International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 10 (2017), pp. 2431-2438
© Research India Publications
http://www.ripublication.com

Computational Intelligence Based Sports Success Prediction System Using Functional Pattern Growth Tree – A Case Study

M Sudha

Associate Professor, School of Information Technology and Engineering
Information Technology Department Vellore Institute of Technology University - India

Abstract

It is ɑ known fact that ɑ country like India doesn't perform as well in sports as it would like to. This is evident from the medal count of the country at the highest levels of sports such as the Olympics. Countries with far lesser population and GDP have often won more accolades at the same competition than India. However, Countries like the U.S.A. and China have consistently won medals across disciplines. A country's sporting prowess is ɑ matter of both pride for its people and reputation in front of other countries. The intent of this study is to find any possible relations (if they exist) between various factors which are indirectly related to and their influence on the "Success of Sports" in ɑ particular country using the proposed computational intelligence based sports success prediction system (CISP).

INTRODUCTION

Now a day's Sports has a more commercial insight and as a whole depends upon audience viewing it. An educated person has a better sense of decision making and since sport hinges on the decisions made by players their education is also very important. A country participating in sports trade generally can be seen as a country which gives importance to the sport and thus by giving importance thrives to succeed in sports. Frequent Pattern Analysis of Data mining helps in finding the frequently occurring patterns in a given data sets. By taking all these attributes along with the number of medals won by a country we can create a custom data set which mimics the

Market Basket analysis format. Then by using required algorithms patterns can be obtained from such non intuitive factors of sports and a new understanding of a sport as a whole occurs.

In [1] the current data regarding sports has been analysed and data mining techniques was applied to improve the performance in sports. They applied data mining technique to predict the number of points made, SNM, and mean and SD to minimise the data. Thereby to visualize the data (performance) they used parallel axes and histograms. In [2] they suggested a new approach for pattern analysis and data classification. They adopted the network to modify the existing methods which helps for better efficiency. In [3] author overviewed the data mining techniques that are existing over the decades and also how to extract the different kind of patterns i.e. sequential patterns, maximum pattern mining, closed pattern mining and short patterns. Author proposed an algorithm to find the efficiency of pattern to overcome low-frequency problem in data mining. This can be done by two process pattern deploying and pattern evolving.

In [4] the authors discussed the brief outlook of the present status and also the future approaches to determine the frequent patterns. They discussed in detail about the current application of frequent patterns like DNA reorganization, for building better classification relations and deep understanding and interpretations of patterns help us in semantic notations. The importance of data mining and its applications regarding finding patterns in domains like forecasting, business and transportation etc. [5]. In [6] the authors analyses about the problems existing in physical training and sports and suggested how to overcome those difficulties by using data mining techniques. They developed a data mining algorithm to identify existing in practical problems and to overcome them by applying the critical neural networks for forecasting the performance of the sportsmen. Data-Information-Knowledge-Wisdom (DIKW) hierarchy. The DIKW framework then sets the stage for disambiguating data from knowledge and sets definitional boundaries for what data, information and knowledge are. Applying this to the sports domain, certain activities and techniques operate at the data level [7].

In [8] the authors introduce the Open source development, which has become more prominent in recent years in a multitude of software areas. In the domain of data mining tools, several solutions have gained significant acceptance such as Weka and Rapid Miner. Both tools share the same underlying learning algorithms, however, their approach to displaying results, are very much different. A novel text mining problem, which is refer to as Comparative Text Mining (CTM). A set of comparable text collections, the task of comparative text mining is to discover any latent common themes across all collections as well as summarize the similarity and differences of these collections along each common theme[8]. In [10] the author current world sports and its statistical information Since the relationships between sports results and various data elements are directly affected by several factors such as type of sports, the environment, and the objectives of players, several methods have been suggested to predict the results based on available data. The research paper reviewed a number of game result prediction systems based on data mining techniques. It included data collection and feature selection methods, classification techniques and the results, advantages and disadvantages of the selected systems.

BACKGROUND AND DATA SET Description

A lot of research, has always been in sports but the domain of research is restricted to sports personnel, managers and teams but never has it taken into consideration about the amount of people attending sporting events, how education plays a part in sport, how gender distribution in sports contributes to success and to how sports contribution to the GDP of a country helps in the country being a successful. European Union keeps collection of all such data and by analysing the patterns for EU countries new understanding of sports can be obtained. Data for the research was referred from the European Union Sports Statistics from *Knoema.com*

CISP - MODULES

The data obtained from the above source is raw data, it contains missing data, redundancy, lacking in certain behaviour and trends and conflicting values and attributes etc.

