Seventh International Conference on Intelligent Systems Design and Applications

A Comparative Study on Neural Network based Soccer Result Prediction

Burak Galip Aslan[1]

[1]*Izmir Institute of Technology*
Dept. of Computer Engineering
35430 Urla, Izmir, Turkey
bgaslan@ieee.org

Mustafa Murat Inceoglu[2]

[2]*Ege University*
Dept. of Computer Education and Instr. Tech.
35040 Bornova, Izmir, Turkey
mustafa.inceoglu@ege.edu.tr

Abstract

This study mainly remarks the efficiency of black-box modeling capacity of neural networks in the case of forecasting soccer match results, and opens up several debates on the nature of prediction and selection of input parameters. The selection of input parameters is a serious problem in soccer match prediction systems based on neural networks or statistical methods. Several input vector suggestions are implemented in literature which is mostly based on direct data from weekly charts. Here in this paper, two different input vector parameters have been tested via learning vector quantization networks in order to emphasize the importance of input parameter selection. The input vector parameters introduced in this study are plain and also meaningful when compared to other studies. The results of different approaches presented in this study are compared to each other, and also compared with the results of other neural network approaches and statistical methods in order to give an idea about the successful prediction performance. The paper is concluded with discussions about the nature of soccer match forecasting concept that may draw the interests of researchers willing to work in this area.

1. Introduction

Soccer games are played in a paired fashion and every match can end up as home win, away win or draw. There are expected and unexpected factors that have influence on the result of a match. The more forecasting system can transform unexpected factors into expected factors, the more useful data available for the prediction of the result. Any data can be valuable for prediction; however the key issue is building the forecasting system on proper input data types. There is no direct answer to the question of what should be used as training data for the system.

There have been several studies on the effective use of neural network approaches in forecasting the results of soccer matches. Considering the problem as a classification problem where there is no mathematical model present, the use of neural networks for building forecasting system about soccer matches could be an interesting approach. Although statistical approaches [2], [3], [4], [5] are having been implemented widely on this problem domain, the black-box modeling capability of neural networks is proven to be effective or even better when compared to well-known statistical approaches [1].

Two similar learning vector quantization architectures based on different input vectors have been tested in this study and both of the results are compared with the related neural network and statistical approaches with the same dataset [6] in order to open up a discussion about the selection of input parameters.

2. Related work

There are several statistical approaches proposed for forecasting the results of the soccer matches. The common point of these studies is the usage of *if-then* combinations based on statistical data as data for their forecasting system. The following details have been given for explaining the mechanisms of statistical approaches while considering team A playing against team B, where; team A is home team, and team B is away team.

- Elo system: had been firstly introduced to be used in chess games; however the model has been improved to calculate the probability of the outcome of a soccer match [2].

0-7695-2976-3/07 $25.00 © 2007 IEEE
DOI 10.1109/ISDA.2007.12

Probability of (A wins against B): P_{AB}
Score: Total points of a team

$$P_{AB} = 44.8\% + 0.53\% * | Score_A - Score_B | \quad (1)$$

- The goal-ratio compare model: had been proposed for the prediction of soccer match results. This model relies on the nested *if-else* combinations over the statistical data (goal-ratio data), to produce a result for the possible match up of two teams [3].
GR: Goals scored per match by a team

```
IF | GR_A - GR_B | ≥ 0.3 THEN higher
wins
ELSE IF 0.1 < | GR_A - GR_B | < 0.3
        IF GR_A > GR_B THEN team A wins
        ELSE team A wins or draws
     ELSE team A wins or draws
( | GR_A - GR_B | ≤ 0.1 )
```

- Latest six matches comparison model: is based on the scores of teams at the match date [1]. (Score: Total points of a team)

```
IF | Score_A - Score_B | ≥ 6 THEN higher
wins
ELSE IF | Score_A - Score_B | = 5
        IF Score_A > Score_B THEN team
A wins
        ELSE team A wins or draws
     ELSE IF | Score_A - Score_B | ≥ 2
THEN higher wins
          ELSE team A wins or draws
( | Score_A - Score_B | ≤ 1 )
```

There is also another forecasting approach based on statistical methods tested over the data of Israeli soccer league in the study of Mehrez et al. which is not given in detail in this study [4].

- The study of Cheng et al.: proposes a hybrid (back-propagation + learning vector quantization) neural network approach to the problem as shown in Figure 1. The main idea is exploiting the non-linear mapping capability of neural networks in the domain of forecasting the results of soccer matches [1].

