© 2017 IJSRSET | Volume 3 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Survey Paper on Manufacturing Defect Analysis and Prediction for Inspecting a Product

Amit Hombal, Pattabiraman V
School of Computing Science and Engineering, Vellore Institute of Technology, VIT University Chennai Campus, Chennai, Tamil Nadu, India

ABSTRACT

Finding defects in the manufactured product before shipment by comparing client defined specification and on the basis of past history report in which defect were found by inspectors. Through this, a new guide line is to be produced for inspectors that where they have to stress while inspecting a products. And to predict the possible future defects in the product which is produced by a particular company. Based on the number of defects client has to decide whether to give future orders or next orders to the manufacturing company.

Keywords : Defect prediction, Risk analysis, JSON schema

I. INTRODUCTION

Predictive models are models of the relation between the specific performance of a unit in a sample and one or more know attributes or features of the unit. The objective of the model is to access the likelihood that a similar unit in a different sample will exhibit the performance. Feasible approach for assessing the quality of product is to predict the defect between releases. In present industries inspections will be done by automatic defect sensors, and by image processing. These methodologies accepted everywhere in textile industries. But now a days industry people taking analysis into the next level that is pre-shipment inspection. This pre-shipment inspection is done after all the defect detection process has been completed and final product will be ready for delivery with acceptance quality level.

This inspection will give inference about the product's finishing style with respect to predefined standard specifications.

Predictive analytics encompasses a variety of statistical techiques from predictive modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events. In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities.

Models capture relationship among many factors to allow assessment of risks or potential associated with a particular set of conditions, guiding decision making for candidate transactions. The defining functional effect of these technical approaches is that predictive analytics provides a predictive score (probability) for each individual in order to determine, inform, or influence organizational processes that pertain across large numbers of individuals, such as in marketing, credit risk assessment, fraud detection, manufacturing, healthcare, and government operations including law enforcement.

Predictive modeling is an area of data minig that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown events of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs. The core of predictive analytics relies on capturing relationship between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome.

Existing predictive algorithms are briefed below :

II. METHODS AND MATERIAL

Naive Bayes Classifier

It is a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. It is a simple technique for constructing classifiers : models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular future is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is as apple, regardless of any possible correlations between the color, roundness, and diameter features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

K-Nearest Neighbors

It is a non-parametric method used for classification and regression. In both cases, the input consists of the k closet training examples in the feature space. The output depends on whether k-NN is used for classification or Regression:

- In k-NN classification, the is a class membership. An object is classified by a majority vote of its neighbors,
- with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer,
- typically small). If k=1, then the object is simply assigned to the class of that single nearest neighbor.
- In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

Both for classification and regression, it c an be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For examples, a common weighing scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

SUPPORT VECTOR MACHINE

These are supervised learning models with associated learning algorithm that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New example are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides as improvement to the support vector machines is called support vector clustering and is often used in industrial applications either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.

RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision tree's habit of overfitting to their training set.

DECISION TREE

Decision tree learning uses a decision tree as a predictive model which maps observation about an item(represented in the branches) to conclusions about the item's target value(represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable cant take a finite set of values are called classification trees: in these tree structures, leaves represents class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values(typically real numbers) are called regression trees. Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target given the values of the input variables represented by the path from the root to the leaf.

GENERALIZED LINEAR MODEL

It is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. Ordinary linear regression predicts the expected value of a given unknown quantity (the response variable, a random variable) as a linear combination of a set of observed values (predictors). This implies that a constant change in a predictor leads to a constant in the response variable (that is a linear-response model). This is appropriate when the response variable has a normal distribution (intuitively, when a response variable can vary essentially indefinitely in either direction with no fixed "zero value", or more generally for any quantity that only varies by a relatively small amount, e.g., human heights).

