www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 7 July 2015, Page No. 13319-13323

Probabilistic Determination of Class for an Unbalanced Dataset

Samadhan B. Patil[1], Vaishali Ghate[2], Dhanashree Tarodmalle[3]

[1] Department of Information Technology, Shah and Anchor Kutchhi Engineering College, Mahavir Education Trust Chowk, W.T.Patil Marg, Chembur, Mumbai-400088, India.
samadhanpatil81@gmail.com

[2] Department of Information Technology, Shah and Anchor Kutchhi Engineering College, Mahavir Education Trust Chowk, W.T.Patil Marg, Chembur, Mumbai-400088, India.
vaishalighate@yahoo.co.in

[3] Department of Information Technology, Shah and Anchor Kutchhi Engineering College, Mahavir Education Trust Chowk, W.T.Patil Marg, Chembur, Mumbai-400088, India.
sakec.dhanashreet@gmail.com

Abstract: ***Poker is one of the world's most popular and widely played card games. In Poker, there is a fixed set of winning conditions and the player with the highest winning condition wins the game. The main part of the game is to bet strategically and in a calculated manner so that there is less chance of risk and the opponents are not able to guess the cards in the hand. To help players understand when and how to bet smartly, this application will be developed. This system provides knowledge to the users about their probability of winning based on the cards available to them. The system which has been developed is lightweight and easy-to-use so that all types of players can use it. The aim of this system is to help gamblers bet better thereby increasing their winnings, addiction to Poker gambling and also generate greater revenue collections for gaming consortiums.. The most important point of this paper is to show how we have used data mining and statistical probabilities to formulate an algorithm which gives out correct predictions of the winning hand. We formally define the system and outline the challenges that arose while developing technology to support it. We hope that this paper will encourage more research by the gaming consortiums and the gambling community in this exciting area of winning by probability calculations and card counting.***

Keywords: Data mining, Machine learning, Poker, Winning Probability, Naïve Bayes algorithm.

.

1. Introduction

Data mining is the computational process of discovering patterns in large data sets can also be defined as the extraction of hidden predictive information from large databases. The overall goal of the data mining process is to extract information from a data set and transform it into an easily understandable structure for further use by various skilled users which involves database and data management aspect, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, online updating, but also visualization.

Data mining is a powerful technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions without large dependence on older methods like focus groups. Poker is a game that caught the interest of the AI research community in the last decade. Poker game state is hidden: each player can only see his cards or the community cards. It is only at the end of each game that opponents may show their cards, thus being much more difficult to understand how the opponent plays. Poker is also a stochastic game, i.e., it admits the element of chance since the player cards are randomly dealt. The following are the most important properties of poker:-

1) Imperfect information - This property creates a necessity for using and coping with deception and ensures a theoretical advantage of using randomized mixed strategies.

2) Non-deterministic dynamics - This means that the cards we get are stochastic.

3) Partial observable - Players can't always know the opponent's hole cards, even when a game is over.

4) Multi-players- There are at least two players.

There are 6 popular types of Poker that are played world-wide [1]:-

Omaha, 7-Card Stud, 5-Card Draw, High / Low Chicago, Follow the Queen and *Texas Hold 'em*.

We made use of *Texas Hold'em* Poker as the game for our system. Played in the World Series of Poker, Texas Hold 'Em is easily the most popular poker game. In Texas Hold 'Em, players are dealt two "pocket" or "hole cards" then wait for 5 community cards to be revealed. Betting takes place in four rounds: once after the hole cards are dealt, once after the first

three community cards are revealed (referred to as “the flop”), once after the fourth community card is revealed (“the turn”) and lastly after the fifth community card is flipped (“the river”). A showdown occurs after the river where the remaining players reveal their hole cards and the player with the best hand wins all the wagers in the pot. If two or more players have the same best hand then the pot is split amongst the winners. Players must make their best hands with any combination of 5 cards (their hole cards and the communal).

