International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Mapping of Advertiser codes to Advertisers using Machine Learning

Ashay Vivek Sant and S. R. Khonde

Modern Education Society's College of Engineering Pune, India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

As a business requirement, the advertiser codes need to be mapped to the full forms using Machine Learning Approach. This needs to be performed on production using a dedicated microservice. This will aid us in reducing the database lookup calls, and also be helpful for auto-suggestions for the user when creating new data. The data related to campaigns, the agreements, the flight time, date, amount, brand mapping is already present in a oracle based relational schema and is already normalized. This data needs to be combined and treated as a training data-set for learning a suitable model. The model learnt should also provide the functionality for auto-completion on the User Interface. The model learnt should be so generic that it could be tuned for any dataset in the near future, thereby making it potential candidate for prediction in advertisement world. For learning models an ensemble of classifiers could be run in parallel and then majority voting could be performed to decide the best classification. Simple Algorithms like that Naive Bayes, hyperplane learning algorithms like SVM and regression could be used along with unsupervised approaches like that of clustering.

Keywords: *Machine, Ensemble Of Classifiers, Bayes Theorem, Support Vector Machines, SVM, Linear Regression, Lookup, Agency, Sellers, Clustering, Unsupervised, Supervised, Reinforcement*

Introduction

Advertisement World consists of 3 entities

1) Advertisers
2) Sellers
3) Media Agencies

A. Advertisers

Advertisers are players who create add and want to display the advertisement somewhere in the media world. Advertisers would like to target specific audiences via different modes like that of Radio, Television or Internet. These advertisers have advertisements and they get in touch with Media Agencies for add placements. This way the advertisement flows from actual advertiser to Media Agency. E.g. Nike, Coke

B. Sellers

Sellers are the actual owners of the advertisement slots. Consider for example Google, Facebook, Yahoo and many more. These industries have dedicated web pages and the slots for displaying the advertisement. The sellers can directly sell the media slots to the advertisers or route the advertisements Via Media Agencies. Thus sellers are most important part of advertisement workflow who help the advertisements reach out to the audiences

C. Media Agencies

These are intermediators in between the Advertisers and the Sellers. IPG and Havas are major media agency players in the advertisement world. These agencies handle the entire work flow including the money spent, money left and the no. of slots which are yet to be buyed etc. These behold the transactions on behalf of Advertisers to the Seller side.

Now the aim is to learn a model to map advertiser codes to actual advertisers and make auto-suggestions to the user on the User Interface when the user wishes to create an advertisement campaign on the UI on behalf of an ad agency. This could be done by the production data-set present in the oracle databases. There are ad agencies that have run the campaign earlier. This already has mapping to the advertiser and further these advertisers are mapped to advertiser codes. This data will be fed as training data-set to the model. Thus the model would receive campaign data as training data and then when given campaign name and other campaign details dynamically from the UI would then return the probable advertisers as suggestions to the user creating campaign.

Literature Survey

[1]Bayes classifier is popular algorithm used for text classification. In this paper, the authors have predicted the song performer based on lyrics only. The precision which was achieved was 93% and Recall of 95%. Also F1 measure of 94% was achieved. Bayes rule states that P ($\theta|D$) = P (θ) P ($D|\theta$) / P (D) Wherein θ Is a class

- D is a document
- P (θ) is a class probability
- P (D) is the probability of document
- P (D| θ) is conditional probability of class for given document
- P (θ|D) is conditional probability that document D belongs to class c

The problems stated include creation of the dataset, guest appearances of music performers needed to be handled with care. [2]The authors have described multiple machine learning algorithms as part of this Paper. The authors focus on predicting the future using machine learning. Here by future we mean unknown data-set. The algorithms described by author include

1) *Decision-Tree*
2) *Support Vector Machine*
3) *K Nearest Neighbor Classier*
4) *Naive Bayes Algorithm*
5) *Linear Regression*