The proposed system first classifies the match into three different categories depending on LVQ's input data which is; $[X_1, X_2]^T$ where;
- AverageScore: Points per match
- AverageNetGoals: (scored –conceded) per match

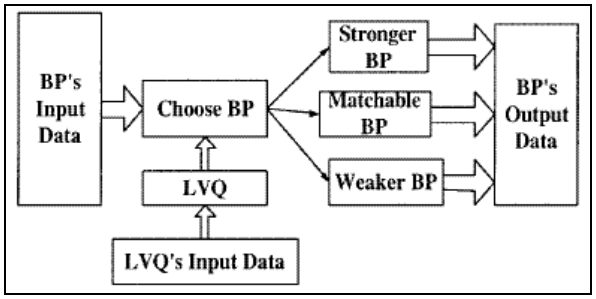

Fig. 1. The forecasting system proposed by Cheng et al. [1]

$$X_1 = AverageScore_A - AverageScore_B \quad (2)$$

$$X_2 = AverageNetGoals_A - AverageNetGoals_B \quad (3)$$

The resultant categories are whether;
- Team A is stronger than team B
- Team A is match able to team B
- Team A is weaker than team B

Depending on the result of the first classification stage, this information is combined with the BP's input data in order to achieve a final prediction for the match. The details of BP's input data are depicted in Table 1.

Table. 1. BP's input vector of the forecasting system proposed by Cheng et al. [1]

$$\begin{bmatrix} Team\ (A) \\ Team\ (*) \end{bmatrix} = \begin{bmatrix} win - ratio\ (A) \\ draw - ratio\ (A) \\ lose - ratio\ (A) \\ average - goal\ (A) \\ average - lose\ (A) \\ morale\ (A) \\ hom\ e - away\ (A) \\ win - ratio\ (*) \\ draw - ratio\ (*) \\ lose - ratio\ (*) \\ average - goal\ (*) \\ average - lose\ (*) \\ morale\ (*) \\ hom\ e - away\ (*) \end{bmatrix}$$

The explanation of morale rating of n^{th} match is shown below in expression 4. The result is total points achieved after a match.

Home – away (X) is also 1.0 for home team, and 0.5 for away team.

$$\text{Morale of } n^{th} \text{ match} = 3 * \text{result } (n-1) + 2 * \text{result } (n-2) + 1 * \text{result } (n-3) \quad (4)$$

All of the studies mentioned above are organized and tested over league-type competitions. There is also a study of Silva et al. for forecasting the results of tournament-type soccer competitions [5].

3. Methodology

Two different learning vector quantization approaches based on different input vectors are introduced in this study, and they are named as LVQ_A and LVQ_B. Both of the methods use the same dataset [6]. Hence their input data preparation section is the same for both of the methods. However, the input vector of LVQ_B is more plain then LVQ_A. The input vector of LVQ_B uses only 2 parameters while the input vector of LVQ_A uses 4 parameters.

3.1. Input data preparation

Italian Serie A, season 2001-2002 dataset [6] has been used for evaluating the performance of networks used in this study. The same dataset has been used in the study of Cheng et al. [1] in order to be compare the results of this study with other neural network and statistical approaches with assumptions and selections below;

- The league consists of 18 teams which also mean that approximately 9 matches are played each week and there are a total of 34 weeks a season.
- The match results of the first 6 weeks are not included in the training dataset considering that random factors play significant role at the start of the season and can distort the training procedure.
- The first half of the season has been used only as training data and every match of second half of the season is predicted with trained neural network. In other words, each week (W) of the second half of the season is tested by using the results up to that week as training dataset;

Most of the forecasting approaches tend to use *visible* statistical data directly as training data. (For example; points of the team, position of the team, etc.) In this study, the previous match results are transformed into a novel formats as explained below.

Each team has two attributes; namely *home rating* (H) and *away rating* (A). These ratings are calculated using a simple increment or decrement over the team rating from the results of matches the team has played. There are 18 teams so J value changes between 1 and 18, and J is integer.

(H_J=0 and A_J=0 as initial condition – season starts)
repeat each week for each team where ($0 \leq J \leq 19$)
  case: (team J) has a draw then no change
  case: (team J) has won at home then H_J++
  case: (team J) has won away then A_J++
  case: (team J) has lost at home then H_J--
  case: (team J) has lost away then A_J--
until the season is over.

