A learning-based system for predicting sport injuries

Guangying Liu[1], *Hua* Sun[2,*], *Wanjian* Bai[3], *Hongmei* Li[3], *Zhigang* Ren[3], *Zhongde* Zhang[4], and *Lingxia* Yu[2]

[1]State Grid Corporation of China, Beijing, China
[2] Shandong Luneng Sports & Culture Company, State Grid Shandong Company, Shandong, China
[3]State Grid Shangdong Electric Power Company, Shandong, China
[4]Shandong Luneng Software Technology Co., Ltd, Shandong, China

Abstract. In the big data era, learning-based techniques have attracted more and more attentions in many industry areas. The sport injury prediction is one of the most critical issues in data analysis of soccer teams. However, learning-based methods have not been widely used due to the poor data quality and computational capacity. In this paper, we propose a learning-based model to forecast sport injuries according to the data from various information systems. We first reduce the attributes that have significant impact on the injury risk by using learning-based methods. Then, we provide an algorithm based on the random forest method to prevent the over-fitting problem. We have evaluated the proposed model with the real-world data. The experimental results show that our model works efficiently and achieves low error rates.

1 Introduction

Sports injuries can affect any and all parts of the body depending on the particular repetitive movement performed just like any repetitive motion injury. While there are factors that raise the risk of injury, there are also elements that predispose athletes to sports injuries. Rehabilitation and preventative efforts should be centered on a thorough knowledge of risk factor etiology as well as knowledge of how such factors contribute to sports injuries.

Many different methods are used to describe the stimuli of motor injury. These include interviews with injured athletes, video analysis of actual injuries, clinical studies (studying the clinical manifestations of joint injuries, understanding the mechanisms of injury, mainly by X-ray, magnetic resonance imaging, arthroscopy, CT scans), and in recent studies (Measurements to understand ligament loading patterns are ligament strain or strength), cadaver studies and simulations in the case of injury, measured / estimated from "close to injury."

In this paper, we propose a learning-based model to forecast sport injuries according to the data from various information systems. We first reduce the attributes that have significant impact on the injury risk by using learning-based methods. Then, we provide an algorithm based on the random forest method to prevent the over-fitting problem. We

[*] Corresponding author: ls_sunht@163.com

evaluate the proposed model and algorithms with the real-world data. The experimental results show that our model works efficiently and achieves low error rates.

The remainder of this paper is organized as follows. Section 2 reviews the recent studies related to the topic. A comprehensive framework for injury risk analysis in smart grid is provided in Section 3. Section 4 presents useful feature analysis methods for decision making and multi-variables reduction and classification methods. The paper is concluded in Section 5.

2 Related Work

Please follow these instructions as carefully as possible so all articles within a conference have the same style to the title page. This paragraph follows a section title so it should not be indented.

In today's rapidly developing world, thanks to new technologies, the generated data volume is growing rapidly. In term of both software and hardware, provide data collection from different resources. Consequently, traditional data analysis methods are inapplicable. There is a major need to process data in a more intuitive and more effective way. In this case, data mining [1] is proposed as a new data processing method. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases[2] . Fayyad defines data mining as ``a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database" [3]. While Giudici defines it as ``a process of selection, exploration and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database "[4]. Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Some common tools for prediction include: neural networks, regression, Support Vector Machine (SVM), and discriminant analysis. [5].

There are lots of papers exist in the field of utilizing data mining algorithms to achieve the purpose of prediction, for example, Mr. Arning et al. presented a system which utilize the data mining satisfies the need of prediction, a computer program product, and an associated method, including a user friendly interface. In this paper, I think the place that is worthy of our reference is that the data mining method to be employed for the purpose of prediction is selected automatically. The automatic selection is performed depending on the data type of the column that is selected for the prediction[6]. And the article of Folriz Amooee applied various prediction algorithms for data analysis. Different kinds of trees such as CHAID, C\&R, and QUEST along with other prediction algorithms including neural networks, Bayesian, logistic regression, and SVM has been applied on data. Investigating and comparing each algorithm's accuracy[7]. In the another paper of Sellappan Oalaniappan and Rafiah Awang, they presented an Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, *Decision Trees*, *Naive Bayes* and *Neural Network*. IHDPS can discover and extract hidden knowledge associated with heart disease from a historical heart disease database[8]. There is a paper presented a data mining procedure based on association rule mining for extracting relationships among climate parameters, and the procedure can be applied to extract the intense summer day (hot day) patterns during summer months. By anticipating the extreme summer temperature, the day to day practice will be planned in advance based on human comfort [9]. Mr. Alzghoul et al. used three classification algorithms to investigate the performance and products availability of each algorithm [9]. In addition to this paper, other researches also used data stream mining for machine monitoring and reliability analysis[10], online failure prediction[11] and tool condition monitoring[12].

3 System Architecture

The text of your paper should be formatted as follows:

The popular learning-based method have not been widely used due to poor data quality and computational capacity. In this paper, we provide a comprehensive learning-based model for sport injury prediction. A five-stage damage risk classification workflow is proposed, but the main concern is the implementation phrase of the clustering model. In this paper, we construct a seven phases of a comprehensive model, including data, data quality assessment, data processing, feature analysis, classification model establishment, prediction and results shown in Figure 1.

