CLASSIFICATION OF DATA IN ERP SYSTEMS USING DECISION TREE ALGORITHMS

Farrukh Suyarov*

*Master Student,
Fergana branch of Tashkent University of Information Technologies,
Named after Muhammad al Khwarizmi,
Fergana, UZBEKISTAN
Email id: suyarov.f@gmail.com

DOI: 10.5958/2278-4853.2022.00010.6

ABSTRACT

ERP systems has become indispensable for business intelligence, varying from small enterprises to multinational corporations. Its significance as a hub for enterprise data analysis will develop exponentially throughout the next years. In terms of the recent developments in machine learning and artificial intelligence, using intelligent classification techniques for data analysis will hereby be the supreme exciting area. As the extensive usage of ERP systems increase, great volume of business information gets generated and managing such large information automatically becomes a challenging task. The accurate classification of data also becomes an additional multifarious task. It is an essential part for operative ERP applications and classical machine learning problems. This paper presents an overview of machine learning principles using decision trees to classify data in ERP systems, which can greatly benefit legacy systems, not powered with machine learning tools.

KEYWORDS: *Machine Learning, Data Analytics, Decision Trees, Algorithms, Classifier.*

INTRODUCTION

Classification is a major problem in data analysis; hence it is a primary focus of machines learning. In recent decades, substantial research and practical development were made in this field. Numerous algorithms and frameworks were developed to achieve more reliable and accurate classification models. All these discoveries have wide range applications in miscellaneous fields and problem domains. Despite this rapid growth in machine learning, ERP systems until recently were largely slow and conservative in integrating its capabilities, mainly because of adherence to existing models and difficulty in bringing rapid change to underlying business processes. However, it has become so obvious, that by using machine learning ERP systems would gain substantial improvement in terms of data analysis and forecasting. **[1]**

Small scale to giant scale sized businesses who use ERP systems spend a substantial part of their time, money, and struggle - in categorizing their data. Such categorization in traditional ERP systems is made by system administrators and operators manually. However, with the growth of data and categories this task becomes time consuming and error prone. A lot of modern versions of popular ERP systems come with some sort of machine learning capabilities, but their

functions are generally standardized and applicable to certain predefined scenarios. Gradient boosted decision trees are well known for accuracy in classifying structured data. Therefore, it is very important if one can shed a light for using algorithms based on this technique in categorizing data of general ERP. **[2]**

MAIN PART

Decision trees are very adaptable prediction models that may quickly classify, categorize, or value things based on a variety of criteria. They're also a fantastic technique to quickly depict a multi-factor choice. **[3]**

Decision trees are sets of "if then" rules that you may envisage connecting the tree's branches. If you have a lot of input data, you may use a variety of 'if/then' rules to proceed down multiple branch splits until you reach a leaf, which is an outcome or plausible prediction of a target variable. **[4]**

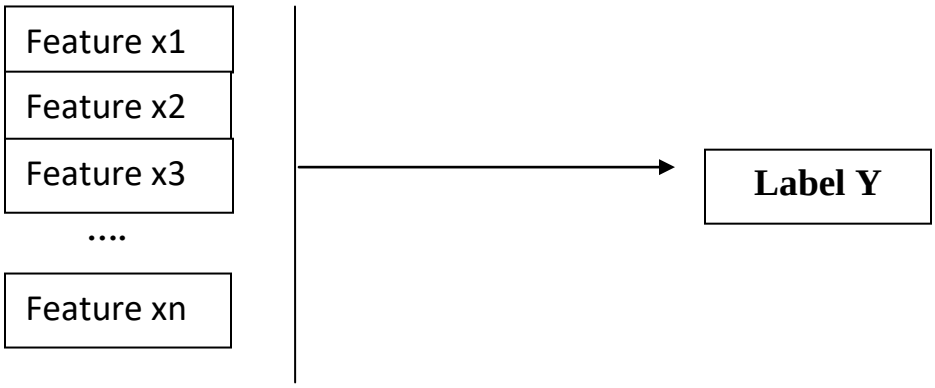
Decision trees are made up of three elements: nodes, branches and leaves.