.

2. Literature Survey

Machine learning [2] investigates how computers can learn (or improve their performance) based on the data. Machine learning focuses on prediction, based on known properties learned from the training data. Poker [3] is usually played with a standard deck of 52 cards. Each card is marked with one of 13 face values and one of 4 suits. In a common version of poker, a player receives a hand of five cards. Hands that match certain combinations, or patterns, have specific names like "FULL HOUSE" or "ROYAL FLUSH". BetOnline[4] is the most popular of all real- money poker sites those are available to U.S. players. PokerTrackerSoftware [5] LLc is the name of a poker tool software company that produces the popular PokerTracker line of pokertracking and analysis software. Pokertracker’s software imports and parses the hand histories that poker sites create during online play and stores the resulting statistics/information about historical play into a local database library for self-analysis, and for in-game opponent analysis using a real-time Head-up display. It calculates and graphs statistics such as hands per hour, winnings per hand, wins per hour, cumulative profit and loss, and individual game profit and loss across multiple currencies. A poker sites calculator is an application that lets you run any scenario that you see at a poker table. Once you say what cards you have, and what cards other players have, the poker calculator will go to work and, in a matter of seconds, tell you what your odds of winning are. There are no guarantees but, in the long run, using the kind of statistical information you get from a poker odds calculator can give you a real edge over players that don't realize what they're missing out on. PokerListings.com's Odds Calculator[6] is the fastest, most accurate and easy-to-use poker odds calculator on the Web! Know exactly what your chances of winning are at any point in a hand and make your decisions easier.

Poker Hand	# of hands	Probability	# of combinations
Royal Flush	4	0.00000154	480
Straight Flush	36	0.00001385	4320
Four of a kind	624	0.0002401	74880
Full house	3744	0.00144058	449280
Flush	5108	0.0019654	612960
Straight	10200	0.00392464	1224000
Three of a kind	54912	0.02112845	6589440
Two pairs	123552	0.04753902	14826240
One pair	1098240	0.42256903	131788800
Nothing	1302540	0.50117739	156304800
Total	2598960	1.0	311875200

Figure 1: Probability Statistics [7]

The number of combinations represents the number of instances in the entire domain.

3. Proposed System

The analysis and prediction of the best possible winning hand combination depending on the cards the user has in his hands has been calculated by the system. The user can enter 2-5 cards and the best possible winning hand will be displayed. The winning hand displayed will be according to the winning hand ranking combinations. They are shown below:

Figure 2 : The winning hand ranking in a decending order

We have taken the dataset (training and test data) and used it to predict the class (i.e. the winning hand rank) in which they fall under. The final model of our system was decided after the pre-processing done before experimentation.

The *poker hand dataset* is obtained from machine learning site. This dataset is pre-processed before applying any suitable algorithm of classification. The *pre-processing block* involves: Adjusting Dataset and manipulating data as their priority in a deck of cards. Class-wise separation of data and forming different card combinations from the given set of card attributes. Class-wise separation of data means segregating out instances of each class value. Different card combinations dataset of 2 cards, 3 cards and 4 cards are prepared. The next step involves *determination of co-occurrence of classes*.

This can be elaborated as: for a given set of card combination, find out the different classes which occur simultaneously. This is done for every card combination. The co-occurrence of each class is shown in the matrices for 2, 3, 4 cards combinations.

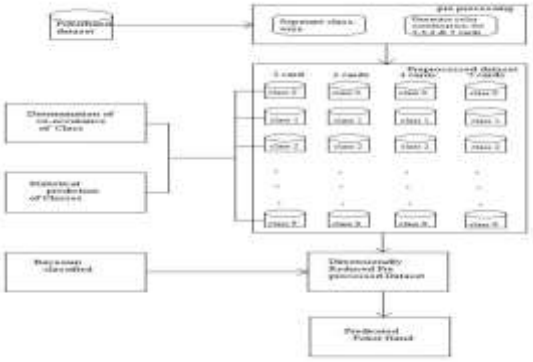


Figure 3: System block representation of our model

Statistical prediction of classes: Hardcode conditions are implemented. This step helps to select classes from all the classes. These are conditions developed from the reasons and logic we understand while playing the game. *Bayesian classifier* along with statistical probabilities is used. The pre-processed dataset along with the modified Bayesian classifier predicts the class for a given card combination.