3.2. LVQ_A method

As shown in Fig. 2., LVQ_A neural network has an input vector with four attributes where the first pair is the current home rating (H_M), and away rating (A_M) of the home team; and second pair is the current home rating (H_N), and away rating (A_N) of the away team where $0 < \{M, N\} < 19$ and both M, N are integers.

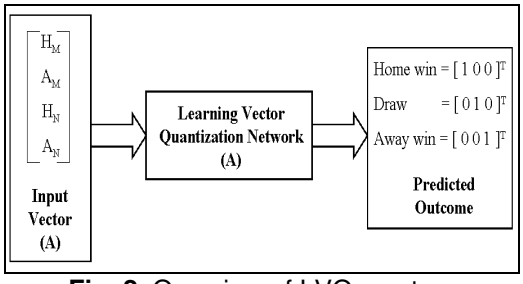

Fig. 2. Overview of LVQ_A system

3.3. LVQ_B method

LVQ_B neural network has an input vector with only two attributes where the first one is the current home rating (H_M) of the home team and current away rating (A_N) of the away team where ($0 < [M, N] < 19$ and both M, N are integers). The overview of the system is given in Fig. 3.

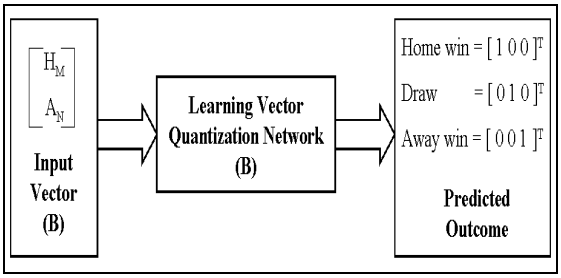

Fig. 3. Overview of LVQ_B system

3.4. Training procedure

The predictions of each week (approximately 9 matches per week) are calculated by taking the results of the past weeks as training data. As the second half of the season consists of 17 weeks, there are 17 different learning vector quantization networks generated due to training. In order to achieve convergence, the number of neurons is chosen as 125 in every LVQ_A network, and 25 in LVQ_B network. The squared error values (training goals) are tried to be kept as low as possible, depending on the training data for each week. The purpose of trying to have a lower training goal is to achieve a better trained network where possible. The contribution of each week to training performance can have direct effect on the convergence performance, so the training goal is decreased whenever possible.

The proposed system in this study also puts in practice a date-based approach rather than a week-based approach used by Cheng et al. [1]. The postponed matches and matches that have been played in different days of the week are all taken into account. It should be remarked that there are a total of 154 matches that have been played in the second half of the season (the matches to be predicted) in reality rather than 153 matches processed by Cheng et al. [1]. The system proposed in this study takes care of the postponed matches and properly transforms the relevant data into input vector attributes.

3.5. Testing

Considering that the weeks are independent from each other, no data for validation is used. Because briefly, having a system train to have minimum error on the predictions last week does not mean that it will predict the matches of next week better.

When a week is evaluated and correct prediction ratio is calculated, that data from that week is added into training pool. As end of the season approaches, the training set consists of more matches. Week by week total successful predicted matches LVQ_A, LVQ_B and [1] are shown below in Fig. 4.

4. Results and discussion

The overall correct prediction performance of proposed LVQ_A network is 51.29%, while overall performance of proposed LVQ_B network is 53.25% even using a simpler input vector. The hybrid neural network proposed by Cheng et al. has a correct prediction rate of 52.29% [1]. The dataset [6] had also been applied to several statistical forecasting approaches and following

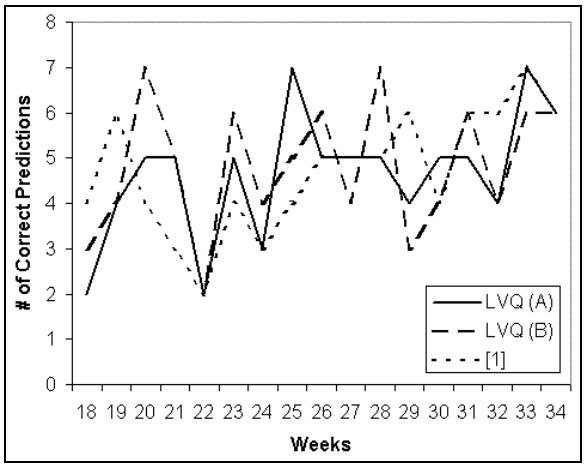

Fig. 4. Test results showing the total number of correct predicted results per week.

results are represented; Elo [2] has a successful prediction rate of 47.71% while goal-ratio compare model [3] has a correct prediction rate of 49.02%.

