ISSN 2319-2720
Volume 7, No.2, April – June 2018
International Journal of Computing, Communications and Networking
Available Online at http://www.warse.org/ijccn/static/pdf/file/ijccn16722018.pdf
https://doi.org/10.30534/ijccn/2018/16722018

Efficient and Private Scoring of Decision Trees, based on Pre-Computation Technique with Support Vector Machines and Logistic Regression Model

Rajeev kumarsah[1], Priyadarshini G[2] , Pallavi M[3], Navya k[4], Chetana Srinivas[5]

[1]Student, EWIT, India, rajeevsah244@gmail.com
[2]Student, EWIT, India, priyupriya1996@gmail.com
[3]Student, EWIT, India,pallum1996@gmail.com
[4]Student, EWIT, India, navya789k@gmail.com
[5]Assoc professor EWIT, India, chetansrinivas@ewit.edu

ABSTRACT

Numerous information driven customized administrations require that private information of clients is scored against a prepared machine learning model. In this paper we propose a novel convention for security protecting order of choice trees, a famous machine learning model in these situations. Our answers is made out of building squares, to be specific a safe correlation convention, a convention for negligently choosing inputs, and a convention for increase. By joining a portion of the building hinders for our choice tree order convention, we additionally enhance beforehand proposed answers for characterization of help vector machines and calculated relapse models. Our conventions are data hypothetically secure and, dissimilar to already proposed arrangements, don't require secluded exponentiations. We demonstrate that our conventions for protection saving arrangement prompt more proficient outcomes from the perspective of computational and correspondence complexities. We introduce exactness and runtime comes about for 7 characterization benchmark datasets from the UCI archive.

Key words: Private order, choice trees, bolster vector machines, calculated relapse, security safeguarding calculation, multiparty calculation.

1. INTRODUCTION

Information driven machine learning can limitlessly enhance the nature of our day by day lives and is as of now doing as such from various perspectives. Medicinal services suppliers utilize frameworks in view of machine figuring out how to analyse patients; wearable gadgets are associated with wellness following applications that utilization machine figuring out how to influence individual wellbeing proposals; to web search tools and web-based social networking destinations depend on machine figuring out how to choose which bstance to show to every individual client, including which notices; internet business organizations use machine figuring out how to figure out which items or motion pictures to prescribe to clients in light of their earlier buy conduct; web based dating administrations utilize machine learning trying to interface individuals with the affection for their lives. . . the rundown continues endlessly. To profit by any of these customized administrations, the individual information of clients –, for example, individual inclinations, perusing conduct or restorative lab comes about – should be scored against a prepared machine learning model. In this paper we propose systems to play out this scoring in a protection safeguarding way so people don't need to impart their own information to anybody "free" yet may at present advantage from these kinds of customized administrations. All the more particularly, we manage situations where a man holding information (Alice) needs to score her information against a model possessing another gathering (Bob) with the end goal that, toward the finish of the convention, Bob adapts nothing about Alice's information and Alice learns as meagre as conceivable about Bob's model.

2. PROPOSED SYSTEM

We have proposed a novel convention for protection saving arrangement of choice trees, and enhanced the execution of already proposed conventions for general hyper plane-based classifiers and for the two particular instances of help vector machines and strategic relapse. To infer the parameters for our choice tree order convention, we again enhance the Support Vector Machines and Logistic Regression idea for grouping. The Client and Server can pre-process this information independent from anyone else with the assistance of understood computationally secure plans with a trusted expert isn't accessible. We exhibit exactness and runtime comes about for 2 order benchmark datasets from the UCI store.

2.1 ADVANTAGES

The proposed method achieves high efficiency of privacy protection. Our solutions are very efficient and use solely

modular addition and multiplications. Our classification data shows that we have scored a new data with accuracy and the performance evaluation is highly efficient. Our results for privacy-preserving machine learning classification are highly secured. Support Vector Machines and Hypervisor based Classifiers Hyper plane Based Classifiers and Support Vector Machines hyper plane-based classifiers are parametric, discriminative classifiers. For a setting with t features3 and k classes, the model comprises of k vectors w = (w1,...,wk) with wi∈Rt and the classification result is gotten by deciding, for Alice's element vector x ∈Rt, the file k∗ = argmaxi∈[k]hwi,xi, where h•,•i is the inward item.