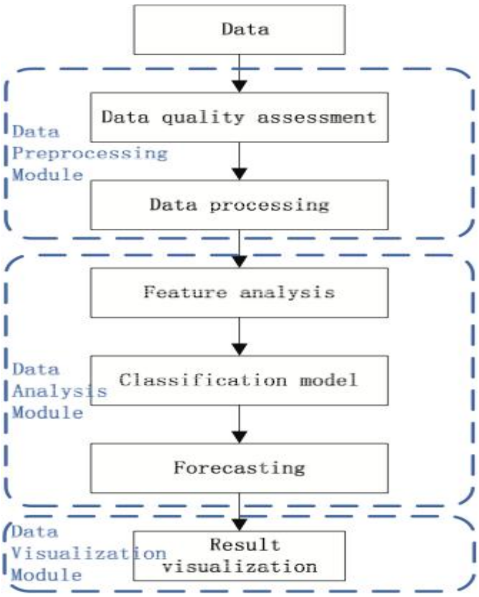

Fig. 1. The process of the sport injury prediction.

Our seven-stage workflow can be categorized into three sub-modules, such as data preprocessing, data analysis and results visualization. For data preprocessing, its first step is data collecting from various information systems. In a more realistic situation, the quality of the real data is poor, so data quality assessment is necessary. Currently, we only check the missing value rate of data set. In the future, a load data quality assessment model will be provided. High-quality data often performs well in a learning-based model. This article provides detailed data processing statements. There are three main operations, including discrete attributes to nominal values, filling in missing values and sample values.

For data analysis submodule, it includes feature analysis, building classification model and forecasting. The sport data contains a large number of attributes. However, many attributes are irrelevant or redundant to the classification. They think that the performance of the classification algorithm is subtly reduced: (a) the efficiency is greatly reduced; and (b) the prediction error is caused.

Compared the proposed association rule mining (arm) technology with existing feature selection and analysis methods, there are three reasons for this: (1) ARM uses frequent itemsets, which are efficient and easy to use for big data. (2) We want to extract the multi-attribute from the association between different information systems. (3) Rule-based mining rules support decision makers and visualization.

With respect to building classification model and forecasting, we compare the learning-based methods with traditional time series and intelligent methods, the former is more suitable for our data. Then we further compare the well-known learning algorithms: SVM, bagging and Random Forest though experiments. Finally, we choose Random Forest method in this paper to prevent over-fitting of imbalanced HOL data and gains high accurate.

In recent years, the data visualization submodule has attracted a lot of attention in the big data era. The knowledge discovered from the original data is required to be presented in a proper way to users, especially for decision makers. Visualization performs well of making large data sets better accessible using techniques like selecting and zooming.

4 The methods

4.1 System realization method

The system that we presented in this paper mainly utilized the MVC (*Model-View-Controller*) architecture, it divides the system into three types of modules: Controller, Model and View. Concretely, the Model deals with the logic and data of application, and it is also responsible for updating the information of View and receiving the commands from the Controller. To put it simply, model can transfer the data from database according to the request from the Controller and display the data in View; the View is responsible for the presentation of data information; the Controller is in charge of input from users including the keyboard or mouse events in system interfaces, and notifies Model using events.

4.2 Feature extraction method

Sections should be numbered with a dot following the number and then separated by a single space:

The Apriori algorithm is mainly mining the frequent patterns, associations, and correlations of the data set. In this paper, we need to extract all the features of the training data set which have strong association with the injury risk by utilizing the Apriori algorithm. The pseudo code of Apriori algorithm and mining association rules from the frequent itemsets are shown in the Algorithm 1.

Algorithm 1. The algorithm for mining association rules of the dataset.

Input: D, a database of transactions; min_sup, the minimum support count threshold; min_conf, the minimum confidence threshold.
Output: Ruleset, asscociation rules of database D.

```
$L_1$=find_frequent_1-itemset($D$);
for ($k = 2; L_{k-1}! = \emptyset; k$++)
  $C_k$=$L_{k-1} \propto L_{k-1}$;        // join step: generate candiates
  for each $c \in C_k$           // prune step: remove unfruitful candidates
    for each subset $s$ of $c$
      if $c \notin L_{k-1}$ then delete $c$ from $C_k$;
    end for
    get counts of each itemset $c \in C_k$by scanning $D$;
  end for
  $L_k = \{c \in C_k | c.count \geq min\_sup\}$;
for each frequent itemset $l \in L$
  $S$=subset($L$, $l$); // get the subsets of $l$ that are frequent itemsets
  for each subset $l \in S$ and
    conf = support_count($l$) / support_count($s$);
    if (conf ≥ min_conf)
      print the rule $r$:$s \Longrightarrow (l - s)$;
```

```
        add $r$ to RuleSet;
    end for
return RuleSet;
```

4.3 Classification model based on random forest

Injury classification is to partition various patterns into groups. There are many different models which can be categorized into traditional time series methods and Intelligent methods. The drawbacks of traditional methods such as linear regression, time series model (ARIMA, exponential smoothing) are that it only takes time sequence features into consideration, and it is difficult to deal with non-linear nature of patterns. Thus the Intelligent methods have been practiced, such as expert system and artificial neural networks (ANN). The ANN forecasting model gives better performance with nonlinearity issue, but most of the NN models adopt the gradient descent based back-propagation learning scheme to minimizes the mean square error during training process. The error on the training data set is good, but performs badly when out-of-sample data is presented to the network, which yields limited generalization capability.