The author also describes the advantages and disadvantages of each and every algorithm in detail. Author states that Linear Regression could not be applied to non-linear dataset, however it is much easy to understand and design. Bayes algorithm is simple, fast and scalable. It is best suitable for text classification. It assumes the concept of class conditional independence. [3]
A multiTree algorithm is generated for Intrusion detection system. Alongside multiple algorithms are used like that of Decision Tree, KNN, DNN, Random forests etc. to learn an ensemble classifier and perform majority voting. Steps involved in the algorithm include

1) *The training data-set is fed*
2) *Standardization of data by preprocessing module*
3) *Ensemble training of classifiers*
4) *Use of cross-validation for training all classifiers. Also boosting is performed for increased accuracy*

4] The author makes a mention for the TrResampling whereby a new training data-set is created from the existing dataset. The size of this new data-set is same as that of original data-set. The data in original data-set is weighted and selection is based on these weights. The weights are arranged such that more frequently the data appears in the new pseudo training-data set, the less likely it is to be misclassified. TrAdaBoost strategy is used to influence the inclusion of the training-data-set into the pseudo training data-set. The author also used Bagging and Boosting algorithms in conjunction with TrResampling. These relate to the ensemble model.

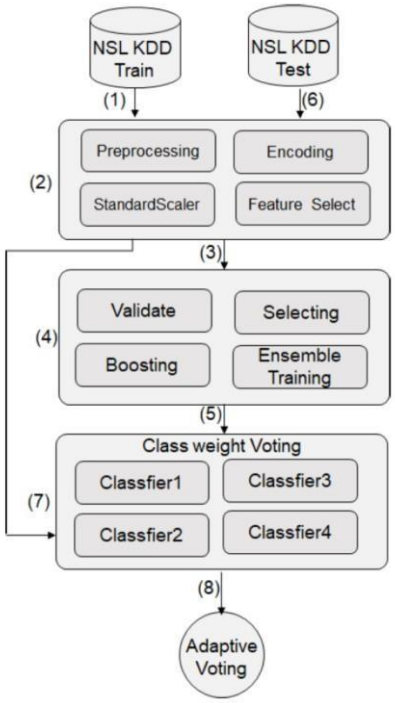


Fig. 1. Ensemble Classification

[[5] Meaning of High availability on oracle clusters is explained by the authors and the paper is from Oracle Corporation USA directly. According to the authors high availability means that when one instance on a node is down, due to some hardware failure or some software failure then there is process named instance recovery that runs. This high availability concept is used in our oracle data-store. [6] Time series data is explained in this white paper. Consider a sensor reading the temperatures of a 2-wheeler engine. Thereby the sensor records the timestamp as an additional parameter along with the said temperature. This may increase the numerosity of the data that is collected. All this data is not necessary for the warehousing 9 and data mining tasks. [7] The literature tells us about undo and redo operations for database recovery in high availability environment. Checkpoints are like restore points in Microsoft windows wherein the systems state is stored just like a snapshot and it is restored as and when needed. The paper suggests an approach of Roll Forward like that of Roll back. We have to save the entire state of the VM in traditional methods. This costly transaction is addressed by the authors which gets rid of pausing the VMs for saving their entire state. The authors stress on storing only the next instruction to be executed and all the data structures that help in restoring any failed transaction. [8] Here the authors have described highlight the below techniques

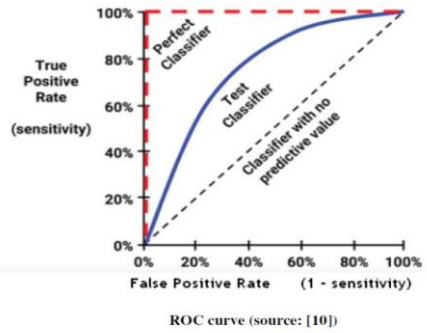
1) *Data Training and Testing on same data*

2) Dataset separation into training and testing parts
3) Cross Validation
In k fold cross validation the test is carried out k times. If k is set to 7 then evaluation process will be done 7 times using the different data each time. Also the sampled data is such that at the end of k runs entire dataset would have been evaluated. The machine learning algorithms investigated in this literature involve
1) Logistic Regression
2) k Nearest Neighbor
3) Decision Tree
4) Naive Bayes