LOGISTIC REGRESSION

It is a regression model where the dependent variable is categorical. This article covers the case of binary dependent variable, that it where it can take only two values, such as pass/fail, win/lose,alive/dead or healthy/sick. Cases with more than two categories are referred to as multinomial logistic regression, or if the multiple categories are ordered, as ordinal logistic regression. Logistic regression can be binomial, ordinal or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types (for example, dead vs alive or win vs loss). Multinomial logistic regression deals with situations where the outcome can have three or more possible types (e.g., disease_A vs disease_B vs disease_C) that are not ordered. Ordinal logistic regression deals with dependent variables that are ordered. In binary logistic regression, the outcome is usually coded as '0' or '1', as this leads to the most straightforward interpretation. If a particular observed outcome for the dependent variable is the noteworthy possible outcome (referred to as a 'success' or a 'case') it is usually coded as '1' and the contrary outcome(referred to as a 'failure' or a 'noncase') as '0'. Logistic regression is used to predict the odds of being a case based on the values of the independent variables(predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a noncase. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using cumulative normal distribution curve instead. Equivalently, in the latent variable interpretation of these two methods, the first assumes a standard logistic distribution of errors and the second standard normal distribution of error.

ENSEMBLE LEARNING

Ensemble methods use multiple learning algorithms to obtain better predictive performance that could be obtained from any of the constituent learning algorithms alone. Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine learning ensemble refers only to a concrete finite set of alternative models, but typically allows for much more flexible structure to exist among those alternatives. An

ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The trained ensemble, therefore, represents a single hypothesis. This hypothesis, however, is not necessarily contained within the hypothesis space of the models from which it is built. Thus ensembles can be shown to have more flexibility in the functions they can represent. This flexibility can, in theory, enable them to over-fit the data more than a single model but in practice, some ensemble techniques (especially bagging – involves having each model in the ensemble vote with equal weight. In order to promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set.) tend to reduce problems related to over-fitting of the training data. Supervised learning algorithms are commonly described as performing the task of searching through a hypothesis space to find a suitable hypothesis that will make goof predictions with a particular problem. Even if the hypothesis space contains hypothesis that are very well-suited for a particular problem, it may be very difficult to find a good one. Ensemble combine multiple hypothesis to form a (hopefully) better hypothesis. The term ensemble is usually reserved for methods that generate multiple hypothesis using the same base learner. The broader term of multiple classifier systems also covers hybridization of hypotheses that are not induced by the same base learner.

GRADIENT BOOSTING

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Like other boosting methods, gradient boosting combines weak learners into a single strong learner, in an iterative fashion. It is easiest to explain in the least-squares regression setting, where the goal is to learn a model F that predicts values y` = F(x), minimizing the mean squared error (y`-y)2 to the true values y (averaged over some training set).

Approach :

1) Collecting the products specification data from the clients and inspection report data from inspectors: This data will be very much standard for the manufacturers and as well for clients also. Data is very much categorized according to product specifications. Each and every time data is coming in zip file.

2) The zip file will be extracted and loaded .sql file into database (mongodb) and created the .db file. Then the sql queries are fired to get the data through java and the results are stored in the JSON file.

3) The json file is then used for creating the document and report.

4) Compare the data with standard product specification, find how much variatioin is present between them based on given values: Data is preprocessed and used in the defect analysis and prediction. Based on the resultant, product's acceptance quality level will be cross verified by taking sample out of whole shipment stock. And the data has to be generated according to the inspector's guideline and submitted to the client.

5) Find the frequently occurring defects on a particular product and instruct the same to inspectors to give more importance while inspecting: The manufacturers past history will play major role in this stage. Because it is based on the past history data, as frequently found defects will be highlighted and the same is instructed to inspectors to give keen observation on those areas.

6) Based on the number of defects client should decide whether to give orders to him or not. For everything there will be standard difference, apart from that the product will be rejected and hence the order too.

7) Algorithm Used : *Gradient Boosting Model*

Boosting is another technique for collecting predictors. Ensembles : Weighted combination of predictors.

Boosting :

In which learners learn sequentially with early learners fitting simple models to the data and analyzing the data for errors. Those errors identify problems on particular instances of the data. These are difficult or hard to fit examples and later models focus primarily on those examples trying to get them right. In the end all the models are given weights and this set combined into some overall predictors. So boosting is a method of converting a sequence of weak learners into a very complex predictors. Initial learners tend to be simple and then weighted combination can grow more and more complex when the learners added. Gradient boosting is a instantiation of this idea for regression. The idea is to repeatedly follow this procedure to learn simple predictor of the given data. Then compute the errors residual to make the errors per data point in the predictions.

GBM Algorithm:

Make a set of predictors yˆ[i] (y had one for each data point).

