times. Except for ANN (using Neurolab Library), the methods used in our paper are mainly implemented by scikit-learn [18], an excellent machine learning package using Python.

When these approaches were applied to the data, the following four steps were carried out:

- Data preprocessing and selection;
- Application of training and learning data to produce machine learning models;
- Application of the machine learning models to predict outcomes of the testing data set;
- Evaluating of performance.

4.1. **Data preprocessing.** Different athletes have different swimming times at the same age. Thus, we aggregated the performances of each age i to t_{i_min}, $t_{i\_average}$, t_{i_max} and t_{i_std}, which indicated the performance at that age. The variables t_{i_min}, $t_{i\_average}$, t_{i_max} and t_{i_std}, represent the best performance time, the average time, the worst time and the standard deviation of the time, respectively, at age i. The times of a single swimmer can be described as follows.

$$Record = \left\{time_i = (t_{i\_min}, t_{i\_average}, t_{i\_max}, t_{i\_std})\right\}, \text{and i} \in \left[10, 15\right] \cup 18. \quad (5)$$

To determine the degree to which performance at older ages depends on performance at younger ages, times of performances from ages 10 to 15 were used to represent each swimmer's performance at a young age. Based on the performance at age 18, we divided the athletes into three groups from fastest to slowest, according to the USA Swimming 2013-2016 National Age Group Motivational Times

(http://www.usaswimming.org/DesktopDefault.aspx?TabId=1465&Alias=Rainbow&Lang=en). For example, the 100M freestyle swimming time standards are shown in Table 3.

TABLE 3. Definition of male swimming time standards in 100M freestyle levels

Time Standards/Cuts	Mean time in 18-year-old male (sec.)
AAAA Min	$time <= 54.09$
AAA Min	$54.09 < time <= 56.59$
Slower than AAA Min	$time > 56.59$

4.2. **Training and learning model.** Similar to most machine learning approaches, the data set is divided into two portions, a training set and a testing set by classical 10-fold cross-validation. We transformed the input features by scaling each feature to a given range [0, 1.0]. The swimming times for young ages were used as the inputs to produce models for machine learning methods, as seen in Figure 8. A list of different parameters used for each machine learning method are shown in Table 4. Our simulations mainly focused on two events with the highest Pearson coefficient, 100M/100Y breaststroke, and the lowest Pearson coefficient, 100M/100Y freestyle. We then predicted the swimming performance time and classified their level at age of 18.

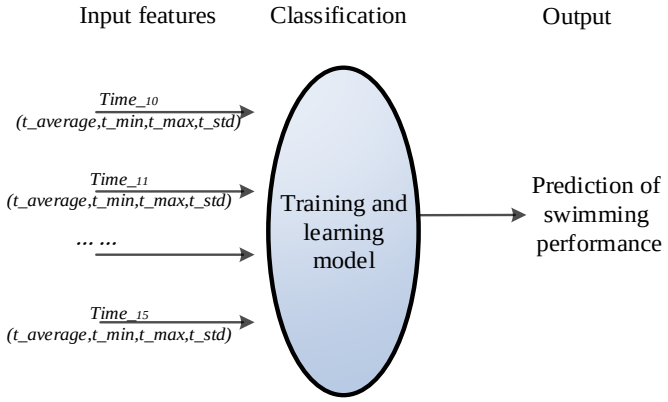

FIGURE 8. Classification model.

4.3. **Experiment results.**

4.3.1. *Time prediction and error measurements.* Here, we applied SVR and ANN methods to predict performance time at age 18 using performance times (t_{i_min}, $t_{i\_average}$, t_{i_max} and t_{i_std}) at the younger ages. The QPR method was used to predict times using the average times ($t_{i\_average}$) at each age. In our experiments, the mean absolute deviation (MAD) is used to measure how close the forecasts were to the real swimming times. The results in Table 5 and Table 6 show the MAD of the different predictors for different strokes and for males and females.