The datasets tested by the authors include
1) Fire Dataset (Air quality, temperatures for classification)
2) Blood Transfusion Service Center Dataset (Time since donation, Donation Frequency, total blood donated)
3) Iris Dataset (Iris flower data set. Iris species of plants are investigated)
4) Lenses Data Set (hard contact lenses, soft contact lenses,
no contact lenses)

The author concludes that in each of the data-set different models perform differently and these is no one model that could be used for each and every dataset. This boosts our decision for using ensemble of classifiers so that we could leverage the benefit from all algorithms. [9] This paper describes optimization techniques to reduce the training time. Also shallow learning networks that do not have as many layers as deep Neural networks are not suitable for the large and multidimensional data sets that we have nowadays. As an example consider the task of recognizing the image using CNN (Convolution Neural Network). The low level pixel details from an image are taken and analyzed. Then mid-level features including the Shapes and edges along with patterns observed are further analyzed. DNN is implemented using CNN. CNN is primarily used in Image Processing. CNN takes input of 2D image and multiple layers of Neural Networks call filters or kernels are present. The mapping is to only specific neurons included in the spatial locality. This reduces chances of over fitting. The final layers in the CNN are fully connected and thus are responsible for classification. [10] The ultimate aim in Machine Learning is to learn a model from the dataset. This dataset is called as training data-set which is stored in relational databases. This can come from highly available database as well. When we are considering the application of highly available databases, Machine Learning is a strong source. The dataset is stored in relational databases and Hyper-Box approach tries searching the database and analyzing it using queries. Consider the example suggested by the authors of detecting knearest neighbors. In this case we need to identify the k nearest neighbors. This takes a scan through the data after it has been retrieved. Why not accelerate it at the time of querying itself? If we put the distance metrics in the where clause of the SQL Query we could just filter out the neighbors at the relational database level, thereby accelerating the Machine Learning process using KNN. [11] Database queries are quite complex for new users. Hence using a natural language processing approach these can be made simpler. Consider the query spelled as a sentence. After this there are following steps involved
1) Token Analyzing
2) Spelling Correction
3) Ambiguity elimination
4) Tagger
5) Morpheme
6) Syntax analysis and Context Free Grammar analysis using Tree
7) Translation to XML
8) Translation to Database Query

This way the end-user just needs to type a sentence and then automatically the query processing engine will translate the sentence to the query using intermediatory XML syntax. This way we could use relational highly available database and highly available micro services to translate Natural Language to Query as requested by the end-user



ROC curve (source: [10])

Fig. 2. ROC Curve

Proposed Methodology

A. *Problem Definition*

To learn a suitable model from the data-set that is already present on production and staging environments to infer the suggestions and mapping from advertiser codes to full forms. The mappings are already present in relational database but unless queried for, those are not returned. To help suggest users for close matches for preferences on production, we need to derive a model from existing data-set.

B. *The Framework*

Using an ORM framework like that of Spring Data JPA in java or Django framework in python could be used for rapid web development. Since Micro service architecture is present, the common method to exchange information will be Rest API's. Hence a full

MVC framework needs to be used. This is supported in Spring (Java) and Python (Django)

C. Preprocessor

The preprocessor handles missing data and outlier detection primarily. Outliers can be analyzed by simple deleted flag, or by particular pattern. Say a particular entry occurs in database many times but it is deprecated. This need not be considered for learning the model. Also the entries that skew the data could be identified like that of very high plan cost or very low plan cost could be marked outliers.

D. Ensemble of Classifier

The ensemble of classifiers consists of simple machine learning models to the likes of Support Vector Machine, Bayes Classifier and Linear Regression. These models work together as majority voting partners. These algorithms can be executed in parallel in separate threads. This will enhance the throughput of the system as well. Also kernel methods could be employed if the separation is not possible in low dimensional space.