"Support Vector Machine" (SVM) is a regulated machine learning calculation which can be utilized for both characterization and relapse challenges. Be that as it may, it is for the most part utilized as a part of characterization issues. In this calculation, we plot every datum thing as a point in n-dimensional space (where n is number of highlights you have) with the estimation of each component being the estimation of a specific organize. At that point, we perform grouping by finding the hyper-plane that separate the two classes extremely well (take a gander at the underneath depiction). Bolster Vectors are essentially the co-ordinates of individual perception. Bolster Vector Machine is a wilderness which best isolates the two classes (hyper-plane/line).

3. Logistic Regression

Logistic Regression is the suitable relapse investigation to direct when the reliant variable is dichotomous (paired). Like all relapse investigations, the strategic relapse is a prescient examination. Strategic relapse is utilized to depict information and to clarify the connection between one ward paired variable and at least one ostensible, ordinal, interim or proportion level autonomous factors. Once in a while calculated relapses are hard to decipher; the Intellects Statistics instrument effectively enables you to lead the investigation, at that point in plain English translates the yield.

4. Decision Tree

Decision Trees are an exemplary administered learning calculations. A choice tree is a choice help device that uses a tree-like chart or model of choices and their conceivable results, including chance-occasion results, asset expenses, and utility. The choice tree calculation can be utilized for tackling the relapse and order issues too. The principle objective of choice tree is to accomplish idealize grouping with least number of choice and it isn't generally conceivable because of irregularities of information.

Choice Trees Decision trees are non-parametric, discriminative classifiers2. Alice holds an info vector x = (x1,...,xt) ∈Rt comprising of t highlights. The classification calculation comprises of a mapping C: Rt→{c1,...,ck} on x. The aftereffect of the classification C(x) is one of the k conceivable classes c1,...,ck. The model is a tree structure and is held by Bob Bounce's model is D = (d,G,H,w), where d is the profundity of the tree, G: {1,...,2d}→{1,...,k} is a mapping from the files of the leaves to the records of the classes, H: {1,...,2d−1}→{1,...,t}isamappingfromtheindices of the inward hubs (constantly considered in level-arrange) to the lists of Alice's info highlights and w = (w1,...,w2d−1) with wi∈ R contains the limits relating to each inner hub. For each interior hub vi with 1 ≤ I ≤ 2d −1, let zi be the Boolean variable signifying the consequence of contrasting xH(i) with wi, which is one if xH(i) ≥ wi and zero generally. The classification procedure goes as takes after: • Starting from the root hub, for the current interior hub vi, assess zi. On the off chance that zi = 1, take the left branch; generally, the correct branch. • The calculation ends when a leaf is come to. On the off chance that the j-th leaf is achieved, at that point the yield is cG(j).

4.1.Sample Example

Give us a chance to think of you as are intending to go out for feasting, as your companions are going to however you are reluctant in settling on which eatery to pick. At whatever point you need to go out for eating you inquire as to whether he supposes you will like that place or not. You give him a rundown of eateries that you have gone by and let him know whether you loved every eatery or not (giving a named preparing dataset). Bobby, pose couple of inquiries like, regardless of whether you like rooftop top seating? Does eatery serve Indian sustenance? Is eatery open till midnight? Does eatery have unrecorded music et cetera to answer your inquiry? It puts forth a few educational inquiries to give the answer whether you will like that eatery or not. In this Bobby is a choice tree for finding your eatery inclinations.

4.2.Types of Decision Tree

i.Classification Trees

ii.Regression Trees

4.2.1.Classification Trees

It is the default sort of choice tree used to isolate the dataset into various classes. The reaction variable is clear cut in nature. (2 classes or various categories)Example: We have two factors age and weight .Based on this we will decide if the individual will join rec center or not.

4.2.2 Regression Trees

It is utilized when the reaction variable is ceaseless or numerical in nature. This is again characterized into direct relationship and nonlinear connection between the indicators and response. Example: Predicted cost of a buyer decent. How to use Decision Tree? The few scenarios where we can use decision tree algorithm are,

1. The choice trees are suited if the preparation information contains mistake. Since they are vigorous to blunders.

2. It is utilized when the preparation information has missing esteems. Since they can deal with missing esteems by looking the information into different segments.

