A SVM-based Committee Machine for Prediction of Hong Kong Horse Racing

Wai-Chung Chung, Chuan-Yu Chang
National Yunlin University of Science and Technology
Department of Computer Science and Information Engineering
chuanyu@yuntech.edu.tw

Chien-Chuan Ko
National Chiayi University
Department of Computer Science and Information Engineering
kocc@mail.ncyu.edu.tw

Abstract—**In Hong Kong and Macao, horse racing is the most famous gambling with a long history. This study proposes a novel approach to predict the horse racing results in Hong Kong. A three-years-long race records dataset obtained from Hong Kong Jockey Club was used for training a support-vector-machine-based committee machine. Bet suggestions could be made to gamblers by studying previous data though machine learning. In experiment, there are 2691 races and 33532 horse records obtained. Experiments focus on accuracy and return rate were conducted separately through constructing a committee machine. Experimental results showed that the accuracy and return rate achieve 70.86% and 800,000% respectively.**

Keywords—Machine Learning, Predict, Horse Racing

I. Introduction

The use of machine learning techniques to predict the results of sporting events has been an important research topic in the field of artificial intelligence. There are many factors that can influence the outcome of races such as field condition, parentage, previous records, jockey, and draw number. There are many researchers have been conducted. Zanjan *et al.* used machine learning techniques to predict New York's Aqueduct Race [7]. Robert *et al.* predicted the result of Greyhound Races [5]. From literatures above, a few observations can be made: (1) It's feasible to predict horse racing results by using machine learning. (2) The higher number of horses participated in a race, the lower of the prediction accuracy. (3) Accuracy and the return rate was an interchangeable relationship. Therefore, this paper proposed a Support-Vector-Machine-based committee machine to predict the future racing results based on the past records in Hong Kong.

Reasons of selected "Hong Kong's horse racing" as a subject of this study are: (1) There are abundant amount of data to support the result: In Hong Kong, there are 83 race days in a year. In each race day contains at least 8 horse races with 14 participant horse in maximum. There are 664 races data and results in a year, near 2000 race records in a three years period. In addition, distortions produced by extreme cases can be reduced by large amount of data. (2) An intact database system managed by the HKJC which is the official race event organizer. Race record included results, win odds and dividend are recorded and organized by the HKJC. The better quality of the data may produce a better model. (3)High return: Since horse racing in Hong Kong provided a diversity of ways of betting. Some of them are extremely high in dividend. Using an example in 1st Jan 2015, a $10 bet on triple trio can yield a $521,391 dividend. (4) High difficulties: assume there are 14 participants in a race, the chance of winning a triple trio is only 1 over 42 million by random guess.

To demonstrate the capability of the proposed method, we compared the forecasting ability of Random Forest and Support Vector Machine respectively. As the experimental results suggested, support vector machine tends to have a better forecasting power.

II. ProposeD Method

A. Data acquisition

In this study, all data are obtained from the official website of HKJC, http://racing.hkjc.com/. Data dated from 1st Jan 2012 to 30th June 2015 excluded all international race events. The dataset contains 2691 race records and 33532 horse records. Table 1 shows the result of the 1st race on 1st Jan 2012. Table 2 shows a historical record of a horse named DRAGON CHOICE.

B. Data featuring

From race records and historical records of horses, 19 features can be extracted. (1) Class, (2)Distance, (3) Horse's win rate, (4)Horse's place rate, (5) Horse's show rate (6) Jockey's win rate, (7) Jockey's place rate, (8) Jockey show rate, (9) Trainer's win rate, (10) Trainer's place rate, (11) Trainer's show rate, (12) Actual weight, (13) Declared horse weight, (14) Draw, (15) Average break position, (16) Fastest finish time, (17) Win-odds, (18) place of last races, (19) Average place of last 3 races. All features undergo normalization can be represented as:

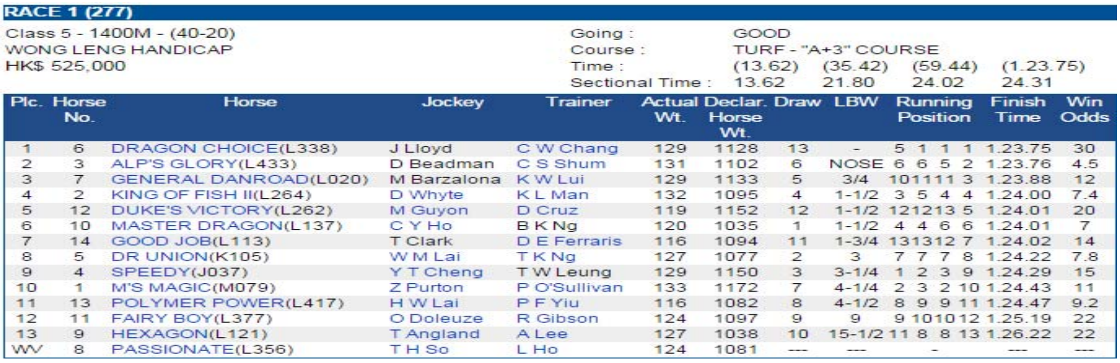
$$\mathbf{v}_i = [f_1, f_2, f_3, \cdots, f_{19}] \tag{1}$$