TABLE 4. Parameters of each method

Methods	Parameters
KNN	k=5
Linear SVM	$kernel$=linear
RBF SVM	$kernel$=rbf
DT	max_depth=10
RF	max_depth=10, $n\_estimators$=10, $max\_features$=1
AdaBoost	default parameters
NB	default parameters
LDA	default parameters
QDA	default parameters
ANN	2 layers with 24 inputs 10 neurons in hidden layer
SVR	$kernel$=linear,C=1.0

The experimental results of the predictions for male 100M freestyle times at age 18 over the known historical swimming data are shown in Figures 9.

TABLE 5. MAD of swimming time predictions for breaststroke

Methods	male 100M	male 100Y	female 100M	female 100Y
	(103 Records)	(179 Records)	(143 Records)	(210 Records)
QPR	8.00s	7.70s	8.98s	8.42s
ANN	1.20s	2.44s	2.97s	2.65s
SVR	1.01s	1.90s	2.43s	1.67s

TABLE 6. MAD of swimming time predictions for freestyle

Methods	male 100M	male 100Y	female 100M	female 100Y
	(340 Records)	(572 Records)	(548 Records)	(743 Records)
QPR	6.70s	8.00s	8.21s	6.05s
ANN	1.50s	1.24s	1.50s	1.17s
SVR	1.18s	1.02s	1.28s	0.95s

The regression model predictor cannot generate swimming times from past average times, unlike ANN or SVR. Thus, it is difficult to make a prediction of an athletes performance at age 18 based on his/her earlier ages using QPR. SVR has demonstrated success in swimming time prediction and MAD is approximately 1.0 seconds off in some strokes, implying that SVR is able to analyze performance times. However, the forecast value of the SVR analysis is too conservative compared with the ANN method, which is able to predict certain peak values.

4.3.2. *Level classification and evaluating performance.* Next we introduced a Wisdom of Crowd Classifier (WoCC) approach by aggregating the prediction lists of the nine classification methods. In WoCC, we applied the twelve popular machine learning approaches to the multi-label prediction problems; we then combined the level-prediction lists simply by re-ranking the level list according to the new predicted rank using the median value. Specifically, first, swimmers were given prediction labels, including AAAA Min, AAA Min and slower than AAA Min, as shown

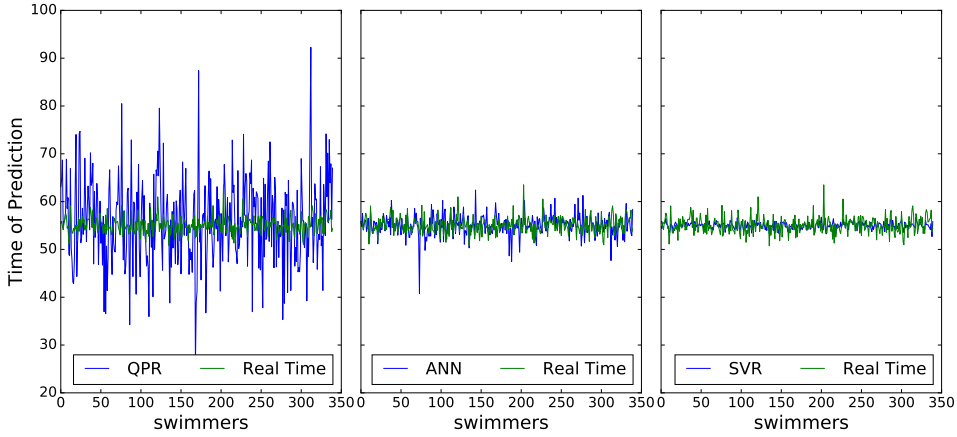

FIGURE 9. Predictions of swimming times.

in Table 4, which are forecasted by different machine learning methods, labeled 1, 2 and 3, respectively, and sorted into a prediction list. Then, the median value of the list was chosen as the WoCC prediction. We then compared the median value with the average value and the most common prediction label. The experiments indicated that the median value was the best, as the WoCC predicted.