Therefore in order to feed the data in the algorithm and to get the required results first we need to process the data.

CISP- Data Pre-Processing

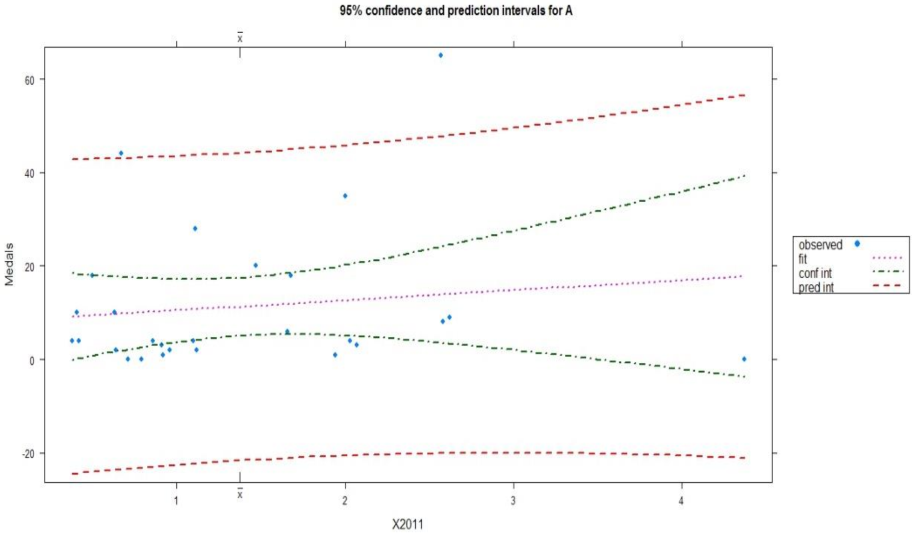
Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. Scikit Learn was used to pre-process the data in python.

Missing Data was filled by taking the ‘average value’ of the available data. This process is called Data Cleaning in which the Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data. Redundant data was ‘eliminated’ by the process of Data Discretization involving the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals. Few Data Attributes were merged and conflicts were resolved by ignoring the conflict values using Data Integration in which Data with different representations are put together and conflicts within the data are resolved.Data Minimization was done by dividing by means involving the method of Data Reduction which step aims to present a reduced representation of the data in a data warehouse. Data binning that involves grouping number of more or less continuous values into a smaller number of "bins" and Data classification into categories for its most effective and efficient use done by Quartile Deviation.

CISP- Data Visualisation

Using R programming language we plotted the linear regression of the processed data. Sample plotted graphs are shown below.

- Percentage of Employment 15-29 age **VS** Medals



CISP- Finding Relation between the Patterns

In this study Frequent Pattern (FP) Growth tree is applied for the pattern analysis**,** this is a modelling technique to find the associations and connections between the attributes. Python libraries have been used to implement the FP Growth Algorithm on the processed dataset obtained after pre-processing data. Basically FP-tree is a compact data structure that represents the data set in tree form. Each transaction is read and then mapped onto a path in the FP-tree. This is done until all transactions have been read. Different transactions that have common subsets allow the tree to remain compact because their paths overlap.

RESULT AND DISCUSSIONS

1. ['10-64 medals', 'From 1 to 6 times->Males->15.3% to 10%']

Countries have won a high amount of medals in Olympics in 2012 when 15 to 10 % of the male population of the country attend sporting events around 6 times annually.

2. ['10-64 medals', 'More than 6 times->Females->15.3% to 10%']

Countries have won a high amount of medals in Olympics in 2012 when 15 to 10 % of the female population of the country attend sporting events around 6 times annually.

3. ['4-10 medals', 'At least once->Females->15.3% to 10%']

Countries have won around 4 to 10 medals in Olympics in 2012 when 15 to 10 % of the female population of the country attend sporting events at least one.

4. ['4-10 medals', 'At least once->Total->15.3% to 10%']

Countries have won around 4 to 10 medals in Olympics in 2012 when 15 to 10 % of the total population of the country attend sporting events at least one.

