Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.60 – 66

A Novel Methodology for Classification Using Shapley Values of Cooperative Game Theory

P.RAJA[1], V.PASHYANTHI[2], U.NANDINI[3], L.S. CHAKRAVARTHY[4]

[1]Department of CSE, VITS College of Engineering, Sontyam, Visakhapatnam, India

[2]Department of CSE, VITS College of Engineering, Sontyam, Visakhapatnam, India

[3]Department of CSE, VITS College of Engineering, Sontyam, Visakhapatnam, India

[4]Assoc. Prof., Department of CSE, VITS College of Engineering, Sontyam, Visakhapatnam, India

raja.p1995@gmail.com [1]; pashyanthived@gmail.com [2]; nandini.uggini@gmail.com [3]; chakri.ls@gmail.com [4]

ABSTRACT- Classification is a well known task in data mining and machine learning .The general meaning of classification is categorizing or grouping according to their characteristics or values. A new classification approach based on cooperative game theory is proposal in this project. Cooperative game consists of a set of players and a characteristic function which specifies the values created by different subsets of the players in the game. If we want to find classes in classification process, objects can be imagined as players in a game and according to the values obtained by these players classes will be separated.

According to dataset Shapley values are calculated .These Shapley values put forward to calculate the weights for each data set of which will be useful for classification of the objects. This concept is applicable to different types of data. So, we will be successful if we assign the Shapley values accurately and determine which class the objects belong to according evaluated values. We planned to apply this classification technique on benchmark data set or live data set for analysis purpose.

1. INTRODUCTION

Classification is one of the most powerful subjects in data mining which exactly tries to predict the class of each instance. It aims to assigns a class label to an object or an event. The assignment is always based on measurement that is obtained from that object.

Game Theory: The project is processed by co-operative Game Theory using Shapley values.. The branch of mathematics concerned with the analysis of strategies for dealing with competitive situations where the outcome of a participant's choice of action depends critically on the actions of other participants.

To each cooperative game it assigns a unique distribution (among the players) of a total surplus generated by the coalition of all players, using co-operative game theory we get the Shapley values. Cooperative game consists of a set of players and a characteristic function which specifies the values created by different subsets of the players in the game.

R Language: R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology. R can be provided as an open source. Here objects which are with closest shapley values are assigned to its closest centroid. This results in a partitioning of the data. Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions. The algorithm converges when the assignments no longer change. The algorithm execution is visually depicted. Note that each iteration needs N × k comparisons, which determines the time complexity of one iteration. We used R-language for code implementation and used R-studio software for code execution.

We have taken samples and generated Shapley values by converting numerical values to binary values. We have converted these values into matrix form (m*n) by using R language, from this matrix form Shapley values are obtained. After calculating these Shapley values we divided all the nearest Shapley values into one class and vice versa. By this we divided the values into classes. This is how classification is done using Shapley values in co-operative game theory.

2.RELATED WORK

The k-means algorithm: The k-means is a simple iterative method to partition a given dataset into a user specified number of clusters. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster [1].

Naive Bayes: Given a set of objects, each of which belongs to a known class, and each of which has a known vector of variables, our aim is to construct a rule which will allow us to assign future objects to a class, given only the vectors of variables describing the future objects. Problems of this kind, called problems of supervised classification, are ubiquitous, and many methods for constructing such rules have been developed. One very important one is the naive Bayes method—also called idiot's Bayes, simple Bayes, and independence Bayes [2].

Bayesian Tree: Function to create a quantity called the posterior probability of trees using a model of evolution, based on some prior probabilities, producing the most likely phylogenetic tree for the given data. This dissertation presents the statistical framework of Bayesian analysis of tree models with various applications. Prior specification for such models and development of algorithms for sampling from the posterior distributions are both challenging problems. This addresses each of these issues and extends the Bayesian tree model in several ways, including data resampling (Dirichlet Process Prior), random threshold in the splitting rules, and autoregressive processes for modeling nonlinear structure in time series data [3].

Decision tree: A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The path from root to leaf represents classification rules.

Game Theory: The branch of mathematics concerned with the analysis of strategies for dealing with competitive situations where the outcome of a participant's choice of action depends critically on the actions of other participants. Game theory has been applied to contexts in war, business, and biology. Game theory is "the study of mathematical models of conflict and cooperation between intelligent rational decision-makers. Game theory is mainly used in economics, political science,

and psychology, as well as logic, computer science, biology and poker. Originally, it addressed zero-sum games, in which one person's gains result in losses for the other participants. Today, game theory applies to a wide range of behavioral relations, and is now an umbrella term for the science of logical decision making in humans, animals, and computers [4].