Root nodes contain the overarching question or the decision your tree is seeking to answer. Branches, represent options, criteria or courses of action. Leaf nodes which appear at the end of branches represent either a further question to be asked (decision node), outcome (end node) or probability / uncertainty of an outcome (chance / change node). **[5]**

Decision trees can be used to make two different kinds of predictions using classification (grouping and typifying) or regression (valuing and weighting).The main difference between a classification tree and regression tree is whether the prediction that you are making is a category as in the case of a classification tree or whether the outcome is a continuous value as in the case of a regression tree. **[6]**

Decision tree based machine learning algorithms are quite accurate in classifying data in relatively small datasets. Though their performance on large datasets can be improved with various techniques. Main advantage of such algorithms besides accuracy is explainability, which means that model predictions can be explained through visualized decision tree. Therefore, decision trees are not “black boxes”, where there are no clues about decision making process and this fact strongly supports ethical use of such models. **[7]**

The type of decision tree algorithms used for classification is called CART → (Classification and Regression Tree). Initially, any machine learning classification model requires pre-classified data to be trained on. This data is considered an input. Input data is divided into two types: label and features. Features are properties that serve to correlate with label. The label would be an outcome associated with given features. Basic classification dataset might look like this: **[8]**



Machine learning models operate only on certain types of data, mainly floating-point numbers and their vectors. Therefore, data types like text should be converted into floating point numbers before use in training. The process of preparing raw data for training is often called data transformation. Basically, before training our model, the dataset should contain a vector of features $\mathbf{x_n}$ and label **Y**. **[9]**

$$(\mathbf{x}, Y) = (x_1, x_2, x_3, \ldots, x_k, Y)$$

In training machine learning models using decision tree algorithms the ideal tree would be the shortest tree possible, with the fewest splits, that can classify all data points appropriately. This appears to be an easy problem; however it is actually an NP-hard problem. The time it takes to build the optimal tree is polynomial, and it grows exponentially as the dataset grows. **[10]**

For a dataset with only ten data points and an algorithm of quadratic complexity, O (n2), the algorithm builds the tree in 10*10 = 100 iterations. When you increase the size of the dataset to 100 data points, the number of iterations the algorithm will do jumps to 10,000.

The program tries to partition the dataset into the lowest subset possible at each split [**11].** The purpose of this approach, like any other Machine Learning algorithm, is to reduce the loss function as much as feasible. However, because of separating data points into classes, the loss function should calculate a split based on the proportion of data points in each class before and after the split. In other words, a loss function that evaluates the split based on the purity of the resulting nodes is required. Gini Impurity and Entropy are loss functions that compare the class distribution before and after the split **[12].**

There are several methods to measure the efficiency of decision trees. One of them is Gini Impurity which is the measure of variance across the different classes **[13].**

$$\mathrm{G(node)} = \sum_{k=1}^{c} p_k \overline{(1 - p_k)}$$

Probability of *not* picking a data point from class *k*

Probability of picking a data point from class *k*

$$p_k = \frac{\text{number of observations with class k}}{\text{all observations in node}}$$

Another one is Entropy. It is like Gini Impurity, is a measure of chaos within a node. In the context of decision trees, chaos is defined as a node in which all classes are equally represented in the data. **[14]**

$$\text{Entropy(node)} = -\sum_{i=1}^{c} p_k \log(p_k)$$

Probability of picking a data point from class k

$$p_k = \frac{\text{number of observations with class k}}{\text{all observations in node}}$$

In classification of data in ERP systems a big part of effort should be directed at evaluating feature importance, because the model predictions can vary depending on set of features used in training. A single feature can be used in the different branches of the tree, feature importance then is its total contribution in reducing the impurity. **[15]**

The decrease in node impurity is weighted by the likelihood of accessing that node to compute feature significance. The number of samples that reach the node divided by the total number of samples is the node probability. The higher the value the more important the feature.