4.2.3 Advantages

The few advantages of decision trees are

i. Easy to explain.
ii. Data type is not constraint as they can handle both categorical and numerical values.

5. BUILDING BLOCKS

Secure Distributed Comparison For performing secure conveyed bitwise examination we utilize the convention of Garay et al. [30] with mystery partaking's in the field Z2. That convention has dlog`e + 1 adjusts and utilizes 3`−blog`c−2 augmentations. The convention will be signified by πDC and it safely actualizes the appropriated examination usefulness FDC.

A. Secure Argmax
Suppose that the parties P1,...,Pn have bitwise shares of a tuple of values (v1,...,vk) and want one of them, let's say P1, to learn all the arguments m ∈{1,...,k} such that vm ≥ vj for all j ∈{1,...,k}, but no party should learn any vj or the relative order between the elements. I.e., the parties just want P1 to learn

B. Secure Bit-Decomposition
In this we deal with the problem of converting from sharesJxKq of a value x in a large field Zq to shares ofJxiK2 in the field Z2, where x`···x1 is the binary representation of x.

C. Oblivious Input Selection In our applications there are also circumstances in which Alice holds a vector of inputs x = (x1,...,xn) and Bob holds an index k, and they want to obtain bitwise secret sharing's of xk for further uses in the protocol, but without revealing any information about the inputs or k.

6. BINOMIAL LOGISTIC REGRESSION

A binomial strategic regression, predicts the likelihood that a perception falls flat into one of the two classifications of a dichotomous ward variable in light of at least one free factors that can be either constant of straight out. This is regularly called as straightforward Logistic relapse.
Example: Let us predict, whether students will pass or not (i.e. the dependent variables are pass and fail) in their final exam based on internal marks, assignment submission and few other independent variables.

6.1 EXPLANATION

1. The logistic regression outputs probabilities based on the following equationlogit(pi)=log(pi/(1−pi))=β0+β1x1+...+βkxkthe co-efficient refers to each βi.

2. Odds ratios are simply the exponential of the weights i.e. The first co-efficient we have outlook=sunny:-

3.5821.Calculation of exp(-3.5821) gives 0.0278 that is the corresponding value in the odds ratio table.

log(Odds(outlook=sunny)/Odds(outlook=¬sunny))

It will also display the correctly classified Instances and incorrectly classified Instances.

With that data, we can understand the accuracy of the algorithm.

7. K-MEANS ALGORITHM

The k-means algorithm is used for the clustering of the data. The main idea is to define k center, one for each cluster. These center should be placed in a cunning way because of different location causes different result. This calculation goes for limiting a target work know as squared mistake work given by: Where, '||xi - vj||' is the Euclidean separation amongst xi and vj. 'ci' is the quantity of information focuses in ith bunch. 'c' is the quantity of group focuses.

7.1 Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of center.

1) Randomly select '*c*' cluster center.

2) Calculate the distance between each data point and cluster center.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster center.

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'ci' speaks to the quantity of information focuses in ith bunch. Recalculate the separation between every datum point and new got group focuses. If no information point was reassigned then stop, generally rehash from stage 3).

Advantages

1) Fast, vigorous and less demanding to get it.

2) Relatively proficient: O(tknd), where n is # objects, k is # groups, d is # measurement of each protest, and t is # cycles. Ordinarily, k, t, d << n.

3) Gives best outcome when informational index are unmistakable or all around isolated from each other.

