ACCURACY PREDICTION USING ANALYSIS METHODS AND F-MEASURES

Ms. Caroline El Fiorenza J1, Prateek Parmar2, Shivam Tiwari3, Ayush Dixit4, Arjun Gupta5

1Assistant Profesor, Computer Science and Engineering Department, SRM Institute of Science and Technology, Chennai, India
2Computer Science and Engineering Department, SRM Institute of Science and Technology Chennai, India
1caro.fiorenza@gmail.com, 2prateekparmar0@gmail.com, 3shivam949@gmail.com
4ayush.dixit9@gmail.com, 5arjungupta0321@gmail.com

Abstract— Accuracy prediction is basically used in machine learning for evaluating the accuracy of data to get better results during analysis of data for various purposes like financial analysis, credit card fraud detection and sales prediction. Predicting the accuracy of data is necessary for making better decisions in field of business, engineering, medical science and analytics. We introduce a methodology for analysis that improves the accuracy of data while ensuring that the performance of the algorithm also improves so that it improves decision making so that it can be used in real world applications. The analysis involves three phases, first is product analysis phase which involves product analysis and SWOT analysis. Then comes analysis phase where we use various techniques like Straight line method of depreciation, moving average technique, simple linear regression and multiple linear regression. These methods are used for analyzing the trend in data and for comparison. Then comes the next phase where we calculate accuracy and find optimal value. For that we first add more data, then we select essential features for getting accurate results. For that we use multiple algorithms. Multiple algorithms basically consists of algorithms that are used for clustering, classification and comparison. These algorithms are used for creating a better machine learning model by using ensemble method. Ensemble method is basically a method of combining various weak algorithms to create a more accurate algorithm that gives better performance. For checking and performance and getting an accurate value we use Algorithm tuning. Algorithm tuning is used for getting an improved algorithm that gives less error percentage is assists in making predictions. This gives an accurate and optimized model for training the data.

Keywords—Accuracy analysis, Algorithms, computation, feature selection, machine learning, classification, predictions.

1. Introduction

Analysis is an important part of every company whether it’s an IT company, a consulting company or a bank. Analysis is done for examining the current scenarios and drawing a comparison with the previous scenarios to come up with conclusions on the basis of the comparisons done between the present and past data. Analysis involves examining historical knowledge to realize info concerning the present and future state of an organization. Analysis is done in various fields like business for analysing the sales of the products, defected products and making decisions regarding the amount of products to be manufactured to avoid over production. It’s also used in stock market analysis for analysing the trends in stock market to make better decisions while investing the money. It’s also done by E-commerce and credit card companies for fraud detection. Any undetected fraud in transactions can result in the company losing its customers which will be a huge loss for the company. Hence it’s very important to have an accurate data during analysis to avoid any complications that result in huge loss. Many companies these days use machine learning algorithm to generate financial reports regarding their sales, their services and production to increase the profit margin. Hence it is important that the machine considers the factors that give better results and are cost effective. Hence it’s important to select those features for analysis that determine the profits
and loss. Hence machine learning algorithm should be trained in such a way that it not only gives accurate data but also gives better performance and considers cost effective features to minimise the overall cost. Hence we are proposing an analysis, which gives an accurate data and training the algorithm to provide better performance overall. Our main aim is to increase the accuracy rate. We will use previous data for future prediction of product. We will use some analysis methods and algorithms such as straight-line technique, moving average technique, simple linear regression and multiple linear regression. Both product analysis and SWOT analysis are used for the analysis. Product analysis lets us grasp the necessary materials, processing and economic selections that area unit needed before any product are often factory-made. An understanding of those selections will facilitate us in planning and creating for ourselves. A SWOT analysis is a structured method that checks strengths, weaknesses, opportunities and threats and it also analyses the potential of a business venture. The method is important for assessing the decision-making method and check the viability of a business, project, or product. The product analysis and SWOT analysis is carried out and the results are compared. Then we move on to optimizing the data to get better accuracy rate. For that we have to select the best features and neglect the features that are irrelevant. For that we use multiple algorithms. The multiple algorithms help in creating a machine learning model that gives less error percentage and it’s done by combining various machine learning models. For that we use ensemble mode which basically combines various weak models to come up with a model that gives better accuracy and improves the overall performance. The ensemble mode basically boosts the algorithm and combines the different classifiers. The algorithm is then checked for optimal value for getting less error percentage and for that we use algorithm tuning. These techniques combined together give rise to a machine learning model that can be used for training the data to obtain less error percentage and better performance. Hence it can be used in various fields by industries for analysis, prediction and decision making.