The study also consists of the performance evaluation of *latest-six matches* approach which has a correct prediction rate of 44.77% [1]. LVQ_B forecasting system proposed in this study has the highest correct prediction ratio with 53.25% which has a better overall prediction performance than the study of Cheng et al. [1], while LVQ_A system again proposed in this study is just 1% behind the study of Cheng et al. [1] with 51.29%. All of the neural network approaches have better correct prediction rates over statistical forecasting methods.

Another interesting point is that even overall correct prediction performance of LVQ_B network is better than both LVQ_A and [1], its success is mostly dependant on the elevated performance in correct home win prediction rates. While LVQ_B network dominates the home win prediction rates, its draw prediction and away win prediction rates are significantly lower than rates in LVQ_B network. A detailed week-by-week comparison of LVQ_A and LVQ_B networks depending on the separate home win, draw, and away win prediction performances are shown in Fig. 5., Fig 6. and Fig 7.

As depicted in Fig. 5.; LVQ_A system has a total of 52 out of 70 correct home win predictions with 74.29% success rate. LVQ_B system has 62 out of 70 correct home win predictions with 88.57% success rate.

As shown in Fig. 6.; LVQ_A system has a total of 11 out of 47 correct draw predictions with 23.40% success rate. LVQ_B system has 8 out of 47 correct draw predictions with 17.02% success rate.

Given in Fig. 7.; LVQ_A system has a total of 16 out of 37 correct draw predictions with 43.24% success

rate. LVQ_B system has 12 out of 37 correct draw predictions with 32.43% success rate.

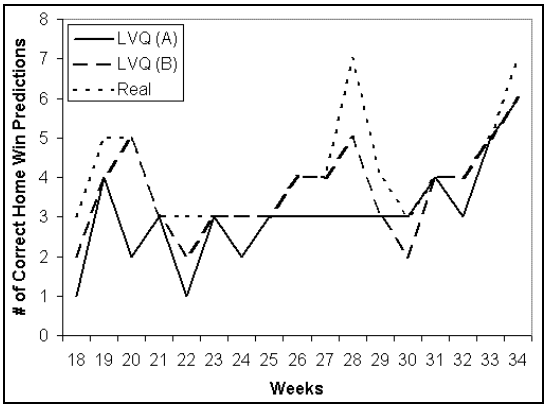

Fig. 5. Test results showing the total number of correct home win predictions

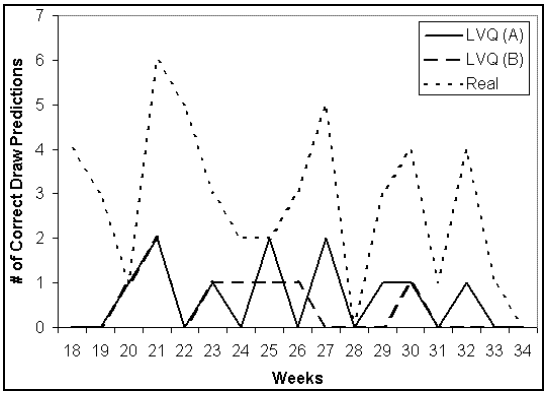

Fig. 6. Test results showing the total number of correct draw predictions

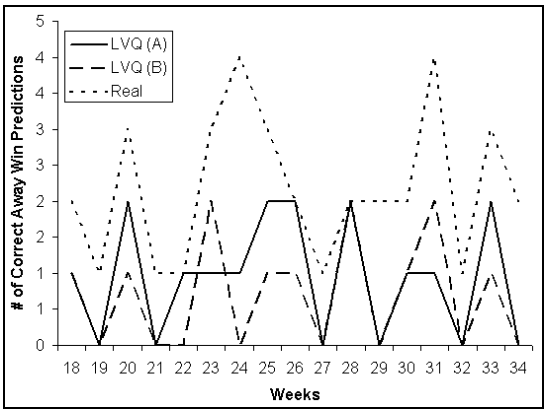

Fig. 7. Test results showing the total number of correct away win predictions

It can be seen from Fig. 8. that while LVQ_B network has a slightly better performance with home win prediction rates; LVQ_A network has a more balanced prediction distribution on all of the matches.