Cooperative Game Theory: The branch of mathematics concerned with the analysis of strategies for dealing with competitive the outcome of a participant's choice of action depends critically on the actions of other participants. Game theory has been applied to contexts in war, business, and biology. A cooperative game is given by specifying a value for every coalition. Formally, the game (coalitional game) consists of a finite set of players N, called the grand coalition, and a characteristic function $v : 2^N \rightarrow \mathbb{R}$ from the set of all possible coalitions of players to a set of payments that satisfies υ(Ø) = 0. The function describes how much collective payoff a set of players can gain by forming a coalition and the game is sometimes called a value game or a profit game. The players are assumed to choose which coalitions to form, according to their estimate of the way the payment will be divided among coalition members. A cooperative game is given by specifying a value for every coalition. Formally, the game (coalitional game) consists of a finite set of players N, called the grand coalition, and a characteristic function $v : 2^N \rightarrow \mathbb{R}$ from the set of all possible coalitions of players to a set of payments that satisfies υ(Ø) = 0. The function describes how much collective payoff a set of players can gain by forming a coalition and the game is sometimes called a value game or a profit. The players are assumed to choose which coalitions to form, according to their estimate of the way the payment will be divided among coalition members [5].

Shapley value: In game theory, the Shapley value, named in honor of Lloyd Shapley, who introduced it in 1953, is a solution concept in cooperative game theory. To each cooperative game it assigns unique distribution (among the players) of a total surplus generated by the coalition of all players. To each cooperative game it assigns a unique distribution (among the players) of a total surplus generated by the coalition of all players. The Shapley value is characterized by a collection of desirable properties [6].

3. METHODOLOGY

In this method classification is performed by using two steps:

(i) Calculating Shapley values:

Algorithm for Calculating Shapley Values:

INPUT: Binary matrix of size ‘m*n’

OUTPUT: Shapley values of size ‘m’

Algorithm Shapley Values (m, n, r, c, w, shap_val)

r←no of rows; //calculate number of rows

c←no of columns; //calculate number of columns

For j←1 to c do

w[j]←∑B[j]; //calculation of sum and column storing in w

End For

For j←1 to c do

For i←1 to r do

If (B[i, j]==1)

sh[i, j]←1/w[j]

End For

End For

For i←1 to r do

shap _val[i] ←1/c ∑ sh[i] //calculate row sum & store in shap_val array

End For

Algorithm Description: The input of the algorithm is binary matrix (m*n) and the output is shapley values of size 'm'. Initially store the number of rows in variable 'r' and number of columns in variable 'c'. Create an array(w) of size 'c', now create an array(shap_val) of size r. calculate sum of each column and store the resultant values in array(w).For each row replace 1 by reciprocating its respective column sum. Calculate sum of each row and store the values in array (shap_val). The values in shap_val array are the required Shapley values for classification.

(ii)Classification using Shapley values obtained from above step:

Algorithm For Classification:

INPUT: Shapley values of size 'm'

OUTPUT: N Classes

Algorithm Classification (shap_max, norm_shap,n)

shap_max←max(shap_val); //Finding the maximum value

norm_sort_shap←shap_val/shap_max; //Dividing all values with max value

Segragation //perform segration based on how many classes we require

Steps for performing clustering:

The Shapley values obtained from the Shapley value algorithm is taken as input. By taking the shapley values as input we need to find the maximum value in the given Shapley values. Divide every Shapley value so that their range exists between 0 and 1. Perform segression on the obtained values. The values that are similar are grouped as clusters.

Sample Data 1:

	1	2	3	4	5	6	7
gene 1	1	0	1	0	0	0	0
gene 2	1	1	0	0	0	0	0
gene 3	0	0	0	1	1	1	1
gene 4	0	1	1	0	0	0	0
gene 5	1	0	1	0	0	0	0
gene 6	1	0	1	0	0	0	0
gene 7	0	1	1	0	0	0	0
gene 8	1	0	1	0	0	0	0

Result:

Shapley values
0.05238095
0.07619048
0.57142857
0.07142857
0.05238095
0.05238095
0.07142857
0.05238095

Sample Data 2:

	1	2	3	4	5	6	7
gene 1	1	0	0	0	0	0	1
gene 2	0	0	0	1	0	1	0
gene 3	0	0	0	1	1	1	1
gene 4	0	0	0	0	1	1	0
gene 5	0	0	1	1	0	0	0
gene 6	0	0	0	0	1	1	0
gene 7	0	1	0	1	0	0	0
gene 8	0	0	0	0	0	1	1

Result:

Shapley values
0.19047619
0.06428571
0.15952381
0.07619048
0.17857143
0.07619048
0.17857143
0.07619048