2. Literature survey
In this section, we intend to provide an insight into different works done in the field of AI, machine learning and data analysis to understand their significance in the relation with financial analysis. M. Liu, C. Xu, Y. Luo, C. Xu, Y. Wen and D. Tao [1], mentioned cost-sensitive feature selection as their major domain of concentration and appropriate inferences were derived on that so as to extend the efficiencies, algorithms like F-measures and due to the neglect of class imbalance issue, conventional feature selection methods usually select the features that are biased towards the majority class. The selected features will be more representative of all classes. Algorithms like support vector machines(SVM) are often targeted for prediction and classification. The result from the model is biased because all the individual algorithms have additional or less similar accuracy. [2] Abuh Adah examined the statistical results using logistic regression. The four items in the questionnaire reveal significant level results, meaning that the straight-line method of depreciation can significantly affect the financial information quality of service companies. [3]. L. Miao, M. Liu and D. Zhang examined the cost sensitive feature selection’s application in software defect prediction. A cost-sensitive feature choice framework’s planned and 3 cost-sensitive feature choice algorithms, i.e., CSVS, CSLS, CSCS area unit developed that think about unequal misclassification prices and imbalance downside at the same time in feature choice part. Compared with ancient feature choice, value-sensitive feature choice aim to reduce the entire cost instead of the entire error rate. Shameem A. Puthiya Parambath S., Usunier, N. and Grandvalet, Y. examined [4] f-measures by investing the property of pseudo-linearity of a number of them to get a powerful non-asymptotic reduction to cost-sensitive classification. Musicant, David R, Vipin Kumar and Aysel Ozgur [5] provided new theoretical evidence about heuristic techniques that are very popular in the data mining community. A feature selection method using f-measure optimization is provided by using a Support Vector Machine (SVM). Aistis Raudys, S., Usunier, N. and Grandvalet, Y [6], used the moving average technique in practical trading to determine the most suitable methodology. A large-scale study by testing all the MAs on 1850 real-world daily time series from the following domains: Stock, ETF, Futures and Forex.
3. Inference

From the paper, “Cost Sensitive Feature Selection by optimizing F-measures”[1], we discovered that F-measures optimization using Cost sensitive feature selection involves assigning different costs to each class and after solving the cost sensitive feature selection problems, the most cost effective features will be selected. However execution time ought to be reduced. In the paper, “Cost-Sensitive Feature Selection with Application in software defect prediction”[3], quantitative evaluation of the impact of choosing three strategies on warranty costs, and the profitability of companies is studied. The cost-sensitive strategies have larger sensitivity than their counterpart strategies. The larger sensitivity of cost-sensitive feature choice strategies demonstrates the effectiveness of cost-sensitive feature choice strategies on imbalance datasets. The larger sensitivity indicates that value-sensitive strategies attain tiny total cost by preventing high-cost errors. Here in, “Predicting Stock Market using Regression Technique”[7], the 50 factors were divided into 4 factors and then multiple linear regression method is used and hence it is a good model for stock market prediction. A comparative analysis of 19 most popular moving averages techniques used in practical trading and determine the most suitable according the criteria “smoothens vs. lag ratio in the given paper, “Moving Averages for Financial Data Smoothing”[6], and the best smoothens/leg ratio is achieved by Exponential Hull Moving Average and Triple Exponential Moving Average. This helps in financial analysis and stock market prediction as it analysis forecasting accuracy and minimizing risks.

4. Proposed methodology

We propose a methodology that analyses the data in such a way that it focuses on both accuracy and performance of the algorithm. The method is divided into three phases where each phase focuses on filtering the data and improving the accuracy.