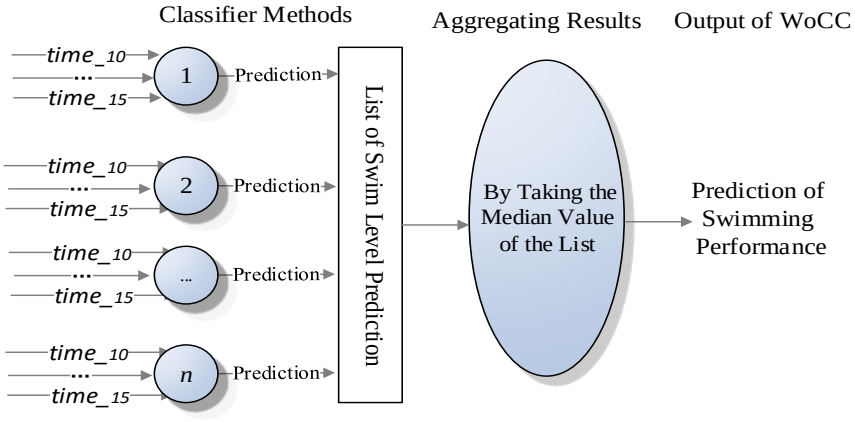

FIGURE 10. Illustration of a Wisdom of Crowd Classifier(WoCC).

For multi-label classification problems, accuracy was used to measure the ability of the classifiers. Accuracy is represented by the ratio $tp/(tp + fp)$ where tp is the number of true positives and fp the number of false positives. The best value is 1.0, and the worst value is 0.

Table 7 and Table 8 list the accuracy of various machine learning methods. Some methods performed well with respect to certain strokes although not with other strokes. For example, the predictions of the Linear SVM are more accurate for the male 100M freestyle than for the 100M breaststroke. The RF performs

well with respect to the female 100M freestyle but not the male 100M breaststroke. For different stroke, one classifier may have a larger difference in accuracy. On the whole, Table 7 and Table 8 show that several methods including KNN, Linear SVM, RF and NB predict with high confidence compared with others. However, the WoCC had the highest accuracy for different strokes for both males and females, with the exception of the male 100Y breastroke.

Nevertheless, the WoCC was more effective than the other methods with one exception in 100Y breaststroke for which NB and DT have slightly better accuracy. The WoCC method may be the preferred method for predicting performance levels, as it exhibited greater accuracy and was more stable with respect to larger data sets.

TABLE 7. Prediction accuracy for freestyle

Methods	male 100M (340 Records)	male 100Y (572 Records)	female 100M (548 Records)	female 100Y (743 Records)
KNN	0.55	0.60	0.59	0.58
Linear SVM	**0.60**	**0.62**	0.63	**0.64**
RBF SVM	0.58	0.61	0.66	0.63
DT	0.48	0.54	0.62	0.58
RF	0.59	0.60	**0.67**	**0.64**
AdaBoost	0.53	0.59	0.59	0.60
NB	0.53	0.49	0.56	0.55
LDA	**0.60**	0.60	0.64	0.63
QDA	0.52	0.56	0.58	0.53
WoCC	**0.61**	**0.64**	**0.67**	**0.65**

TABLE 8. Accuracy of prediction for breaststroke

Methods	male 100M (103 Records)	male 100Y (179 Records)	female 100M (143 Records)	female 100Y (210 Records)
KNN	0.50	0.46	**0.75**	0.64
Linear SVM	0.47	0.46	0.70	**0.66**
RBF SVM	0.37	0.36	0.66	0.54
DT	0.44	0.50	0.69	0.64
RF	0.46	0.49	**0.75**	0.63
AdaBoost	0.47	0.45	0.57	0.56
NB	**0.53**	**0.53**	0.66	0.65
LDA	0.45	0.45	0.67	0.64
QDA	0.41	0.41	0.63	0.56
WoCC	**0.61**	**0.49**	**0.75**	**0.66**