C. Support Vector Machine

Support vector machine is a supervised learning algorithm that seeks for a hyper-plane which separates training samples

978-1-5386-2761-7

TABLE III. RESULT OF THE 1ST JAN 2012.



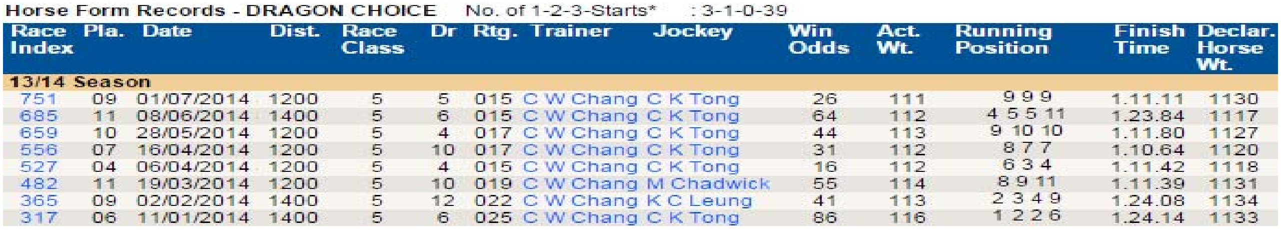
RACE 1 (277)

Class 5 - 1400M - (40-20)
WONG LENG HANDICAP
HK$ 525,000

Going :	GOOD			
Course :	TURF - "A+3" COURSE			
Time :	(13.62)	(35.42)	(59.44)	(1.23.75)
Sectional Time :	13.62	21.80	24.02	24.31

Plc.	Horse No.	Horse	Jockey	Trainer	Actual Wt.	Declar. Horse Wt.	Draw	LBW	Running Position	Finish Time	Win Odds
1	6	DRAGON CHOICE(L338)	J Lloyd	C W Chang	129	1128	13	-	5 1 1 1	1.23.75	30
2	3	ALP'S GLORY(L433)	D Beadman	C S Shum	131	1102	6	NOSE	6 6 5 2	1.23.76	4.5
3	7	GENERAL DANROAD(L020)	M Barzalona	K W Lui	129	1133	5	3/4	101111 3	1.23.88	12
4	2	KING OF FISH II(L264)	D Whyte	K L Man	132	1095	4	1-1/2	3 5 4 4	1.24.00	7.4
5	12	DUKE'S VICTORY(L262)	M Guyon	D Cruz	119	1152	12	1-1/2	121213 5	1.24.01	20
6	10	MASTER DRAGON(L137)	C Y Ho	B K Ng	120	1035	1	1-1/2	4 4 6 6	1.24.01	7
7	14	GOOD JOB(L113)	T Clark	D E Ferraris	116	1094	11	1-3/4	131312 7	1.24.02	14
8	5	DR UNION(K105)	W M Lai	T K Ng	127	1077	2	3	7 7 7 8	1.24.22	7.8
9	4	SPEEDY(J037)	Y T Cheng	T W Leung	129	1150	3	3-1/4	1 2 3 9	1.24.29	15
10	1	M'S MAGIC(M079)	Z Purton	P O'Sullivan	133	1172	7	4-1/4	2 3 2 10	1.24.43	11
11	13	POLYMER POWER(L417)	H W Lai	P F Yiu	116	1082	8	4-1/2	8 9 9 11	1.24.47	9.2
12	11	FAIRY BOY(L377)	O Doleuze	R Gibson	124	1097	9	9	9 101012	1.25.19	22
13	9	HEXAGON(L121)	T Angland	A Lee	127	1038	10	15-1/2	11 8 8 13	1.26.22	22
WV	8	PASSIONATE(L356)	T H So	L Ho	124	1081	---	---	-	---	---