5. ['4-10 medals', 'From 1 to 6 times->M->Y25-34->15.3% to 10%']

Countries have won around 4 to 10 medals in Olympics in 2012 when 15 to 10 % of the age group 24 to 34 of males of the country attend sporting events on an average 6 times.

6. ['4-10 medals', 'From 1 to 6 times->M->Y25-44->15.3% to 10%']

Countries have won around 4 to 10 medals in Olympics in 2012 when 15 to 10 % of the age group 24 to 44 of males of the country attend sporting events on an average 6 times.

7. ['Tertiary education (levels 5-8)->high employment', '4-10 medals']

Countries have won around 4 to 10 medals in Olympics 2012 when sports men have had tertiary education and the country has high employment in sports.

8. ['Upper secondary and post-secondary non-tertiary education (levels 3 and 4)->mediumemployment,4-10 medals']

Countries have won around 4 to 10 medals in Olympics 2012 when sports men have had Upper secondary and post-secondary education and the country has a moderate amount of employment in sports.

9. ['Exports->trade with WORLD->'MediumPercentOfTrade', '10-64 medals']

Countries won in between 10 to 64 medals in 2012 Olympics when they have had dedicated Moderate amount of their GDP in exports trade of sports goods with the world.

10. ['Imports->trade with INT_EU28->'mediumPercentofTrade', '10-64 medals']

Countries won in between 10 to 64 medals in 2012 Olympics when they have had dedicated Moderate amount of their GDP in imports of sports goods within the EU.

CONCLUSION

The investigation is an experimental study and the previous research done in this field was either for a particular sport, or the attributes were selected without mining. The data-set which was used, yielded results with 60% support. The results were non-conclusive. It is most encouraged research and the conclusive results depend on the availability of more qualitative and quantitative data.

References

[1] De Marchi, L. (2005). Aangústia do formato: umahistória dos formatosfonográficos. *revista e-Compós*, (2).

[2] Angulo, P. (2002). Nonalcoholic fatty liver disease. *New England Journal of Medicine*, *346*(16), 1221-1231.

[3] Zhong, N., Li, Y., & Wu, S. T. (2012). Effective pattern discovery for text mining. *IEEE transactions on knowledge and data engineering*, *24*(1), 30-44.

[4] 4.Harinarayan, V., Rajaraman, A., & Ullman, J. D. (1996, June). Implementing data cubes efficiently. In *ACM SIGMOD Record* (Vol. 25, No. 2, pp. 205-216). ACM.

[5] Rathod, D. K., & Valmik, N. Overview of Data Mining Techniques.

[6] Meng, F. H., & Li, Q. L. (2013). Application of data mining in the guidance of sports training. In *Advanced Materials Research* (Vol. 765, pp. 1518-1523). Trans Tech Publications.

[7] Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). Sports knowledge management and data mining. *Annual review of information science and technology*, *44*(1), 115-157.

[8] Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). Open Source Data Mining Tools for Sports. In *Sports Data Mining* (pp. 89-92). Springer US.

[9] 9 Zhai, C., Velivelli, A., & Yu, B. (2004, August). A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth*

ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 743-748). ACM.

[10] 10. Haghighat, M., Rastegari, H., &Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal*, *2*(5), 7-12.

[11] https://www.researchgate.net/.../What_are_the_applications_of_data_mining_in_the_s

[12] www.springer.com/in/book/978144196729

[13] https://ww2.coastal.edu/kingw/statistics/R-tutorials/simplelinear.html

[14] https://www.r-bloggers.com/simple-linear-regression-2/

[15] www.statmethods.net/stats/regression.htm

[16] https://www.analyticsvidhya.com/blog/tag/market-basket-analysis/machinelearningmastery.com/

[17] www.albionresearch.com/data_mining/market_basket.php

[18] https://vlebb.leeds.ac.uk/bbcswebdav/orgs/SCH_Computing/.../de_marchi.pdf

[19] https://arxiv.org/abs/1211.5723

[20] https://vvvvw.aaai.org/Papers/Symposia/Spring/2002/SS-02-06/SS02-06-013.pdf

[21] https://vvvvw.aaai.org/Papers/Symposia/Spring/2002/SS-02-06/SS02-06-013.pdf