A. Product Analysis phase

Here we have a product data and our aim is to obtain an accurate data after the analysis of the product. We first input the data and then categorizes the products on the basis of their features. Then two analysis methodologies, product analysis and SWOT analysis for identifying the areas that must be analyzed for determining the factors which affect the cost and the amount of products that should be produced cause over-production can lead to wastage of production cost.

![Fig. 1: Product Analysis phase](image-url)
B. Analysis and classification phase

We then use various methods that can be used for analyzing the sales, trends in stock market, and production. For that we propose methods like straight line method of depreciation for forecasting the cost charged by an asset throughout its useful life, moving average technique for getting an idea from a given trend of data as it helps in filtering out unwanted noise from data and helps us understand the overall trend based on past prices. Then we use simple linear regression and multiple linear regression for generating details on behavior of the consumers, understanding the business and factors that influence profitability. It has many applications in real time environment like credit card industry, analysis of sales data etc. The data we obtain after the analysis is compared and we receive an output based on the comparison.

C. Feature selection and optimization

The analysis can sometimes contain error and hence we use machine learning algorithms to check error percentage. But we must ensure that it does not affect the overall performance of the algorithm and cause imbalance of class. Hence we use certain methods that can be used for improving the accuracy of data such as adding more data to reduce relying on assumptions. Then we move to feature selection engineering where effective features are selected out of the given features in the data. This improves computation speed and accuracy. But that doesn’t ensure accuracy so we have to use multiple algorithms to improve the accuracy. We use different algorithms for analysis of the input data to train our machine learning model.
Fig. 3: Checking the accuracy by comparing the output obtained from different methods

Then we use ensemble method which is basically combining weak methods to come up with an accurate model for prediction and analysis. It consists of two methods bagging and boosting. We basically take in data and break it down into multiple datasets and then take multiple classifiers which are combined to form a single classifier which reduces the error rate. Then we use algorithm tuning for optimizing the parameters to find an optimum value. Then we compare the data obtained from the different methods used and give an output after the comparison and by using an optimized F-measure model we show results that gives us a model with less error percentage and better performance. This machine learning model can be used to train data for getting accurate results and making better decisions.

5. Methodology

Here we introduce the methods that we are going for analysis and prediction. We use four methods for analysis of data like straight line method of depreciation for analyzing the value of an asset over a time period, moving average technique for analyzing the trend in dataset, simple and multiple linear regression for analyzing the relationship between one or more dependent and independent variables. Then we apply F-measure optimization for getting accurate data.

A. Straight line method of depreciation

Straight line depreciation is default methodology accustomed for analysing usefulness of a set asset equally over its lifetime. It is used when there is no default pattern in which an asset has to be used over time. Straight-line method is used in many places for analysis, since it is the easiest depreciation method to calculate, shows few calculation errors. For instance, if a product has a value of $1000 (V) and its residual value is estimated to be $100 (RV) while its useful life is 9 years (Y), then you’d depreciate the asset at $100 a year ($1000 V - $100 RV = $900 / 9 Y = $100).

B. Moving Average method

A moving average is a way to obtain an overall plan of the trends from an information set entered; it’s a median of any set of numbers. The moving average is very helpful for forecasting long-term trends. We can calculate it for any amount of your time for example, if you have got the information regarding sales for twenty-year, you’ll calculate a five-year moving average, a four-year moving average, a three-year moving average so on. Exchange analysts can usually use moving average to assist them see trends within the exchange and (hopefully) forecast wherever the
stocks are headed. We take a sample dataset and apply moving average on the dataset to find the trend of sales of a product within a certain time period and use excel sheet to obtain a graph based on the input data. This helps us in analysing the trends in the sale of a product. Hence we can come to a conclusion about the total sales made by the product and number of products sold in that period in the year. Similarly this method can also be applied for calculating the number of products to manufacture and the market reputation of the product according to the total sales in an area which will help the companies make decisions about production cost and time.