5. **Conclusions.** In this paper, we used statistical analysis and machine learning methods to study correlations between swimming performance and age, stroke, and gender. In particular, using the Pearson correlation, we investigated the linear correlation between performances at different ages. We also applied the nonlinear regression (QPR) to the times. By applying machine learning methods, we found

that the ANN (a high-dimension nonlinear classification method) and SVM (a high-dimension linear classification method) are able to predict swimming times with low MAD. For example, SVR could successfully predict swimming times with an error within one second for some strokes. The prediction accuracy of the twelve existing methods was found to vary depending on the choice of stroke. Therefore, we used the Wisdom of Crown Classifier (WoCC) to predict swimming performance for various strokes and different genders. By comparing the WoCC with the nine other popular classification methods, we demonstrated that the WoCC is the most reliable and robust prediction method, which may have broader application.

Our study and approach may be used to predict swimming times using different data sets. This study will be particularly helpful for coaches and swimmers who want to make predictions on their swimming times for the near future. Moreover, enhanced and targeted training programs might be developed through analysis of a small group of swimmers for certain age groups to optimize performance results. For example, we found that female athletes usually achieve stable performances around the age of 14 through both CV analysis and the Pearson correlation analysis. Another example is the 100 meter, or yard, breaststroke held a greater Pearson correlation coefficient than what was observed with other strokes. This suggests that enhancement of breaststroke training earlier in a swimmer's career will be beneficial. This enhancement would allow swimmers to have a higher possibility of remaining in the top group further along in their careers. This is because it is observed that swimmers retain a high Pearson coefficient as they move from a younger age group to an older age group with breaststroke.

Our current study is based on the data collected from the best freestyle swimmers whose times were recorded in the USA Swimming website. It would be interesting to examine whether this trend applies to other stroke specialists. In general, the study would be deepened if to investigate specific groups of swimmers, strokes, distances, and to analyze the similarities and differences among these groups in various swimming events. Comparing the overall performance features between different stroke specialists might provide many more insights on swimming performance. A more focused study on younger swimmers might also provide new knowledge on training young swimmers and the impact on their future performance. The results in this study can give both parents and swimmers the opportunity to be proactive in future swimming training.

Acknowledgments. We would like to thank Tao Peng for his helpful discussions on the machine learning methods. Celine Nie would like to thank Coach Chris Duncan, Coach Rod Hansen, and Coach Andi Kawamoto-Klatt for their encouragment and support during her swimming career.

REFERENCES

[1] M. Bächlin and G. Tröster, Swimming performance and technique evaluation with wearable acceleration sensors, *Pervasive and Mobile Computing*, **8** (2012), 68–81.

[2] R. C. Barros, M. P. Basgalupp, A. C. De Carvalho and A. Freitas et al., A survey of evolutionary algorithms for decision-tree induction, *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, **42** (2012), 291–312.

[3] D. Basak, S. Pal and D. C. Patranabis, Support vector regression, *Neural Information Processing-Letters and Reviews*, **11** (2007), 203–224.

[4] C. Cai, G. Wang, Y. Wen, J. Pei, X. Zhu and W. Zhuang, Superconducting transition temperature t c estimation for superconductors of the doped mgb2 system using topological index

via support vector regression, *Journal of Superconductivity and Novel Magnetism*, **23** (2010), 745–748.

[5] D. Cai, X. He and J. Han, Semi-supervised discriminant analysis, in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, IEEE, 2007, 1–7.

[6] J. Cao, S. Kwong and R. Wang, A noise-detection based adaboost algorithm for mislabeled data, *Pattern Recognition*, **45** (2012), 4451–4465.

[7] J. J. Cheh, R. S. Weinberg and K. C. Yook, An application of an artificial neural network investment system to predict takeover targets, *Journal of Applied Business Research (JABR)*, **15** (2013), 33–46.

[8] J. L. Dye and V. A. Nicely, A general purpose curve fitting program for class and research use, *Journal of chemical Education*, **48** (1971), 443.