4. Experiments Performs

Experiment 1: Pre-processing and Decision tree Analysis.

In the dataset we have combined the first two numbers to show the suit number and card number.

*Suit *100 + Card value= Card (Attribute)*

Interchanging the suit and card attributes, we have interchanged the positions of the suit number and card number. The card number is written first and then the suit number is written.

*Card value*100 + Suit = Card (Attribute)*

Altering the priority level for suits

i. Spade : 1 → 4
ii. Hearts: 2 → 3
iii. Diamonds : 3 → 2
iv. Clubs : 4 → 1
v. Ace : changed from 1 to 14 being the highest order

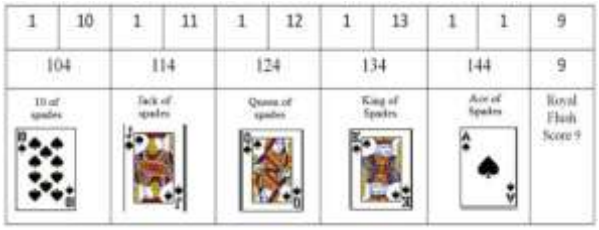


Figure 4 : Representation of final transformation

The number at the end of each line depicts the class value given to a set of i.e. the ranking of cards according to winning hand order.

Class wise division of cards (0-9) Combinations: Dataset is divided into following combinations of 2-Cards, 3-Cards and 4-Cards. It is done to formulate.

2 Cards	3 Cards	4 Cards
1-2, 1-3, 1-4, 1-5	1-2-3, 1-2-4, 1-2-5, 1-3-4,1-3-5, 1-4-5	1-2-3-4
2-3, 2-4, 2-5	2-3-4, 2-3-5, 2-4-5	1-2-3-5
3-4,3-5	3-4-5	1-2-4-5
4-5		1-3-4-5 2-3-4-5

Figure 5 : 4 Class wise Division

various strategies that are applicable to specific conditions. The combinations are such that they can be operated on easily due to size reduction of the entire dataset.

Experiment 2: Judgmental Analysis and Dimensionality reduction.

We used the knowledge of the game to make a set of conditions which were used to find the co-occurrence of classes; thus reducing the number of classes by neglecting the rest.

Figure 6 : Co-occurrence matrix for 4 cards

Similar matrix was made for 2 and 3 cards.

Experiment 3: Applying Bayesian Classifier

The selected classes were given as input to Bayesian Classifier Algorithm. The Algorithm gave probabilities of each class. Bayesian Algorithm was modified by equation 1.

P(CC,C1) = P(CC1/TC1)*P(C1) …….eq(1)

Bayesian Formula to find probability

1. CC= no. of times the input has occurred in class 1 / total instances in class 1
2. TC1= total no. of class 1 instances in type of card combination(e.g. 4 cards)
3. P(C1) : count of class 1 instances in original dataset/ total instances

This probabilities were sorted in decending order and the class with the best probability is the predicted class.

5. Result and Discussion

In experiment 1, we first transformed the attributes of dataset according to the decision model requirement. The suit and card no. attributes are combined into a single attribute. The biggest upside to combining was that the number of attributes was reduced to 6 from 11. Conversion of numeric data set to nominal data set was done to make the dataset workable in Weka. The values were adjusted as per needed. The data set was also disintegrated into 4 parts they are:

i. Column of two cards and score.

ii. Column of three cards and score.

iii. Column of four cards and score.

iv. Column of five cards and score.