FeatureImportance = number_of_samples_at_parent_where_feature_is_used*impurity_at_parent-left_child_samples*impurity_left-right_child_samples*impurity_right

Impurity is the gini/entropy value, then

Normalized importance = feature importance/number_of_samples_root_node(total num of samples)

It is advisable to keep number features used in training optimal, because less or more features can lead to underfitting or overfitting.Selected features should be the most relevant ones to prediction. **[16]**

CONCLUSION

Though effectiveness of data classification in ERP systems based on a single type of algorithms, like decision trees, might not yield high efficiency due to variety and complexity of data, there are many hybrid approaches, like boosting, that can significantly improve model performance. However, simple implementation of decision trees with basic principles described above can be a good start for enhancing many legacy ERP systems with machine learning analytics. **[17,18]**

REFERENCES*:*

1. James G, Witten D, Hastie T, Tibshirani R. an Introduction to Statistical Learning: With Applications in R. New York: Springer; 2013.
2. Tan P, Steinbach M, Kumar V. Introduction to Data Mining. Addison Wesley; 2005.
3. Muthuraman S, Vidyapeetham AV. Data Classification Using Machine Learning Approach. Conference Paper. 2018.

4. Qayyum1 S, Abbasi A. Framework to Enhance Erp Usability By Machine Learning Based Requirements Prioritization. Journal of Software. 2017.

5. Akbarov DE, Umarov SA, Toychiboyev AEU. Algorithm of the Electronic Digital Subscript On The Basis of the Composition of Computing Complexities. The American Journal of Engineering and Technology. 2021;3(4):102-107.

6. Abdurakhmonova MM, Ugli Mirzayev MA, Karimov UU, Karimova GY. Information Culture And Ethical Education In The Globalization Century. The American Journal of Social Science and Education Innovations, 2021;3(03):384-388.

7. Akbarov DE, Umarov ShA. The Application of Logical Operations and Tabular Transformation in the Base Assents of Hash-Function Algorithms. Computer Reviews Journal. 2020;6:11-18.

8. Karimov UU, Karimova GY. The Importance Of Innovative Technologies In Achieving Educational Effectiveness. Журнал Естественных Наук, 2021;1(1).

9. Umarov SA, Akbarov DE. Working out the New Algorithm Enciphered the Data with a Symmetric Key. Journal of Siberian Federal University. Engineering & Technologies. 2016;9(2):214.

10. Abdurakhmonova M, Karimova G, Karimova M. Role of Ethical Culture in Preventing Violence among Spupils. Интернаука, 2021;(11-2):50-51.

11. Bakhtiyor A, Shuxratjon U. View of Models of Multiple Valuable Boolean Functions As Well As Implementation in Cryptographic Reflections. Proceedings of International Conference on Application of Information and Communication Technology and Statistics in Economy and Education, 2016. Pp.91-95.

12. Khusanova MK. Analysis of Discrete Convolution in the Matlab Program. Scientific Progress, 2021;2(4):1023-1028.

13. Karimov U, Abdurakhmon A. Innovative Information Technology in Education. Форум Молодых Ученых, 2017;(5):9-12.

14. Karimov A, Muxammadjonov X. Information Technologies: Information Education and Informatics. Экономика И Социум. 2020;8:40-43.

15. Suyumov JY. Theoretical Basis of Active Teaching Technology on The Basis of Computer Imitation Models. Academicia: An International Multidisciplinary Research Journal, 2021;11(7):205-210.

16. Каримов УУ. Рольсредствмассовойинформациивпроцессеглобализации. In Перспективные Информационные Технологии 2017; pp. 1189-1192.

17. Каримов Ў, Каримова Г. Ахборот Оқими Ва Ахборот Маданиятининг Шаклланиш Тенденциялари. Scientific Progress, 2021;2(3):743-750.

18. Suyumov J, Madaliyeva G, Xakimova K. Imitation Modeling Technologies in Higher Educational Process. Теория И Практика Современной Науки, 2021;(5):18-21.