[9] M. A. Friedl and C. E. Brodley, Decision tree classification of land cover from remotely sensed data, *Remote Sensing of Environment*, **61** (1997), 399–409.

[10] K. Fukunaga and P. M. Narendra, A branch and bound algorithm for computing k-nearest neighbors, *Computers, IEEE Transactions on*, **100** (1975), 750–753.

[11] A. Garg and K. Tai, Comparison of regression analysis, artificial neural network and genetic programming in handling the multicollinearity problem, in *Modelling, Identification & Control (ICMIC), 2012 Proceedings of International Conference on*, IEEE, 2012, 353–358.

[12] Z. Guo, W. Zhao, H. Lu and J. Wang, Multi-step forecasting for wind speed using a modified emd-based artificial neural network model, *Renewable Energy*, **37** (2012), 241–249.

[13] I. Hmeidi, B. Hawashin and E. El-Qawasmeh, Performance of knn and svm classifiers on full word arabic articles, *Advanced Engineering Informatics*, **22** (2008), 106–111.

[14] Y. Jiang, J. Lin, B. Cukic and T. Menzies, Variance analysis in software fault prediction models, in *Software Reliability Engineering, 2009. ISSRE'09. 20th International Symposium on*, IEEE, 2009, 99–108.

[15] A. Liaw and M. Wiener, Classification and regression by randomforest, *R news*, **2** (2002), 18–22.

[16] B. Liu and G. Qiu, Illuminant classification based on random forest, in *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*, IEEE, 2015, 106–109.

[17] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano and G. Stolovitzky, Revealing strengths and weaknesses of methods for gene network inference, *Proceedings of the National Academy of Sciences*, **107** (2010), 6286–6291.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, **12** (2011), 2825–2830.

[19] M.-T. Puth, M. Neuhäuser and G. D. Ruxton, Effective use of pearson's product–moment correlation coefficient, *Animal Behaviour*, **93** (2014), 183–189.

[20] G. Rätsch, T. Onoda and K.-R. Müller, Soft margins for adaboost, *Machine Learning*, **42** (2001), 287–320.

[21] J. F. Reis, F. B. Alves, P. M. Bruno, V. Vleck and G. P. Millet, Oxygen uptake kinetics and middle distance swimming performance, *Journal of Science and Medicine in Sport*, **15** (2012), 58–63.

[22] B. Scholkopft and K.-R. Mullert, Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX*, **1** (1999), p1.

[23] C. Schüldt, I. Laptev and B. Caputo, Recognizing human actions: A local svm approach, in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, IEEE, **3** (2004), 32–36.

[24] A. J. Smola and B. Schölkopf, A tutorial on support vector regression, *Statistics and Computing*, **14** (2004), 199–222.

[25] M. Vaso, B. Knechtle, C. A. Rüst, T. Rosemann and R. Lepers, Age of peak swim speed and sex difference in performance in medley and freestyle swimming. a comparison between 200 m and 400 m in swiss elite swimmers, *Journal of Human Sport and Exercise*, **8** (2013), 954–965.

[26] Q. Wang, G. M. Garrity, J. M. Tiedje and J. R. Cole, Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy, *Applied and Environmental Microbiology*, **73** (2007), 5261–5267.

[27] S.-C. Wang, Artificial neural network, in *Interdisciplinary Computing in Java Programming*, Springer, 2003, 81–100.

[28] C.-H. Wu, J.-M. Ho and D.-T. Lee, Travel-time prediction with support vector regression, *Intelligent Transportation Systems, IEEE Transactions on*, **5** (2004), 276–281.

[29] J. Wu, Z. Cai, S. Zeng and X. Zhu, Artificial immune system for attribute weighted naive bayes classification, in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, IEEE, 2013, 1–8.

Received July 01, 2016; Accepted August 05, 2016.

E-mail address: jiangx@shu.edu.cn
E-mail address: xujunfu@shu.edu.cn
E-mail address: nie.celine@gmail.com
E-mail address: qnie@math.uci.edu