C. Simple Linear regression

Simple linear regression is a method to understand the relationship between two continuous dependent variables. One variable X against one variable quantity Y are plotted against each other. Technically, in regression analysis, the variable is typically referred to as the predictor variable and therefore the dependent variable is termed the criterion variable. However, many of us simply call them the independent and dependent variables. More advanced regression techniques (like multiple regression) use multiple freelance variables. Any line in two-dimensional area are often painted by this equation

\[ y = a + bx \]

Where y is that the variable on the vertical axis, x is the variable on the horizontal axis, a is the y-value wherever the line crosses the vertical axis (often referred to as the intercept), and b is the quantity of modification in y equivalent to a one-unit increase in x (often known as the slope).

D. Multiple linear regression

Multiple linear regression is the most common regression analysis. The multiple linear regression is utilized to clarify the link between one continuous variable and a pair of or extra independent variables. The independent variables are usually continuous or categorical (dummy coded as appropriate). There are many real world uses for multiple linear regression analysis. It would be used to confirm the strength of the impact that the independent variables have on a dependent variable. It's going to be used to forecast effects or impacts of changes. Hence, multiple linear regression analysis helps to understand what amount will the dependent variable change when we modify the independent variables. For instance, a multiple linear regression can tell you ways extensive grade point average is foreseen to increase (or decrease) for every one point increase (or decrease) in ratio. Multiple linear regression analysis is used for prediction of the trends in the data and also the future values so it can be used in business analysis and weather forecasting. The multiple linear regression analysis is also would to get point estimates.

E. Feature Selection Engineering

Feature selection is very important for factor in machine learning as it improves the accuracy by filtering out unnecessary data and selecting features which are necessary for the analysis. Having unwanted features in analysis can result in giving inaccurate result as it will also include the features that are not essential for determining the outcome and hence mix up with other features and give inaccurate data. Hence feature selection is important. But feature selection often ignores the class imbalance problem. Class imbalance problem occurs when one class instance outnumbers the other instance. During feature selection it can consider one feature while ignore other features which are necessary for determining the cost. This leads to cost sensitivity problems. For that we can use f-measure optimization as it basically takes care of the cost sensitive part while giving an accurate data.
F. Multiple algorithms

To obtain an accurate data sometimes we have to use multiple algorithms for training the machine to get more accurate data. Adding more data and using multiple algorithms for accuracy can help in better performance of the machine and the error percentage decreases. We can use regression, clustering and classification algorithms for obtaining an accurate data. We can use various algorithms like k-means clustering for forming clusters, random forest model for classification, and regression model for setting up a base value and then accordingly we can obtain an accurate data by comparing the results of the output with the previous value.

G. Ensemble method

Ensemble method is basically combining various algorithms to come up with a strong algorithm that gives proper accuracy. It consists of two techniques bagging and boosting. Bagging is basically a method of classifying the data and then using different classifiers on each class of the data. Then the classifiers are combined to form a single classifier that gives an improved algorithm for a proper analysis. Boosting is basically a method of improving an algorithm that is a weak learner by improving its performance to create a better performing algorithm. Hence ensemble method is necessary for improving the overall algorithm to filter out the data and find an optimal value that has less error percentage.

6. CONCLUSION

We propose a methodology for analysis and prediction using techniques such as straight line method of depreciation, moving average, simple linear regression and multiple linear regression. By using these methods on a dataset, we can analyze the product value, trends in sale, and forecast the sales on the basis of previous data. The data obtained after using the four techniques is basically analyzed using machine learning algorithms for checking error percentage and performance. If the error is high then we use various methods for reducing the error percentage such as adding more data, feature engineering and feature selection for creating an accurate hypothesis and selecting essential features according to it. Then we use multiple algorithms for checking the overall performance and use algorithm tuning for finding optimum value for the parameters that we consider. Then we combine various weak models to create a strong model that reduces error percentage and gives better performance. The accuracy prediction model can be used in many fields such as production line where we have to predict the total number of goods we need to manufacture to avoid overproduction of goods. So by analyzing the trend of overall sales and overall goods sold over a period of time, we use it for predicting the number of goods to manufacture and also predict the sales. It can also be used in stock market analysis and by investment firms to analyze the trends in stock or sales which helps them in making decisions regarding investment in the market.