7.1.1. Experiments:

For our experimentation, we used R's base glm function, setting the family parameter to binomial(link="logit") to obtain a logistic regression model. The following datasets were chosen for our experimentation:

1) Breast Cancer Wisconsin (Diagnostic): The goal with this dataset is to classify 568 different tumours as malignant or benign. Each tumour is characterized by 30 different continuous features derived from an image of the tumour (i.e. perimeter, area, symmetry, etc.).
2) Pima Indians Diabetes: This dataset includes 767 females of at least 21 years of age, all with Pima Indian.
3) Parkinson's: Here, the task is to differentiate between patients with and without Parkinson's

CONCLUSION

We have proposed a novel protocol for privacy-preserving classification of decision trees, and improved the performance of previously proposed protocols for general hyper plane-based classifiers and for the two specific cases of support vector machines and logistic regression. The pre-distributed data can be distributed during a setup phase by a trusted authority to Alice and Bob. In the case a trusted authority is not available or desirable, Alice and Bob can pre-compute this data by themselves, during a setup phase, with the help of well-known computationally secure schemes. Our solutions are very efficient and use solely modular addition and multiplications. We present accuracy and runtime results for 2 classification benchmark datasets from the UCI repository.

FUTURE ENHANCEMENT

One open problem is improving the performance of the bit-decomposition protocol using techniques similar to the proposed system while keeping the shares. It is a good feature for saving communication and for optimizing an eventual pre-processing phase using OT extension.

REFERENCES

1. P. Pullonen. Actively secure two-party computation: Efficient beaver triple generation. Master's thesis, University of Tartu, May 2013.

2. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2015.

3. R. L. Rivest. Unconditionally secure commitment and oblivious transfer schemes using private channels and a trusted initializer. Preprint avail-able at http://people.csail.mit.edu/rivest/Rivest- commitment.pdf, 1999.

4. B. Schoenmakers and P. Tuyls. Efficient binary conversion for Paillier encrypted values. In S. Vaudenay, editor, EUROCRYPT 2006, volume 4004 of LNCS, pages 522–537, St. Petersburg, Russia, May 28 – June 1, 2006. Springer, Heidelberg, Germany. https://doi.org/10.1007/11761679_31

5. T. Therneau, B. Atkinson, and B. Ripley. Part: Recursive Partitioning and Regression Trees, 2015. R package version 4.1-10.

6. T. Toft. Primitives and Applications for Multi-party Computation. PhD thesis, Aarhus University, 2007.

7. T. Toft. Constant-rounds, almost-linear bit-decomposition of secret shared values. In M. Fischlin, editor, CT-RSA 2009, volume 5473 of LNCS, pages 357–371, San Francisco, CA, USA, Apr. 20–24, 2009. Springer, Heidelberg, Germany. https://doi.org/10.1007/978-3-642-00862-7_24

8. T. Toft. Solving linear programs using multiparty computation. In R. Dingledine and P. Golle, editors, FC 2009, volume 5628 of LNCS, pages 90–107, Accra Beach, Barbados, Feb. 23–26, 2009. Springer, Heidelberg, Germany. https://doi.org/10.1007/978-3-642-03549-4_6

9. R. Tonicelli, A. C. A. Nascimento, R. Dowsley, J. M¨ uller-Quade, H. Imai, G. Hanaoka, and A. Otsuka. Information-theoretically secure oblivious polynomial evaluation in the commodity-based model. Inter-national Journal of Information Security, 14(1):73–84, 2015. https://doi.org/10.1007/s10207-014-0247-8

10. T. Veugen. Linear round bit-decomposition of secret-shared values. Information Forensics and Security, IEEE Transactions on, 10(3):498– 506, March 2015.

11. D. J. Wu, T. Feng, M. Naehrig, and K. E. Lauter. Privately evaluating decision trees and random forests. IACR CryptologyePrint Archive, 2015:386, 2015

Rajeev KumarSah: Pursing B.E in CSE, EWIT (VTU), Bengaluru. His areas of interests are Networking, Cyber Security, Programming the web, IoT, Database

Navya K: Pursing B.E in CSE, EWIT (VTU), Bengaluru. Her areas of interests are Computer Networks, Data structure, Java, Programming the Web

Priyadarshini G: Pursing B.E in CSE, EWIT (VTU), Bengaluru. Her areas of interests are Computer Network, Java, DBMS, Big Data

Pallavi M: Pursing B.E in CSE, EWIT (VTU), Bengaluru. Her areas of interests are Computer Network, DBMS, Software Engineering, Programming the Web

Prof. Chetana Srinivas: Associate Professor, Department of Computer Science, East West Institute of Technology (VTU), Bengaluru. Qualifications: M.Tech(Ph.D). Her areas of Interest are Big Data and Image Processin