The "error" in predictions is " J(y,yˆ) " (To calculate error in prediction. J just relates the accuracy of yˆ in modeling the y).

For Mean Squared Error(MSE):
J(.) = Σ (y[i] – yˆ[i])2

The gradient of J with respect to predictions yˆ is just solved by taking derivative with respect to all the predictions. And gradient descent on the predictions would look like adjusting. Yˆ has to be its old value for some step size alpha times the gradient ' f '. "adjust" yˆ to try to reduce the error :

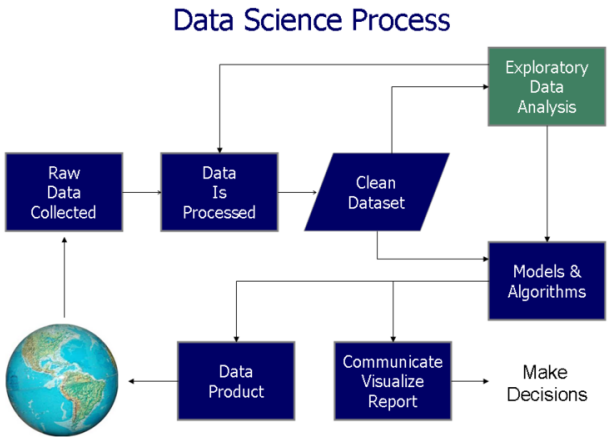
yˆ[i] = yˆ[i] + alpha f[i]
f[i] ≈ ▼J(y,yˆ) = (y[i] – yˆ[i]) for MSE (error residual)

Each learners is estimating the gradient and take a step of size 'alpha' in reducing that gradient by adding the new predictor of the gradient to the yˆ.

Gradient descent : take sequence of steps to reduce J, sum of predictors, weighted by step size alpha.

SYSTEM ARCHITECTURE

The universally accepted data science flow will be as followed.



Data flow includes:

• Collecting the generated report in CSV file format
• Converting those report into required JSON schema
• Testing JSON structure for generalization
• Loading the JSON file into MongoDb
• Finding Best fit algorithm for the problems solution

JSON file format:

JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the Java programming language. JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C family of languages, including C, C++, C#, Java,JavaScript, Perl, Python, and many others. These properties make JSON an ideal data-interchange language.

JSON is built on two structures:

- ✓ A collection of name/value pairs. In various languages, this is realized as an object, record, struct, dictionary, hast table, keyed list, or associative array.
- ✓ An ordered list of values. In most languages, this is realized as an array, vector, list, or sequence.

These are universal data structures. Virtually all modem programming langauges support them in one form or another. It makes sense that a data format is interchangeable with programming languages also be based onn these structures.

H2O : It is nothing but an open source software for data analysis. It is used for exploring and analyzing datasets for data analysis. It does analysis of data using their machine learning classification and clustering algorithm.

MongoDB : It is free and open-source cross-platform document_oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schema.

III. RESULTS AND DISCUSSION

Implementation

Working of the proposed approach is as follows :

The client, textile brand owner is going to give order for their product to different manufacturers based on the risk factor of product satisfaction. Client will collect past history of manufacturer's defects in the product and the report has been collected from inspectors about the sampled product. This process will be done before pre-shipment. Once the product was manufactured according to the clients specification, it has be cross verified by inspectors. Inspector will be carrying clients standard specification and check some random samples from large stock.

Then the difference in specification will be noticed and recorded. If the data is falling in ig difference then acceptance quality level will not be satisfied, product may reject. If not its going to be delivered to the clients. This entire data will be very much useful in measuring risk factor of whether that client can give order for this manufacturer in future or not. And also can measure the inspector's in defect detection based on there inspection report. This efficiency will be found by thier way of taking smaples (means number of samples) and finding only same defects in the product every time (number of defect found).

IV. CONCLUSION

The perspective of project t find a defect in a manufactured product (Acceptance Quality Level), to analyze and predicting model for inspecting a product by a inspector. The risk factor will give proper satisfaction over the orders.

V. REFERENCES

[1]. Markov Random Fields and Karhunen-Loeve Transforms for Defect Inspection of Textile Products-Serhat Ozdemir Bogaziqi University, Dept. of Computer Engineering, Bebek, Istanbul, Turkey 80815.