TABLE IV. HISTORICAL RECORD OF A HORSE NAMED DRAGON CHOICE



Horse Form Records - DRAGON CHOICE No. of 1-2-3-Starts* : 3-1-0-39

Race Index	Pla.	Date	Dist.	Race Class	Dr	Rtg.	Trainer	Jockey	Win Odds	Act. Wt.	Running Position	Finish Time	Declar. Horse Wt.
13/14 Season													
751	09	01/07/2014	1200	5	5	015	C W Chang	C K Tong	26	111	9 9 9	1.11.11	1130
685	11	08/06/2014	1400	5	6	015	C W Chang	C K Tong	64	112	4 5 5 11	1.23.84	1117
659	10	28/05/2014	1200	5	4	017	C W Chang	C K Tong	44	113	9 10 10	1.11.80	1127
556	07	16/04/2014	1200	5	10	017	C W Chang	C K Tong	31	112	8 7 7	1.10.64	1120
527	04	06/04/2014	1200	5	4	015	C W Chang	C K Tong	16	112	6 3 4	1.11.42	1118
482	11	19/03/2014	1200	5	10	019	C W Chang	M Chadwick	55	114	8 9 11	1.11.39	1131
365	09	02/02/2014	1400	5	12	022	C W Chang	K C Leung	41	113	2 3 4 9	1.24.08	1134
317	06	11/01/2014	1400	5	6	025	C W Chang	C K Tong	86	116	1 2 2 6	1.24.14	1133

of different class as far as possible. In terms of regression, assume there are sample *x* and *x'* denoted as feature vectors of a sample space. Its kernel functions can be expressed as

$$K(x,x')=\exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right) \qquad (2)$$

where σ is a free parameters.

Training a support vector machine is equal to project samples into corresponding kernel space and to solve:

$$\min_{\gamma,w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\xi_i \qquad (3)$$

$$s.t.\ y^{(i)}\left(w^T x^{(i)}+b\right)\geq 1-\xi_i,\ i=1,\cdots,m$$

$$\xi_i\geq 0,i=1,\cdots,m \qquad (4)$$

Margin must be less than 1. If a sample a having margin value larger than 1, it will receive a punishment of $C\xi$ where $1 - \xi_i$ ($\xi > 0$). *C* is a cost function to control weight w, making $\|w\|^2$ small enough but most of the samples having a margin value less than 1. Table 4 shows the result obtained by the support vector machine. Support vector machine seems to be having a better performance than random forest by comparing Table 3 and Table 4. In this study, we want to keep the performance of support vector machine and the ensemble feature of random forest. Therefore, a committee machine is constructed by using multiple SVMs.

D. Committee Machine

Each SVM is trained with a similar training set. The final result is produced by voting after all SVM were trained. A better regressor is formed by combining multiple weaker regressors. Figure 1 shows the flow chart of the proposed SVM-based committee machine.

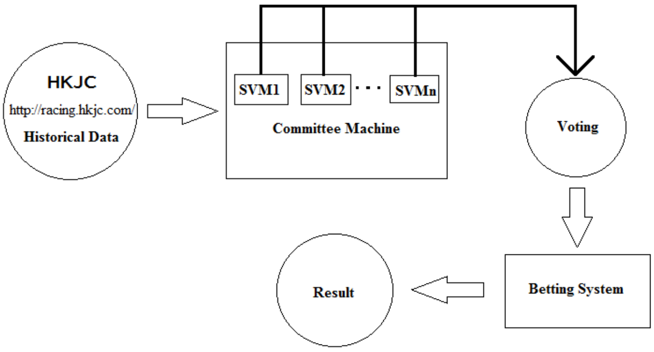


Fig. 1. Flow-chart of committee machine

III. EXPERIMENTAL RESULTS

A. Training and Testing

All race records dated from 30th June 2012 to 30th June 2015 will be tested by a committee machine which is trained with a training set of race records dated from 1st Jan 2012 to 30th June 2012. Result of accuracy and return rate are calculated after the testing.

B. *Comparing on different methods*

In this study, accuracy of different methods of regression has been tested, including random forest (RF), support vector machine (SVM) and committee machine (CM). Table I shows the results of accuracy with single wager. Committee machine is having a higher chance of predicting to winning horse compared to random forest and support vector machine.

TABLE III. ACCURACY OF DIFFERENT METHODS

	Win	Place	Show
RF	25.00%	43.53%	58.10%
SVM	33.52%	55.20%	65.02%
CM	35.84%	54.99%	67.46%

C. *Accuracy on predictions*

In this part, accuracy on predictions is the only consideration. Training data were selected by randomly picking 80% (~360 races) of the total data between the periods from 1st Jan 2012 to 30th June 2012. All data from 30th June 2012 to 30th June 2015 were used as testing data. When T=0.0, testing set contains around 2200 race records, nearly the total amount of races in three years. When T=0.05, acceptable races reduced to nearly 500 races. When T=0.1, there are only 150 races. Table 5 shows the accuracy on predictions, all parameters denoted as follow：

- Win: selected horse resulted in win.
- Place: selected horse resulted in win/place.
- Show: selected horse resulted in win/place/show.
- T: threshold for numeric different of 1st horse and 2nd horse.