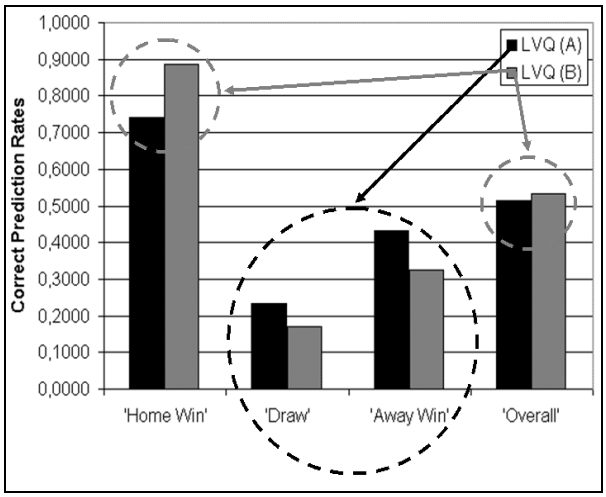

Fig. 8. Home win, draw, and away win performance comparisons of LVQ_A and LVQ_B

5. Conclusion

This study proposes two alternative approaches for the prediction of the results of league based soccer matches. One of the proposed systems (LVQ_B) has achieved a better prediction performance over the study of Cheng et al. [1]. The systems proposed in this study, namely LVQ_A and LVQ_B, both based on learning vector quantization networks using different input parameter vectors. Several interesting contributions of this study can be remarked as follows:

- Both LVQ_A and LVQ_B methods proposed in this study supports the idea that neural network approaches can be used for forecasting the results of the soccer matches effectively. It has been shown that the overall prediction performance of neural network approaches can be better than statistical forecasting methods.
- The selection of input data to be used in forecasting systems is a critical issue. The available data should be transformed into alternative formats in order to be used in the neural nets for better prediction performance. It may not be necessary to apply any available data in the form of input parameters, because masses of data in input parameters might lead to a problem in convergence.
- The handling of postponements and exercising a data-based approach is a very important issue that should not be overlooked considering that even one or two matches can make difference on the overall prediction performance. Plus, without using a date-based approach the values of input parameters can be

distorted as well, leading the system to learn misplaced results.

- This study also remarks that; better overall correct prediction performance does not directly mean a balanced distribution of performance over three possible outcomes (home win, draw, away win). A system that has a better overall performance (e.g. LVQ_B) can even have significantly poor performances in some type(s) of match results. Considering the dominance of home wins in soccer games; the main idea of soccer match result forecasting studies in future may be focused on the evaluation of *well-balanced prediction performance*, rather than only evaluating the overall performance.
- The last but maybe the most important contribution of this study is; although several applications based on neural networks for forecasting the results of soccer matches are available on the internet, there is serious lack of scientific papers about this issue. Of course the scientific importance of soccer match result forecasting could be another debate, but it is clearly out of the scope of this study. However if a researcher considers the problem as a valuable area to work on, the results achieved in this study might have been useful. Because there are only few numbers of scientific studies over this issue.

It would not be surprising to emphasize that a lot of work should be done in this area to explore further steps in this complex domain of soccer match result forecasting. Different leagues, different input parameters and of course different network structures should be tested in order to achieve a well-balanced generic forecasting system.

6. References

[1] Cheng, T., Cui, D., Fan, Z., Zhou, J., Lu, S.: "A New Model to Forecast the Results of Matches Based on Hybrid Neural Networks in the Soccer Rating System", *Computational Intelligence and Multimedia Applications, ICCIMA 2003 Proceedings* (2003) 308-313

[2] Elo, A.E.: "The Rating of Chess Players, Past and Present", *Arco Publishing*, New York (1978)

[3] Jackson, D.: "The Parameter Gaming in Sports", *Proceedings of the International Gaming Conference* (1990)

[4] Mehrez, A., Hu, M.Y.: "Predictors of the Outcome of a Soccer Game - a Normative Analysis Illustrated for the Israeli Soccer League", *Mathematical Methods of Operations Research, Volume 42, Issue 3* (1995) 361-372

[5] Silva, C.F., Garcia, E.S., Saliby, E.: "Soccer Championship Analysis Using Monte Carlo Simulation", *Proceedings of the 2002 Winter Simulation Conference*, (2002) 2011-2016

[6] http://www.soccerway.com/national/italy/serie-a/2001-2002/round-1/results (March 2007)