Recently, the machine learning methods are applied into this field. The widely utilized algorithm are Support Vector Machine(SVM) and decision trees. SVM is to find a maximum-margin hyperplane which separates the n-dimensional data perfectly into its multi classes, when dealing with nonlinearly separable problems, it is often computationally expensive and easily over-fitting. The tree model, such as CART and ID3, is over-fitting easily when dealing with unbalanced data set. Because it only builds a finite generalization of the decision tree. Therefore, an integrated learning method with good generalization ability and precise ability is proposed. Integrated learning such as Boosting, Bagging and Random Forest is a combination of a variety of tree predictor. On the basis of adding trees, the forecaster noticed the point at which the early predictors mispredicted, and the points of the mispredictions were given the additional weight of continuous tree training. Finally, the weighted voting to predict. In bagging, successive trees constructed based on a bootstrap sample of data independently instead of depending on earlier trees. And the majority vote is taken for prediction in the end.

While, Random Forest is another ensemble learning method with little difference. Firstly, bootstrap samples from the original data as bagging. For each of the bootstrap samples, it changes how the classification trees are built, at each node, instead of using the best split among all variables as standard trees, using the best among a randomly chosen subset of predictors at that node. It turns out that this counterintuitive strategy has better performance than classifiers such as support vector machines and neural networks, and is robust to over-fitting. Finally, use majority voting to classify or average the regression as a prediction of bagging. In addition, the generalization error of the forest converges to a limit because the number of trees in the forest becomes larger.

5 Conclusion

In this paper, we designed a system which can conveniently manage and maintain the data information of soccer players, and also can predict the injury risk based on the collected historical data sets by using the data mining tools. We propose a learning-based model to forecast sport injuries according to the data from various information systems. In the framework, we first reduced the attributes that have significant impact on the injury risk by using learning-based methods. Then, we provided an algorithm based on the random forest method to prevent the over-fitting problem.

Although this system has achieved good results, but there are still many problems which need to be improved in future work. For example, we should try multiple different data mining algorithms and compare the accuracy and efficient of these algorithms to choose the most efficient algorithm to predict. Besides, there are lots of issues need to be improved in the design of the system, for example, we should improve the database that supporting the system, because the correlation between the tables in the database is not strong enough, and the relevance among these tables in the reality is not reflected well.

References

1. Jiawei Han and Micheline Kamber. *Data Mining Concept and Techniques*. (2012.)
2. Bhavani Thuraisingham. A primer for understanding and applying data mining. *It Professional*, 2(1):28–31, (2000).
3. Usama M. Fayyad. Data mining and knowledge discovery in databases: Implications for scientific databases. In International Conference on Scientific and Statistical Database Management, 1997. *Proceedings*, p 2, (1997).
4. Richard J. Cleary. Applied data mining: Statistical methods for business and industry. paolo giudici. J. of the American Statistical Association, 101(September):1317–18, (2006).
5. E. W. T. Ngai, Li Xiu, and D. C. K. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications An International Journal, 36(2):2592–602, (2009).
6. Andreas Arning, Martin Keller, Christoph Lingenfelder, and Gregor Meyer. *System and method of using data mining prediction methodology*,(2006).
7. Golriz Amooee, Behrouz Minaeibidgoli, and Malihe Bagheridehnavi. A comparison between data mining prediction algorithms for fault detection (case study: Ahanpishegan co.). Int. J. of Computer Science Issues, abs/1201.6053(6), (2012).
8. Sellappan Palaniappan and Rafiah Awang. Intelligent heart disease prediction system using data mining techniques. In Ieee/acs International Conference on Computer Systems and Applications, pp 108–15, (2008).
9. Ahmad Alzghoul and Magnus Lædˉ Žfstrand. Increasing availability of industrial systems through data stream mining. *Computers & Industrial Engineering*, 60(2):195–205, (2011).
10. S. Cem Karacal. Mining machine data streams using statistical process monitoring techniques. (2007).
11. Roger K. Youree, Jeffrey S. Yalowitz, Aaron Corder, and Teng K. Ooi. A multivariate statistical analysis technique for on-line fault prediction. In International Conference on Prognostics and Health Management, pp 1–5, (2008).
12. C. Karacal, Sohyung Cho, and W. Yu. Sensor stream mining for tool condition monitoring. In International Conference on Computers & Industrial Engineering, pp 1429–33, (2009).