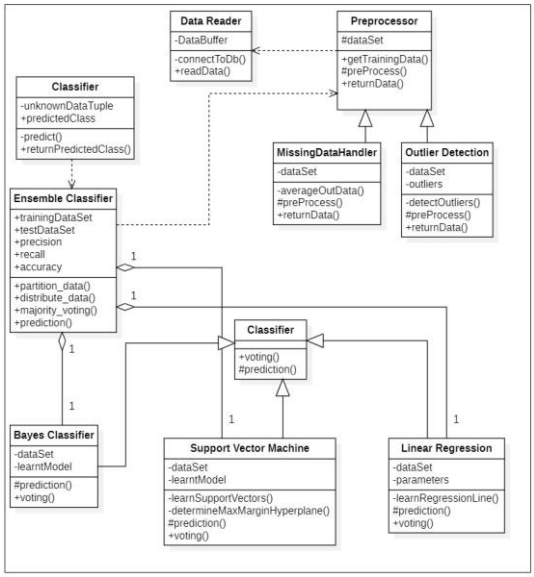


Fig. 3. Class Diagram for Learning Ensemble Model

E. Data Partitioning

The training data need to be portioned and distributed to different classifiers so as to perform the classification task in parallel. This can be done using sampling and k-fold cross validation. Data can be partitioned into k disjoint sets and each model could be fed different partitions, wherein the data-set can be duplicated for multiple models. The reserved testing data-set could also be exchanged between different models.

F. Data-Set and Inference Logic

The advertisers like Coke, Air New Zealand want to create advertisements for selling their products. They may target various types of customers using different advertisement mediums like that of Newspaper, Online on websites or bill boards. These advertisers like to run advertisement campaigns. A campaigns is a set of buys for the slots that an advertiser aims at. These campaigns are created with particular names, Flight Start Date, Flight End Date and also the total budget for the campaign. This campaign may be executed by an Ad Agency. Consider the sample Data

Table i. Campaign data

Campaign Name	***Flight Start Date***	***Flight End Date***	***Total Budget Allocated***	***Advertiser Code***
Coca-Cola Fancy Colors Campaign	21-1-2019	21-6-2019	3000$	COK
Diet For All	21-1-2016	21-3-2017	3200$	COK
Flights For Comfort	18-8-2016	18-8-2017	300$	ANZ
Fly With Air New Zealand	1-8-2017	1-8-2020	90000$	ANZ
Nike Sports For All	19-2-2018	19-4-2018	2000$	NIK

Table i. Advertiser code mapping

Advertiser Name	***Advertiser Code***
Coke	COK
Air New Zealand	ANZ
Nike	NIK
Adidas	ADD

Now from this data when the user wants to create a new campaign we need to suggest user the client codes that he probably wishes to associate with. Say the screen has fields for

1. Campaign Name
2. Campaign Start Date
3. Campaign End Date
4. Campaign Budget
5. The Medium for advertisement (Print, Television, Radio, Telecom, Online targeted)
6. Advertisers for which the campaign is to be created

Now as user fills in details one by one, we need to analyze the dataset and make probable suggestions for Advertisers which are likely to be picked up for the said campaign. This feature will enhance the overall usability and also play a major role in creating pre-populated mapping when each advertiser has got different brands to sell. This adds a business value to the product making the advertisement product more client focused.

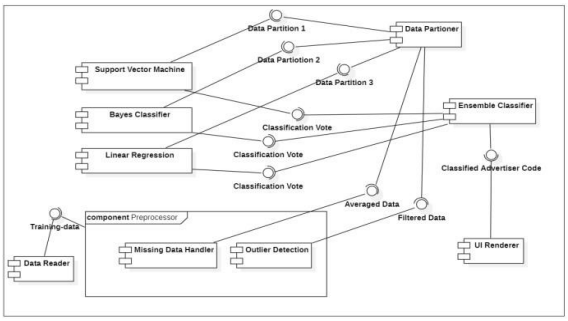


Fig. 4. Component Diagram for Learning Ensemble Model

Advantages

1) The model is learnt using ensemble classifier technique.
Hence even if one of the classifier performs incorrectly, the classification will still be accurate.
2) Enhancing the user experience by auto-suggestions in applications.
3) Parallelism can be exploited. Due to ensemble learning, multiple independent classifiers can work in parallel and asynchronously.
4) The classifier learnt is to be made generic. Hence can adjust according to the training data. It can easily be finetuned to other applications.
5) The code which needs to be designed will be scalable and production level.
6) Version control tools make it easy to maintain. Also bugs and features could be added incrementally.
7) Agile Technique is followed. Continuous delivery using small features is assured.