7. Acknowledgment

We thank our project guide Ms. Caroline El Fiorenza J, Assistant Professor, SRM Institute of Science and Technology for assisting us with the methodologies involved in our project and providing us with guidance. We would like to thank our project co-ordinator for providing us the opportunity to work on this project.

8. References

[1] M. Liu, C. Xu, Y. Luo, C. Xu, Y. Wen and D. Tao, "Cost-Sensitive Feature Selection by Optimizing F-Measures," in IEEE Transactions on Image Processing, vol. 27, no. 3, pp. 1323-1335,
[2] ADAH, Abuh. Straight Line Method of Depreciation and Financial Information Quality of Nigerian Service Companies. International Journal of Finance and Accounting Studies, [S.I.], v. 2, n. 2, p. 1-7, nov. 2014. ISSN 2203-4706.

[3] L. Miao, M. Liu and D. Zhang, "Cost-sensitive feature selection with application in software defect prediction," Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, 2012, pp. 967-970.

[4] Puthiya Parambath, S., Usunier, N. and Grandvalet, Y. (2019). Optimizing F-Measures by Cost-Sensitive Classification. [online] Papers.nips.cc. Available at: https://papers.nips.cc/paper/5508-optimizing-f-measures-by-cost-sensitive-classification.pdf [Accessed 2 Mar. 2019].

[5] Musicant, David R., Vinip Kumar and Aysel Ozgur. “Optimizing F-Measure with Support Vector Machines.” FLAIRS Conference (2003).

[6] Puthiya Parambath, S., Usunier, N. and Grandvalet, Y. (2019). Optimizing F-Measures by Cost-Sensitive Classification. [online] Papers.nips.cc. Available at: https://papers.nips.cc/paper/5508-optimizing-f-measures-by-cost-sensitive-classification.pdf [Accessed 2 Mar. 2019].

[7] A. Shah and Bhavsar, C. (2015). Predicting Stock Market using Regression Technique. Research Journal of Finance and Accounting, [online] Vol.6(, No.3,), p.8. Available at: https://pdfs.semanticscholar.org/e6fd/f038446d249888bfc9794e8bdbc3920d17d2.pdf [Accessed 2 Mar. 2019].

[8] Statistics How To. (2019). Moving Average: What it is and How to Calculate it - Statistics How To. [online] Available at: https://www.statisticshowto.datasciencecentral.com/moving-average/ [Accessed 2 Mar. 2019].

[9] Statistics How To. (2019). Linear Regression: Simple Steps and Video - Find the Equation, Coefficient and Slope -StatisticsHowTo[online]Available at: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/ [Accessed 2 Mar. 2019].

[10] Wikipedia contributors. "F1 score." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 16 Dec. 2018. Web. 2 Mar. 2019. Available at :: https://en.wikipedia.org/wiki/F1_score

[11] Statistics Solutions. (2019). What is Multiple Linear Regression? -Statistics Solutions. [online] Available at: https://www.statisticssolutions.com/what-is-multiple-linear-regression/ [Accessed 2 Mar. 2019].

[12] Excel-easy.com. (2019). Moving Average in Excel. [online] Available at: https://www.exceleasy.com/examples/moving-average.html [Accessed 2 Mar. 2019].

[13] Jan, O. (2019). Straight Line Depreciation Method | Formula | Example. [online] Accountingexplained.com. Available at: https://accountingexplained.com/financial/non-current-assets/straight-line-depreciation [Accessed 2 Mar. 2019].

[14] Bragg, S. and Bragg, S. (2019). Straight line depreciation. [online] AccountingTools. Available at: https://www.accountingtools.com/articles/2017/5/15/straight-line-depreciation [Accessed 2 Mar. 2019].

[15] Mv Ngo Tien HoA, High Speed And Reliable Double Edge Triggered D-Flip-Flop For Memory Applications", Journal of VLSI Circuits And Systems, 1 (01), 13-17,2019

[16] MN BORHAN,"Design Of The High Speed And Reliable Source Coupled Logic Multiplexer", Journal of VLSI Circuits And Systems 1 (01), 18-22,2019