After the combinations were done, they were then given as input to Weka and various decision trees were applied to it. Decision stump, FT tree, M5P, J48, etc. were applied to it. They did not avail the necessary output as the accuracy of the generated decision tree should be high (>60%).Moreover, the rules obtained were insufficient as not all the values were classified correctly. Also, the accuracy was so less that the result could not be used for prediction in any manner. Therefore, it can be easily stated that this experiment was unsuccessful in obtaining the desired accuracy and so a new approach for prediction had to be generated and adopted.

.

Algorithm	Accuracy
J48	49.95%
MP5	48.34%
REP	49.92%
Random forest	54.18%
Random Tree	50.25%
BF Tree	No answer
Decision stump	49.91%
LAD	49.90%
Naïve Bayesian	56.68%

Figure 7: Decision Trees With The Accuracy Achieved

In the 2nd experiment, we formulated algorithms for each class on basis of judge mental analysis. The analysis includes basic knowledge of the game for predicting the class of your cards. It helped to generate hardcode selection of classes from the set of the 10. This indirectly helped to increase the probability. For 10 classes probability of one class is 0.1 by reducing class probability increases to 0.25(if down to 4 classes).

In the 3rd experiment, the class with the highest probability is the predicted by our algorithm. The Accuracy of Bayesian Algorithm was increased as we used only the selected datasets.