Disadvantages

1) Ensemble classifier is used. Hence the algorithm has to go through multiple stages.
2) Since there is constant read, we don't need to worry about the dead-locks. However there will be resource contention among multiple threads.
3) The processor time is handled by Operating System or the JVM. Hence workload needs to be even.
4) Since the model is made generic, it may result in over engineering in early stages.
5) The model would depend on training data. If the outlier detection is not handled we may end up into incorrect
classifiers being trained

Conclusions

We will be creating and training a model using the existing dataset that exists in the database. This data-set is incremental and live and keeps on changing on production. Also the relevance in various attributes is not accounted to by chance but is a mere fact of how the data needs to be created. The data has to be relevant so that it actually forms a context altogether. There are various fields some of which may skew the data. To overcome these, we will introduce only attributes that are true representative of the context.
The values of campaign names, the budget and the flight patterns help us in learning a model. The approach used for learning the model is an ensemble of classifiers which could be run in parallel.

References

[1]. Dalibor Bui, Jasminka Doba, "Lyrics classification using Naive Bayes",
[2]. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, Opatija, Croatia, 02 July 2018, PP 1011-1015
[3]. O. Obulesu, M. Mahendra, M. ThrilokReddy, "Machine Learning Techniques and Tools: A Survey", 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, Coimbatore, India, 03 January 2019, PP 605-611
[4]. Xianwei Gao, Chun Shan, Changzhen Hu, Zequn Niu, Zhen Liu, "An Adaptive Ensemble Machine Learning Model for Intrusion Detection", IEEE Access(Volume 7), IEEE, 19 June 2019, PP 82512-82521
[5]. Xiaobo Liu, Zhentao Liu, Guangjun Wang, Zhihua Cai, Harry Zhang, "Ensemble Transfer Learning Algorithm", IEEE Access(Volume 6), IEEE, 13 December 2017, PP 2389-2396
[6]. Anjan Kumar Amirishetty, Yunrui Li, Tolga Yurek, Mahesh Girkar, Wil- son Chan, Graham Ivey, Vsevolod Panteleenko, Ken Wong, "Improving Predictable Shared-Disk Clusters Performance for Database Clouds", 2017 IEEE 33rd International Conference on Data Engineering (ICDE), IEEE, San Diego, CA, USA, 18 May 2017, PP 237 -2
[7]. White Paper from Oracle, "Oracle NoSQL Database For Time Series Data", December 2017
[8]. Chetan Jaiswal, Vijay Kumar, "DbHAaaS: Database High Availability as a Service", 2015 11th International Conference on Signal-Image Technology 32 MES College of Engineering, Pune and Internet-Based
[9]. Systems (SITIS), IEEE, Bangkok, Thailand, 08 February 2016, PP 725-732
[10]. Vasileios Tsoukas, Konstantinos Kolomvatsos, Vasileios Chioktour, Athana- sios Kakarountas, "A Comparative Assessment of Machine Learning Algo-rithms for Events Detection", 2019 4th South-East Europe Design Automa- tion, Computer Engineering, Computer Networks and Social Media Confer- ence (SEEDA-CECNSM), IEEE, Piraeus, Greece, 21 November 2019, PP 1-4
[11]. Ajay Shrestha And Ausif Mahmood, "Review of Deep Learning Algorithms and Architectures", IEEE Access (Volume 7), IEEE, 22 April 2019, PP 53040-53065
[12]. Stelios E. Papadakis, Vangelis A. Stykas, George Mastorakis1 and Constandinos X. Mavromoustakis, "A hyper-box approach using relational databases for large scale machine learning", 2014 International Conference on Telecommunications and Multimedia (TEMU), IEEE, Heraklion, Greece, 09 October 2014, PP 69-73
[13]. Hanane Bais,Mustapha Machkour, Lahcen Koutti , "Querying Database using a universal Natural Language Interface Based on Machine Learning", 2016 International Conference on Information Technology for Organizations Development (IT4OD), IEEE, Fez, Morocco, 26 May 2016, PP 1-6