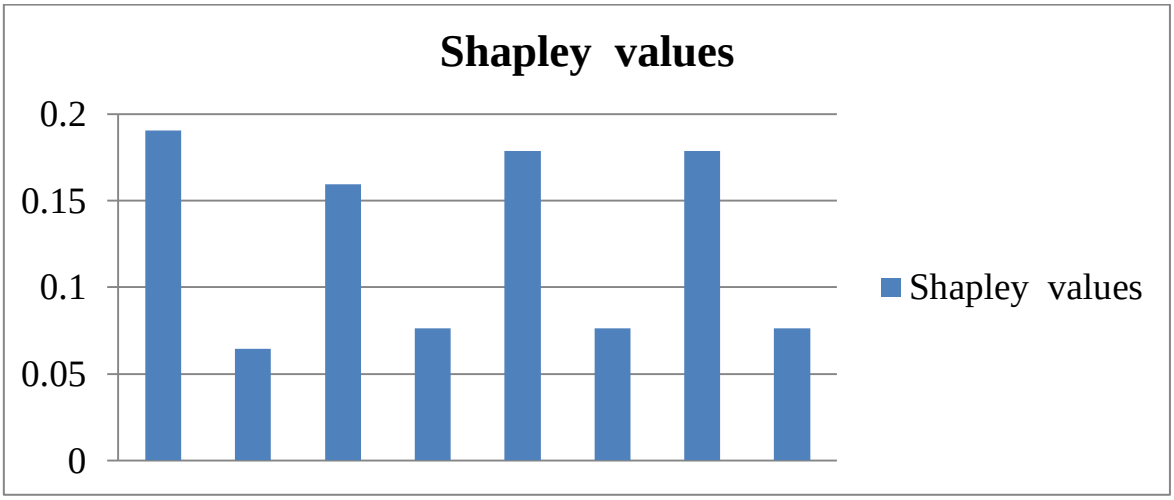

Fig1: Graphical representation of Shapley values of sample data 1

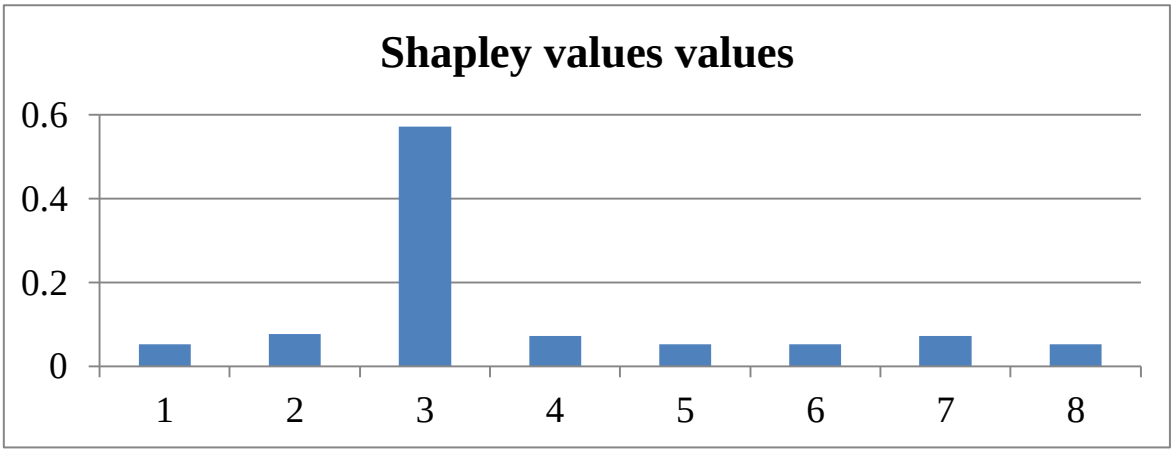

Fig2: Graphical representation of Shapley values of sample data 2

From the above graphs, we observed that the genes oh higher potential values are grouped into one class. Lower potential values are grouped into another class.

4. CONCLUSION

The purpose of this project is to classify the data. Classification is done through Shapley values by using cooperative game theory. Numerical values are converted into binary values. Binary values are converted into matrix form. By applying formal methodology Shapley values are obtained. We have done classification. We obtained optimum values.

Future Enhancement

In this project we have used binary values for classification based on the shapley values of co-operative game theory. But in future this project can be expended by taking numerical values as well. Hence this would be our future enhancement.

References

[1] Lloyd SP (1957) Least squares quantization in PCM. Unpublished Bell Lab Tech. Note, portions presented at the Institute of Mathematical Statistics Meeting Atlantic City, NJ, September 1957. Also, IEEE Trans Inform Theory (Special Issue on Quantization) volume IT-28, pp 129]
[2] Domingo's, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 29: 103–130]
[3] Hastings W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications.Biometrika, 57, 97–109]
[4] Dutta, Prajit K. (1999), Strategies and games: theory and practice, MIT Press, ISBN 978-0-262-04169-0. Suitable for undergraduate and business students.
[5] Bilbao, Jesus Mario (2000), Cooperative Games on Combinatorial Structures, Kluwer Academic Publishers
[6] Lloyd S. Shapley. "A Value for n-person Games". In Contributions to the Theory of Games, volume II, by H.W. Kuhn and A.W. Tucker, editors. Annals of Mathematical Studies v. 28, pp. 307–317. Princeton University Press, 1953.