TABLE IV. ACCURACY OF PREDICTIONS

		Wager on 1 Horse	Wager on 2 Horses	Wager on 3 Horses
T = 0.0	Win	35.84%	50.77%	61.11%
	Place	54.99%	72.18%	82.21%
	Show	67.46%	83.75%	91.65%
T = 0.05	Win	56.36%	68.29%	76.71%
	Place	71.23%	83.17%	89.82%
	Show	81.99%	91.19%	96.28%
T = 0.1	Win	70.86%	80.13%	85.43%
	Place	82.11%	88.74%	94.03%
	Show	88.74%	93.37%	96.68%

D. *Betting based on predictions*

The return rate was calculated by betting based on predictions. In this study, the initial balance was set at $500. The amount of wages was set in two ways, a static one and a dynamic one. Then the result is compared with a betting strategy which only bet on the horse with lowest win odds. The strategy will be considered as infeasible if the account balance drops lower than zero during the testing. After each test completed, the committee machine will be re-trained. Except the one only bet on the horse with lowest win odds, all of the betting strategies are feasible within 10 successive trials. When T increases, there are less acceptable races. Meanwhile, the system becomes more stable because less bets were made on high risk races.

TABLE V. RETURN RATE

	Avg Static	Avg Dynamic	Lowest Win odds
T = 0.0	8625.2%	840164.1%	-1929.6%
T = 0.05	1856.15%	13692.2%	605.6%
T= 0.1	509.35%	2494.8%	333.6%

E. Result analysis

Since each race has 12.4 horses participated on average, there is only 8% chance to bet on the winning horse if picked randomly. The method suggested in this study is having a better chance of guessing the right horse. Even when T=0.0, which means all races are being bet, there is 35.84% of accuracy. In the case of T=0.1, accuracy is increased to 70.86%. It has been improved by over 700% compared to picking randomly.

For return rate, this study use betting on the horse with the lowest win odds as a control data set. “Theoretically speaking”, lower win odds may imply a higher winning chance. In reality, this study suggested a loss in return rate when T=0.0 (Table 6) by using this strategy. Average return of strategy of static wager value is nearly 8000%. Average return of dynamic wager value is even higher nearly 800,000%.

No matter which strategy that we choose, the average return rates are higher than the control group, particularly in dynamic wager. When picking a higher value of *T*, which means the first horse in predictions have a greater advantage over other horses, it is easier for it to win. Therefore, it resulted in a higher accuracy. But it is a rare situation to happen. Since horse racing in Hong Kong is a handicap and graded race. Horse with better performance is required to carry more lead block. Horses with similar performance will be put together in a race. These two factors inhibit the chance for a horse getting a high advantage.

IV. CONCLUSIONS

In this study, the accuracy of the committee machine for predicting the winner horse in Hong Kong Horse Racing could be as high as 70.86%. This study also suggested a similar tradeoff between accuracy and return rate likely to the study by Robert with his research in Greyhound Races [5].

In terms of the return rate, different strategies may lead to a different result. There would be a lower but more stable return rate of 8000% by using the static wager strategy in a 3 year period. There would be a higher return rate and a higher risk by using the dynamic wager strategy. On the average return rate is nearly 800,000%.

ACKNOWLEDGMENT

Part of this work was supported by Ministry of Science and Technology, Taiwan, under the grants NSC 104-2815-C-224-009-E.

REFERENCES

[1] Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). Cambridge, Mass.: MIT Press.

[2] Breiman, Leo (2001). "Random Forests". Machine Learning, 45 (1), 5-32.

[3] Kecman Vojislav (2001), Learning and Soft Computing, Support Vector Machines, Neural Networks and Fuzzy Logic Models, The MIT Press, Cambridge, MA, 608 pp., 268.

[4] Mitchell, T. (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7

[5] Robert, P. S., Osama, K. S., Hsinchun, C. (2010). Greyhound Racing Using Support Vector Machines: A Case Study. Sports Data Mining, pp 101-108.

[6] Smola, Alex J.; Schölkopf, Bernhard (2004). "A tutorial on support vector regression", Statistics and Computing, 14 (3): 199–222.

[7] Zanjan, G. (2010). Horse Racing Prediction Using Artificial Neural Networks. Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing, 306(13), 155-160.