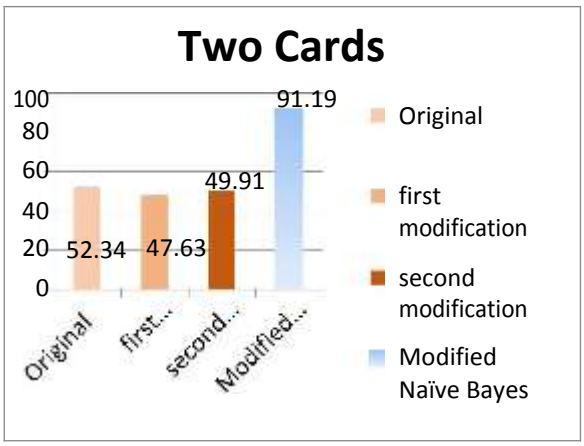


Figure 8 : Accuracy of prediction for 2 cards

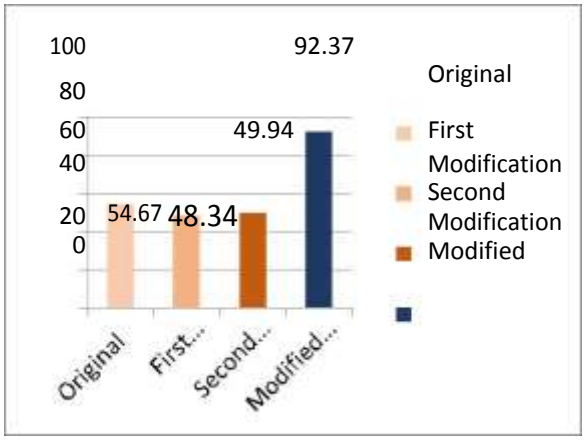


Figure 9 : Accuracy of prediction for 3 cards

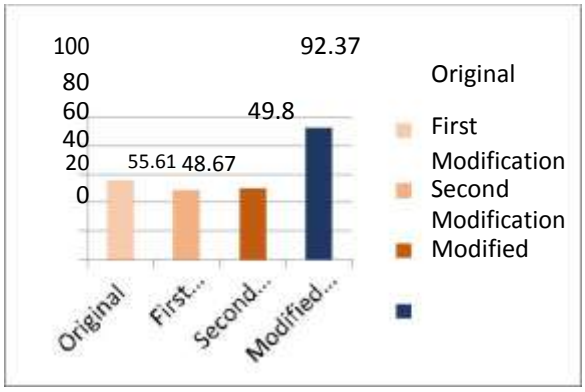
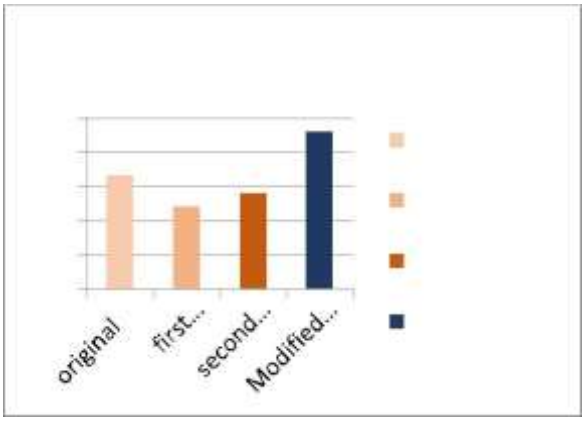


Figure 10 : Accuracy of prediction for 4 cards



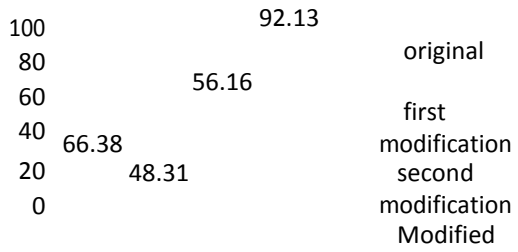


Figure 11 : Accuracy of prediction for 5 cards

6. Conclusion

Thus by performing classification based on statistical analysis and modified Naïve Bayes, the accuracy of this algorithm was found out to be 92.13%.

Importance of data mining techniques for predicting the winning hand possibility in poker has been clearly outlined in this paper. This paper depicts a clear view of the accuracy that we have achieved versus the accuracy that has been achieved by only using a statistical approach by directly using the dataset. It also examines the comparison of different transformations needed to achieve optimal accuracy. Also that the importance of transformation of the datasets necessary to achieve the highest accuracy using both the probabilistic statistical formulae and the cross-referencing of datasets to it for obtaining correct winning hand predictions has been effectively stated and reasoned.

We hope that this paper will gain momentum amongst poker companies as well as data mining research enthusiasts to study the necessity of using statistics for probability related data mining for various games.

Our future research will address improvisation of the system by considering all the 7 cards in the game. Currently the system demands user to enter the best possible 5 cards combination from 7 cards. This can be made better by enabling the system to accept all 7 cards from the table and cards in hand. The best possible five cards combination with highest winning probability will be produced as the outcome. This will reduce the user's efforts of applying tactic in choosing the appropriate cards.

References

[1] TeachingPoker.com (2012).Types of Poker. Available:http://teachingpoker.com/Types.html.

[2] Han, Kamber, "Data Mining concepts and techniques", Morgan Kauffman 2 edition. Amsterdam, NED: Illinious University, March 2006.

[3] Suquamish Clearwater Casino Resort(2014).6 POPULAR TYPES OF POKER[Blog]. Available: http://www.clearwatercasino.com/6-popular-types-of-poker/

[4] BetOnline. (2013). Online Promotions[Advertisement]. Available: http://www.betonline.ag/poker

[5] StevenM.(2010, Sept. 24). Poker Tracker [Blog]. Available: http://www.pokertracker.com/

[6] Cards Chat A worldwide poker community (2004). Easy-to-use Poker Odds Calculator[Blog]. Available: http://www.cardschat.com/poker-odds-calculator.php

[7] Robert Cattral. Carleton University(2002).Poker hand Data Set [Data]. Available: http://archive.ics.uci.edu/ml/machine -learning-databases/poker/poker-hand.names.

[8] Han and Kamber. Data Mining concepts and techniques, 3rd ed. USA: Morgan Kauffman, 2012.

Author Profile

Samadhan B. Patil received the B.E degrees in Computer Engineering from K.C.E.S's C.O.E.I.T., in 2010.