[2]. Defect Prediction using Combined Product and Project Metrics A Case Study from the Open Source "Apache" MyFaces Project Family-Dindin Wahyudin, Alexander Schatten, Dietmar Winkler, A Min Tjoa, Stefan Biffl Institute for Software Technology and Interactive Systems Vienna University of Technology, Vienna, Austria.

[3]. Ajay Kumar ,"Neural network based detection of local textile defects",Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong Received 5 April 2002; accepted 28 October 2002.

[4]. Henry Y.T. Ngan, Grantham K.H. Pang, Nelson H.C. Yung, "Automated fabric defect detection—A review", Industrial Automation Research Laboratory, Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, Laboratory for Intelligent Transportation System Research, Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong.

[5]. Meryem Ouahilal 1 , Mohammed El Mohajir 2 , Mohamed chahhou 2 , Badr Eddine El Mohajir 1, "A Comparative Study of Predictive Algorithms for Business Analytics and Decision Support systems: Finance as a Case Study", Faculty of Science, Abdelmalek Essaadi University Tetuan, Morocco.

[6]. Bin Zhangh, Abhinav Sethi, Tara N. Sainath2, Bhuvana Ramabhadran2, "A PPLICATION SPECIFIC LOSS MINIMIZATION USING GRADIENT BOOSTING", lUniversi ty of Washington, Department of Electrical Engineering, Seattle, WA 981252 IBM T. J. Waston Research Center, Yorktown Heights, NY 10598.

[7]. Nicolas Chauffert, Jonathan Israël, Bertrand Le Saux, "BOOSTING FOR INTERACTIVE MAN-MADE STRUCTURE CLASSIFICATION", Onera - The French Aerospace Lab F-91761 Palaiseau, France.

[8]. Chaitanya Kaul,Ashmin Kaul,Saurav Verma, "Comparative Study on Healthcare Prediction systems using Big Data", Mukesh Patel School of Technological Management and Engineering Narsee Monjee Institue of Management Studies Mumbai, Maharashtra. IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS'15.

[9]. Stefano Cabras, María Eugenia Castellanos, and Ernesto Staffetti, "Contact-State Classification in Human- Demonstrated Robot Compliant Motion Tasks Using the Boosting Algorithm", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 40, NO. 5, OCTOBER 2010.

[10]. Qu-Tang Cai, Yang-Qui Song, Chang-Shui Zhang, "COST-SENSITIVE BOOSTING ALGORITHMS AS GRADIENT DESCENT", State Key Laboratory on Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Automation, Tsinghua University, Beijing, China. 1- 4244-1484-9/08/$25.00 ©2008 IEEE

[11]. Deepshikha Bhargava, Ramesh C. Poonia, Upma Arora, "Design and development of an Intelligent agent based framework for Predictive Analytics", Amity Institute of Information Technology, Amity University Rajasthan, Jaipur, India. 978-9-3805-4421-2/16/$31.00 #2016 IEEE.

[12]. U. Surya Kameswari, Prof. I. Ramesh Babu "Sensor Data Analysis and Anomaly Detection using Predictive Analytics for Process Industries", Dept. of Computer Science and Engineering Acharya Nagarjuna University Andhra Pradesh, India.

[13]. Kosemani Temitayo Hafiz, Dr. Shaun Aghili, Dr. Pavol Zavarsky, "The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada", Department of Information Systems Security and Assurance Management, Concordia University of Edmonton,Edmonton, Canada.

[14]. Chensheng Sun, 2 Sanyuan Zhao, 1 Jiwei Hu, 1 Kin-Man Lam, "TOTALLY-CORRECTIVE BOOSTING USING CONTINUOUS-VALUED WEAK LEARNERS", Center for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China, School of Information and Electronic Engineering, Beijing Institute of Technology, Beijing, China. 978-1-4673-0046-9/12/$26.00 ©2012 IEEE.

[15]. A.Rishika Reddy, P. Suresh Kumar, "Predictive Big Data Analytics in Healthcare", Computer Science and Engineering Kakatiya Institute of Technology & Science Warangal, India. 2016 Second International Conference on Computational Intelligence & Communication Technology. 978-1-5090-0210-8/16 $31.00 © 2016 IEEE DOI 10.1109/CICT.2016.129